ESTADÍSTICA II

CUADERNO DE TRABAJO

CIENCIAS BÁSICAS

textosuniversitarios

Javier Bech Vertti

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

ISBN 978-607-8285-62-4

ESTADÍSTICA II

CUADERNO DE TRABAJO

ISBN 978-607-8285-62-4

ESTADÍSTICA II

CUADERNO DE TRABAJO

Javier Bech Vertti

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

ISBN 978-607-8285-62-4

ESTADÍSTICA II

CUADERNO DE TRABAJO

ISBN 978-607-8285-62-4

Hecho en México / *Made in Mexico*

ISBN 978-607-8285-62-4

PRÓLOGO

El "Cuaderno de Trabajo**: ESTADÍSTICA II",** es una obra elaborada durante mi año sabático para **apoyar en particular al estudiante de la carrera de Mercadotecnia y a otros estudiantes de carreras afines adscritas al Centro de Ciencias Económicas y Administrativas de la U.A.A**., en el estudio de los conceptos estadísticos durante el desarrollo de su **curso "Estadística II" (Estadística Inferencial en otras carreras)** por medio de una gran variedad de **ejemplos ilustrativos resueltos en detalle, actividades de aprendizaje, autoevaluaciones, ejercicios de refuerzo y sesiones de inducción a los software estadísticos Excel y Minitab**, así como un sin número de **ayudas didácticas** que incluyen ejercicios complementarios, autoevaluaciones con reactivos de falso o verdadero y de opción múltiple, un amplio glosario de términos, simbología utilizada, fórmulas claves utilizadas en cada capítulo y las respuestas a los ejercicios planteados en todo **el "C**uaderno de Trabajo**: ESTADÍSTICA II".**

Este **"C**uaderno de Trabajo**: ESTADÍSTICA II",** fue diseñado en principio como **auxiliar para el estudiante** y en forma **opcional para el profesor** que lo desee utilizar como **guía del curso** en el planteamiento y resolución de los ejercicios aquí planteados **coadyuvando con esto en el proceso enseñanza-aprendizaje** durante el desarrollo del curso de **"Estadística II" y/o "Estadística Inferencial".** Debo hacer hincapié en que **NO pretende suplir en forma alguna la labor del docente en el aula**, sin embargo la forma de abordar los contenidos del programa de la materia permite **ampliar y reforzar los temas tratados en el curso así como estimular nuevas prácticas pedagógicas y herramientas**, para construir un **aprendizaje significativo en el aula de clase** y que en forma muy puntual **debe llevar al estudiante a ser capaz de**: identificar **las características, principios, elementos, supuestos y propósitos del diseño de experimentos, del análisis de regresión lineal simple y múltiple** para aplicarlos en **problemas de su área de estudio** en el ámbito de los fenómenos económicos, financieros, comerciales y administrativos, **demostrando capacidad para analizar e interpretar resultados numéricos estadísticos en contextos específicos.**

ISBN 978-607-8285-62-4

El "**C**uaderno de Trabajo**: ESTADÍSTICA II**" está **organizado** en **tres capítulos** que corresponden a las **tres unidades del programa de la materia de Estadística II para la carrera de Mercadotecnia**: **Análisis de experimentos, Regresión Lineal Simple y Regresión Lineal Múltiple**. En cada uno de ellos se tratan contenidos relevantes del programa y por eso, **todos se inician con la descripción de los aprendizajes esperados que debe lograr el estudiante**. Cada contenido se estructura en las siguientes secciones:

1. **Síntesis de los conceptos básicos**: es un resumen de los conceptos centrales involucrados en los aprendizajes en el aula de clase. Asimismo, se encuentran las principales fórmulas y relaciones numéricas que sustentan la Estadística.

2. **Ejemplos ilustrativos resueltos**: en esta sección se plantean ejercicios representativos de la clase y se resuelven en detalle.

3. **Actividades de aprendizaje**: esta sección le permitirá al estudiante fijar las ideas y están diseñados para ser resueltos primero en forma convencional, es decir a mano, y posteriormente utilizando un software estadístico para comparar sus resultados, fomentando así la retroalimentación correspondiente.

4. **Autoevaluaciones**: esta sección le permitirá al estudiante ejercitar los aprendizajes en el aula de clase y podrá autoevaluar su desempeño y darse cuenta en que puntos o áreas se encuentra más débil y en cuales más fuerte para así enfocar sus esfuerzos en los puntos más débiles, ahorrándole tiempo en la preparación de su examen departamental.

5. **Ejercicios de refuerzo**: en esta sección, se presentan varios ejercicios para su resolución, orientados a la preparación del estudiante para el examen departamental del capítulo.

6. **Ejemplos ilustrativos resueltos en Excel y/o Minitab**: esta sección utiliza pantallas de captura, cuadros de dialogo, gráficos y salidas de resultados que le permitirá al estudiante familiarizarse con los comandos básicos necesarios para lograr buenos resultados con estos software estadísticos.

7. **Notas al margen izquierdo**: esta sección le permitirá advertir al estudiante de algún aspecto a remarcar o alertar de lo que se dice en el texto desarrollando su capacidad de análisis al tener que comprender y examinar el texto minuciosamente.

Uso de calculadora y software estadístico: Para trabajar con el presente **"Cuaderno de trabajo: Estadística II",** el estudiante debe **usar calculadora y algún software estadístico**. En cuanto al **uso de la calculadora**, en particular en el desarrollo del **ejemplo ilustrativo del Capitulo 2**

ISBN 978-607-8285-62-4

correspondiente al Análisis de Regresión Lineal Simple, se incluye un cada apartado un **pequeño tutorial sobre el uso del módulo de regresión lineal simple de la calculadora Casio fx.82MS**, sólo como un ejemplo de su uso ya que cada estudiante dispone de calculadoras diferentes, y que en forma opcional el estudiante podrá utilizar para agilizar los cálculos del ejemplo ilustrativo, así como de los demás ejercicios planteados en el cuaderno.

El uso de un **software estadístico como Excel y/o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero se resuelva el ejercicio en forma manual** y posteriormente se utilice un software para comparar los resultados. Es importante mencionar **que pueden existir diferencias en las respuestas debido a la cantidad de dígitos** que se utilizan en los cálculos manuales. Dado que en algunos contenidos se utiliza álgebra matricial, **se sugiere utilizar en general aproximaciones de al menos 5 dígitos.**

Al final de cada capítulo se presenta el siguiente **conjunto de ayudas didácticas**:

- **Ejercicios complementarios de recapitulación** que el estudiante podrá resolver a mano o mediante el uso de un software estadístico orientados a la preparación del examen departamental del capítulo.
- **Autoevaluaciones con reactivos de falso o verdadero y de opción múltiple** que permitirá reforzar el aprendizaje del alumno.
- **Un glosario de términos** de los conceptos expuestos.
- **Simbología** utilizada.
- **Fórmulas clave** que apoyan la resolución de los ejercicios propuestos en cada sección.
- **Ejemplo ilustrativo de uso de la calculadora CASIO fx.82MS** (solo en el Capítulo 2).

Al final del **"Cuaderno de trabajo: Estadística II",** se presenta:

- **Apéndice con Tablas.**
- **Bibliografía.**

Y en un **anexo** aparte al **"Cuaderno de trabajo: Estadística II",** se presentan las:

- **Respuestas a los ejercicios** planteados en las actividades de aprendizaje, autoevaluaciones, ejercicios de refuerzo y complementarios, así como de las autoevaluaciones con reactivos de de falso o verdadero y de opción múltiple de cada capítulo.

ISBN 978-607-8285-62-4

Finalmente**...** el viejo refrán que dice que **"La práctica hace al maestro"** tiende a ser mas cierto de lo que muchas veces pensamos. **La práctica, y no la genética**, es lo que hace **al maestro**. Y es **la perseverancia**, a lo largo del tiempo, lo que verdaderamente saca adelante a los **triunfadores**, por eso este "**C**uaderno de Trabajo**: ESTADÍSTICA II**" se ha elaborado con la finalidad de que **el estudiante ejercite los procedimientos que se trabajarán a lo largo del curso de "Estadística II",** pensando que entre **más se practique**, mucho mayor será **la comprensión que se tenga de ellos.** Este material pude ser de **estudio independiente o combinado** con las instrucciones del docente en caso de opte por utilizarlo como guía de estudio. Se recomienda **utilizarlo como material de apoyo, para reforzar conocimientos, para autoevaluarse, o como preparación para la evaluación presencial**. Lo importante es que **le sirva al estudiante para que identifique cuáles son los temas que necesita reforzar y para darse cuenta de los logros que ha alcanzado.** **Espero que sea de utilidad.**

ISBN 978-607-8285-62-4

PROYECTO DE AÑO SABÁTICO 2011 ARQ. Y M. EN ADMÓN. JAVIER BECH VERTTI

P.1 PROBLEMÁTICA Y NECESIDADES.

Ante la carencia de un **Cuaderno de Trabajo** para el **estudio y la enseñanza de las metodologías estadísticas** necesarias para entender **el comportamiento de algunos fenómenos aleatorios relacionados con el área de Mercadotecnia**, que permita **desarrollar en el alumno habilidades** para **aplicar dichas metodologías** asociadas al análisis de datos y dado que se requiere de la **consulta de una vasta bibliografía** , se ha observado, a través de 19 años de experiencia docente, que los **alumnos suelen limitarse a los apuntes que el profesor les ofrece**, **eventualmente consultan algún texto** y en **pocas ocasiones acuden a asesoría**; lo que **redunda en un bajo nivel de aprovechamiento**, evidente en los trabajos y exámenes que realizan.

Por lo que hace **a los profesores**, también se ha observado que **eligen como apoyo algún texto base**, que **puede no apegarse completamente al programa de la materia** y **complementan con otros textos** los temas faltantes **para la elaboración de apuntes y trabajos**, con lo que dejan **puntos del programa parcialmente cubiertos o sin cubrir** además de la **falta de uniformidad en la nomenclatura** al utilizar varios textos **lo que crea conflicto en los alumnos** que por alguna razón, deben repetir el curso.

En la **mayoría de los textos no se incluye ningún apartado para realizar actividades extra clase ni de autoevaluación**, con el que **los alumnos** puedan en **forma personal reafirmar y evaluar sus conocimientos a medida que avanza el semestre y retroalimentarse para preparar sus exámenes**, por lo que **no se puede fomentar en forma efectiva el proceso de auto aprendizaje** que marca tanto el programa de la materia como el **modelo educativo adoptado por la Institución** donde **el alumno debe ser un agente activo con una orientación constructivista.**

ISBN 978-607-8285-62-4

P.2 JUSTIFICACIÓN DEL PROYECTO

De acuerdo a la **problemática y necesidades planteadas**, se justifica el esfuerzo para remediarlas mediante **la elaboración de un "Cuaderno de Trabajo para la materia de Estadística II y /o Estadística Inferencial"** principalmente para la **carrera de Mercadotecnia,** con posibilidad de utilizarse **en carreras afines adscritas al Centro de Ciencias Aconómicas y Administrativas.** La materia de **Estadística II en la carrera de Mercadotecnia** se cursa en el **cuarto semestre** y el **cuaderno de trabajo conducirá a los alumnos en su estudio** permitiéndoles **ampliar de forma importante su horizonte de conocimientos.** Asimismo, **servirá de apoyo para el desarrollo del trabajo de los profesores** uniformizando por un lado tanto los **criterios, niveles y alcances establecidos por la Academia de Métodos Estadísticas Básicos Nivel B-2,** (Carreras con 1 o 2 cursos curriculares de Matemáticas) para cada contenido, **como la nomenclatura que deberá utilizar el profesor**, **evitando así la confusión de los alumnos que por algún motivo tuvieran que recursar la materia**; además **el profesor podrá contar como apoyo** con una **serie de actividades individuales y/o grupales** así como **reactivos o ítems** para que los alumnos **desarrollen y/o realicen autoevaluaciones** ya sea **en forma presencial en el aula de clase o bien mediante alguna plataforma ó Software de cómputo** estadístico como pudieran ser **Moodle, Minitab y/o Excel**, que ya tiene habilitado el Departamento de Estadística como apoyo para algunos cursos.

P.3 OBJETIVO GENERAL Y OBJETIVOS PARTICULARES

a) General:

Elaborar un **documento** que sirva como **"Cuaderno de Trabajo: ESTADÍSTICA II"** para la materia de **"Estadística II" de la Carrera de Licenciado en Mercadotecnia** y **carreras afines adscritas al Centro de Ciencias Administrativas de la U.A.A.**, que contenga material útil para estudiar y entender el comportamiento de fenómenos aleatorios relacionados con el área de la Mercadotecnia y carreras afines **dentro del programa vigente de Estadística II** para la carrera de **Mercadotecnia**, ofreciendo **a alumnos y profesores** información **diversa y actualizada**, así como **herramientas para construir un aprendizaje significativo en el aula de clase**, que contribuya a **mejorar la calidad de la educación**, cercanos a la realidad que viven los estudiantes de nuestro Estado y por ende del país.

ISBN 978-607-8285-62-4

b) Específicos:

- Que el **"C**uaderno de Trabajo**: ESTADÍSTICA II"** impacte en los **procesos educativos y de enseñanza-aprendizaje** por medio de la interacción de los alumnos con los contenidos pedagógicos incorporados en el mismo.
- Que el **"C**uaderno de Trabajo**: ESTADÍSTICA II"** se pueda **utilizar como una herramienta de apoyo docente** en el tratamiento de los temas y contenidos de los libros de Texto, **con base en el programa vigente de la materia de Estadística II de Mercadotecnia** y carreras afines, con la finalidad de **ampliar y reforzar los temas que en ellos se traten**, así como **estimular nuevas prácticas pedagógicas y herramientas**, para construir un **aprendizaje significativo en el aula de clase.**
- Que el **"C**uaderno de Trabajo**: ESTADÍSTICA II" defina al profesor como guía y mediador del proceso de debate, reflexión y participación** que se genere en el aula y le **sugiera estrategias didácticas e innovadoras para el tratamiento de los contenidos curriculares,** a fin de integrarlas a sus experiencias y métodos propios.
- Que el **"C**uaderno de Trabajo**: ESTADÍSTICA II"** indique **la posible incorporación de las TIC´S en los procesos educativos**, a fin de establecer un **puente natural entre la forma tradicional de presentar los contenidos curriculares y las posibilidades que brindan las nuevas tecnologías.**

P.4

METAS Y/O ACTIVIDADES

Desarrollar las unidades marcadas en el programa vigente de la **materia de Estadística II de la carrera de Mercadotecnia**, mediante:

a) La revisión bibliográfica para la selección y compilación del material idóneo para el programa de la materia.

b) Estructuración del material.
- Breve introducción de conceptos básicos.
- Elaboración de ejemplos ilustrativos.
- Actividades para desarrollar en clase y/o laboratorio, en forma individual y/o grupal.
- Sesiones de inducción y/ o prácticas de laboratorio para ser realizadas con algún paquete de cómputo, con ejemplos ilustrativos.
- Propuestas de problemas, ejercicios y/o tareas para reforzar los temas con y sin respuestas.

c) Autoevaluaciones para resolver en clase o para que el profesor pueda capturarlas y aplicarlas en algún medio digital como la plataforma Moodle que el Departamento de Estadística ya tiene a disposición de los docentes.

ISBN 978-607-8285-62-4

d) **Apéndices.**
e) **Glosario de términos.**
f) **Formularios.**

P.5 CALENDARIZACIÓN DE ACTIVIDADES

Calendarización.

a) Agosto 2011 a Diciembre 2011: Desarrollo de la unidad I, y 50% de la Unidad II.
b) Diciembre 2011 a Enero de 2012: Presentación del Informe semestral de avance.
c) Enero a Julio de 2012: Desarrollo del complemento de la Unidad II; y de la totalidad de la unidad III.
d**)** Agosto de 2012: Presentación del Informe Final.

P.6 CRITERIO DE EVALUACIÓN

Las actividades del proyecto se evaluarán por medio de la comparación de los trabajos reportados con el cumplimiento de las metas propuestas en los tiempos programados en la calendarización.

P.7 OBSERVACIONES GENERALES

El "Cuaderno de Trabajo**: ESTADÍSTICA II"** aunque va dirigido específicamente a la **carrera de Mercadotecnia** podrá ser utilizado perfectamente como apoyo durante el desarrollo de la materia de "**Estadística II" ó también llamada "Estadística Inferencial"** en **carreras adscritas al Centro de Ciencias Económicas y Administrativas de la U.A.A**. y/o equivalentes a la **Academia de Métodos Estadísticos Básicos Nivel B-2. Del Departamento de estadística de la U.A.A.**

ISBN 978-607-8285-62-4

REVISIÓN BIBLIOGRÁFICA

RB.1 RESUMEN

Con el fin de poder **estructurar y desarrollar** el **"Cuaderno de trabajo: ESTADÍSTICA II"** se procedió con anterioridad a realizar una exhaustiva **revisión bibliográfica** de diversos libros de texto que tuvieran relación con los **contenidos que marca el programa de la materia de Estadística II para la carrera de Mercadotecnia adscrita al Centro de Ciencias Económicas y Administrativas de la U.A.A.,** así como material igualmente relacionado en Internet para **identificar** lo qué se conoce del tema, lo que se ha investigado a la fecha así como qué aspectos aún permanecen desconocidos. Se aplicó el siguiente **proceso** encaminado a **localizar, procesar y reconstruir información relevante** de acuerdo a su fuente, al proceso de análisis implicado y al resultado esperado.

RB.2 ETAPAS DE LA REVISIÓN BIBLIOGRÁFICA

Introducción:
- Definición de objetivos.

Método:
- Búsqueda bibliográfica.
- Criterios de selección.
- Recuperación de la información. Fuentes documentales.
- Evaluación de la calidad de los artículos seleccionados.
- Análisis de la variabilidad, fiabilidad y validez de los artículos.

Desarrollo:
- Organización y estructuración de los datos.
- Elaboración del mapa mental.
- Combinación de los resultados de diferentes originales.

Bibliografía

ISBN 978-607-8285-62-4

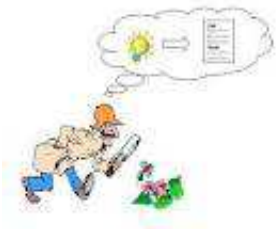
RB.3

INTRODUCCIÓN: DEFINICIÓN DE OBJETIVOS

- **R**esumir información sobre los contenidos especificados en el programa de la materia de Estadística II para la carrera de Mercadotecnia y carreras similares adscritas al Centro de Ciencias Económicas y Administrativas de la U.A.A.
- Identificar los aspectos relevantes conocidos, los desconocidos y controvertidos sobre los contenidos revisados.
- Identificar las aproximaciones teóricas elaborados sobre los contenidos revisados.
- Identificar las variables asociadas a los contenidos en estudio.
- Proporcionar información amplia sobre el tema.
- Ahorrar tiempo en la lectura de documentos primarios
- Mostrar evidencia disponible.
- Sugerir aspectos o temas de estudio más profundo.

RB.4

MÉTODO Ó PROCEDIMIENTO



- **Búsqueda bibliográfica**: Se acudió a fuentes documentales primarias básicamente libros de texto, la mayoría de ellos especificados como bibliografía en el programa de la materia de Estadística II para la carrera de Mercadotecnia. De fuentes secundarias se acudió a búsqueda por Internet.
- **Criterio de selección**: Se eligieron los descriptores o palabras y frases claves para comenzar la búsqueda. Asimismo se utilizaron términos alternativos (sinónimos) para los conceptos o variables para usarlos como palabras claves.
- **Recuperación de la información**: El criterio de selección de los libros y artículos estuvo en función de su calidad metodológica y que cumplieran con los criterios de calidad científica que se busca.
- **Evaluación de la calidad de los artículos seleccionados**: En una primera fase el aspecto que se tomó en cuenta fue: el título, los autores, el contenido y los resultados. En cuanto al título se buscó que fuera útil y relevante para los contenidos que se buscan, de los autores se identificó la credibilidad y/o experiencia en el tema y del contenido se analizó que fuera el correcto y si los resultados son aplicables al tema de estudio.
- **Análisis de la variabilidad, fiabilidad y validez de los artículos:** En una segunda etapa se procedió a la lectura crítica de los documentos y/o información evaluando la confiabilidad, la precisión de sus resultados y la pertinencia o aplicabilidad de los resultados al tema.

ISBN 978-607-8285-62-4

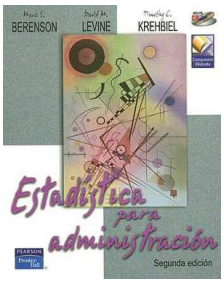
RB.5

DESARROLLO PARA ORGANIZAR Y ESTRUCTURAR LA INFORMACIÓN

- **Organización y estructuración de los datos**: En primer lugar se redujo la información eliminando todo aquello no esencial mediante un proceso de segmentación de la información básica (en pequeñas notas de papel) y ordenando dicha información por grupos para que éstos comenzaran a adquirir características comunes. En segundo lugar se procedió a asignarle un nombre a cada grupo. En tercer lugar se procedió a integrar los grupos que se parecían bastante, de manera que algunos quedaran aislados y otros integrados. En cuarto y último lugar se priorizaron los grupos para identificar la información que fuera más relevante dentro de la organización alcanzada.
- **Elaboración del mapa mental**: Para el proceso de organización descrito anteriormente se recurrió a un mapa conceptual o un mapa mental que incluyo un solo concepto alrededor del cual se estructuró toda la información jerarquizando diferentes niveles de generalidad e inclusividad conceptual y se conformaron conceptos, proposiciones y palabras enlace.
- **Combinación de los resultados de diferentes originales:** Con los resultados de los diferentes originales se procedió a combinarlos para estructurar y desarrollar el "Cuaderno de trabajo: ESTADÍSTICA II" para la carrera de Mercadotecnia y carreras afines adscritas al Centro de Ciencias Económicas y Administrativas de la U.A.A., suministrando la información siguiendo un proceso lógico y paulatino de forma que primero se redactaron las ideas que son antecedentes y posteriormente se desarrollaron las ideas consecuentes.

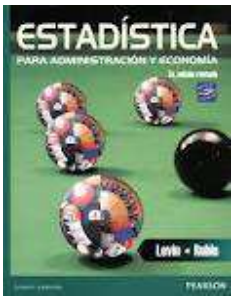
RB.6

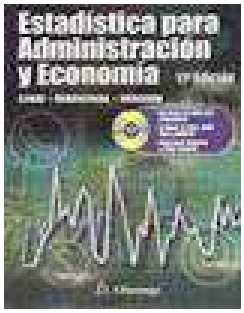
BIBLIOGRAFÍA



1. **B**erenson, Mark L. y Levine, David M. **ESTADÍSTICA PARA ADMINISTRACIÓN.** Editorial Pearson. Cuarta Edición, 2006. Formato: Rústico. Idioma: español. País: México. ISBN: 9702608023. No. de páginas: 648.
2. Carlberg, Conrad. **ANÁLISIS ESTADÍSTICO CON EXCEL**. Editorial Anaya multimedia-Anaya interactiva. Edición, 2011. Formato: Rústico. Idioma: Español. País: España. ISBN: 9788441530263. No. de páginas: 528.
3. Carrascal Arranz, Ursicino. **ESTADÍSTICA DESCRIPTIVA CON MS MICROSIFT EXCEL 2010: VERSIONES 97 A 2010**. Editorial Alfaomega grupo editor. Edición, 2012. Formato: Rústico. Idioma: español. País: México. ISBN: 9786077071969. No. de páginas: 288.
4. Dawson-Saunders, Beth y G. Trapp, Robert. **BIOESTADÍSTICA MÉDICA**. Editorial Manual modern. Segunda edición, 1997. Formato: Rústico. Idioma:

ISBN 978-607-8285-62-4





español. País: México. ISBN: 9684267517. No. de páginas: 403.

5. Gutiérrez Pulido, Humberto y De la Mara Salazar, Román. **ANÁLISIS Y DISEÑO DE EXPERIMENTOS**. Editorial Mc Graw Hill (México). Edición, 2008. Formato: Libro electrónico. Idioma: español. País: México. Código producto: 9786071501394. Tamaño: 12.20 MB.
6. Hildebrand, David K y Ott R., Lyman. **ESTADISTICA APLICADA** a la administración y a la economía. Editorial Addison–Wesley Iberoamericana, 1997. Formato: Rústico. Idioma: español. País: México. ISBN: 0201625520. No. de páginas: 943.
7. Johnson, Robert. **ESTADÍSTICA ELEMENTAL**. Editorial Cengage Learnin. Décima edición, 2008. Formato: Rústico. Idioma: español. País: México. ISBN: 9789706868350. No. de páginas: 725.
8. Kazmier, Leonard y Díaz Mata, Alfredo. **ESTADÍSTICA APLICADA** a la administración y a la economía. Editorial Mc Graw-Hill Interamericana. Edición, 2006. Formato: Rústico. Idioma: español. País: México. ISBN: 9701059182. No. de páginas: 406.
9. Levin, Richard I. **ESTADÍSTICA PARA ADMINISTRADORES**. Editorial Prentice- Hall hispanoamericana, S.A. Segunda edición, 1988. Formato: Rústico. Idioma: español. País: México. ISBN: 9688801526. No. de páginas: 940.
10. Levin, Richard I. **ESTADÍSTICA PARA ADMINISTRACIÓN Y ECONOMIA**. Editorial Pearson. Séptima edición, 2010. Formato: Rústico. Idioma: español. País: México. ISBN: 9786074429053. No. de páginas: 952
11. Levine, David M, Berenson, Mark I. **ESTADÍSTICA PARA ADMINISTRACIÓN**. Editorial Pearson. Cuarta Edición, 2008. Formato: Rústico. Idioma: español. País: México. ISBN: 9702608023. No. de páginas: 648.
12. Lind, Douglas A., Marchal, William G. y Wathen, Samuel A. Wathen. **ESTADÍSTICA APLICADA** a los Negocios y a la Economía. Editorial Mc Graw –Hill Interamericana. 12ª. Edición, 2008. Formato: Rústico. Idioma: español. País: México. ISBN: 9701048342. No. de páginas: 800.
13. Lind, Douglas A., Marchal, William G. y Mason, Robert D. **ESTADÍSTICA PARA ADMINISTRACIÓN Y ECONOMÍA**. Editorial Alfaomega. Onceava edición, 2004. Formato: Rústico. Idioma: español. País: México. ISBN: 9701509749. No. de páginas: 830.
14. **MANUAL DE MINITAB 15**. Versión en español para Windows. Edición, 2007. Formato: electrónico. Idioma: español. País: México.
15. Márquez, Felicidad. **ESTADÍSTICA DESCRIPTIVA** a través de Excel. Editorial Alfaomega grupo editor. Edición, 2009. Formato: Rústico. Idioma: español. País: México. ISBN: 9786077686989. No. de páginas: 288.
16. Montgomery, Douglas C. **DISEÑO Y ANÁLISIS DE EXPERIMENTOS**. Editorial Limusa. Primera edición, 2008. Formato: Rústico. Idioma: español. País: México. ISBN: 9681861566. No. de páginas:
17. Pérez López, Cesar. **ESTADÍSTICA APLICADA** a través de Excel. Editorial Pearson-Prentice Hall. Edición, 2002. Formato: Rústico. Idioma: español. País: México. ISBN: 8420535362. No. de páginas: 596.
18. Velasco Sotomayor, Gabriel. **ESTADÍSTICA CON EXCEL**. Editorial Trillas. Primera edición, 2005. Formato: Rústico. Idioma: español. País: México. ISBN: 9682406269. No. de páginas:
19. Walpole, Ronald E. **PROBABILIDAD Y ESTADÍSTICA**. Editorial Mc Graw-Hill Interamericana. Edición, 1992. Formato: Rústico. Idioma: español. País: México. ISBN: 9684229925. No. de páginas:

ISBN 978-607-8285-62-4

CONTENIDO

ISBN 978-607-8285-62-4

ISBN 978-607-8285-62-4

ISBN 978-607-8285-62-4

CAPÍTULO 3
ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

ISBN 978-607-8285-62-4

ISBN 978-607-8285-62-4

ISBN 978-607-8285-62-4

ESTADÍSTICA II

CUADERNO DE TRABAJO

CAPÍTULO 1
ANÁLISIS DE EXPERIMENTOS

Javier Bech Vertti

ISBN 978-607-8285-62-4

ESTADÍSTICA II CAPÍTULO 1

Av. Universidad No. 940
Ciudad Universitaria
C.P. 20131, Aguascalientes, Ags.
http://www.uaa.mx/direcciones/dgdv/editorial/

ISBN 978-607-8285-62-4

Hecho en México / *Made in Mexico*

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

CAPÍTULO 1. ANÁLISIS DE EXPERIMENTOS

OBJETIVO 1.1. El alumno podrá comprender los conceptos básicos del diseño de experimentos y su análisis

ANTECEDENTES

CONCEPTOS DE:

Variables aleatorias. La media de una población. Media de una muestra. Varianza y desviación estándar para datos. La distribución normal estándar. La distribución *t* de Student. La distribución *F*. Muestreo aleatorio simple. Error de muestreo. Distribución muestral de las medias. Estimadores puntuales e intervalos de confianza. Pruebas de hipótesis. Pruebas de hipótesis para dos medias

1.1.1 INTRODUCCIÓN, PRINCIPIOS BÁSICOS, ETAPAS EN EL DISEÑO DE EXPERIMENTOS Y ANOVA.

CONCEPTOS BÁSICOS DISEÑO DE EXPERIMENTOS

El diseño de un experimento es la secuencia completa de pasos tomados de antemano para asegurar que los datos apropiados se obtendrán de modo de modo que permitan un análisis objetivo que conduzca a deducciones válidas con respecto al problema establecido.

El **diseño experimental** es una **técnica estadística** que permite **identificar y cuantificar las causas de un efecto** dentro de un **estudio experimental**. En un **diseño experimental** se **manipulan deliberadamente una o más variables**, vinculadas a las causas, para medir el efecto que tienen en otra variable de interés. El **diseño experimental** prescribe una serie de **pautas relativas** a **qué variables** hay que manipular, **de qué manera**, **cuántas**

ISBN 978-607-8285-62-4

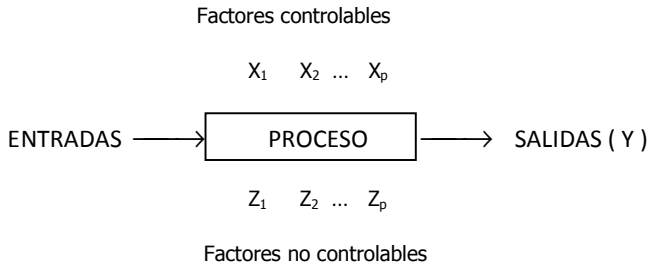
veces hay que repetir el experimento y **en qué orden** para poder establecer con un **grado de confianza** predefinido la necesidad de **una presunta relación de causa-efecto**.

El diseño experimental encuentra aplicaciones en la **mercadotecnia,** la **industria**, la **agricultura,** la **medicina**, las **ciencias de la conducta**, etc. constituyendo una **fase esencial en el desarrollo de un estudio experimental**

Las variables predictoras y de respuesta son las variables de interés en un experimento (las que se miden u observan) se denominan variables dependientes o de respuesta. Otras variables en el experimento que afectan la respuesta y que el experimentador puede establecer o medir se denominan variables predictoras, explicativas o independientes. A una variable predictora continua a veces también se denomina covariable y una variable predictora categórica a veces es mencionada como un factor

El **proceso o sistema** bajo estudio puede representarse por medio del siguiente modelo:

MODELO GENERAL DE UN PROCESO O SISTEMA



Por ejemplo un ingeniero quiere estudiar la resistencia de una pieza plástica sometida a tres temperaturas cambiantes. La pieza puede ser elaborada con tres tipos de plástico distintos.

Factor	Plástico	Temperatura
Nivel	A	Bajo (-20C)
Nivel	B	Medio (20C)
Nivel	C	Alto (60C)

En este experimento la resistencia de una pieza plástica es la variable de respuesta, la temperatura cambiante (aunque es una variable continua, sólo se usan tres temperaturas, por lo tanto es medida en forma categórica) y el tipo de plástico distinto (medida en forma categórica) son los factores; cada categoría de ambos factores son los tratamientos ó niveles de los factores.

En un **diseño de experimentos** existen tres principios básicos que son:

1. Reproducción.
2. Aleatorización.
3. Control Local.

1.- Reproducción.

La ***reproducción ó réplica*** se refiere a una repetición del experimento básico. Este concepto tiene dos propiedades importantes, la primera es que permite al experimentador obtener una estimación del **error experimental** y en segundo lugar, el uso de réplicas permite calcular **una estimación mas precisa del efecto de un factor** en el experimento si se usa la media de la muestra como una estimación de dicho efecto.

La **Unidad Experimental** es la unidad a la cual se le aplica un solo tratamiento (que puede ser una combinación de muchos factores) en una reproducción del experimento.

El **Error Experimental** es el que describe la situación de no llegar a resultados idénticos con dos unidades experimentales tratadas idénticamente y refleja:

- Errores de experimentación
- Errores de observación

ISBN 978-607-8285-62-4

- Errores de medición
- Variación del material experimental (esto es, entre unidades experimentales)
- Efectos combinados de factores extraños que pudieran influir las características en estudio, pero respecto a los cuales no se ha llamado la atención en la investigación.

El **error experimental** puede reducirse:

- Usando material experimental más homogéneo o por estratificación cuidadosa del material disponible.
- Utilizando información proporcionada por variables aleatorias relacionadas
- Teniendo más cuidado al dirigir y desarrollar el experimento
- Usando un diseño experimental muy eficiente.

La **Confusión** se presenta cuando dos o más efectos se confunden en un experimento si no es posible separar sus efectos, cuando se lleva a cabo el subsecuente análisis estadístico.

La asignación aleatoria es el uso de métodos aleatorios para designar pacientes a tratamientos distintos o viceversa.

2.- Aleatorización.

La ***aleatorización*** fundamenta el uso de los métodos estadísticos en el diseño de experimentos, entendiéndose por aleatorización al hecho de que tanto la asignación del material experimental como el orden en que se realizarán las pruebas individuales ó ensayos se determinen aleatoriamente. Al aleatorizar adecuadamente el experimento se ayuda a "cancelar" los efectos de factores extraños que pudieran estar presentes.

3.- Control local.

El ***control local ó análisis por bloques*** es una técnica que se utiliza para incrementar la precisión del experimento. Podemos decir que un bloque es una porción del material experimental más homogéneo que el total del material y al realizarse un análisis por bloques se hacen las comparaciones entre las condiciones de interés del experimento dentro de cada bloque.

En el **diseño de experimentos** se deben seguir los siguientes **pasos, etapas ó directrices**:

1.- Enunciado o planteamiento del problema.
2.- Elección de factores y niveles.
3.- Selección de la variable de respuesta.
4.- Formulación de hipótesis.
5.- Proposición de la técnica experimental y el diseño.

ISBN 978-607-8285-62-4

6.- Ejecución del experimento.
7.- Aplicación de las técnicas estadísticas a los resultados experimentales.
8.- Conclusiones con medidas de la confiabilidad de las estimaciones generadas y recomendaciones. Deberá darse cuidadosa consideración a la validez de las conclusiones para la población de objetos o eventos a la cual se van a aplicar.

Un **diseño experimental** sirve, generalmente, para **comparar** las medias de dos o más **tratamientos (niveles de factor)** a través del **análisis de varianza,** propuesto por **Ronald A. Fisher** a principios del Siglo XX, de los datos experimentales. Como es conocido, **un experimento** consiste en **una manipulación intencional y controlada de una o más variables** para **evaluar su (supuesto) efecto** en la variable **dependiente o variable-respuesta**.

1.1.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 1.1.1.1 DISEÑO DE EXPERIMENTOS

Comúnmente en ANOVA y diseño de experimentos, los investigadores seleccionan factores que varían sistemáticamente durante un experimento para determinar su efecto sobre la variable de respuesta. Los factores sólo pueden asumir un número limitado de valores posibles, conocidos como tratamientos ó niveles de factor.

Un ingeniero quiere estudiar la resistencia de una pieza plástica (Variable de respuesta) sometida a temperaturas cambiantes (Factor 2). La pieza puede ser elaborada con tres tipos de plástico distintos (niveles del Factor 1) . De ahí que se plantee las siguientes preguntas:

- ¿Qué efecto tienen la composición de la pieza (Factor 1) y la temperatura (Factor 2) en la resistencia de la pieza (Variable de respuesta)?
- ¿Existe algún material con el que la pieza resulte más resistente que con cualquiera de los otros dos independientemente de la temperatura?

Para darles respuesta, el ingeniero se plantea realizar una batería de experimentos. Cada uno de ellos consiste en tomar una pieza de un material dado, someterla a una temperatura prefijada y aplicarle una presión hasta que la pieza se quiebre. El grado de presión necesario será la medida de resistencia de la pieza.

Por fijar ideas, selecciona tres temperaturas, -20 °C, 20 °C y 60 °C (niveles ó tratamientos del Factor 2). Por lo tanto, puede realizar 9, es decir, 3x3, pruebas distintas. Además, decide repetir cada una de las 9 pruebas 4 veces cada una. Finalmente, decide aleatorizar las pruebas, es decir, desordenarlas aleatoriamente en el tiempo. Tras realizar los experimentos, obtiene 36, es decir, 4x9, medidas de resistencia distintas. A partir de ese momento, realiza un estudio cuantitativo utilizando técnicas estadísticas, como la ANOVA, que ya no forman parte propiamente de la fase del diseño experimental.

ISBN 978-607-8285-62-4

Los factores pueden ser una variable categórica o basarse en una variable continua, pero sólo se deben utilizar algunos valores controlados en el experimento.

Por ejemplo, Usted estudia los factores que podrían afectar la resistencia del plástico durante el proceso de elaboración. Usted decide incluir los dos factores siguientes en su experimento:

Factor	Plástico	Temperatura
Nivel	A	Bajo (-20C)
Nivel	B	Medio (20C)
Nivel	C	Alto (60C)

El tipo de plástico es una variable categórica. Puede ser sólo de tipo A , B o tipo C. Por su parte, la temperatura es una variable continua, pero aquí es un factor, porque sólo tres valores de temperatura de -20C, 20C y 60C se prueban en el experimento.

Cuestionario:

1) ¿ Se trata de un experimento uni o multifactorial?
   R: Se trata de un diseño multifactorial. Hay dos factores.
2) ¿Cuáles son los factores?
   R: El Factor 1 es el tipo de plástico. El Factor 2 es la temperatura.
3) ¿Cuáles son los tratamientos?
   R: Los tratamientos son los niveles de cada factor. El Factor 1 tiene tres tratamientos ó niveles que son los tipos de plástico. El Factor 2 tiene tres tratamientos ó niveles que son las tres temperaturas.
4) ¿Cuál es la variable de respuesta?
   R: La variable de respuesta es la resistencia de la pieza plástica.
5) ¿Cuál es la unidad experimental?
   R: Cada pieza sometida al experimento.
6) ¿Cuántas replicas haríamos?
   R: En este caso se decidió hacer 4 réplicas de cada combinación.
7) ¿ Cuántas medidas necesitamos?
   R: En este caso se necesitan 36 mediciones.
8) ¿Cómo haríamos el diseño?
   R: Se somete cada pieza de plástico a las distintas temperaturas y se mide su resistencia. Finalmente se aleatorizan cada una de las cuatro réplicas en cada combinación.

1.1.1.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 1.1.1.1 DISEÑO DE EXPERIMENTOS

Un investigador desea estudiar el tiempo de conexión en minutos que pasan los alumnos en una dirección de internet desde cuatro puntos geográficos de una región y en tres horas determinadas. Para cada punto geográfico y cada hora determinada se tomaron cuatro lecturas. De ahí que se plantee las siguientes preguntas:

1) ¿ Se trata de un experimento uni o multifactorial?
   Respuesta a la pregunta
   1._____
2) ¿Cuántos y qué factores está considerando el investigador?.
   Respuesta a la pregunta
   2._____
3) ¿Cuántos niveles ó tratamientos está considerando el investigador en cada factor?.
   Respuesta a la pregunta
   3._____

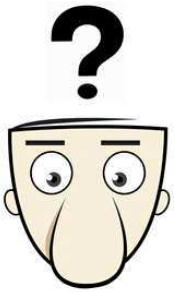
ISBN 978-607-8285-62-4

4) ¿Qué variable de interés eligió el investigador?.
   Respuesta a la pregunta
   4._____
5) ¿ Cuál es la Unidad experimental?
   Respuesta a la pregunta
   5._____
6) ¿ Cuántas réplicas haría el investigador?
   Respuesta a la pregunta
   6._____
7) ¿Cuántas lecturas distintas obtuvo el investigador?.
   Respuesta a la pregunta
   7._____
8) ¿Cómo haría el diseño el investigador?
   Respuesta a la pregunta
   8._____

1.1.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 1.1.1.1
DISEÑO DE EXPERIMENTOS



La junta de educación de un estado desea estudiar las diferencias en el número de alumnos de las clases entre las escuelas primaria, secundaria y preparatoria, en varias ciudades. Se seleccionó una muestra aleatoria de tres ciudades. Se eligieron dos escuelas al mismo nivel dentro de cada ciudad y se registró el número de alumnos promedio de clase para la escuela con los resultados siguientes. De ahí que se plantee las siguientes preguntas:

1) ¿Cuántos y qué factores está considerando la junta de educación?.
   Respuesta a la pregunta
   1._____
2) ¿Cuántos niveles ó tratamientos está considerando la junta de educación en cada factor?.
   Respuesta a la pregunta
   2._____
3) ¿Qué variable de interés eligió la junta de educación?.
   Respuesta a la pregunta
   3._____

ISBN 978-607-8285-62-4

4) ¿Cuántas lecturas distintas obtuvo la junta de educación?.
Respuesta a la pregunta
4._____

5) ¿Qué técnica estadística se presume que utilice la junta de educación?
Respuesta a la pregunta
5._____

1.1.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 1.1.1 DISEÑO DE EXPERIMENTOS

1.1.1.1 En la ciudad de Villagrande, una cadena de comida rápida está adquiriendo una mala reputación debido a que tardan mucho en servirle a los clientes. Como la cadena tiene cuatro restaurantes en esa ciudad, se tiene la preocupación de si los cuatro restaurantes tienen el mismo tiempo promedio de servicio. Uno de los dueños de la cadena ha decidido visitar cada uno de los locales y registrar el tiempo de servicio para cinco clientes escogidos al azar.

1) ¿ Se trata de un experimento uni o multifactorial?.
2) ¿Cuáles son los factores?.
3) ¿Cuáles son los tratamientos?.
4) ¿Cuál es la variable de respuesta?.
5) ¿Cuál es la unidad experimental?.
6) ¿Cuántas replicas haríamos?.
7) ¿Cuantas medidas necesitamos?.
8) ¿Cómo haríamos el diseño?.

1.1.1.2 En una investigación se evaluaron diferentes campañas publicitarias sobre las ventas con tres técnicas diversas de promoción: descuentos, precios bajos y regalos así como con tres distintos personajes para un mismo producto. Además, decide repetir 4 veces cada prueba para cada combinación.

1) ¿Se trata de un experimento uni o multifactorial?

ISBN 978-607-8285-62-4

2) ¿Cuáles son los factores?
3) ¿Cuáles son los tratamientos?
4) ¿Cuál es la variable de respuesta?
5) ¿Cuál es la unidad experimental?
6) ¿Cuántas replicas haríamos?
7) ¿Cuantas medidas necesitamos?
8) ¿Cómo haríamos el diseño?

1.1.1.3 En una investigación se analizó el efecto de tres tipos de descuento sobre las ventas, se tomó en cuenta que dos de las ubicaciones de los anuncios de los descuentos en la tienda podrían afectar las ventas. Se decidió repetir tres veces cada prueba en cada combinación.

1) ¿Se trata de un experimento uni o multifactorial?
2) ¿Cuáles son los factores?
3) ¿Cuáles son los tratamientos?
4) ¿Cuál es la variable de respuesta?
5) ¿Cuál es la unidad experimental?
6) ¿Cuántas replicas haríamos?
7) ¿Cuantas medidas necesitamos?
8) ¿Cómo haríamos el diseño?

ISBN 978-607-8285-62-4

OBJETIVO 1.2. El alumno aplicará el diseño completamente aleatorizado y el análisis de varianza para realizar pruebas de hipótesis entre *k* medias de tratamiento. Utilizará el método *T* de Tukey de comparación múltiples para determinar cuales de las *k* medias son significativamente diferentes entre sí

ANTECEDENTES

CONCEPTOS DE:

Experimento. Unidad experimental. Medidas. Variable de respuesta. Ensayos ó réplicas. Aleatorización. Agrupamiento. Bloqueo. Balanceo. Factores controlados. Factores no controlados. Tratamientos ó niveles de un factor. Error experimental. Efectos del tratamiento. Variación total. Variación entre tratamientos. Variación dentro de tratamientos. Análisis de varianza (ANOVA).

1.2.1 ELEMENTOS Y SUPUESTOS DEL DISEÑO COMPLETAMENTE ALEATORIZADO. ANÁLISIS DE VARIANZA.

CONCEPTOS BÁSICOS ANÁLISIS DE VARIANZA

El Análisis de varianza (ANOVA) prueba la hipótesis de que las medias de dos o más poblaciones son iguales. Los ANOVA evalúan la importancia de uno o más

En la **estructura de un diseño de experimentos** se deben considerar los siguientes **elementos**:

1.- El conjunto de **tratamientos** incluidos en el estudio.
2.- El conjunto de **unidades experimentales** utilizadas en el estudio.
3.- Las reglas y procedimientos por los cuales los **tratamientos son asignados a las unidades experimentales** (o viceversa).
4.- Las **medidas** o evaluaciones que se hacen a las **unidades experimentales** luego de aplicar los **tratamientos**.

En un **diseño de experimentos ó ANOVA** de **un factor ó una vía** existen los siguientes **supuestos básicos**:

1.- Cada una de las observaciones de la variable dependiente son independientes de las demás.
2.- Las fuentes de variación en el experimento deberán permanecer constantes o ser iguales.
3.- Los datos deben distribuirse normalmente.
4.- Se supone que un modelo aditivo, es el que mejor explica el comportamiento

ISBN 978-607-8285-62-4

factores al comparar las medias de la variable de respuesta en los diferentes niveles de factores. La hipótesis nula establece que todas las medias de la población (medias de los niveles de factores) son iguales mientras que la hipótesis alterna establece que al menos una es diferente.

Para ejecutar un ANOVA, debe tener una variable de respuesta continua y al menos un factor categórico con dos o más niveles. Los ANOVA requieren datos de poblaciones normalmente distribuidas con varianzas aproximadamente iguales entre los niveles de factores.

de la variable dependiente.
Si se tiene cuidado en el aspecto de **manejar un mismo número de datos** para cada uno de los niveles del factor, los supuestos anteriormente mencionados, podemos considerar que se cumplen a pesar de no haberse verificado. Al hecho de que se maneje **el mismo número de datos en cada uno de los niveles del factor se le llama *caso balanceado.***

Cuando las **medidas numéricas en *k* grupos o niveles son continuas** y se cumplen ciertas **suposiciones** se emplea una metodología conocida como **ANÁLISIS DE VARIANCIA (ANOVA)** para **comparar los valores medios de los grupos o tratamientos**. Aunque el término **"análisis de variancia"** podría parecer un nombre equivocado dado que el objetivo es analizar las diferencias entre las medias de los grupos, mediante un análisis de la variación de los datos, tanto ***entre*** como ***dentro*** de los ***k*** tratamientos, es viable derivar conclusiones sobre posibles diferencias en las medias de los grupos. En el método de **ANOVA (ANDEVA),** se subdivide la variación total en las mediciones de los resultados en lo que se puede atribuir a diferencias ***entre*** los ***k*** grupos y los que se deben al azar o que es atribuible a la variación inherente ***dentro*** de los ***k*** grupos. La variación "dentro del grupo" se considera un **error experimental;** mientras que la variación "entre grupos" se atribuye a los **efectos del tratamiento.**

1.2.2 EL ANÁLISIS DE VARIANZA DEL DISEÑO UNIFACTORIAL Ó DE UNA VÍA COMPLETAMENTE ALEATORIZADO. COMPARACIONES MÚLTIPLES. EL MÉTODO T DE TUKEY.

CONCEPTOS BÁSICOS DISEÑO DE UN FACTOR

El nombre "análisis de la varianza" se basa en la manera en la cual el procedimiento utiliza las varianzas para determinar si las medias son diferentes. El procedimiento funciona comparando la varianza

En la técnica del **Análisis de Varianza (ANOVA) del diseño unifactorial ó de una vía** existe una sola variable no métrica como **factor**, la cual se analiza para conocer su efecto sobre una **variable dependiente (métrica).**

Por medio de esta técnica se determina la relación que existe entre la **variable independiente** y la **dependiente,** lo cual se hará analizando si existe una variación o desviación significativa entre la variable dependiente de los diferentes niveles del factor.

Al hablar de ***variación*** , nos referimos a las diferencias entre los datos observados y los promedios de los mismos datos, por lo que se pueden considerar tres tipos de variaciones en este diseño:

1.- VARIACIÓN TOTAL. Suma de las diferencias elevadas al cuadrado entre cada observación y la media total **(SCT).**

2.- VARIACIÓN DE TRATAMIENTO Ó ENTRE TRATAMIENTOS. Suma de las diferencias elevadas al cuadrado entre la media de cada tratamiento y la media total o general (**$SC_{tratamientos}$**).

ISBN 978-607-8285-62-4

entre las medias de los tratamientos ó niveles de un factor y la varianza dentro de los tratamientos como un método para determinar si los grupos son todo parte de una población más grande o poblaciones separadas con características diferentes.

Por ejemplo, Usted diseña un experimento para evaluar la durabilidad de cuatro productos de fibra esponja experimental para lavar utensilios. Usted coloca una muestra de cada tipo de fibra esponja en diez hogares y mide la durabilidad después de 60 días. Debido a que está examinando un factor (tipo de fibra esponja), usted utiliza un ANOVA de un solo factor.

3.- VARIACIÓN ALEATORIA O DENTRO DE TRATAMIENTOS. Suma de las diferencias elevadas al cuadrado entre las observaciones y sus medias de tratamiento (**SCE)**.

La siguiente tabla representa la **matriz de las observaciones** al efectuar el experimento:

Tratamientos o niveles del Factor 1	**Observaciones**						
	1	**2**	**j**	**...**	**N**	**Total**	**Media**
1	X_{11}	X_{12}	X_{1j}		X_{1n}	$X_{1.}$	$\overline{X}_{1.}$
2	X_{21}	X_{22}	X_{2j}		X_{2n}	$X_{2.}$	$\overline{X}_{2.}$
i			X_{ij}		X_{in}	$X_{i.}$	$\overline{X}_{i.}$
...							
K	X_{k1}	X_{k2}			X_{kn}	$X_{k.}$	$\overline{X}_{k.}$
						$X_{..}$	$\overline{X}_{..}$

En el caso de un **Único Factor (Experimento Unifactorial ó de una vía)** el modelo de análisis de la varianza es:

$$\boldsymbol{X_{ij} = \mu + \tau_i + \varepsilon_{ij} \begin{cases} i = 1, 2, \dots, k \\ j = 1, 2, \dots, n \end{cases}}$$

Donde:

X_{ij} = observación j de la variable dependiente bajo los efectos del nivel i del factor manejado en el experimento.
μ= promedio general de la variable dependiente.
τ_i= efecto del nivel i del factor manejado en el experimento
ε_{ij} = error aleatorio de cada una de las observaciones de la variable dependiente. Es la cantidad de variación no explicada por el Factor, también se conoce como Error del experimento ó variación residual.

Como los efectos de los tratamientos se consideran como desviaciones de la media general por lo tanto:

$$\sum_{1=1}^{k} \tau_i = 0$$

El análisis de varianza consiste en descomponer o subdividir la suma de cuadrados total de la siguiente manera:

$$\mathbf{SCT = SC_{tratamiento} + SCE}$$

La prueba de hipótesis es un procedimiento que

ISBN 978-607-8285-62-4

evalúa dos enunciados mutuamente excluyentes sobre una población. Una prueba de hipótesis utiliza datos de muestra para determinar a cuál enunciado respaldan mejor los datos. Estos dos enunciados se denominan hipótesis nula e hipótesis alternativa. Siempre son enunciados sobre los atributos de las poblaciones, tales como el valor de un parámetro, la diferencia entre parámetros correspondientes de múltiples poblaciones o el tipo de distribución que describe mejor a la población.

Alfa(α) es utilizado en pruebas de hipótesis y es el máximo nivel de riesgo aceptable para rechazar una hipótesis nula verdadera (error de tipo I) y se expresa como una probabilidad cuyos valores se encuentran entre 0 y 1. Alfa con frecuencia es denominado nivel de significancia. Debe establecerse antes de comenzar el análisis y una vez realizada la prueba comparar los valores críticos de la región de rechazo con Alfa (α) para determinar la significancia de la prueba. Si se concluye que al menos una media es diferente y para explorar más las diferencias entre las medias específicas, se puede utilizar un método de comparación múltiple como el método *T* de Tukey

La suma de cuadrados es la cantidad calculada en el análisis de varianza y usada para obtener cuadrados medios para la prueba ***F***.

Cuando se desea probar la **igualdad** de las **medias** de los **niveles o tratamientos de un solo factor**, el juego de hipótesis es:

$$\boldsymbol{H_0: \mu_1 = \mu_2 = \cdots = \mu_k}$$

$$\boldsymbol{H_1: al\ menos\ una\ \mu_k\ es\ diferente}$$

Con el fin de determinar si las medias de los diversos tratamientos son todas iguales, se pueden examinar dos estimadores diferentes de la varianza de la población. Uno de los estimadores se basa en **la suma de los cuadrados dentro de los tratamientos *(SCt);*** el otro se basa en **la suma de los cuadrados entre los tratamientos *(SCE*)**. Si la hipótesis nula es cierta, estos estimadores deben ser aproximadamente iguales; si es falsa, el estimador basado en la suma de los cuadrados entre grupos debe ser mayor.

En el Análisis de Varianza, el estimador de la varianza entre los tratamientos ***(CMt)*** se calcula dividiendo la suma de los cuadrados de los tratamientos entre los grados de libertad entre los tratamientos ***(k-1).*** La varianza dentro de los tratamientos, ***(CME),*** se estima dividiendo la suma de los cuadrados dentro de los tratamientos entre los grados de libertad dentro de los tratamientos ***(N-k).*** Si en realidad hay una diferencia entre los tratamientos, el ***(CMt),*** será significativamente ***mayor*** que el ***(CME).*** La prueba estadística se basa en la razón de las dos varianzas, ***CMt/CME***. La distribución de esta razón se conoce como la **distribución *F*,** por lo que el estadístico de prueba es:

$$F_{CALC.} = \frac{CM_{tratamiento}}{CME} = \frac{SC_{tratamientos}/_{g.l.}}{SCE/_{g.l.}}$$

La regla de decisión es rechazar la hipótesis nula de que no hay diferencia entre los tratamientos si al nivel de significancia α

$$\boldsymbol{F_{calc} \gg F_{\alpha,(k-1),(N-k)}}$$

Para obtener la **Suma de Cuadrados** en un **diseño balanceado** se usan las siguientes fórmulas:

$$SCT = \sum_{I=1}^{k}\sum_{j=1}^{n} X_{ij} - \bar{X}..)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}^2 - \frac{X_{..}^2}{N}$$

$$SC_{Tratamientos} = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_i - \bar{X}_{..})^2 = n\sum_{i=1}^{k}(X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k}\frac{X_{i.}^2}{n} - \frac{X_{..}^2}{N}$$

$$SCE = SCT - SC_{Tratamiento}$$

ISBN 978-607-8285-62-4

Como verificación:

$$SCE = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$$

$$CM_{trat} = \frac{SC_{trat}}{k-1} \qquad CME = \frac{SCE}{N-k}$$

Para obtener la **Suma de Cuadrados** en un **diseño desbalanceado** se usan las siguientes fórmulas:

$$SCT = \sum_{I=1}^{k}\sum_{j=1}^{n_i} X_{ij} - \bar{X}..)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{X_{..}^2}{N}$$

$$SC_{Tratamientos} = \sum_{i=1}^{k}\sum_{j=1}^{ni}(X_i - \bar{X}_{..})^2 = n_i\sum_{i=1}^{k}(X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k}\frac{X_{i.}^2}{n_i} - \frac{X_{..}^2}{N}$$

$$SCE = SCT - SC_{Tratamiento}$$

Como verificación:

$$SCE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

Nota importante: El obtener la ***SCE*** por diferencias puede dar lugar a tener un error si en los cálculos anteriores para obtener la ***SCT*** ó la ***SCtratamientos*** existe algún error, por lo que se recomienda obtener por separado la ***SCE.***

$$CM_{trat} = \frac{SC_{trat}}{k-1} \qquad CME = \frac{SCE}{N-k}$$

El cuadrado medio es el cociente entre la suma de cuadrados y los grados de libertad.

Debido a que en el cálculo de varianzas entre y dentro de tratamientos hay varios pasos, el grupo completo de resultados se puede organizar en una tabla de análisis de varianza ***(ANOVA***) cuya estructura es la siguiente:

ISBN 978-607-8285-62-4

El cuadrado medio dentro del grupo o tratamientos es la estimación de la variación en el análisis de varianza. Se usa en el denominador de la prueba estadística *F*.

El cuadrado medio entre grupos o denominado error es la estimación de la variación en el análisis de varianza. Se usa en el numerador de la estadística *F*.

TABLA DE ANOVA:

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	$F_{calculada}$
Factor 1 Tratamientos (entre tratamientos)	$k-1$	SC_{trat}	$CM_{t=SC_t/g.l.}$	$F_{calc} = CM_t/CME$
Error (dentro de tratamientos)	$N-k$	SCE	$CME = SCE/g.l.$	
Total	$N-1$	SCT		

El método *T* de Tukey se utiliza en el análisis ANOVA para construir intervalos de confianza para todas las diferencias en parejas entre medias de los niveles de factor mientras controla el nivel de significancia por familia en un nivel que usted especifique. Es importante considerar el nivel de significancia por familia al realizar múltiples comparaciones, porque las posibilidades de cometer un error de tipo I para una serie de comparaciones son mayores que el nivel de significancia para cualquier comparación individual. Para contrarrestar este nivel de significancia más alto, el método de Tukey ajusta el intervalo de confianza para cada intervalo individual (sobre todo en el caso de diseños desbalanceados), de manera que el nivel de confianza simultáneo resultante sea igual al valor que usted especifique.

COMPARACIONES MÚLTIPLES: EL MÉTODO T DE TUKEY

Con la finalidad de determinar **cuáles de las *k* medias son significativamente diferentes** de las otras podemos utilizar el procedimiento de **Tukey**. Este método es un ejemplo de un procedimiento de comparación ***post hoc*** (o ***a posteriori),*** pues las hipótesis de interés son formuladas ***después*** de que los datos han sido inspeccionados.

Para usar el procedimiento de **Tukey,** simplemente se ordenan en forma descendente las medias de los tratamientos y se comparan las diferencias observadas entre cada par de promedios con el valor correspondiente al **rango ó alcance crítico**. Si $|\overline{X}_{i.} - \overline{X}_{i.'}| \geq$ **rango ó alcance crítico**, se concluye que las medias poblacionales $\mu_{i.}$ y $\mu_{j.}$ son diferentes. El **rango ó alcance crítico** se obtiene entonces de la cantidad dada en la ecuación siguiente:

$$\boldsymbol{rango} \text{ ó } \boldsymbol{alcance\ crítico} = q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}}$$

Si en uno ó más grupos hay tamaños de muestra desiguales, se reemplaza ***n*** de la ecuación anterior por la llamada media armónica:

$$n_h(\boldsymbol{media\ armónica}) = \frac{k}{\sum_{i=1}^{k}\frac{1}{n_i}}$$

Con el **método de Tukey** se puede establecer también un conjunto de intervalos de confianza estimados simultáneamente para las **verdaderas diferencias entre cada par de medias**. Lo anterior se logra sumando y restando el alcance o rango crítico a las diferencias en cada par de medias muestrales.

$$(\overline{X}_{i.} - \overline{X}'_{i.}) - q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}} \leq (\mu_{i.} - \mu'_{i.}) \leq (\overline{X}_{i.} - \overline{X}'_{i.}) + q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}}$$

ISBN 978-607-8285-62-4

Si en uno ó más grupos hay tamaños de muestra desiguales, se reemplaza ***n*** de la ecuación anterior por la llamada media armónica para cada comparación por parejas de las medias de muestra:

$$n_h(media\ arm\acute{o}nica) = \frac{k(grupos\ comparados)}{\sum_{i=1}^{k} \frac{1}{n_{i(grupos\ comparados)}}}$$

El valor $q_{\alpha(k,N-k)}$ se obtiene de la tabla de puntos porcentuales del rango studentizado del apéndice buscando en $\alpha = 0.05$ ó 0.01 $seg\acute{u}n\ se\ indique\ en\ el\ problema$, ***k=*** Número de grupos ó tratamientos en general y ***g.l.= N-k*** (Número total de observaciones menos el número de grupos). Si en la tabla no hay ninguna entrada que corresponda exactamente a los grados de libertad especificados se puede tomar el más cercano al especificado o hacer una interpolación con los valores que se encuentren con los grados de libertad entre los cuales se encuentre el especificado.

1.2.2.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 1.2.2.1 DISEÑO DE UN FACTOR BALANCEADO

Un experimento se lleva a cabo para investigar la posible influencia de la altura en que se muestra un producto y su efecto sobre las ventas (en miles de pesos). Para este experimento fueron manejados tres niveles de altura; inferior, medio y parte superior del estante. Se tomaron lecturas durante 8 días consecutivos y los resultados fueron los siguientes:

Altura del estante	**Observaciones ó repeticiones (días)**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Inferior	77	82	86	78	81	86	77	81
Media	88	94	93	90	91	94	90	87
Superior	85	85	87	81	80	79	87	93

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre los valores promedios de ventas para las tres alturas en que se muestra un producto.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿En cual o cuales alturas en las que se muestra un producto se vende más y cuanto más?.

ISBN 978-607-8285-62-4

Prueba de hipótesis para k medias

Solución al inciso a.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

La hipótesis nula es que las ventas medias son las mismas para los tres niveles de altura del estante donde se muestra un producto.

$$H_0:\ \mu_{1.} = \mu_{2.} = \ \mu_{3.}$$

La hipótesis alternativa es que las ventas promedio no son iguales para los tres niveles de altura del estante donde se muestra un producto.

$$H_1: \text{No todas las ventas promedio son iguales}$$

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución ***F***

Una razón *F* es aquella que se utiliza en el análisis de varianza, entre otras pruebas, para comparar la magnitud de dos estimaciones de la varianza de la población y determina si ambas estimaciones son aproximadamente iguales; en el análisis de varianza, se emplea la razón de la varianza entre tratamientos con la varianza dentro de los tratamientos.

$$F_{Calculada} = \frac{CMt}{CME} = \frac{SCt/_{g.l.}}{SCE/_{g.l.}}$$

Donde:

$$SCT = SCt + SCE$$

Empiece por calcular las sumas y medias para cada tratamiento así como la suma y media total:

Altura del estante	Observaciones o repeticiones (días)								Total filas	Medias filas
	1	2	3	4	5	6	7	8		
Inferior	$X_{11} = 77$	$X_{12} = 82$	$X_{13} = 86$	$X_{14} = 78$	$X_{15} = 81$	$X_{16} = 86$	$X_{17} = 77$	$X_{18} = 81$	$X_{1.} = 648$	$\bar{X}_{1.} = 81.000$
Media	$X_{21} = 88$	$X_{22} = 94$	$X_{23} = 93$	$X_{24} = 90$	$X_{25} = 91$	$X_{26} = 94$	$X_{27} = 90$	$X_{28} = 87$	$X_{2.} = 727$	$\bar{X}_{2.} = 90.875$
Superior	$X_{31} = 85$	$X_{32} = 85$	$X_{33} = 87$	$X_{34} = 81$	$X_{35} = 80$	$X_{36} = 79$	$X_{37} = 87$	$X_{38} = 93$	$X_{3.} = 677$	$\bar{X}_{3.} = 84.625$
							Totales de filas:		$X_{..} = 2052$	$\bar{X}_{..} = 85.5$

ISBN 978-607-8285-62-4

Si utilizamos la fórmula abreviada, primero debe sumar el cuadrado de totas las observaciones y al final restar el promedio del cuadrado de la suma total de la siguiente manera:

$$SCT = \sum_{I=1}^{3}\sum_{J=1}^{8} X_{ij}^2 - \frac{X_{..}^2}{N} = (77^2 + 82^2 + \cdots + 93^2) - \frac{2052^2}{24} = 176{,}134 - 175{,}446 = \mathbf{688}$$

Posteriormente determine la SCt ó la suma de los cuadrados de los errores debido a los tratamientos. Ésta es la suma de las diferencias al cuadrado que existen entre cada media de tratamiento $(\bar{X}i.)$ y la media total $(\bar{X}..)$.

Si usamos la fórmula abreviada primero debe sumar el cuadrado de cada total de tratamiento entre el tamaño de la muestra correspondiente a dicho tratamiento y al final restar el promedio del cuadrado de la suma total de la siguiente manera:

$$SCt = \sum_{i=1}^{3} \frac{X_{i.}^2}{n_i} - \frac{X_{..}^2}{N} = \left(\frac{648^2}{8} + \frac{727^2}{8} + \frac{677^2}{8}\right) - \frac{2052^2}{24} = \mathbf{399.25}$$

Para calcular el término SCE, encuentre la desviación que existe entre cada observación y su media de tratamiento.

Si usamos la fórmula abreviada, primero debe sumar el cuadrado de todas las observaciones y al final reste la suma del cuadrado de cada total de tratamiento entre el tamaño de la muestra correspondiente a dicho tratamiento de la siguiente manera:

$$SCE = \sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}^2 - \sum_{i=1}^{k} \frac{X_{i.}^2}{n} = 77^2 + 82^2 + \cdots + 93^2 - (\frac{648^2}{8} + \frac{727^2}{8} + \frac{677^2}{8})$$
$$= 176{,}134 - 175{,}845.25 = \mathbf{288.75}$$

Para encontrar el valor calculado de ***F***, trabaje con la tabla de ANOVA. El término de cuadrado de la media es otra expresión que se utiliza para un cálculo de la varianza.

El cuadrado de la media para los tratamientos es ***SCt*** dividida entre sus grados de libertad. El resultado es el cuadrado de la media para los tratamientos y se escribe ***CMt.***

ISBN 978-607-8285-62-4

Una tabla de ANOVA es la tabla donde se recogen todos los datos necesarios para realizar el contraste en el análisis de varianza.

Tabla de ANOVA

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada}$***
Tratamientos	$v_1 = k - 1$ $= 3 - 1$ $= 2\ g.l.$	$SCt = 399.25$	$CMt = {SCt}/{g.l.}$ $= {399.25}/{2}$ $= 199.625$	$F_{Calc.} = {CMt}/{CM}$
Error	$v_2 = N - k$ $= 24 - 3$ $= 21\ g.l.$	$SCE = 288.75$	$CME = {SCE}/{g.l.}$ $= {288.75}/{21}$ $= 13.75$	$= {199.625}/{13.75}$ $= 14.52$
Total	$N - 1$ $= 24 - 1$ $= 23\ g.l.$	$SCT = 688$		

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

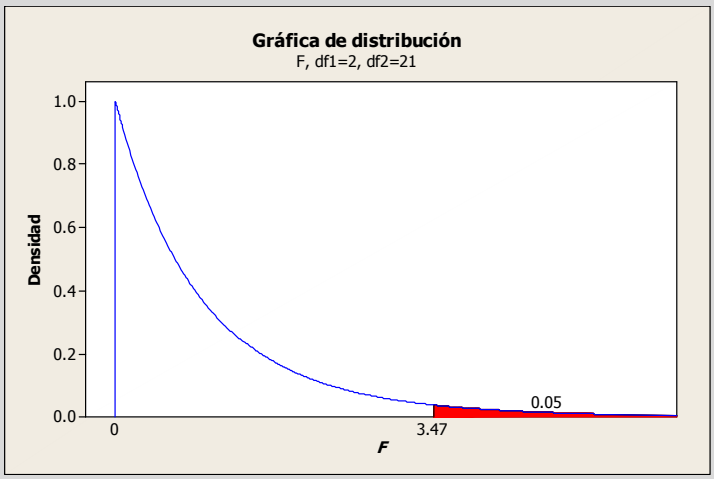
La región crítica es el rango (o conjunto de valores) donde debe ocurrir una prueba estadística para rechazar la hipótesis nula.

Para determinar la región de rechazo, se necesita el valor crítico. El valor crítico en el estadístico ***F*** se encuentra en las tablas ***F*** del apéndice (). Para utilizar esta tabla se necesita conocer los grados de libertad en el numerador y en el denominador. Los grados de libertad en el numerador son iguales al número de tratamientos, designados como k, menos 1. Los grados de libertad en el denominador son el número total de observaciones, N, menos el número de tratamientos. Para este problema existen 3 tratamientos y un total de 24 observaciones, por lo tanto los grados de libertad en el numerador son: k-1=3-1= 2 g.l. y los grados de libertad del denominador son: N-k=24-3=21 g.l.

Como existen tablas para niveles de Alfa diferentes, busque la que corresponda al nivel de significancia solicitada, en este caso 0.05, y desplácese horizontalmente sobre la parte superior de la página hasta llegar a los 2 grados de libertad del numerador. Luego descienda en esa columna hasta llegar a la fila que presenta 21 grados de libertad. El valor en esta intersección es **3.47** que en este caso es el valor crítico.

Grados de libertad en el denominador	**Grados de libertad en el numerador**			
	1	2	3	4
16	4.49	3.63	3.24	3.01
17	4.45	3.59	3.2	2.96
18	4.41	3.55	3.16	2.93
19	4.38	3.52	3.13	2.9
20	4.35	3.49	3.10	2.87
21	4.32	**3.47**	3.07	2.84
22	4.30	3.44	3.05	2.82
23	4.28	3.42	3.03	2.80

ISBN 978-607-8285-62-4



Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

La regla de decisión es rechazar H_0 si el valor calculado de ***F*** es mayor a 3.47.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como el valor calculado de ***F*** es de 14.52, que es mayor al valor crítico de 3.47; por lo tanto la hipótesis nula se rechaza y llegamos a la conclusión de que no todas las medias de la población son iguales, es decir al menos una de ellas es diferente.

Administrativa: Existe evidencia suficiente para concluir que estadísticamente el volumen de ventas promedio no es el mismo en al menos uno de los tres niveles de altura donde se muestra el producto.

NOTA: En este punto sólo podemos llegar a la conclusión de que existe una diferencia en al menos una de las medias de tratamiento. No podemos determinar qué grupo o grupos de tratamiento difieren ni por qué cantidad difieren unos de otros.

Prueba *T* de Tukey de comparaciones múltiples.

<u>Solución al inciso b.</u>

El método ***T* de Tukey** permite examinar, en forma simultánea, comparaciones entre todos los pares de grupos mediante los siguientes pasos:

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

$$\bar{x}_{2.} = 90.875; \;\; \bar{x}_{3.} = 84.625; \;\; \bar{x}_{1.} = 81$$

ISBN 978-607-8285-62-4

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y sus diferencias.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

$$\bar{x}_{2.} - \bar{x}_{3.} = 90.875 - 84.625 = 6.25$$
$$\bar{x}_{2.} - \bar{x}_{1.} = 90.875 - 81.000 = 9.875$$
$$\bar{x}_{3.} - \bar{x}_{1.} = 84.625 - 81.000 = 3.625$$

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$$Rango\ crítico = q_{0.05,3,21}\sqrt{\frac{13.75}{8}}$$

$$n_h(media\ armónica) = \frac{k}{\sum_{i=1}^{k}\frac{1}{n_i}} = \frac{3}{\frac{1}{8}+\frac{1}{8}+\frac{1}{8}} = 8$$

$$Rango\ crítico = 3.56\sqrt{\frac{13.75}{8}} = \mathbf{4.67}$$

Nota: el valor de q de **3.56** se obtuvo de la tabla de puntos porcentuales del rango studentizado con $\alpha = 0.05; k = 3\ y\ g.l. = 21$

Grados de libertad del error	K=número de niveles ó medias de tratamiento			
	2	**3**	4	5
18	2.97	3.61	4.00	4.28
19	2.96	3.59	3.98	4.25
20	2.95	3.58	3.96	4.23
21	2.94	**3.56**	4.21	4.42
22	2.93	3.55	3.93	4.20
23	2.93	3.54	3.91	4.18
24	2.92	3.53	3.90	4.17

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

$\bar{x}_{2.} - \bar{x}_{3.} = 90.875 - 84.625 = 6.25 > 4.677$; la prueba es (S) y $\mu_{2.} > \mu_{3.}$
$\bar{x}_{2.} - \bar{x}_{1.} = 90.875 - 81.000 = 9.875 > 4.677$; la prueba es (S) y $\mu_{2.} > \mu_{1.}$
$\bar{x}_{3.} - \bar{x}_{1.} = 84.625 - 81.000 = 3.625 < 4.677$; la prueba es (NS) y $\mu_{3.} = \mu_{1.}$

ISBN 978-607-8285-62-4

Paso 5. Construir la gráfica de medias.

Paso 5.- Construir la gráfica de medias[1]:

```
Nivel  N   Media    ----+---------+---------+---------+-----
1      8  81.000   (-----*------)
2      8  90.875                    (------*------)
3      8  84.625          (------*-----)
                      ----+---------+---------+---------+-----
                      80.0      84.0      88.0      92.0
```

Paso 6. Construir los intervalos de confianza de cada par de medias.

Paso 6.- Establecer el conjunto de intervalos de confianza

Un intervalo de confianza es un rango de valores, derivado de estadísticas de muestra, que probablemente incluya el valor de un parámetro desconocido de la población. Debido a su naturaleza aleatoria, es poco probable que dos muestras de una población dada generen intervalos de confianza idénticos. Sin embargo, si repitió muchas veces su muestra, un determinado porcentaje de los intervalos de confianza resultantes incluiría el parámetro desconocido de la población.

$$(\bar{X}_{i.} - \bar{X}_{i'}) - q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}} \leq (\mu_{i.} - \mu_{i'.}) \leq (\bar{X}_{i.} - \bar{X}_{i'}) + q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}}$$

$$(\bar{X}_{2.} - \bar{X}_{3.}) - 3.56\sqrt{\frac{13.75}{8}} \leq (\mu_{2.} - \mu_{3.}) \leq (\bar{X}_{2.} - \bar{X}_{3.}) - 3.56\sqrt{\frac{13.75}{8}}$$

$$n_{h(media\ armónica)} = \frac{k(grupos\ comparados)}{\sum_{i=1}^{k}\frac{1}{n_i(grupos\ comparados)}} = \frac{2}{\frac{1}{8}+\frac{1}{8}} = 8$$

$$6.25 - 3.56\sqrt{\frac{13.75}{8}} \leq (\mu_{2.} - \mu_{3.}) \leq 6.25 + 3.56\sqrt{\frac{13.75}{8}}$$

$$6.25 - 4.67 \leq (\mu_{2.} - \mu_{3.}) \leq 6.25 + 4.67$$

$$\mathbf{1.58} \leq (\mu_{2.} - \mu_{3.}) \leq \mathbf{10.92}$$

El porcentaje de estos intervalos de confianza que incluyen el parámetro es el nivel de confianza del intervalo. A las cotas de un intervalo, que se conocen como límites de confianza.

Conclusión: Como ambos límites del intervalo son positivos podemos decir que estadísticamente el volumen de ventas promedio, cuando el producto se exhibe en la altura media, es mayor que cuando el producto se exhibe en la altura superior por un mínimo de 1.58 (miles de $) y un máximo de 10.92 (miles de $).

$$(\bar{X}_{2.} - \bar{X}_{1.}) - 3.56\sqrt{\frac{13.75}{8}} \leq (\mu_{2.} - \mu_{1.}) \leq (\bar{X}_{2.} - \bar{X}_{1.}) - 3.56\sqrt{\frac{13.75}{8}}$$

[1] Obtenidas con el software estadístico Minitab 15

ISBN 978-607-8285-62-4

$$n_{h(media\ armónica)} = \frac{k(grupos\ comparados)}{\sum_{i=1}^{k} \frac{1}{n_i(grupos\ comparados)}} = \frac{2}{\frac{1}{8}+\frac{1}{8}} = 8$$

$$9.875 - 3.56\sqrt{\frac{13.75}{8}} \leq (\mu_{2.} - \mu_{1.}) \leq 9.875 + 3.56\sqrt{\frac{13.75}{8}}$$

$$9.875 - 4.67 \leq (\mu_{2.} - \mu_{1.}) \leq 9.875 + 4.67$$

$$\mathbf{5.20} \leq (\mu_{2.} - \mu_{1.}) \leq \mathbf{14.54}$$

Conclusión: Como ambos límites del intervalo son positivos podemos decir que estadísticamente el volumen de ventas promedio, cuando el producto se exhibe en la altura media, es mayor que cuando el producto se exhibe en la altura inferior, por un mínimo de 5.20 (miles de $) y un máximo de 14.54 (miles de $).

$$(\bar{X}_{3.} - \bar{X}_{1.}) - 3.56\sqrt{\frac{13.75}{8}} \leq (\mu_{3.} - \mu_{1.}) \leq (\bar{X}_{3.} - \bar{X}_{1}) - 3.56\sqrt{\frac{13.75}{8}}$$

$$n_{h(media\ armónica)} = \frac{k(grupos\ comparados)}{\sum_{i=1}^{k} \frac{1}{n_i(grupos\ comparados)}} = \frac{2}{\frac{1}{8}+\frac{1}{8}} = 8$$

$$3.625 - 3.56\sqrt{\frac{13.75}{8}} \leq (\mu_{3.} - \mu_{1.}) \leq 3.625 + 3.56\sqrt{\frac{13.75}{8}}$$

$$3.625 - 4.67 \leq (\mu_{3.} - \mu_{1.}) \leq 3.625 + 4.67$$

$$\mathbf{-1.04} \leq (\mu_{3.} - \mu_{1.}) \leq \mathbf{8.29}$$

Conclusión: Como el intervalo de confianza pasa por cero, podemos decir que estadísticamente el volumen de ventas promedio cuando el producto se exhibe en la altura superior ó inferior, es el mismo.

ISBN 978-607-8285-62-4

1.2.2.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 1.2.2.1 DISEÑO DE UN FACTOR BALANCEADO

Se recopiló la siguiente información de un experimento:

Tratamiento 1	Tratamiento 2	Tratamiento 3
8	3	3
6	2	4
10	4	5
9	3	4

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre los valores promedios de los tres tratamientos.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿En cual o cuáles tratamientos las medias resultaron mayores y por cuánto más?.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ANOVA de una vía ó unifactorial. Prueba de hipótesis de cinco pasos.

Solución al inciso a.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

Empiece por calcular las sumas y medias para cada tratamiento así como la suma y media total y llene la siguiente tabla:

Trat.	Observaciones o repeticiones				Total columnas	Medias columnas
	1	2	3	4		
1	$X_{11} =$	$X_{12} =$	$X_{13} =$	$X_{14} =$	$X_{1.} =$	$\bar{X}_{1.} =$
2	$X_{21} =$	$X_{22} =$	$X_{23} =$	$X_{24} =$	$X_{2.} =$	$\bar{X}_{2.} =$
3	$X_{31} =$	$X_{32} =$	$X_{33} =$	$X_{34} =$	$X_{3.} =$	$\bar{X}_{3.} =$
Totales					$X_{..} =$	$\bar{X}_{..} =$

Calcule SCT, SCt y la SCE

Suma de Cuadrados Total, Suma de Cuadrados de Tratamientos y Suma de Cuadrados del Error

$$\boldsymbol{SCT = \sum_{I=1}^{k}\sum_{J=1}^{n} X_{ij}^2 - \frac{X_{..}^2}{N} =}$$

$$\boldsymbol{SCt = \sum_{i=1}^{k} \frac{X_{i.}^2}{n_i} - \frac{X_{..}^2}{N} =}$$

$$\boldsymbol{SCE = \sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}^2 - \sum_{i=1}^{k} \frac{X_{i.}^2}{n} =}$$

Para encontrar el valor calculado de ***F***, trabaje con la tabla de **ANOVA**.

Tabla de ANOVA

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada}$***
Tratamientos	$v_1 = k - 1 =$	$SCt =$	$CMt = {SCt}/{g.l.} =$	$F_{Calc.} = {CMt}/{CME}$
Error	$v_2 = N - k =$	$SCE =$	$CME = {SCE}/{g.l.} =$	
Total	*N-1=*	$SCT =$		

ISBN 978-607-8285-62-4

Paso 3. Región de Rechazo	**Paso 3.-** Establecer la región de rechazo de (H_0).
Paso 4. Regla de decisión	**Paso 4.-** Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.
Paso 5. Conclusiones.	**Paso 5.-** Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa). **Estadística**: **Administrativa:**
Prueba post-hoc ó aposteriori. Comparaciones múltiples. Método de Tukey	**<u>Solución al inciso b.</u>**
Paso 1. Ordenar las medias en forma descendente.	**Paso 1. Ordenar las medias en forma descendente:**
Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.	**Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :**

ISBN 978-607-8285-62-4

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$$\boldsymbol{Rango\ ó\ alcance\ crítico = q_{\alpha,k,N-k}\sqrt{\frac{CME}{n_h}}}$$

$$\boldsymbol{n_h(media\ armónica) = \frac{k}{\sum_{i=1}^{k}\frac{1}{n_i}} =}$$

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

Paso 5. Construir la gráfica de medias.

Paso 5.- Construir la gráfica de medias:

Paso 6. Construir los intervalos de confianza de cada par de medias.

Paso 6.- Establecer el conjunto de intervalos de confianza:

$$\boldsymbol{(\bar{X}_{i.} - \bar{X}_{i'}) - q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}} \leq (\mu_{i.} - \mu_{i.)}') \leq (\bar{X}_{i.} - \bar{X}_{i'}) + q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}}}$$

$$\boldsymbol{n_{h(media\ armónica)} = \frac{k(grupos\ comparados)}{\sum_{i=1}^{k}\frac{1}{n_i(grupos\ comparados)}}}$$

Intervalos de confianza de Tukey al 95%

Conclusiones.

CONCLUSIONES:

ISBN 978-607-8285-62-4

1.2.2.2 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 1.2.2. 2 DISEÑO DE UN FACTOR BALANCEADO

Citrus Clean es un limpiador nuevo multiusos que se está probando colocando exhibidores en tres lugares diferentes dentro de varios supermercados. El número de botellas 12 onzas vendidas en cada lugar dentro del supermercado se reporta de la siguiente manera:

Lugar	No. De observación.			
	1	2	3	4
Cerca del pan (1)	18	14	19	17
Cerca de la cerveza (2)	12	18	10	16
Con otros limpiadores (3)	26	28	30	32

Los resultados del paquete de software de estadística son los siguientes:

```
Fuente  GL      SC      MC      F      P
Lugar    2  504.00  252.00  30.65  0.000
Error    9   74.00    8.22
Total   11  578.00

                              ICs de 95% individuales para la media
                              basados en Desv.Est. agrupada
Nivel  N   Media  Desv.Est.   --+---------+---------+---------+-------
1      4  17.000      2.160        (----*-----)
2      4  14.000      3.651   (----*-----)
3      4  29.000      2.582                            (----*-----)
                              --+---------+---------+---------+-------
                             12.0      18.0      24.0      30.0

Desv.Est. agrupada = 2.867


Intervalos de confianza simultáneos de Tukey del 95%
Todas las comparaciones de dos a dos entre los niveles de Lugar

Nivel de confianza individual = 97.91%


Lugar = 1 restado de:

Lugar  Inferior  Centro  Superior  ------+---------+---------+---------+--
2        -8.663  -3.000     2.663           (----*---)
3         6.337  12.000    17.663                          (----*----)
                                   ------+---------+---------+---------+--
                                        -12         0        12        24


Lugar = 2 restado de:

Lugar  Inferior  Centro  Superior  ------+---------+---------+---------+--
3         9.337  15.000    20.663                            (----*---)
                                   ------+---------+---------+---------+--
                                        -12         0        12        24
```


ISBN 978-607-8285-62-4

a) Utilice el proceso de prueba de hipótesis de cinco pasos con un nivel de significancia 0.05 para probar si existe alguna diferencia en el número medio de botellas vendidas en cada lugar dentro del supermercado.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿Cual o cuales de los tres lugares diferentes dentro de varios supermercados vendieron más botellas y cuantas más?.

Prueba dehipótesis

Solución al inciso a.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

ISBN 978-607-8285-62-4

Administrativa:

Prueba post-hoc ó aposteriori. Comparaciones múltiples de Tukey.

Solución al inciso b.

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$$Rango\ ó\ alcance\ crítico = q_{\alpha,k,N-k}\sqrt{\frac{CME}{n_h}}$$

$$n_h(media\ armónica) = \frac{k}{\sum_{i=1}^{k}\frac{1}{n_i}} =$$

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

Paso 5. Construir la gráfica de medias.

Paso 5.- Construir la gráfica de medias:

Paso 6. Construir los intervalos de confianza de cada par de medias.

Paso 6.- Establecer el conjunto de intervalos de confianza:

$$(\bar{X}_{i.} - \bar{X}_{i'}) - q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}} \leq (\mu_{i.} - \mu_{i.)}') \leq (\bar{X}_{i.} - \bar{X}_{i'}) + q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}}$$

$$n_{h(media\ armónica)} = \frac{k(grupos\ comparados)}{\sum_{i=1}^{k}\frac{1}{n_i(grupos\ comparados)}}$$

Conclusiones.

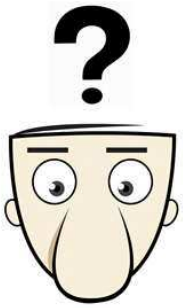
CONCLUSIONES:

ISBN 978-607-8285-62-4

1.2.2.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presentan dos ejercicios de autoevaluación los cuales ponen a prueba su comprensión del material anterior. Las respuestas a estos ejercicios de autoevaluación se encuentran al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlos y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 1.2.2.1
DISEÑO DE EXPERIMENTOS



La siguiente es una tabla de ANOVA parcial:

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrado Medio*	$F_{calculada}$
Tratamientos	2			
Error			20	
Total	11	500		

Complete la tabla y conteste las siguientes preguntas. Utilice un nivel de significancia 0.05.

a) ¿Cuántos tratamientos hay?.
Respuesta al inciso a. _____

b) ¿Cuál es el tamaño total de la muestra?.
Respuesta al inciso b. _____

c) ¿Cuál es el valor crítico de *F*?
Respuesta al inciso c. _____

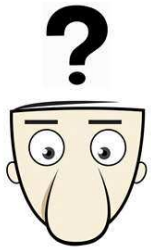
d) Formule las hipótesis nula y alternativa.
Respuesta al inciso d._____

e) ¿A qué conclusión llego en cuanto a la hipótesis nula?
Respuesta al inciso e._____

ISBN 978-607-8285-62-4

1.2.2.2 EJERCICIO DE AUTOEVALUACIÓN

AUTOEVALUACIÓN 1.2.2.2
DISEÑO DE UN FACTOR BALANCEADO



Una organización de consumidores quiere saber si existe alguna diferencia en el precio de un juguete en particular en tres tipos de tiendas diferentes. El precio del juguete se revisó en una muestra de cinco tiendas de descuento, cinco tiendas de artículos diversos y cinco tiendas departamentales.

Tienda	No. de observación				
	1	2	3	4	5
Descuento (1)	12	13	14	12	15
Variedad (2)	15	17	14	18	17
Departamental (3)	19	17	16	20	19

Los resultados de un paquete de software estadístico son los siguientes:

```
Fuente  GL     SC     MC      F      P
Tienda   2  63.33  31.67  13.38  0.001
Error   12  28.40   2.37
Total   14  91.73
                              ICs de 95% individuales para la media
                              basados en Desv.Est. agrupada
Nivel  N   Media  Desv.Est.    -+---------+---------+---------+--------
1      5  13.200      1.304    (------*------)
2      5  16.200      1.643                   (------*------)
3      5  18.200      1.643                             (------*------)
                              -+---------+---------+---------+--------
                             12.0      14.0      16.0      18.0

Desv.Est. agrupada = 1.538

Intervalos de confianza simultáneos de Tukey del 95%
Todas las comparaciones de dos a dos entre los niveles de Tienda

Nivel de confianza individual = 97.94%

Tienda = 1 restado de:

Tienda  Inferior  Centro  Superior --+---------+---------+---------+------
2          0.406   3.000     5.594                (-------*------)
3          2.406   5.000     7.594                       (------*-------)
                                  ---+---------+---------+---------+------
                                    -3.5       0.0       3.5       7.0
Tienda = 2 restado de:

Tienda  Inferior  Centro  Superior --+---------+---------+---------+------
3         -0.594   2.000     4.594             (-------*------)
                                  ---+---------+---------+---------+------
                                    -3.5       0.0       3.5       7.0
```

a) Utilice el proceso de prueba de hipótesis de cinco pasos con un nivel de significancia 0.05 para probar si existe alguna diferencia en el precio medio de un juguete.

b) Según el método *T* de Tukey de comparaciones múltiples, ¿Cual o cuales tiendas tienen el precio de un juguete en particular más alto y por cuanto más?.

ISBN 978-607-8285-62-4

1.2.2 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO
1.2.2
DISEÑO DE UN FACTOR BALANCEADO

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelvas los ejercicios en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. **Se sugiere utilizar aproximaciones de 5 dígitos**.

1.2.2.1 En un estudio se compararon los efectos de cuatro promociones mensuales sobre las ventas. A continuación presentamos las ventas unitarias de cinco tiendas que utilizaron las cuatro promociones en meses diferentes:

Muestra gratis	78	87	81	89	85
Regalo de un paquete	94	91	87	90	88
Descuento	73	78	69	83	76
Reembolso por correo	79	83	78	69	81

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre los valores promedios de ventas de las cuatro promociones mensuales.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor e interprete su resultados. ¿Cuál ó cuales promociones producen mayores ventas y por cuanto más?.

1.2.2.2 Se compararon tres métodos de entrenamiento para ver si conducen hacia una mayor productividad a los empleados que los cursan. Los datos que se presentan a continuación son medidas de la productividad de individuos entrenados por cada método.

Método 1	45	40	50	39	53	44
Método 2	69	53	57	61	49	59
Método 3	41	37	43	40	52	37

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre las medidas de productividad en los tres métodos utilizados.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor e interprete sus resultados. ¿Cuál ó cuáles métodos conducen a una mayor productividad y por cuanto más?.

ISBN 978-607-8285-62-4

1.2.2.3 En la ciudad de Villagrande, una cadena de comida rápida está adquiriendo una mala reputación debido a que tardan mucho en servirle a los clientes. Como la cadena tiene cuatro restaurantes en esa ciudad, se tiene la preocupación de si los cuatro restaurantes tienen el mismo tiempo promedio de servicio. Uno de los dueños de la cadena ha decidido visitar cada uno de los locales y registrar el tiempo de servicio para cinco clientes escogidos al azar. En sus cuatro visitas vespertinas registró los siguientes tiempos de servicio en minutos:

Restaurante1	3	4	5.5	3.5	4
Restaurante 2	5.3	5.5	6.5	6	7.5
Restaurante 3	6	7.5	9	10.5	10
Restaurante 4	3	4	5.5	2.5	3

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre los valores promedios de ventas para las tres alturas en que se muestra un producto.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor e interprete sus resultados. ¿En cual o cuáles de ellos fue menor el tiempo promedio de servicio y por cuánto menos?.

ISBN 978-607-8285-62-4

1.2.2.1 EJEMPLO ILUSTRATIVO EN EXCEL

EJEMPLO ILUSTRATIVO INTEGRAL EN EXCEL 1.2.2.1 DISEÑO DE UN FACTOR BALANCEADO

Prueba de hipótesis para k medias

Un experimento se lleva a cabo para investigar la posible influencia de la altura en que se muestra un producto y su efecto sobre las ventas (en miles de pesos). Para este experimento fueron manejados tres niveles de altura; inferior, medio y parte superior del estante. Se tomaron lecturas durante 8 días consecutivos y los resultados fueron los siguientes:

Altura del estante	**Observaciones ó repeticiones (días)**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Inferior	77	82	86	78	81	86	77	81
Media	88	94	93	90	91	94	90	87
Superior	85	85	87	81	80	79	87	93

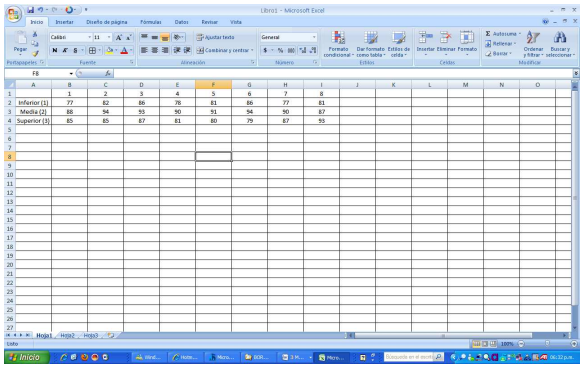
a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre los valores promedios de ventas para las tres alturas en que se muestra un producto.

<u>Solución al inciso a.</u>

Cuando el número de observaciones en cada tratamiento **es extenso y/o existen muchos tratamientos**, los cálculos manuales son tediosos. Existen muchos paquetes de software que pueden mostrar los resultados entre ellos **Excel.**

Comenzamos introduciendo los datos en **la hoja de Excel**, tal y como se muestra a continuación:

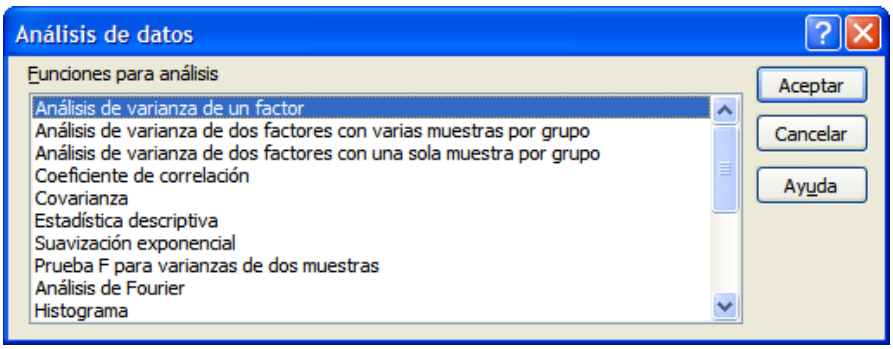
Hoja de Excel.



ISBN 978-607-8285-62-4

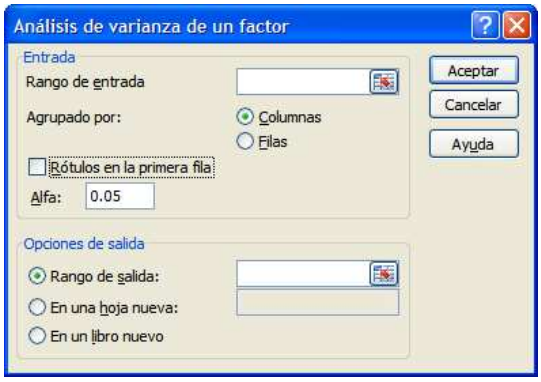
Como tenemos un ***modelo con un solo factor fijo*** seleccionamos la opción ***Análisis de datos*** del menú ***Datos***, utilizaremos la opción ***Análisis de la varianza de un factor***, del cuadro ***Análisis de datos*** de la figura siguiente:

Cuadro de diálogo: Análisis de datos.



En la lista ***Funciones para análisis***, elija la modalidad de ***Análisis de varianza de un factor*** y oprima el botón ***Aceptar*** para obtener el siguiente cuadro de dialogo rellenando su pantalla de entrada:

Cuadro de diálogo: Análisis de varianza de un factor.



En el cuadro ***Rango de entrada*** introduzca, (seleccionando con el cursor las celdas donde están los datos incluyendo los rótulos de la primera columna, **pero no los del primer renglón**), la referencia de celda correspondiente al rango de datos que está analizando. La referencia deberá contener dos o más rangos adyacentes organizados en columnas o filas.

En el campo ***Agrupado por*** haga clic en el botón ***Filas*** para indicar que los datos del rango de entrada están organizados en filas (es posible también organizarlos por columnas si lo desea). Si la primera columna del rango de entrada contiene rótulos, active la casilla de verificación ***Rótulos en la primera columna.*** Esta casilla de verificación debe quedar desactivada si el rango de entrada carece de rótulos; Microsoft Office Excel 2007 generará los rótulos de datos correspondientes para la tabla de resultados. Deje sin cambio el campo Alfa con el valor de 0.05 (nivel con el que desee evaluar los valores críticos de la función estadística *F)*. El nivel ***alfa*** es un nivel de importancia relacionado con la probabilidad de que haya un error de tipo I (rechazar una hipótesis verdadera).

ISBN 978-607-8285-62-4

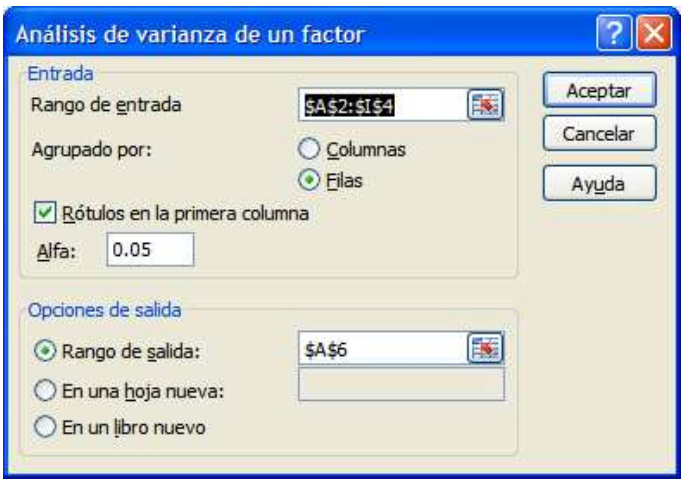
Cuadro de diálogo: Análisis de varianza de un factor.

Alfa(a) es utilizado en pruebas de hipótesis. Alfa (a) es el máximo nivel de riesgo aceptable para rechazar una hipótesis nula verdadera (error de tipo I) y se expresa como una probabilidad cuyos valores se encuentran entre 0 y 1. Alfa con frecuencia es denominado nivel de significancia. Debe establecerse antes de comenzar el análisis y una vez realizada la prueba comparar los valores *p* con Alfa (a) para determinar la significancia utilizando el siguiente criterio:

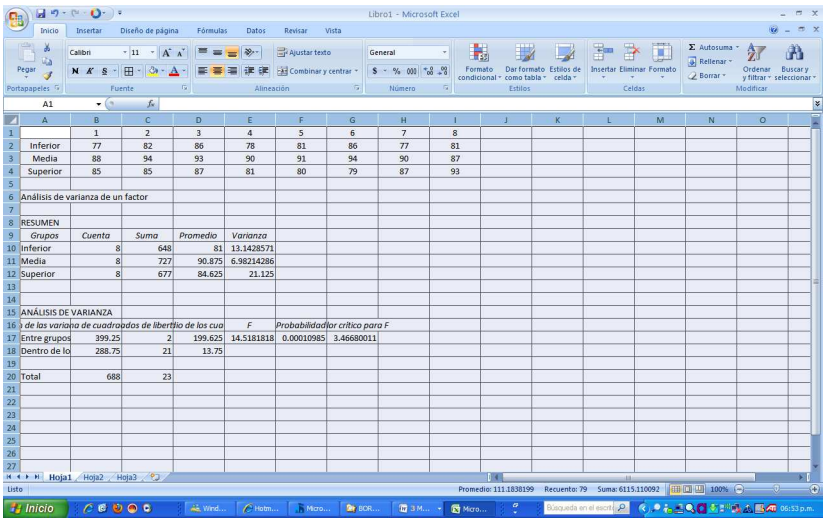
- Si el *valor p* es menor que o igual al nivel a, rechace la hipótesis nula en favor de la hipótesis alternativa.
- Si el *valor p* es mayor que el nivel Alfa (a), no rechace la hipótesis nula.

Los niveles de significancia más utilizado son 0.05 y 0.01. En estos niveles, sus posibilidades de encontrar un efecto que realmente no existe es de sólo 5% y 1%. Mientras menor sea el valor Alfa (a) , menores serán la probabilidades de que rechace de manera incorrecta la hipótesis nula. Sin embargo, un valor menor para Alfa (a) también significa una posibilidad reducida de detectar un efecto si verdaderamente existe un error (menor potencia).

En cuanto a las ***opciones de salida***, en el campo ***Rango de salida*** introduzca la referencia, (dando un clic), correspondiente a la celda superior izquierda de la tabla de resultados, en este caso la celda A6 y oprima el botón ***Aceptar***.



A continuación se muestra la salida del análisis de la varianza de un solo factor:



	A	B	C	D	E	F	G	H	I
1		1	2	3	4	5	6	7	8
2	Inferior	77	82	86	78	81	86	77	81
3	Media	88	94	93	90	91	94	90	87
4	Superior	85	85	87	81	80	79	87	93
5									
6	Análisis de varianza de un factor								
7									
8	RESUMEN								
9	*Grupos*	*Cuenta*	*Suma*	*Promedio*	*Varianza*				
10	Inferior	8	648	81	13.1428571				
11	Media	8	727	90.875	6.98214286				
12	Superior	8	677	84.625	21.125				
13									
14									
15	ANÁLISIS DE VARIANZA								
16	*de las variaiva de cuadrapdos de libertlio de los cua*				*F*	*Probabilidadlor crítico para F*			
17	Entre grupos	399.25	2	199.625	14.5181818	0.00010985	3.46680011		
18	Dentro de lo	288.75	21	13.75					
19									
20	Total	688	23						

Observe que Excel utiliza el término "**Entre grupos"** para **"Tratamientos"** y **"Dentro de los grupos"** para **" Error".** Sin embargo, tienen los mismos significados.

Conclusión: Como el ***p-valor*** del test ***F*** de Fisher es **menor que 0.05**, existen **diferencias significativas** entre las ventas de las diferentes alturas del estante donde se exhibe el producto al **95% de confianza**

NOTA: Excel en este caso no tiene opción para realizar pruebas Pos-hoc ó Aposteriori como la prueba de Tukey.

ISBN 978-607-8285-62-4

1.2.2.1 EJEMPLO ILUSTRATIVO EN MINITAB 15

EJEMPLO ILUSTRATIVO INTEGRAL EN MINITAB 1.2.2.1 DISEÑO DE UN FACTOR BALANCEADO

Prueba de hipótesis

Un experimento se lleva a cabo para investigar la posible influencia de la altura en que se muestra un producto y su efecto sobre las ventas (en miles de pesos). Para este experimento fueron manejados tres niveles de altura; inferior, medio y parte superior del estante. Se tomaron lecturas durante 8 días consecutivos y los resultados fueron los siguientes:

Altura del estante	**Observaciones ó repeticiones (días)**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Inferior	77	82	86	78	81	86	77	81
Media	88	94	93	90	91	94	90	87
Superior	85	85	87	81	80	79	87	93

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre los valores promedios de ventas para las tres alturas en que se muestra un producto.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿En cual o cuales alturas en las que se muestra un producto se vende más y cuanto más?.

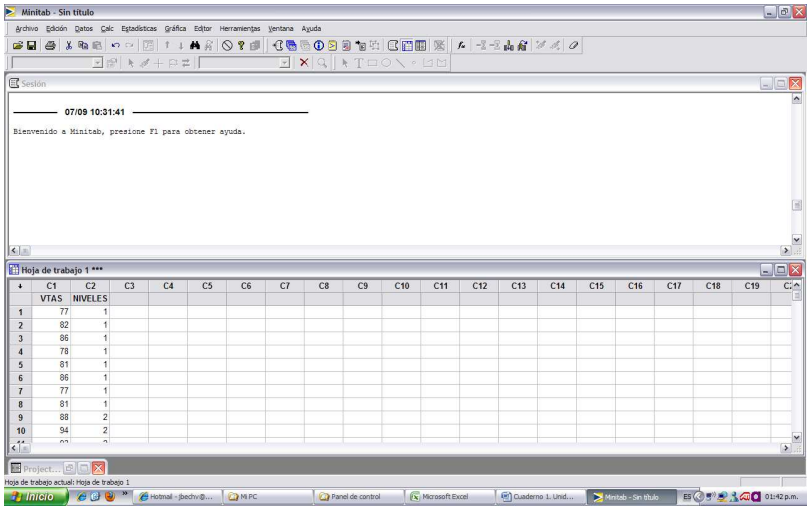
<u>Solución al inciso a.</u>

Cuando el número de observaciones en cada tratamiento **es extenso y/o existen muchos tratamientos,** los cálculos manuales son tediosos. Existen muchos paquetes de software que pueden mostrar los resultados entre ellos **Minitab.**

Comenzamos introduciendo los datos en la hoja de Trabajo 1 de Minitab, tal y como se muestra a continuación:

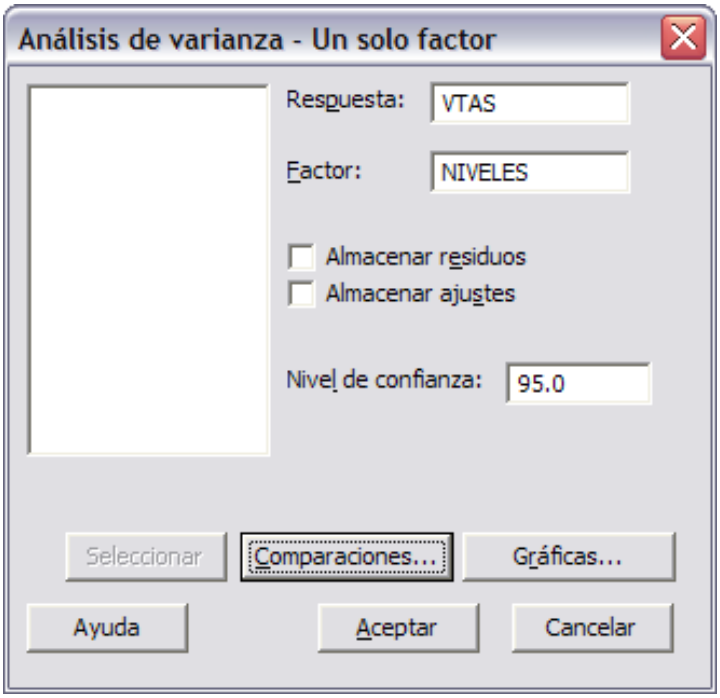
ISBN 978-607-8285-62-4

Hoja de trabajo.



Como tenemos un ***modelo con un solo factor fijo*** seleccionamos la opción ***Anova y un solo factor*** del menú ***Estadísticas,***

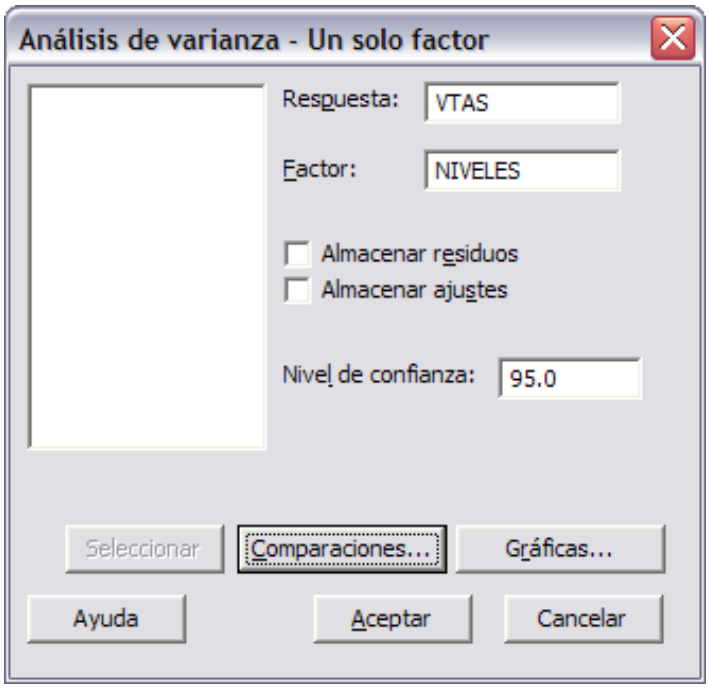
Cuadro de diálogo Análisis de varianza-Un solo factor



En **Respuesta**, ingrese **VTAS**. En **Factor**, ingrese **NIVELES.**

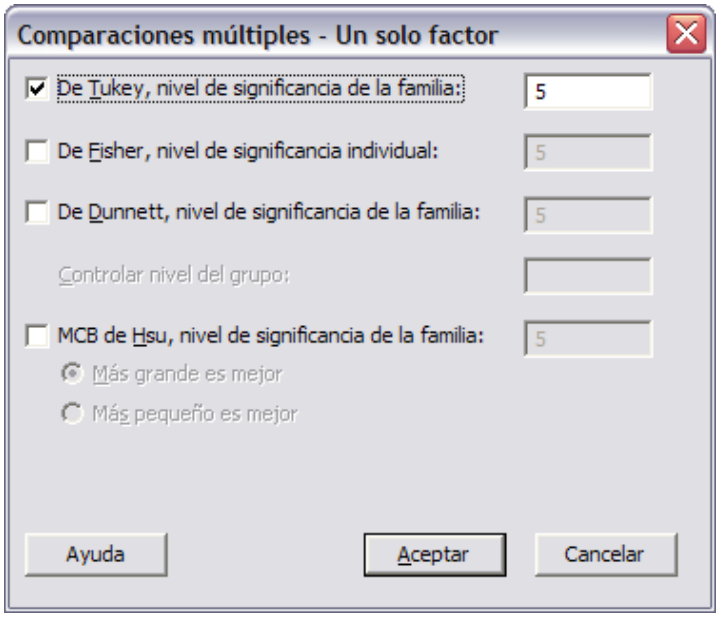
ISBN 978-607-8285-62-4

Cuadro de diálogo Análisis de varianza-Un solo factor



Haga clic en el botón **Comparaciones**. Marque **De Tukey, nivel de significancia de la familia.**

Cuadro de diálogo: Comparaciones múltiples-Un solo factor. Método de Tukey



Haga clic en ***Aceptar*** en cada cuadro de dialogo.

ISBN 978-607-8285-62-4

Salida de del análisis de varianza de un solo factor.

Salida de la ventana Sesión

ANOVA unidireccional: VTAS vs. NIVELES

```
Fuente   GL     SC     MC      F      P
NIVELES   2  399.3  199.6  14.52  0.000
Error    21  288.7   13.7
Total    23  688.0

S = 3.708   R-cuad. = 58.03%   R-cuad.(ajustado) = 54.03%
```

Interpretación

Interpretación de los resultados

En la Tabla de **ANOVA,** el valor ***p (0.000)*** para **NIVELES** indica que hay **suficiente evidencia de que no todas las medias son iguales** cuando **alfa** se establece en **0.05**. Para explorar las diferencias entre medias, examine los resultados de las comparaciones múltiples.

Intervalos de confianza simultáneos de Tukey del 95%

<u>Solución al inciso b.</u>

Salida de la ventana Sesión

Gráfica de medias.

```
                              ICs de 95% individuales para la media
                              basados en Desv.Est. agrupada
Nivel  N   Media  Desv.Est.  ----+---------+---------+---------+-----
1      8  81.000      3.625  (-----*------)
2      8  90.875      2.642                        (------*------)
3      8  84.625      4.596          (------*-----)
                              ----+---------+---------+---------+-----
                               80.0      84.0      88.0      92.0

Desv.Est. agrupada = 3.708
```

Salida de las comparaciones múltiples de Tukey.

```
Intervalos de confianza simultáneos de Tukey del 95%
Todas las comparaciones de dos a dos entre los niveles de NIVELES

Nivel de confianza individual = 98.00%


NIVELES = 1 restado de:

NIVELESInferior  Centro  Superior ------+---------+---------+---------+--
2         5.208   9.875    14.542                          (------*------)
3        -1.042   3.625     8.292                 (-----*------)
                                       ------+---------+---------+---------+---
                                             -7.0       0.0       7.0      14.0


NIVELES = 2 restado de:

NIVELESInferior  Centro  Superior -----+---------+---------+---------+---
3       -10.917  -6.250    -1.583  (------*------)
                                       ------+---------+---------+---------+---
                                             -7.0       0.0       7.0      14.0
```


ISBN 978-607-8285-62-4

Interpretación de Tukey.

Comparaciones de Tukey

De Tukey provee dos conjuntos de Intervalos de confianza de comparaciones múltiples; **1)** la media del nivel 1 (Inferior), restada de las medias de los niveles 2 y 3 (Media y Superior) y **2)** la media del nivel 2 restada de la media del nivel 3;

- El primer intervalo en el primer conjunto de Tukey **(5.208, 9.875, 14.542)** da el intervalo de confianza de la media del nivel 1, restado de la media del nivel 2. **Las medias del nivel 1 y 2 son estadísticamente diferentes**, porque el intervalo de confianza para esta combinación de medias **excluye cero**, es decir son ambos límites positivos, lo que indica que las ventas en el nivel 2 son mayores que en el nivel 1 por un mínimo de **5.208** y un máximo de **14.542 miles de pesos.**

- La **media del nivel 1** restada de la **media del nivel 3 no son estadísticamente diferentes**, porque **el intervalo de confianza incluye cero.**

- La **media del nivel 2** restada de de **la media del nivel 3 son estadísticamente diferentes**, porque el intervalo de confianza para esta combinación **excluye cero,** es decir ambos límites son negativos, lo que indica que las ventas en el nivel 2 son mayores que en el nivel 3 por un mínimo de **1.583** y un máximo de **10.917 miles de pesos.**

ISBN 978-607-8285-62-4

1.2.2.2 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 1.2.2.2 DISEÑO DE UN FACTOR DESBALANCEADO

Una organización de consumidores querría comparar el precio de un juguete en particular, en tres tiendas en un suburbio: jugueterías de descuento, tiendas de departamentos y bazares. Los resultados fueron los siguientes:

Tiendas	**Observaciones ó repeticiones (tiendas)**					
	1	**2**	**3**	**4**	**5**	**6**
Jugueterías de descuento	12	14	15	16		
Tiendas departamenta-les	15	17	14	17	17	15
Bazares	20	19	19	18	18	

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre los precios promedios de un juguete en las tres tiendas donde se venden.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor e interprete sus resultados. ¿En cuál o cuáles tiendas es mayor el precio y por cuánto más?.

Prueba de hipótesis

<u>Solución al inciso a.</u>

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

La hipótesis nula es que los precios medios del juguete son los mismos para los tres tipos de tienda donde se vende el juguete.

$$H_0\colon\ \mu_{1.} = \mu_{2.} = \ \mu_{3.}$$

La hipótesis alternativa es que los precios promedio del juguete no son iguales para los tres tipos de tienda donde se vende el juguete.

$$H_1\colon \text{No todas los precios promedio del juguete son iguales}$$

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución ***F***

$$F_{Calculada} = \frac{CMt}{CME} = \frac{SCt/_{g.l.}}{SCE/_{g.l.}}$$

Donde:

$$SCT = SCt + SCE$$

Empiece por calcular las sumas y medias para cada tratamiento así como la suma y media total:

Tiendas	Observaciones o repeticiones (días)						Total fila	Medias filas
	1		3	4	5	6		
Jugueterías de descuento	$X_{11} = 12$	$X_{12} = 14$	$X_{13} = 15$	$X_{14} = 16$			$X_{1.} = 57$	$\bar{X}_{1.} = 14.25$
Tiendas departamentales	$X_{21} = 15$	$X_{22} = 17$	$X_{23} = 14$	$X_{24} = 17$	$X_{25} = 17$	$X_{26} = 15$	$X_{2.} = 95$	$\bar{X}_{2.} = 15.833$
Bazares	$X_{31} = 20$	$X_{32} = 19$	$X_{33} = 19$	$X_{34} = 18$	$X_{35} = 18$		$X_{3.} = 94$	$\bar{X}_{3.} = 18.80$
					Totales de filas:		$X_{..} = 246$	$\bar{X}_{..} = 16.40$

Si utilizamos la fórmula abreviada, primero debe sumar el cuadrado de totas las observaciones y al final restar el promedio del cuadrado de la suma total de la siguiente manera:

Suma de Cuadrados Total.

$$SCT = \sum_{I=1}^{3}\sum_{J=1}^{8} X_{ij}^2 - \frac{X_{..}^2}{N} = (12^2 + 14^2 + \cdots + 18^2) - \frac{246^2}{15} = 4104 - 4034.40 = 69.60$$

Si usamos la fórmula abreviada primero debe sumar el cuadrado de cada total de tratamiento entre el tamaño de la muestra correspondiente a dicho tratamiento y al final restar el promedio del cuadrado de la suma total de la siguiente manera:

Suma de cuadrados de tratamientos.

$$SCt = \sum_{i=1}^{3} \frac{X_{i.}^2}{n_i} - \frac{X_{..}^2}{N} = \left(\frac{57^2}{4} + \frac{95^2}{6} + \frac{94^2}{5}\right) - \frac{246^2}{15} = 49.22$$

ISBN 978-607-8285-62-4

Si usamos la fórmula abreviada, primero debe sumar el cuadrado de todas las observaciones y al final reste la suma del cuadrado de cada total de tratamiento entre el tamaño de la muestra correspondiente a dicho tratamiento de la siguiente manera:

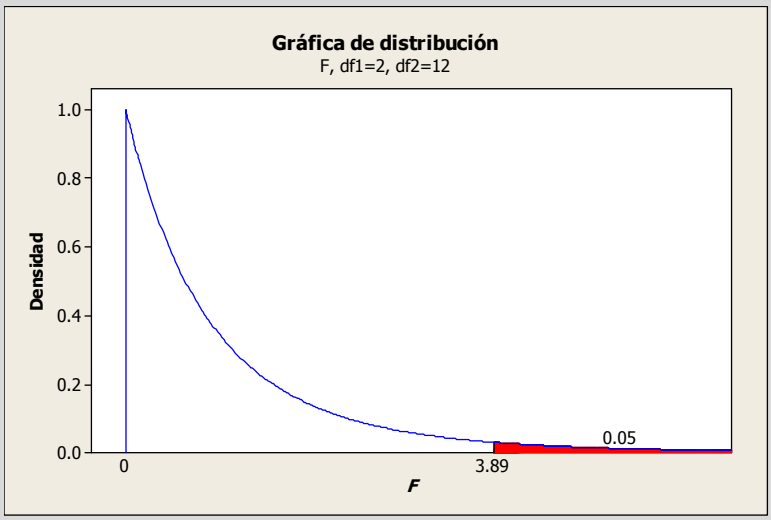
$$SCE = \sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}^2 - \sum_{i=1}^{k}\frac{X_{i.}^2}{n} = 12^2 + 14^2 + \cdots + 18^2 - (\frac{57^2}{4} + \frac{95}{6} + \frac{94^2}{5})$$
$$= 4{,}104 - 4{,}083.62 = 20.38$$

Para encontrar el valor calculado de ***F***, trabaje con la tabla de **ANOVA**. El término de cuadrado de la media es otra expresión que se utiliza para un cálculo de la varianza. El cuadrado de la media para los tratamientos es ***SCt*** dividida entre sus grados de libertad. El resultado es el cuadrado de la media para los tratamientos y se escribe ***CMt.***

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada}$***
Tratamientos	$v_1 = k - 1$ $= 3 - 1 = \mathbf{2\ g.l.}$	$SCt = 49.22$	$CMt = {}^{SCt}/_{g.l.}$ $= {}^{49.22}/_{2} = 24.61$	$F_{Calc.} = {}^{CMt}/_{CM}$
Error	$v_2 = N - k$ $= 15 - 3$ $= 12\ \boldsymbol{g.l.}$	$SCE = 20.38$	$CME = {}^{SCE}/_{g.l.}$ $= {}^{20.38}/_{12} = 1.70$	$= {}^{24.61}/_{1.70}$ $= \mathbf{14.48}$
Total	15-1=14	$SCT = 69.60$		

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de $(H_0$



ISBN 978-607-8285-62-4

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

La regla de decisión es rechazar H_0 si el valor calculado de ***F*** es mayor a 3.89.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como el valor calculado de ***F*** es de 14.48, que es mayor al valor crítico de 3.89; por lo tanto la hipótesis nula se rechaza y llegamos a la conclusión de que no todas las medias de la población son iguales, es decir al menos una de ellas es diferente.

Administrativa: Existe evidencia suficiente para concluir que estadísticamente el precio promedio del juguete no es el mismo en al menos una de las tres tiendas donde se vende el mismo.

Prueba *T* de Tukey de comparaciones múltiples.

<u>Solución al inciso b.</u>

El método ***T* de Tukey** permite examinar, en forma simultánea, comparaciones entre todos los pares de grupos mediante los siguientes pasos:

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

$$\bar{x}_{3.} = 18.80;\ \ \bar{x}_{2.} = 15.833;\ \ \bar{x}_{1.} = 14.25$$

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

$$\bar{x}_{3.} - \bar{x}_{2.} = 18.80 - 15.833 = 2.967$$

$$\bar{x}_{3.} - \bar{x}_{1.} = 18.80 - 14.25 = 4.550$$

$$\bar{x}_{2.} - \bar{x}_{1.} = 15.833 - 14.25 = 1.583$$

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$$Rango\ ó\ alcance\ crítico = q_{\alpha,k,N-k}\sqrt{\frac{CME}{n_h}}$$

$$n_h(media\ armónica) = \frac{k}{\sum_{i=1}^{k}\frac{1}{n_i}} = \frac{3}{\frac{1}{4}+\frac{1}{6}+\frac{1}{5}} = 4.865$$

ISBN 978-607-8285-62-4

$$Rango\ crítico = q_{0.05,3,12}\sqrt{\frac{1.70}{4.865}}$$

$$Rango\ crítico = 3.77\sqrt{\frac{1.70}{4.865}} = 2.229$$

Nota: el valor de q de 3.77 se obtuvo de la tabla de puntos porcentuales del rango studentizado con $\alpha = 0.05; k = 3\ y\ g.l. = 12$

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

$\bar{x}_{3.} - \bar{x}_{2.} = 18.80 - 15.833 = 2.967 > 2.229$; la prueba es (S) y $\mu_{2.} > \mu_{3.}$

$\bar{x}_{3.} - \bar{x}_{1.} = 18.80 - 14.25 = 4.550 > 2.229$; la prueba es (S) y $\mu_{2.} > \mu_{1.}$

$\bar{x}_{2.} - \bar{x}_{1.} = 15.833 - 14.25 = 1.583 < 2.229$; la prueba es (NS) y $\mu_{3.} = \mu_{1.}$

Paso 5. Construir la gráfica de medias.

Paso 5.- Construir la gráfica de medias[2]:

```
Nivel N   Media  Desv.Est.  ----+---------+---------+---------+---
1     4  14.250      1.708  (------*------)
2     6  15.833      1.329          (-----*-----)
3     5  18.800      0.837                          (-----*-----)
                            ----+---------+---------+---------+---
                                14.0      16.0      18.0      20.0
```

Paso 6. Construir los intervalos de confianza de cada par de medias.

Paso 6.- Establecer el conjunto de intervalos de confianza:

$$(\bar{X}_{i.} - \bar{X}_{i.}^{'}) - q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n_h}} \leq (\mu_{i.} - \mu_{i.)}^{'}) \leq (\bar{X}_{i.} - \bar{X}_{i.}^{'}) + q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n_h}}$$

$$(\bar{X}_{3.} - \bar{X}_{2.}) - 3.77\sqrt{\frac{1.70}{5.455}} \leq (\mu_{2.} - \mu_{3.}) \leq (\bar{X}_{2.} - \bar{X}_{3.}) - 3.77\sqrt{\frac{1.70}{5.455}}$$

$$n_{h(media\ armónica)} = \frac{k(grupos\ comparados)}{\sum_{i=1}^{k}\frac{1}{n_i(grupos\ comparados)}} = \frac{2}{\frac{1}{5}+\frac{1}{6}} = 5.455$$

$$2.967 - 3.77\sqrt{\frac{1.70}{5.455}} \leq (\mu_{3.} - \mu_{2.}) \leq 2.967 + 3.77\sqrt{\frac{1.70}{5.455}}$$

$$2.967 - 2.105 \leq (\mu_{3.} - \mu_{2.}) \leq 2.967 + 2.105$$

$$\mathbf{0.862} \leq (\mu_{3.} - \mu_{2.}) \leq \mathbf{5.072}$$

[2] Construida con el software esradístico Minitab 15

ISBN 978-607-8285-62-4

Conclusiones.

Conclusión: Como ambos límites del intervalo son positivos podemos decir que estadísticamente el precio promedio del juguete, cuando el mismo se vende en Bazares, es mayor que cuando el juguete se vende en Tiendas departamentales por un mínimo de 0.862 y un máximo de 5.072 unidades venta.

$$(\bar{X}_{3.} - \bar{X}_{1.}) - 3.77\sqrt{\frac{1.70}{4.44}} \leq (\mu_{3.} - \mu_{1.}) \leq (\bar{X}_{3.} - \bar{X}_{1.}) - 3.77\sqrt{\frac{1.70}{4.44}}$$

$$n_{h(media\ armónica)} = \frac{k(grupos\ comparados)}{\sum_{i=1}^{k}\frac{1}{n_i(grupos\ comparados)}} = \frac{2}{\frac{1}{5}+\frac{1}{4}} = 4.44$$

$$4.55 - 3.77\sqrt{\frac{1.70}{4.44}} \leq (\mu_{3.} - \mu_{1.}) \leq 4.55 + 3.77\sqrt{\frac{1.70}{4.44}}$$

$$4.55 - 2.33 \leq (\mu_{3.} - \mu_{1.}) \leq 4.55 + 2.33$$

$$\mathbf{2.22} \leq (\mu_{3} - \mu_{1.}) \leq \mathbf{6.88}$$

Conclusión: Como ambos límites del intervalo son positivos podemos decir que estadísticamente el precio promedio del juguete, cuando el mismo se vende en Bazares, es mayor que cuando el juguete se vende en Jugueterías De descuento por un mínimo de 2.22 y un máximo de 6.88 unidades venta.

$$(\bar{X}_{2.} - \bar{X}_{1.}) - 3.77\sqrt{\frac{1.70}{4.80}} \leq (\mu_{2.} - \mu_{1.}) \leq (\bar{X}_{2.} - \bar{X}_{1}) - 3.77\sqrt{\frac{1.70}{4.80}}$$

$$n_{h(media\ armónica)} = \frac{k(grupos\ comparados)}{\sum_{i=1}^{k}\frac{1}{n_i(grupos\ comparados)}} = \frac{2}{\frac{1}{6}+\frac{1}{4}} = 4.80$$

$$1.583 - 3.77\sqrt{\frac{1.70}{4.80}} \leq (\mu_{2.} - \mu_{1.}) \leq 1.583 + 3.77\sqrt{\frac{1.70}{4.80}}$$

$$1.583 - 2.243 \leq (\mu_{2.} - \mu_{1.}) \leq 1.583 + 2.243$$

$$\mathbf{-0.660} \leq (\mu_{2.} - \mu_{1.}) \leq \mathbf{3.826}$$

Conclusión: Como el intervalo de confianza pasa por cero, podemos decir que estadísticamente el precio promedio del juguete cuando se vende en las Tiendas departamentales y/o en las Jugueterías de descuento es el mismo.

ISBN 978-607-8285-62-4

1.2.2.3 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 1.2.2.3 DISEÑO DE UN FACTOR DESBALANCEADO

Se recopiló la siguiente información de un experimento:

Tratamiento 1	Tratamiento 2	Tratamiento 3
8	3	3
11	2	4
10	1	5
	3	4
	2	

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre las medias de los tres tratamientos.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor e interprete sus resultados. ¿ En cuál o cuáles tratamientos la media es mayor y por cuánto más?.

NOTA: El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

ANOVA de una vía ó unifactorial. Prueba de hipótesis

Solución al inciso a.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

Empiece por calcular las sumas y medias para cada tratamiento así como la suma y media total y llene la tabla correspondiente:

No. De tratamiento	Observaciones o repeticiones					Total de columnas	Medias de columnas
	1	2	3	4	5		
1	$X_{11} =$	$X_{12} =$	$X_{13} =$	$X_{14} =$	$X_{15} =$	$X_{1.} =$	$\bar{X}_{1.} =$
2	$X_{21} =$	$X_{22} =$	$X_{23} =$	$X_{24} =$	$X_{25} =$	$X_{2.} =$	$\bar{X}_{2.} =$
3	$X_{31} =$	$X_{32} =$	$X_{33} =$	$X_{34} =$	$X_{35} =$	$X_{3.} =$	$\bar{X}_{3.} =$
					Totales	$X_{..} =$	$\bar{X}.. =$

Calcule SCT, SCt y la SCE

Suma de Cuadrados Total.

$$SCT = \sum_{I=1}^{k}\sum_{J=1}^{n} X_{ij}^2 - \frac{X_{..}^2}{N} =$$

Suma de Cuadrados de tratamientos.

$$SCt = \sum_{i=1}^{k} \frac{X_{i.}^2}{n_i} - \frac{X_{..}^2}{N} =$$

Suma de Cuadardao del Error.

$$SCE = \sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}^2 - \sum_{i=1}^{k} \frac{X_{i.}^2}{n} =$$

Para encontrar el valor calculado de ***F***, trabaje con la tabla de **ANOVA**.

Tablas de ANOVA.

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada}$***
Tratamientos	$v_1 = k - 1 =$	$SCt =$	$CMt = {SCt}/{g.l.} =$	$F_{Calc.} = {CMt}/{CME}$
Error	$v_2 = N - k =$	$SCE =$	$CME = {SCE}/{g.l.} =$	
Total	*N-1=*	$SCT =$		

ISBN 978-607-8285-62-4

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

Prueba post-hoc ó aposteriori. Comparaciones múltiples. Método de Tukey

<u>Solución al inciso b.</u>

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$$\boldsymbol{Rango\ ó\ alcance\ crítico = q_{\alpha,k,N-k}\sqrt{\frac{CME}{n_h}}}$$

$$\boldsymbol{n_h(media\ armónica) = \frac{k}{\sum_{i=1}^{k}\frac{1}{n_i}} =}$$

ISBN 978-607-8285-62-4

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

Paso 5. Construir la gráfica de medias.

Paso 5.- Construir la gráfica de medias:

Paso 6. Construir los intervalos de confianza de cada par de medias.

Paso 6.- Establecer el conjunto de intervalos de confianza:

$$(\bar{X}_{i.} - \bar{X}_{i'.}) - q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}} \leq (\mu_{i.} - \mu_{i'.}) \leq (\bar{X}_{i.} - \bar{X}_{i'.}) + q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}}$$

$$n_{h(media\ arm\acute{o}nica)} = \frac{k(grupos\ comparados)}{\sum_{i=1}^{k}\frac{1}{n_i(grupos\ comparados)}}$$

Conclusiones.

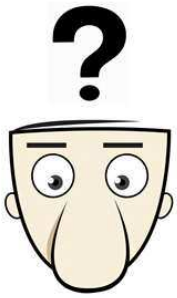
Conclusiones:

ISBN 978-607-8285-62-4

1.2.2.3 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. Las respuestas a este ejercicio de autoevaluación se encuentran al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 1.2.2.3
DISEÑO DE UN FACTOR DESBALANCEADO



Dadas las mediciones de las cuatro muestras que presentamos a continuación,

Muestra 1	26	31	34	3	39	
Muestra 2	39	28	30	29	40	31
Muestra 3	14	15	21	19	28	17
Muestra 4	21	28	20	22	18	

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre las mediciones promedios de las cuatro muestras.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿ Cuál o cuáles mediciones resultaron mayores y por cuánto más?.

NOTA: El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

Solución al inciso a.

Prueba de hipótesis.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

Empiece por calcular las sumas y medias para cada tratamiento así como la suma y media total y llene la tabla correspondiente:

No. De tratamiento	Observaciones o repeticiones					Total de columnas	Medias de columnas
	1	2	3	4	5		
1	$X_{11} =$	$X_{12} =$	$X_{13} =$	$X_{14} =$	$X_{15} =$	$X_{1.} =$	$\bar{X}_{1.} =$
2	$X_{21} =$	$X_{22} =$	$X_{23} =$	$X_{24} =$	$X_{25} =$	$X_{2.} =$	$\bar{X}_{2.} =$
3	$X_{31} =$	$X_{32} =$	$X_{33} =$	$X_{34} =$	$X_{35} =$	$X_{3.} =$	$\bar{X}_{3.} =$
					Totales	$X_{..} =$	$\bar{X}_{..} =$

Calcule SCT, SCt y la SCE

Suma de Cuadrados Total.

$\boldsymbol{SCT} =$

Suma de Cuadrados de los tratamientos.

$\boldsymbol{SCt} =$

Suma de Cuadrados del Error.

$\boldsymbol{SCE} =$

Para encontrar el valor calculado de ***F***, trabaje con la tabla de **ANOVA**.

Tabla de ANOVA.

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	$F_{calculada}$
Tratamientos				
Error				
Total				

ISBN 978-607-8285-62-4

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

Prueba post-hoc ó aposteriori. Comparaciones múltiples. Método de Tukey

Solución al inciso b.

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$\boldsymbol{Rango\ ó\ alcance\ crítico} =$

$\boldsymbol{n_h(media\ armónica)} =$

ISBN 978-607-8285-62-4

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.	**Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:**
Paso 5. Construir la gráfica de medias.	**Paso 5.- Construir la gráfica de medias:**
Paso 6. Construir los intervalos de confianza de cada par de medias.	**Paso 6.- Establecer el conjunto de intervalos de confianza:**
Conclusiones.	**Conclusiones:**

ISBN 978-607-8285-62-4

1.2.2 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 1.2.2 DISEÑO DE UN FACTOR DESBALANCEADO

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

1.2.2.4 Los datos siguientes muestran el número de quejas procesadas diariamente de un grupo de cuatro empleados de compañías de seguros observados durante un cierto número de días:

Empleado 1	15	17	17	12		
Empleado 2	12	10	13	17		
Empleado 3	16	19	17	20	17	
Empleado 4	13	12	12	14	10	9

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas en el número promedio de quejas procesadas diariamente para los cuatro empleados.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor e interprete sus resultados. ¿Cuál o cuales empleados recibieron más quejas y por cuántas más?.

1.2.2.5 Una compañía de investigación ha diseñado tres sistemas distintos para limpiar manchas de aceite. La siguiente tabla contiene los resultados de cada sistema, medidos en que tanta superficie (en metros cuadrados) es limpiada en una hora. Los datos se obtuvieron probando cada método en varias sesiones:

Sistema A	55	60	63	56	59	55
Sistema B	72	68	79	64	77	
Sistema C	66	52	61	57		

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre los valores promedios de de las mediciones de los tres sistemas.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿ Cuál ó cuáles sistemas fueron más efectivos y por cuánto más?.

ISBN 978-607-8285-62-4

1.2.2.2 EJEMPLO ILUSTRATIVO EN EXCEL

EJEMPLO ILUSTRATIVO INTEGRAL EN EXCEL 1.2.2.2 DISEÑO DE UN FACTOR DESBALANCEADO

Prueba de hipótesis para K medias. Diseño desbalanceado.

Una organización de consumidores querría comparar el precio de un juguete en particular, en tres tiendas en un suburbio: jugueterías de descuento, tiendas de departamentos y bazares. Los resultados fueron los siguientes:

Tiendas	Observaciones ó repeticiones (tiendas)					
	1	2	3	4	5	6
Jugueterías de descuento	12	14	15	16		
Tiendas departamentales	15	17	14	17	17	15
Bazares	20	19	19	18	18	

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre los precios promedios de un juguete en las tres tiendas donde se venden.

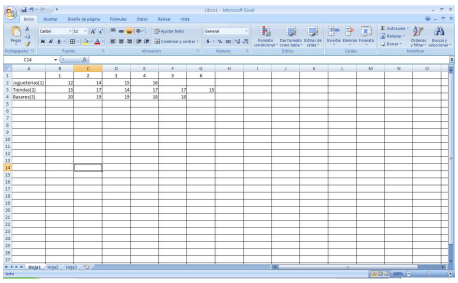
b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor e interprete sus resultados. ¿En cuál o cuáles tiendas es mayor el precio y por cuánto más?.

Solución al inciso a.

Cuando el número de observaciones en cada tratamiento es extenso y/o existen muchos tratamientos, los cálculos manuales son tediosos. Existen muchos paquetes de software que pueden mostrar los resultados entre ellos Excel.

Comenzamos introduciendo los datos en la hoja de Excel, tal y como se muestra a continuación:

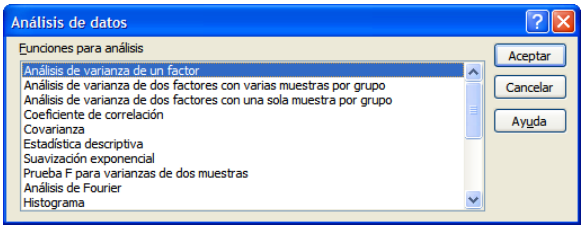
Hoja de trabajo de Excel.



ISBN 978-607-8285-62-4

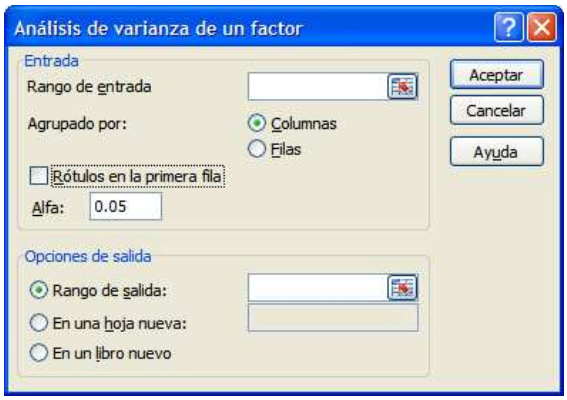
Como tenemos un ***modelo con un solo factor fijo*** seleccionamos la opción ***Análisis de datos*** del menú ***Datos,*** utilizaremos la opción ***Análisis de la varianza de un factor***, del cuadro ***Análisis de datos*** de la figura siguiente:

Cuadro de diálogo: Análisis de datos.



En la lista ***Funciones para análisis,*** elija la modalidad **de *Análisis de varianza de un factor*** y oprima el botón ***Aceptar*** para obtener el siguiente cuadro de dialogo rellenando su pantalla de entrada:

Cuadro de diálogo: Análisis de varianza de un factor.



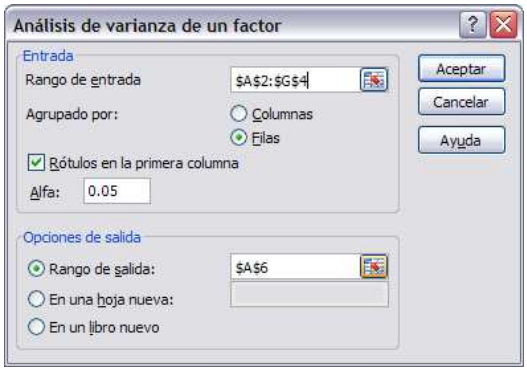
En el cuadro ***Rango de entrada*** introduzca, (seleccionando con el cursor las celdas donde están los datos incluyendo los rótulos de la primera columna, **pero no los del primer renglón**), la referencia de celda correspondiente al rango de datos que está analizando. La referencia deberá contener dos o más rangos adyacentes organizados en columnas o filas.

En el campo ***Agrupado por*** haga clic en el botón ***Filas*** para indicar que los datos del rango de entrada están organizados en filas (es posible también organizarlos por columnas si lo desea). Si la primera columna del rango de entrada contiene rótulos, active la casilla de verificación ***Rótulos en la primera columna***. Esta casilla de verificación debe quedar desactivada si el rango de entrada carece de rótulos; Microsoft Office Excel 2007 generará los rótulos de datos correspondientes para la tabla de resultados. Deje sin cambio el campo Alfa con el valor de 0.05 (nivel con el que desee evaluar los valores críticos de la función estadística *F).* El nivel ***alfa*** es un nivel de importancia relacionado con la probabilidad de que haya un error de tipo I (rechazar una hipótesis verdadera).

ISBN 978-607-8285-62-4

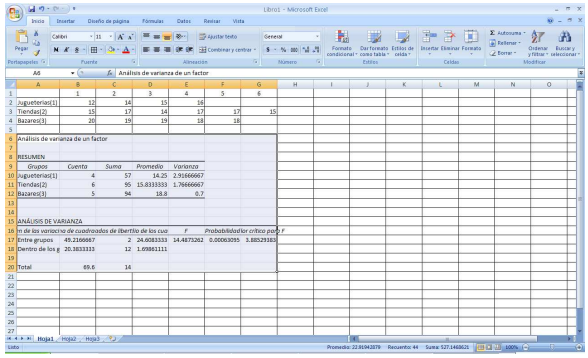
En cuanto a las ***opciones de salida***, en el campo ***Rango de salida*** introduzca la referencia, (dando un clic), correspondiente a la celda superior izquierda de la tabla de resultados, en este caso la celda A6 y oprima el botón Aceptar.

Cuadro de diálogo: Análisis de varianza de un factor.



A continuación se muestra la salida del análisis de la varianza de un solo factor:

Salida del análisis de varianza de un solo factor. Diseño desbalanceado.



Observe que Excel utiliza el término "Entre grupos" para "Tratamientos" y "Dentro de los grupos" para " Error". Sin embargo, tienen los mismos significados.

- Si el *valor p* es menor que o igual al nivel Alfa (a), rechace la hipótesis nula en favor de la hipótesis alternativa.
- Si el *valor p* es mayor que el nivel Alfa (a), no rechace la hipótesis nula.

Los niveles de significancia más utilizado son 0.05 y 0.01.

Conclusión: Como el p-valor del test *F* de Fisher es menor que 0.05, existen diferencias significativas entre los precios del juguete en al menos un nivel de tiendas donde se vende el producto al 95% de confianza.

NOTA: Excel en este caso no tiene opción para realizar pruebas Pos-hoc ó Aposteriori como la prueba de Tukey

ISBN 978-607-8285-62-4

1.2.2.2 EJEMPLO ILUSTRATIVO EN MINITAB 15

EJEMPLO ILUSTRATIVO INTEGRAL EN MINITAB 1.2.2.2 DISEÑO DE UN FACTOR DESBALANCEADO

Análisis de Varianza unifactorial. Diseño desbalanceado. Prueba de hipótesis

Una organización de consumidores querría comparar el precio de un juguete en particular, en tres tiendas en un suburbio: jugueterías de descuento, tiendas de departamentos y bazares. Los resultados fueron los siguientes:

Tiendas	**Observaciones ó repeticiones (tiendas)**					
	1	**2**	**3**	**4**	**5**	**6**
Jugueterías de descuento	12	14	15	16		
Tiendas departamentales	15	17	14	17	17	15
Bazares	20	19	19	18	18	

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre los precios promedios de un juguete en las tres tiendas donde se venden.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor e interprete sus resultados. ¿En cuál o cuáles tiendas es mayor el precio y por cuánto más?.

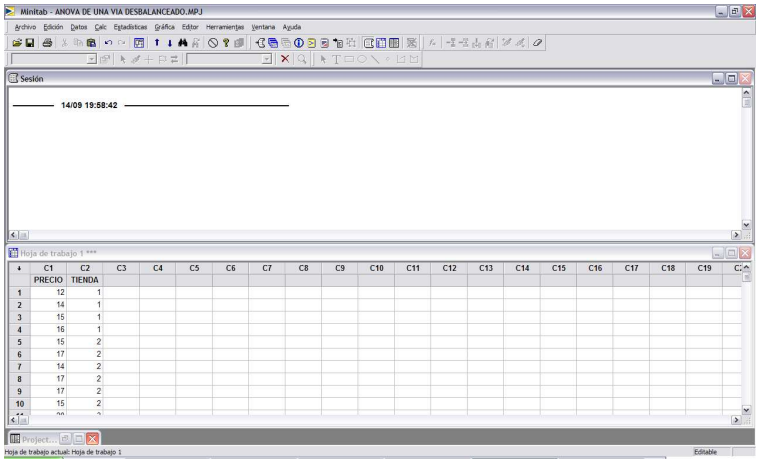
<u>Solución al inciso a.</u>

Cuando el número de observaciones en cada tratamiento es extenso y/o existen muchos tratamientos, los cálculos manuales son tediosos. Existen muchos paquetes de software que pueden mostrar los resultados entre ellos **Minitab (Versión 15).**

Comenzamos introduciendo los datos en la hoja de Trabajo 1 de Minitab, tal y como se muestra a continuación:

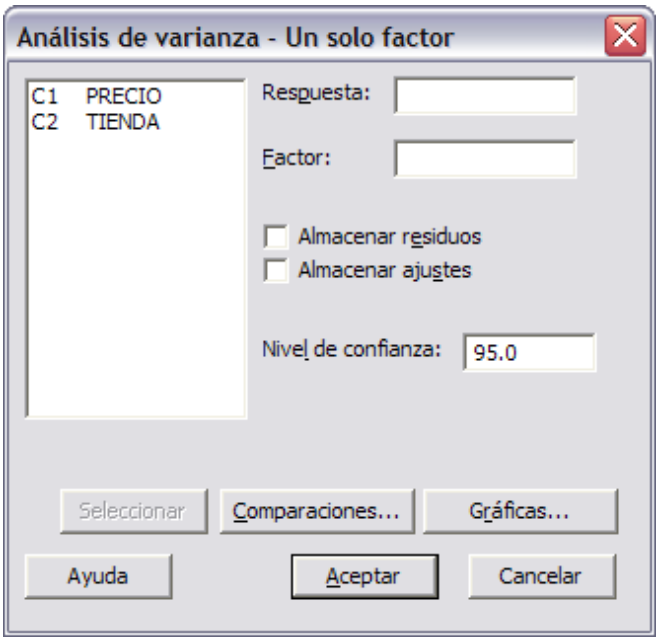
ISBN 978-607-8285-62-4

Hoja de trabajo de Excel



Como tenemos un *modelo con un solo factor fijo* seleccionamos la opción ***Anova y un solo factor*** del menú ***Estadísticas***,

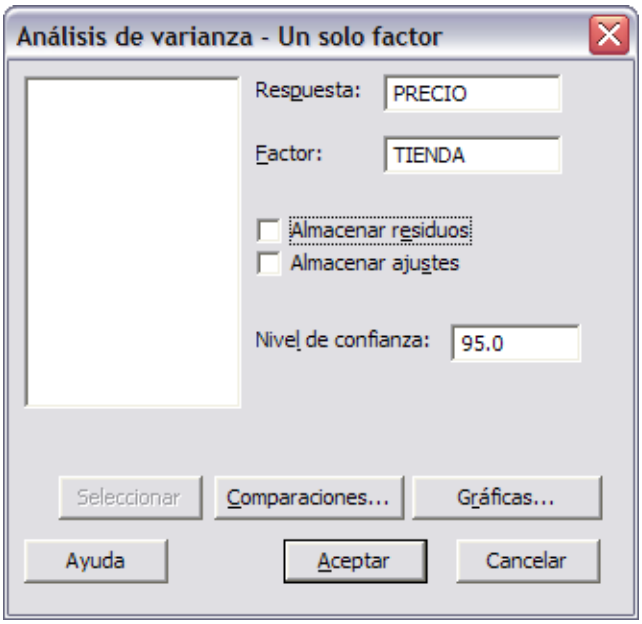
Cuadro de diálogo: Análisis de varianza-Un solo factor.



En Respuesta, ingrese **PRECIO**. En **Factor**, ingrese **TIENDA**

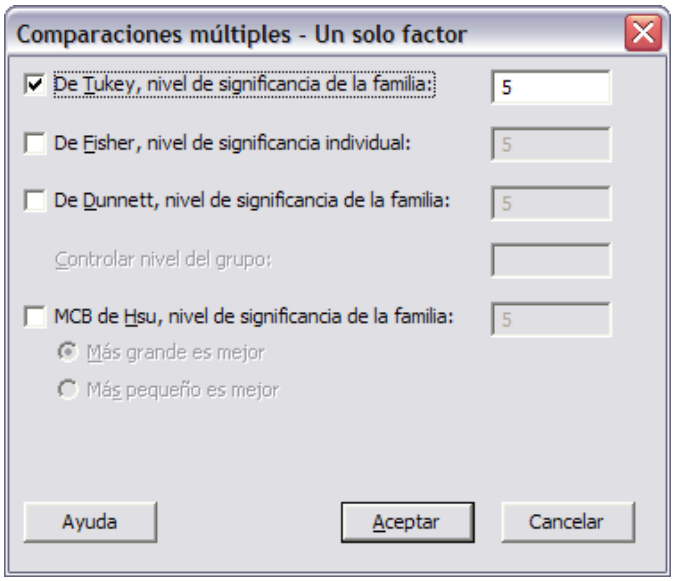
ISBN 978-607-8285-62-4

Cuadro de diálogo: Análisis de varianza-Un solo factor.



Haga clic en el botón **Comparaciones**. Marque **De Tukey, nivel de significancia de la familia.**

Cuadro de diálogo: Comparaciones múltiples-Un solo factor. Prueba de Tukey.



Haga clic en **Aceptar** en cada cuadro de dialogo.

ISBN 978-607-8285-62-4

- Si el *valor p* es menor que o igual al nivel Alfa (a), rechace la hipótesis nula en favor de la hipótesis alternativa. Si $0.01 \leq p \leq 0.05$ se dice que la prueba es significativa (S). Si $p < 0.01$ se dice que la prueba es altamente significativa (AS).

- Si el *valor p* es mayor que el nivel Alfa (a), no rechace la hipótesis nula. Si $p > 0.05$ se dice que la prueba es no significativa (NS).

Los niveles de significancia más utilizado son 0.05 y 0.01.

Prueba post-hoc ó aposteriori. Comparaciones múltiples. Prueba de Tukey.

Gráfica de medias

Intervalos de confianza simultáneos de Tukey del 95%

Conclusiones.

Salida de la ventana Sesión

ANOVA unidireccional: PRECIO vs. TIENDA

```
Fuente  GL     SC     MC      F      P
TIENDA   2  49.22  24.61  14.49  0.001
Error   12  20.38   1.70
Total   14  69.60

S = 1.303   R-cuad. = 70.71%   R-cuad.(ajustado) = 65.83%
```

Conclusión: En la Tabla de **ANOVA**, el valor p (0.001)<0.05 para **TIENDA** indica que hay **suficiente evidencia** de que **no todas las medias son iguales** cuando alfa se establece en 0.05. Para explorar las diferencias entre medias, examine los resultados de las comparaciones múltiples.

Solución al inciso b.

```
                                ICs de 95% individuales para la media
                                basados en Desv.Est. agrupada
Nivel  N   Media  Desv.Est.  ------+---------+---------+---------+-
1      4  14.250      1.708  (------*------)
2      6  15.833      1.329          (-----*-----)
3      5  18.800      0.837                          (-----*-----)
                              ------+---------+---------+---------+-
                                 14.0      16.0      18.0      20.0

Desv.Est. agrupada = 1.303


Intervalos de confianza simultáneos de Tukey del 95%
Todas las comparaciones de dos a dos entre los niveles de TIENDA


Nivel de confianza individual = 97.94%


TIENDA = 1 restado de:


TIENDA Inferior  Centro  Superior  -------+---------+---------+---
2       -0.659   1.583     3.826          (------*-------)
3        2.219   4.550     6.881                  (-------*-------)
                                  -------+---------+---------+----
                                      -3.0       0.0       3.0       6.0


TIENDA = 2 restado de:


TIENDAInferior  Centro  Superior  -------+---------+---------+--+--
3        0.863   2.967     5.070                   (------*------)
                                 -------+---------+---------+---------+--
                                      -3.0       0.0       3.0       6.0
```

Conclusión: De Tukey provee dos conjuntos de Intervalos de confianza de comparaciones múltiples; 1) la media del nivel 1 (Jugueterías de descuento), restada de las medias de los niveles 2 y 3 (Tiendas departamentales y Bazares) y 2) la media del nivel 2 restada de la media del nivel 3;

ISBN 978-607-8285-62-4

- El primer intervalo en el primer conjunto de Tukey (-0.659, 1.583, 3.826) da el intervalo de confianza de la media del nivel 1, restado de la media del nivel 2. El precio del juguete en Las Jugueterías de descuento y en las Tiendas departamentales no son estadísticamente diferentes, porque el intervalo de confianza incluye cero.

- El segundo intervalo en el primer conjunto que muestra la media del nivel 1 restada de de la media del nivel 3 son estadísticamente diferentes, porque el intervalo de confianza para esta combinación excluye cero, es decir ambos límites son positivos, lo que indica que el precio promedio del juguete en los Bazares es mayor que en las Jugueterías de descuento por un mínimo de 2.219 y un máximo de 6.681.

- El primer intervalo en el segundo conjunto de Tukey que muestra la media del nivel 2 restada de de la media del nivel 3 son estadísticamente diferentes, porque el intervalo de confianza para esta combinación excluye cero, es decir ambos límites son positivos, lo que indica que el precio promedio del juguete en los Bazares es mayor que en las Tiendas departamentales por un mínimo de 0.863 y un máximo de 5.070.

ISBN 978-607-8285-62-4

OBJETIVO 1.3. El alumno podrá aplicar el diseño en bloques aleatorios para realizar una prueba de hipótesis para un diseño aleatorizado en bloques, utilizar el método *T* de Tukey de comparaciónes múltiples y determinar la eficiencia relativa de este diseño.

ANTECEDENTES

CONCEPTOS DE:

Experimento. Unidad experimental. Variable de respuesta. Ensayos ó réplicas. Aleatorización. Agrupamiento. Bloqueo. Balanceo. Factores controlados. Factores no controlados. Tratamientos ó niveles de un factor. Error experimental. Efectos del tratamiento. Variación total. Variación entre tratamientos. Variación dentro de tratamientos. Análisis de varianza (ANOVA).

1.3.1 ELEMENTOS Y SUPUESTOS DEL DISEÑO DE BLOQUES AL AZAR

CONCEPTOS BÁSICOS DISEÑO DE BLOQUES AL AZAR

Un bloque Es una segunda fuente de variación, además de los tratamientos. En el análisis de varianza, un protocolo donde los sujetos

En muchos problemas de experimentos, es necesario hacer un diseño de tal manera que la **variabilidad proveniente de fuentes conocidas pueda ser sistemáticamente controlada.**

Se pretende **reducir el efecto de la variabilidad** proveniente de causas propias del experimento pero **independiente del efecto que se desea estudiar.**

Para los fines del análisis de varianza **el bloqueo** introduce **un efecto adicional ficticio**, cuyo objetivo es **separar del error experimental**, **alguna fuente de variabilidad conocida.**

Un **diseño de bloques completamente aleatorizado** de dos factores es un **diseño de dos factores balanceado competo** en el cual los efectos de un factor (el factor de tratamiento) son relevantes, mientras los efectos del otro factor (el factor bloqueado) no. **El factor bloqueado** es incluido para **reducir la incertidumbre** en las estimaciones **del efecto principal del factor de tratamiento.**

ISBN 978-607-8285-62-4

en cada bloque o grupo están asignados a tratamiento diferente

Debido a que el objetivo de un **diseño de bloques completamente aleatorizado** es calcular los **efectos principales del factor de tratamiento**, **no debe haber interacción** entre el factor de tratamiento y el factor bloqueado. Se utiliza un análisis **de varianza de dos sentidos** para estimar los efectos y realizar pruebas de hipótesis sobre los efectos principales del factor de tratamiento.

Un **diseño de bloques completamente aleatorizado** proporciona gran ventaja sobre un diseño completamente aleatorio cuando **el factor bloqueado afecta fuertemente la respuesta** y proporciona una desventaja pequeña cuando el factor bloqueado no tiene poco o nada de efecto. Por tanto, cuando se tiene duda, es una buena idea utilizar un diseño bloqueado.

El **diseño en bloque completo al azar** es un plan en el cual las **unidades experimentales** se **asignan a grupos homogéneos, llamados bloques**, y los **tratamientos** son luego, **asignados al azar dentro de los bloques.**

El **objetivo de agrupamiento** es lograr que las unidades dentro de un **bloque** sean lo mas **uniformes** posibles con respecto a la **variable dependiente**, de modo que las diferencias observadas se deban realmente a los tratamientos. Al controlar la variación dentro de los bloques reducimos la variabilidad del error experimental.

En el **diseño en bloques completos aleatorizados** se divide el material experimental en tantos **bloques** como **números de réplicas a utilizar**. Cada **bloque** es luego dividido en tantas unidades experimentales como tratamientos haya en estudio. Como el **diseño en bloques completos aleatorizados** especifica que todos los tratamientos **deben aparecer una vez** en **cada réplica**, la aleatorización se hace separadamente en cada bloque. La aleatorización es similar al diseño completamente aleatorizado para cada bloque.

Elementos en el diseño de bloques al azar

El la **estructura de un diseño de en bloques** se deben considerar los siguientes **elementos**:
1.- El conjunto de **tratamientos** incluidos en el estudio.
2.- El conjunto de **unidades experimentales** utilizadas en el estudio.
3.- Las reglas y procedimientos por los cuales los **tratamientos son asignados a las unidades experimentales** (o viceversa).
4.- Las **medidas** o evaluaciones que se hacen a las **unidades experimentales** luego de aplicar los **tratamientos**.

Supuestos en el diseño de bloques al azar

En un **diseño de experimentos ó ANOVA** de **bloques al azar** existen los siguientes **supuestos básicos**:

Las pruebas usuales de hipótesis **ANOVA de dos sentidos** son válidas bajo las siguientes **condicione ó supuestos básicos**:

1.- El diseño debe estar completo.
2.- El diseño debe ser balanceado.
3.- La varianza poblacional es igual para todos los tratamientos. Esta varianza se denota mediante σ^2.

ISBN 978-607-8285-62-4

1.3.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 1.3.1.1 DISEÑO DE BLOQUES AL AZAR

Para el ensamble de un artículo se considera comparar 4 máquinas diferentes. Como la operación de las máquinas requiere cierta destreza se anticipa que habrá una diferencia entre los operarios en cuanto a la velocidad con la cual operen la máquina. Se decide que requerirán 6 operarios diferentes en un experimento de bloques aleatorizado para comparar las máquinas.

Entonces el factor de interés es uno sólo, pero se crea otro factor para controlar la variabilidad extraña y excluirla así del error experimental.

Aleatorización: debemos asignar cada tratamiento, M1, M2, M3 y M4 a cada bloque

Operario 1	Bloque 2	Bloque 3	Bloque 4	Bloque 5	Bloque 6
22	75	76	84	5	16
45	31	25	51	79	44
27	70	98	10	36	29
2	86	85	78	95	14
M2	M3	M2	M4	M1	M2
M4	M1	M1	M2	M3	M4
M3	M2	M4	M1	M2	M3
M1	M4	M3	M3	M4	M1

ISBN 978-607-8285-62-4

1.3.2 EL ANÁLISIS DE VARIANZA DEL DISEÑO DE BLOQUES AL AZAR

CONCEPTOS BÁSICOS DISEÑO DE BLOQUES AL AZAR

Esta técnica tiene por objeto **eliminar fuentes de variación** que durante el desarrollo del experimento no pudieron ser controladas y nos hacen suponer que tienen una influencia sobre las mediciones obtenidas en la variable dependiente. En esta técnica, los orígenes de las fuentes de variación se encuentran en las unidades de prueba o, lo que es lo mismo, en los lugares en donde se lleva a cabo el experimento, o con aquellas personas con las que efectuamos el experimento. Por ello, **la variable incontrolable tiene como característica el ser no métrica**. Por su naturaleza propiamente cualitativa o no métrica **a esta variable se le conoce con el nombre de bloque**.

Al hablar de ***variación*** , nos referimos a las diferencias entre los datos observados y los promedios de los mismos datos, por lo que se pueden considerar tres tipos de variaciones en este diseño:

1.- VARIACIÓN TOTAL. Suma de las diferencias elevadas al cuadrado entre cada observación y la media total **(SCT).**

2.- VARIACIÓN DE TRATAMIENTO Ó ENTRE TRATAMIENTOS. Suma de las diferencias elevadas al cuadrado entre la media de cada tratamiento y la media total o general (**$SC_{tratamientos}$**).

3.- VARIACIÓN DE BLOQUE Ó ENTRE BLOQUES. Suma de las diferencias elevadas al cuadrado entre la media de cada bloque y la media total o general (**$SC_{bloques}$**).

4.- VARIACIÓN ALEATORIA O DENTRO DE TRATAMIENTOS. Suma de las diferencias elevadas al cuadrado entre las observaciones y sus medias de tratamiento (**SCE)**.

La siguiente tabla representa la **matriz de las observaciones** al efectuar el experimento:

Matriz de observaciones.

Niveles del Factor 1	**Bloques**					**Total**	**Media**
	1	**2**	**j**	**…**	**b**		
1	X_{11}	X_{12}	X_{1j}		X_{1b}	$X_{1..}$	$\overline{X}_{1.}$
2	X_{21}	X_{22}	X_{2j}		X_{2b}	$X_{2..}$	$\overline{X}_{2.}$
i			X_{ij}		X_{ib}	$X_{i.}$	$\overline{X}_{i.}$
…							
K	X_{k1}	X_{k2}			X_{kb}	$X_{k.}$	$\overline{X}_{k.}$
Total	$X_{.1}$	$X_{.2}$	$X_{.j}$		$X_{.b}$	$X_{..}$	
Media	$\overline{X}_{.1}$	$\overline{X}_{.2}$	$\overline{X}_{.j}$		$\overline{X}_{.b}$		$\overline{X}_{..}$

ISBN 978-607-8285-62-4

Modelo estadístico del diseño de bloques al azar.

El modelo estadístico para este diseño es el siguiente:

$$X_{ij} = \mu + \tau_{i.} + \beta_{.j} + \varepsilon_{ij} \begin{cases} i = 1, 2, \dots, k \\ j = 1, 2, \dots, b \end{cases}$$

Donde:

X_{ij}= observación de la variable dependiente bajo los efectos del nivel i del factor manejado en el experimento medida en la unidad de prueba o bloque j.

μ= promedio general de la variable dependiente.

$\tau_{i.}$= efecto del nivel i del factor manejado en el experimento

$\beta_{.j}$= efecto del bloque j sobre la variable dependiente

ε_{ij}= error aleatorio de cada una de las observaciones de la variable dependiente.

Como los efectos de los tratamientos y de bloque, se consideran como desviaciones de la media general por lo tanto :

$$\sum_{i=1}^{k} \zeta_{i.} = 0 \qquad y \qquad \sum_{j=1}^{b} \beta_{.j} = 0$$

El análisis de varianza consiste en descomponer o subdividir la suma de cuadrados total de la siguiente manera:

La suma de cuadrados es la cantidad calculada en el análisis de varianza y usada para obtener cuadrados medios para la prueba ***F***.

$$\mathbf{SCT = SC_{tratamiento} + SC_{bloque} + SCE}$$

Se desea probar la igualdad de las medias de los niveles o **tratamientos del factor 1**. El juego de hipótesis es:

Los cuadrados medios son el cociente entre la suma de cuadrados y los grados de libertad.

$$H_0: \mu_{1.} = \mu_{2.} = \cdots = \mu_{k.}$$

$$H_1: al\ menos\ una\ \mu_{k.}\ es\ diferente$$

El cuadrado medio del tratamiento es la estimación de la variación en el análisis de varianza. Se usa en el numerador de la prueba estadística ***F***.

Con el fin de determinar si las medias de los diversos tratamientos son todas iguales, se pueden examinar dos estimadores diferentes de la varianza de la población. Uno de los estimadores se basa en **la suma de los cuadrados dentro de los tratamientos *(SCt);*** el otro se basa en **la suma de los cuadrados entre los tratamientos y los bloques *(SCE)***. Si la hipótesis nula es cierta, estos estimadores deben ser aproximadamente iguales; si es falsa, el estimador basado en la suma de los cuadrados entre grupos debe ser mayor.

El cuadrado medio del error es la estimación de la variación en el análisis de varianza. Se usa en el denominador de la estadística ***F***.

En el Análisis de Varianza, el estimador de la varianza entre los tratamientos ***(CMt)*** se calcula dividiendo la suma de los cuadrados de los tratamientos entre los grados de libertad entre los tratamientos ***(k-1).*** La varianza dentro de los tratamientos y los bloques, ***(CME),*** se estima dividiendo la suma de los cuadrados dentro de los tratamientos y bloques entre los grados de libertad dentro de los tratamientos y bloques ***(k-1)(b-1).***

ISBN 978-607-8285-62-4

Si en realidad hay una diferencia entre los tratamientos, el ***(CMt),*** será significativamente ***mayor*** que el ***(CME).*** La prueba estadística se basa en la razón de las dos varianzas, ***CMt/CME***. La distribución de esta razón se conoce como la **distribución *F,*** por lo que el estadístico de prueba es:

$$F_{CALC.} = \frac{CM_{tratamiento}}{CME} = \frac{SC_{tratamientos}/_{g.l.}}{SCE/_{g.l.}}$$

La regla de decisión es rechazar la hipótesis nula de que no hay diferencia entre los tratamientos si al nivel de significancia α

$$\boldsymbol{F_{calc} \geq F_{\alpha,(k-1),(k-1)(b-1)}}$$

Para examinar si resulta ventajoso crear bloques algunos investigadores quisieran **probar la hipótesis nula de <u>no</u> existencia de efectos de bloque:**

$$\boldsymbol{H_0: \mu_{.1} = \mu_{.2} = \cdots = \mu_{.j}}$$

$$\boldsymbol{H_1: al\ menos\ una\ \mu_{.j}.\ es\ diferente}$$

Con el fin de determinar si las medias de los diversos bloques son todas iguales, se pueden examinar dos estimadores diferentes de la varianza de la población. Uno de los estimadores se basa en **la suma de los cuadrados dentro de los bloques *(SCb);*** el otro se basa en **la suma de los cuadrados entre los tratamientos y los bloques *(SCE***). Si la hipótesis nula es cierta, estos estimadores deben ser aproximadamente iguales; si es falsa, el estimador basado en la suma de los cuadrados entre grupos debe ser mayor.

El cuadrado medio del bloque es la estimación de la variación en el análisis de varianza. Se usa en el numerador de la prueba estadística ***F***.

El cuadrado medio del error es la estimación de la variación en el análisis de varianza. Se usa en el denominador de la estadística ***F***.

En el Análisis de Varianza, el estimador de la varianza entre los bloques ***(CMb)*** se calcula dividiendo la suma de los cuadrados de los bloques entre los grados de libertad entre los bloques ***(b-1).*** La varianza dentro de los tratamientos y los bloques, ***(CME),*** se estima dividiendo la suma de los cuadrados dentro de los tratamientos y bloques entre los grados de libertad dentro de los tratamientos y bloques ***(k-1)(b-1).*** Si en realidad hay una diferencia entre los bloques el ***(CMb),*** será significativamente ***mayor*** que el ***(CME).*** La prueba estadística se basa en la razón de las dos varianzas, ***CMb/CME***. La distribución de esta razón se conoce como la **distribución *F,*** por lo que el estadístico de prueba es:

$$F_{CALC.} = \frac{CM_{bloques}}{CME} = \frac{SC_{bloques}/_{g.l.}}{SCE/_{g.l.}}$$

La regla de decisión es rechazar la hipótesis nula de que no hay diferencia entre los bloques si al nivel de significancia α

$$\boldsymbol{F_{calc} \geq F_{\alpha,(b-1),(k-1)(b-1)}}$$

ISBN 978-607-8285-62-4

La suma de cuadrados es la cantidad calculada en el análisis de varianza y usada para obtener cuadrados medios para la prueba ***F***.

Para obtener la **Suma de Cuadrados** en un **diseño de bloques aleatorizados** se usan las siguientes fórmulas:

$$SCT = \sum_{i=1}^{k}\sum_{j=1}^{b} X_{ij}^2 - \frac{X_{..}^2}{bk}$$

$$SC_{tratamientos} = \sum_{i=1}^{k} \frac{X_{i.}^2}{b} - \frac{X_{..}^2}{bk}$$

$$SC_{bloques} = \sum_{j=1}^{b} \frac{X_{.j}^2}{k} - \frac{X_{..}^2}{bk}$$

$$SCE = SCT - SC_{tratamientos} - SC_{bloques}$$

Como verificación:

$$SCE = \sum_{I=1}^{K}\sum_{j=1}^{b} (X_{ij} - \bar{X}i. - \bar{X}_{.j} + \bar{X}_{..})^2$$

Nota importante: El obtener la ***SCE*** por diferencias puede dar lugar a tener un error si en los cálculos anteriores para obtener la ***SCT***, la ***SCtratamientos*** ó la ***SCbloques*** existe algún error, por lo que se recomienda obtener por separado la ***SCE.***

Los cuadrados medios son el cociente entre la suma de cuadrados y los grados de libertad.

Donde:

X_{ij} es cada uno de las observaciones
$X_{i.}$ es la suma de las observaciones de la muestra para el tratamiento i.
$X_{.j}$ es la suma de las observaciones de la muestra para el bloque .j
$X_{..}$ es la suma total de todas las observaciones de la muestra.
k es el número de tratamientos
b es el número de bloques

$$\boldsymbol{CM_{trat} = \frac{SC_{trat}}{k-1}} \qquad \boldsymbol{CM_{bloque} = \frac{SC_{bloque}}{b-1}} \qquad \boldsymbol{CME = \frac{SCE}{(k-1)(b-1)}}$$

ISBN 978-607-8285-62-4

Debido a que en el cálculo de varianzas entre y dentro de tratamientos y bloques hay varios pasos, el grupo completo de resultados se puede organizar en una tabla de análisis de varianza ***(ANOVA***) cuya estructura es la siguiente:

Fuente de variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	**$F_{calculada}$**
Factor 1 Tratamientos	$k-1$	SC_t	$CM_t = SC_t / g.l.$	$F_{calc} = CM_t / CME$
Bloques	$b-1$	SC_b	$CM_b = SC_b / g.l.$	$F_{calc} = CM_b / CME$
Error	$(k-1)(b-1)$	SCE	$CME = SCE / g.l.$	
Total	$bk-1$	SCT		

Para saber sobre **qué efecto tuvo en el análisis la creación de bloques, en comparación con el diseño completamente aleatorizado** se calcula **la eficiencia relativa estimada (RE):**

$$RE = (b-1)\ CM_{bloque} + b(k-1)CME\ /\ (bk-1)CME$$

La **eficiencia relativa** estimada nos indica el **número de réplicas de más** que se deberían usar en cada grupo de tratamiento en un diseño unifactorial para obtener la misma precisión ó sensibilidad al comparar las medias de los grupos de tratamientos, de lo que se necesitaría para el diseño aleatorio por bloques.

COMPARACIONES MÚLTIPLES: EL MÉTODO T DE TUKEY

El método T de Tukey es una prueba *a posteriori o post-hoc* para hacer comparaciones apareadas múltiples entre medias después de obtener una prueba *F* significativa en el análisis de varianza.

Es el método más recomendado por los estadígrafos.

Con la finalidad de determinar cuáles de las *k* medias son significativamente deferentes de las otras podemos utilizar el procedimiento de Tukey. Este método es un ejemplo de un procedimiento de comparación *post hoc* (o ***a posteriori),*** pues las hipótesis de interés son formuladas *después* de que los datos han sido inspeccionados.

Para usar el procedimiento de Tukey, simplemente se ordenan en forma descendente **las medias de los tratamientos** y se comparan las diferencias observadas entre cada par de promedios con el valor correspondiente al **rango ó alcance crítico**. Si $|\overline{X}_{i.} - \overline{X}_{i.'}| \geq$ **rango ó alcance crítico**, se concluye que las medias poblacionales $\mu_{i.}$ y $\mu_{j.}$ son diferentes.

ISBN 978-607-8285-62-4

El **rango ó alcance crítico** se obtiene entonces de la cantidad dada en la ecuación siguiente:

$$\boldsymbol{rango\ ó\ alcance\ crítico = q_{\alpha,k,(k-1)(b-1)}\sqrt{\frac{CME}{b}}}$$

Para el modelo de diseño de bloques aleatorizados, los tamaños de muestra de cada grupo de tratamientos son iguales por lo tanto ***b* será el número de bloques.**

Con el método de Tukey se puede establecer también un conjunto de intervalos de confianza estimados simultáneamente para las verdaderas diferencias entre cada par de medias. Lo anterior se logra sumando y restando el alcance o rango crítico a las diferencias en cada par de medias muestrales.

Un intervalo de confianza es un espacio calculado a partir de los datos de una muestra, que tiene una probabilidad dada de comprender el parámetro desconocido.

$$(\bar{X}_{i.} - \bar{X}_{i'}) - q_{\alpha,k,(k-1)(b-1)}\sqrt{\frac{CME}{b}} \leq (\mu_{i.} - \mu_{i.)}^{'} \leq (\bar{X}_{i.} - \bar{X}_{i'}) + q_{\alpha,k,(k-1)(b-1)}\sqrt{\frac{CME}{b}}$$

Los límites de confianza delimitan a un intervalo de confianza. Se calculan de los datos de la muestra y tienen una probabilidad dada de que el parámetro desconocido se ubique entre éstos.

El valor $q_{\alpha,k,(k-1)(b-1)}$ se obtiene de la tabla de puntos porcentuales del rango studentizado del apéndice buscando en $\alpha = 0.05\ ó 0.01\ según\ se\ indique\ en\ el\ problema$, k= Número de grupos ó tratamientos en general y (k-1)(b-1) =(Número de grupos o tratamientos menos 1 por número de bloques menos 1).

NOTA: Si en la tabla no hay ninguna entrada que corresponda exactamente a los grados de libertad especificados se puede tomar el más cercano al especificado o hacer una interpolación con los valores que se encuentren con los grados de libertad entre los cuales se encuentre el especificado.

ISBN 978-607-8285-62-4

1.3.2.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 1.3.2.1 DISEÑO DE BLOQUES AL AZAR

Una cadena de restaurantes de comida rápida tiene cuatro sucursales en determinada zona geográfica y desea evaluar su servicio. El director de investigación de mercados contrata 24 investigadores (clasificadores) con diversas experiencias en evaluación de servicios alimentarios. Después de las consultas preliminares, los 24 investigadores se dividen en seis bloques de cuatro. Los cuatro investigadores mas experimentados quedan ubicados en el bloque 1, los cuatro siguientes en el bloque 2 y así sucesivamente.

Dentro de cada bloque homogéneo se asignan al azar los cuatro clasificadores para evaluar el servicio de determinado restaurante bajo el siguiente criterio 0(baja) a 100(alta). Los resultados se presentan a continuación.

REST	**CLASIFICADORES**					
	1	**2**	**3**	**4**	**5**	**6**
1	70	77	76	80	84	78
2	61	75	67	63	66	68
3	82	88	90	96	92	98
4	74	76	80	76	84	86

a) Con un nivel de significancia de 0.05 ¿existe diferencia en la evaluación entre los cuatro restaurantes?.
b) Con un nivel de significancia de 0.05 ¿existe diferencia en la evaluación entre los seis clasificadores?.
c) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas ($1-\alpha=0.95$) de los tratamientos ó niveles del factor. ¿Cual o cuales restaurantes obtuvieron mayor puntaje y cuanto más?
d) Determine la eficiencia relativa del diseño aleatorizado en bloques en comparación con el diseño completamente aleatorizado

NOTA:
En este caso el bloque es el clasificador del restaurante, porque de ello nos valemos para efectuar el experimento; es lo que **NO** nos interesa evaluar como efecto para la evaluación del restaurante; lo que nos interesa es cuánto es el la evaluación de los restaurantes al compararse los cuatro restaurantes.

ISBN 978-607-8285-62-4

Prueba de hipótesis de cinco pasos para el Factor 1

Solución al inciso a.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

La hipótesis nula es que la evaluación promedio del servicio en los diferentes restaurantes es el mismo.

$$H_0:\ \mu_{1.} = \mu_{2.} = \ \mu_{3.} = \mu_{4.}$$

La hipótesis alternativa es que la evaluación promedio del servicio en los diferentes restaurantes no es el mismo.

$$H_1: \text{No todas las evaluaciones promedio son iguales}$$

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución F

$$F_{Calculada} = \frac{CMt}{CME} = \frac{SCt/_{g.l.}}{SCE/_{g.l.}}$$

Donde:

$$SCT = SCt + SCb + SCE$$

Empiece por calcular las sumas y medias para cada tratamiento así como la suma y media total:

REST	CLASIFICADORES						TOTAL	MEDIA
	1	2	3	4	5	6		
1	$X_{11} = 70$	$X_{12} = 77$	$X_{13} = 76$	$X_{14} = 80$	$X_{15} = 84$	$X_{16} = 78$	$X_{1.} = 465$	$\bar{X}_{1.} = 77.5$
2	$X_{21} = 61$	$X_{22} = 75$	$X_{23} = 67$	$X_{24} = 63$	$X_{25} = 66$	$X_{26} = 68$	$X_{2.} = 400$	$\bar{X}_{2.} = 66.66$
3	$X_{31} = 82$	$X_{32} = 88$	$X_{33} = 90$	$X_{34} = 96$	$X_{35} = 92$	$X_{36} = 98$	$X_{3.} = 546$	$\bar{X}_{3.} = 91.0$
4	$X_{41} = 74$	$X_{42} = 76$	$X_{43} = 80$	$X_{44} = 76$	$X_{45} = 84$	$X_{46} = 86$	$X_{4.} = 476$	$\bar{X}_{4.} = 79.33$
TOTAL	$X_{.1} = 287$	$X_{.2} = 316$	$X_{.3} = 313$	$X_{.4} = 315$	$X_{.5} = 326$	$X_{.6} = 330$	$X_{..} = 1887$	$\bar{X}_{..} = 78.625$

ISBN 978-607-8285-62-4

Para obtener la ***SCT*** obtenga entonces la desviación de cada observación de la media total, elevamos al cuadrado esas desviaciones y sumamos este resultado para las 24 observaciones. Para simplificar los cálculos podemos utilizar la siguiente fórmula abreviada:

Suma de Cuadrado Total

$$SCT = \sum_{I=1}^{4}\sum_{J=1}^{6} X_{ij}^2 - \frac{X_{..}^2}{bk} = (70^2 + 77^2 + \cdots + 86^2) - \frac{1887^2}{(6)(4)}$$
$$= 150{,}661 - 148{,}365.375 = \mathbf{2{,}295.63}$$

La fórmula abreviada para encontrar la ***SCt*** es:

Suma de cuadrados de tratamientos

$$SCt = \sum_{i=1}^{4} \frac{X_{i.}^2}{b} - \frac{X_{..}^2}{bk} = \left(\frac{465^2}{6} + \frac{400^2}{6} + \frac{546^2}{6} + \frac{476^2}{6}\right) - \frac{1887^2}{24}$$
$$= 36{,}037.5 + 26{,}666.66 + 49{,}686 + 37{,}762.66 - 148{,}365.375$$
$$= 150{,}152.82 - 148{,}365.375 = \mathbf{1787.45}$$

La fórmula abreviada para encontrar la ***SCb*** es:

Suma de cuadrados de bloque

$$SCb = \sum_{i=1}^{4} \frac{X_{.j}^2}{k} - \frac{X_{..}^2}{bk} = \left(\frac{287^2}{4} + \frac{316^2}{4} + \frac{313^2}{4} + \frac{315^2}{4} + \frac{326^2}{4} + \frac{330^2}{4}\right) - \frac{1887^2}{24}$$
$$= 20{,}592.25 + 24{,}964 + 24{,}492.25 + 24{,}806.25 + 26{,}569$$
$$+ 27{,}225 - 148{,}365.375 = \mathbf{283.375}$$

Por último determine la ***SCE*** a través de la resta:

$$SCE = SCT - SC_{tratamientos} - SC_{bloques}$$

$$SCE = 2{,}295.625 - 1{,}787.445 - 283.375 = \mathbf{224.805}$$

Una manera de obtener la ***SCE*** en forma directa y como comprobación sería:

Suma de cuadrados del error

$$SCE = \sum_{I=1}^{4}\sum_{J=1}^{6} (X_{ij} - \bar{X}i. - \bar{X}_{.j} + \bar{X}_{..})^2 = (70 - 77.5 - 71.75 + 78.625)^2$$
$$+ (77 - 77.5 - 79 + 78.625)^2 + \cdots$$
$$+ (78 - 77.5 - 82.5 + 78.625)^2$$
$$+ (61 - 66.67 - 71.75 + 78.625)^2$$
$$+ (75 - 66.67 - 79 + 78.625)^2 + \cdots$$
$$+ (68 - 66.67 - 82.5 + 78.625)^2$$
$$+ (82 - 91 - 71.75 + 78.625)^2 + (88 - 91 - 79 + 78.625)^2$$
$$+ \cdots + (98 - 91 - 82.5 + 78.625)^2$$
$$+ (74 - 79.33 - 71.75 + 78.625)^2$$
$$+ (76 - 79.33 - 79 + 78.625)^2 + \cdots$$
$$+ (86 - 79.33 - 82.5 + 78.625)^2 = 32.6 + 98.7 + 53.3 + 40.2$$
$$= \mathbf{224.8}$$

ISBN 978-607-8285-62-4

Para encontrar el valor calculado de ***F,*** trabaje con la tabla de ANOVA. El término de cuadrado de la media es otra expresión que se utiliza para un cálculo de la varianza. El cuadrado de la media para los tratamientos es SCt dividida entre sus grados de libertad. El resultado es el cuadrado de la media para los tratamientos y se escribe CMt.

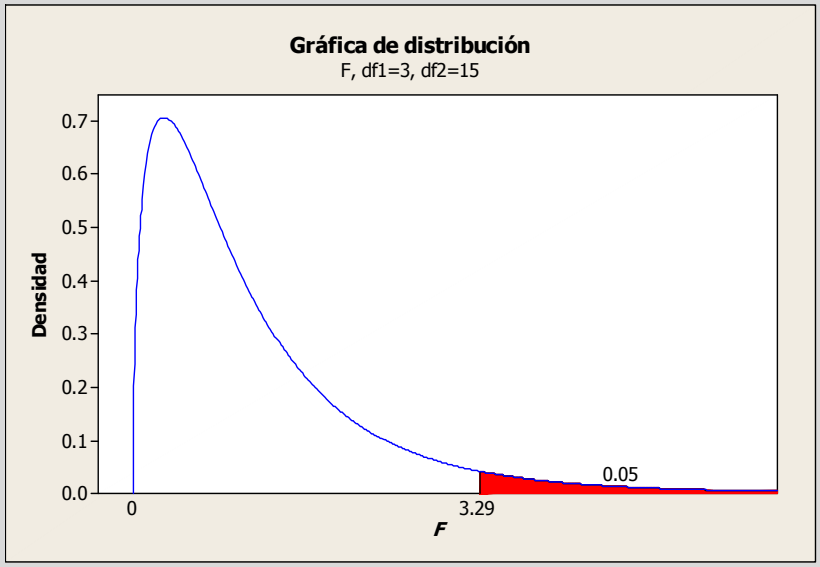
Tabla de ANOVA

TABLA DE ANOVA

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada}$***
Tratamientos	$v_1 = k - 1 = 4 - 1 = 3\ \boldsymbol{g.l.}$	$SCt = 1{,}787.445$	$CMt = {SCt}/{g.l.} = {1{,}787.445}/{3} = 595.815$	$F_{calc.} = \frac{CM_t}{CME} = \frac{595.815}{14.987} = 39.755$
Bloques	$v_1 = b - 1 = 6 - 1 = 5\ \boldsymbol{g.l.}$	$SCb = 283.375$	$CMt = {SCb}/{g.l.} = {283.375}/{5} = 56.675$	$F_{calc.} = \frac{CM_b}{CME} = \frac{56.675}{14.987} = 3.782$
Error	$v_2 = (k-1)(b-1) = (3)(5) = 15\ \boldsymbol{g.l.}$	$SCE = 224.805$	$CME = {SCE}/{g.l.} = {224.805}/{15} = 14.987$	
Total	24-1=23	$SCT = 2{,}295.625$		

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).



ISBN 978-607-8285-62-4

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

La regla de decisión es rechazar H_0 si el valor calculado de F es mayor a **3.29.**

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como el valor calculado de F es de 39.755 que es mayor al valor crítico de 3.29; por lo tanto la hipótesis nula se rechaza y llegamos a la conclusión de que no todas las medias de la población son iguales, es decir al menos una de ellas es diferente.

Administrativa: Existe evidencia suficiente para concluir que estadísticamente al menos el nivel de servicio en uno de los restaurantes no es el mismo.

NOTA: En este punto sólo podemos llegar a la conclusión de que existe una diferencia en al menos una de las medias de tratamiento. No podemos determinar qué grupo o grupos de tratamiento difieren ni por qué cantidad difieren unos de otros.

Prueba de hipótesis de cinco pasos para el Bloque

<u>Solución al inciso b.</u>

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

La hipótesis nula es que la evaluación promedio del servicio en los diferentes bloques es el mismo.

$$H_0:\ \mu_{.1} = \mu_{.2} = \ \mu_{.3} = \mu_{.4} = \mu_{.5} = \mu_{.6}$$

La hipótesis alternativa es que la evaluación promedio del servicio en los diferentes restaurantes no es el mismo.

$$H_1: \text{No todas las evaluaciones promedio son iguales}$$

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

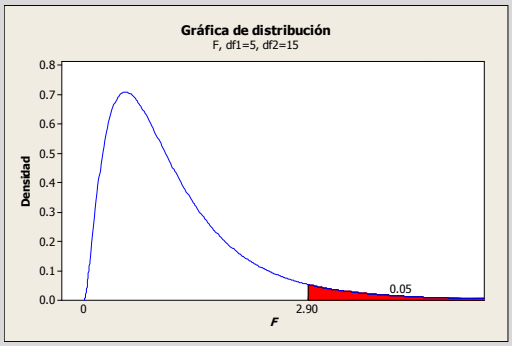
Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución *F*

$$F_{Calculada} = \frac{CMb}{CME} = \frac{SCb/_{g.l.}}{SCE/_{g.l.}} = \mathbf{3.782}$$

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).



Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

La regla de decisión es rechazar H_0 si el valor calculada de F es mayor a **2.90**

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como el valor calculado de F es de 3.782 que es mayor al valor crítico de 2.90; por lo tanto la hipótesis nula se rechaza y llegamos a la conclusión de que no todas las medias de la población son iguales, es decir al menos una de ellas es diferente

Administrativa: Existe evidencia suficiente para concluir que estadísticamente al menos el nivel de servicio en uno de los bloques no es el mismo

NOTA: En este punto sólo podemos llegar a la conclusión de que existe una diferencia en al menos una de las medias de bloques. No podemos determinar qué grupo o grupos de bloques difieren ni por qué cantidad difieren unos de otros.

ISBN 978-607-8285-62-4

Prueba post-hoc ó aposteriori. Prueba de comparaciones múltiples de Tukey.

Solución al inciso c.

El método ***T* de Tukey** permite examinar, en forma simultánea, comparaciones entre todos los pares de grupos mediante los siguientes pasos:

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

$$\bar{x}_{3.} = 91;\ \ \bar{x}_{4.} = 79.33;\ \ \bar{x}_{1.} = 77.5;\ \bar{x}_{2.} = 66.67$$

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

$$\bar{x}_{3.} - \bar{x}_{4.} = 91 - 79.33 = 11.67$$

$$\bar{x}_{3.} - \bar{x}_{1.} = 91 - 77.5 = 13.5$$

$$\bar{x}_{3.} - \bar{x}_{2.} = 91 - 66.67 = 24.33$$

$$\bar{x}_{4.} - \bar{x}_{1.} = 79.33 - 77.5 = 1.83$$

$$\bar{x}_{4.} - \bar{x}_{2.} = 79.33 - 66.67 = 12.66$$

$$\bar{x}_{1.} - \bar{x}_{2.} = 77.5 - 66.67 = 10.83$$

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$$rango\ ó\ alcance\ crítico = q_{\alpha,k,(k-1)(b-1)}\sqrt{\frac{CME}{b}}$$

$$rango\ ó\ alcance\ crítico = q_{0.05,4,(4-1)(6-1)}\sqrt{\frac{CME}{b}}$$

$$rango\ ó\ alcance\ crítico = q_{0.05,4,15}\sqrt{\frac{CME}{b}}$$

Para determinar el rango ó alcance crítico se usa la tabla de puntos porcentuales del rango studentizado para:

$\boldsymbol{\alpha = 0.05, k = 4\ y\ (k-1)(b-1) = 15, el\ valor\ crítico\ superior\ de\ q_{0.05,4,15} es\ 4.08}$. Por lo tanto, al utilizar la formula para el rango ó alcance crítico se tiene,

$$rango\ ó\ alcance\ crítico = 4.08\sqrt{\frac{14.986}{6}} = \mathbf{6.448}$$

ISBN 978-607-8285-62-4

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

$\bar{x}_{3.} - \bar{x}_{4.} = 91 - 79.33 = 11.67 > 6.448$; la prueba es (S) y $\mu_{3.} > \mu_{4.}$

$\bar{x}_{3.} - \bar{x}_{1.} = 91 - 77.5 = 13.5 > 6.448$; la prueba es (S) y $\mu_{3.} > \mu_{1.}$

$\bar{x}_{3.} - \bar{x}_{2.} = 91 - 66.67 = 24.33 > 6.448$; la prueba es (S) y $\mu_{3.} > \mu_{2.}$

$\bar{x}_{4.} - \bar{x}_{1.} = 79.33 - 77.5 = 1.83 < 6.448$; la prueba es (NS) y $\mu_{4.} = \mu_{1.}$

$\bar{x}_{4.} - \bar{x}_{2.} = 79.33 - 66.67 = 12.66 > 6.448$; la prueba es (S) y $\mu_{4} > \mu_{2.}$

$\bar{x}_{1.} - \bar{x}_{2.} = 77.5 - 66.67 = 10.83 > 6.448$; la prueba es (S) y $\mu_{1.} > \mu_{2.}$

Paso 5. Construir la gráfica de medias.

Paso 5.- Construir la gráfica de medias[3]:

```
Rest    Media    -+---------+---------+---------+--------
1     77.5000                  (---*---)
2     66.6667    (---*----)
3     91.0000                                    (---*---)
4     79.3333                     (---*---)
                 -+---------+---------+---------+--------
               64.0      72.0      80.0      88.0
```

Nótese que todos los contrastes excepto $|\bar{X}_{4.} - \bar{X}_{1.}|$ son mayores que el rango ó alcance crítico. Por lo tanto se puede llegar a la conclusión de que hay evidencia de una diferencia importante en la clasificación promedio entre todos los pares de restaurantes, excepto para los restaurantes 1(A) y 4(D). Lo que es más, el restaurante 3 (C) tiene las clasificaciones más altas (es decir, es el preferido), en tanto que el restaurante 2 (B) tiene el más bajo (es decir, es el menos preferido)

Paso 6. Construir los intervalos de confianza de cada par de medias.

Paso 6.- Establecer el conjunto de intervalos de confianza:

1. $(\bar{X}_{i.} - \bar{X}_{i.}^{'}) - q_{\alpha,k,(k-1)(b-1)}\sqrt{\frac{CME}{b}} \leq (\mu_{i.} - \mu_{i.)}^{'} \leq (\bar{X}_{i.} - \bar{X}_{i.}^{'}) + q_{\alpha,k,(k-1)(b-1)}\sqrt{\frac{CME}{b}}$

$$(\bar{X}_{3.} - \bar{X}_{4.}) - q_{0.05,4,15)}\sqrt{\frac{14.986}{6}} \leq (\mu_{3.} - \mu_{4.}) \leq (\bar{X}_{3.} - \bar{X}_{4.}) + q_{0.05,4,15)}\sqrt{\frac{14.986}{6}}$$

[3] Contruida con el software estadístico Minitab 15

ISBN 978-607-8285-62-4

Conclusiones.

$$(91-79.33)-4.08\sqrt{\frac{14.986}{6}} \leq (\mu_{3.}-\mu_{4.}) \leq (91-79.33)+4.08\sqrt{\frac{14.986}{6}}$$
$$11.67-6.448 \leq (\mu_{3.}-\mu_{4.}) \leq 11.67+6.448$$

$$\mathbf{5.22} \leq (\mu_{3.}-\mu_{4.}) \leq \mathbf{18.11}$$

Conclusión: Como ambos límites del intervalo son positivos podemos decir que estadísticamente el restaurante 3 quedó mejor evaluado que el restaurante 4 por un mínimo de 5.22 y un máximo de 18.11 puntos.

2. $(\bar{X}_{3.}-\bar{X}_{1.})-q_{0.05,4,15)}\sqrt{\frac{14.986}{6}} \leq (\mu_{3.}-\mu_{1.}) \leq (\bar{X}_{3.}-\bar{X}_{1.})+q_{0.05,4,15)}\sqrt{\frac{14.986}{6}}$

$$(91-77.5)-4.08\sqrt{\frac{14.986}{6}} \leq (\mu_{3.}-\mu_{1.}) \leq (91-77.5)+4.08\sqrt{\frac{14.986}{6}}$$
$$13.5-6.448 \leq (\mu_{3.}-\mu_{1.}) \leq 13.5+6.448$$

$$\mathbf{7.052} \leq (\mu_{3.}-\mu_{1.}) \leq \mathbf{19.948}$$

Conclusión: Como ambos límites del intervalo son positivos podemos decir que estadísticamente el restaurante 3 quedó mejor evaluado que el restaurante 1 por un mínimo de 7.052 y un máximo de 19.948 puntos.

3. $(\bar{X}_{3.}-\bar{X}_{2.})-q_{0.05,4,15)}\sqrt{\frac{14.986}{6}} \leq (\mu_{3.}-\mu_{2.}) \leq (\bar{X}_{3.}-\bar{X}_{2.})+q_{0.05,4,15)}\sqrt{\frac{14.986}{6}}$

$$(91-66.67)-4.08\sqrt{\frac{14.986}{6}} \leq (\mu_{3.}-\mu_{2.}) \leq (91-66.67)+4.08\sqrt{\frac{14.986}{6}}$$
$$24.33-6.448 \leq (\mu_{3.}-\mu_{2.}) \leq 24.33+6.448$$

$$\mathbf{17.882} \leq (\mu_{3.}-\mu_{2.}) \leq \mathbf{30.778}$$

Conclusión: Como ambos límites del intervalo son positivos podemos decir que estadísticamente el restaurante 3 quedó mejor evaluado que el restaurante 2 por un mínimo de 17.882 y un máximo de 30.778 puntos.

4. $(\bar{X}_{4.}-\bar{X}_{1.})-q_{0.05,4,15)}\sqrt{\frac{14.986}{6}} \leq (\mu_{4.}-\mu_{1.}) \leq (\bar{X}_{4.}-\bar{X}_{1.})+q_{0.05,4,15)}\sqrt{\frac{14.986}{6}}$

$$(79.33-77.57)-4.08\sqrt{\frac{14.986}{6}} \leq (\mu_{4.}-\mu_{1.}) \leq (79.33-77.5)+4.08\sqrt{\frac{14.986}{6}}$$
$$1.83-6.448 \leq (\mu_{4.}-\mu_{1.}) \leq 1.83+6.448$$

$$\mathbf{-4.618} \leq (\mu_{4.}-\mu_{1.}) \leq \mathbf{8.278}$$

Conclusión: Como el intervalo de confianza pasa por cero, podemos decir que estadísticamente el la evaluación de los restaurantes 4 y 1 es la misma.

ISBN 978-607-8285-62-4

5. $(\bar{X}_{4.} - \bar{X}_{2.}) - q_{0.05,4,15)}\sqrt{\frac{14.986}{6}} \leq (\mu_{4.} - \mu_{2.}) \leq (\bar{X}_{4.} - \bar{X}_{2.}) + q_{0.05,4,15)}\sqrt{\frac{14.986}{6}}$

$$(79.33 - 66.67) - 4.08\sqrt{\frac{14.986}{6}} \leq (\mu_{4.} - \mu_{2.}) \leq (79.33 - 66.67) + 4.08\sqrt{\frac{14.986}{6}}$$

$$12.66 - 6.448 \leq (\mu_{4.} - \mu_{2.}) \leq 12.66 + 6.448$$

$$\mathbf{6.212} \leq (\mu_{4.} - \mu_{2.}) \leq \mathbf{19.108}$$

Conclusión: Como ambos límites del intervalo son positivos podemos decir que estadísticamente el restaurante 4 quedó mejor evaluado que el restaurante 2 por un mínimo de 6.212 y un máximo de 19.108 puntos.

6. $(\bar{X}_{1.} - \bar{X}_{2.}) - q_{0.05,4,15)}\sqrt{\frac{14.986}{6}} \leq (\mu_{1.} - \mu_{2.}) \leq (\bar{X}_{1.} - \bar{X}_{2.}) + q_{0.05,4,15)}\sqrt{\frac{14.986}{6}}$

$$(77.5 - 66.67) - 4.08\sqrt{\frac{14.986}{6}} \leq (\mu_{1.} - \mu_{2.}) \leq (77.5 - 66.67) + 4.08\sqrt{\frac{14.986}{6}}$$

$$10.83 - 6.448 \leq (\mu_{1.} - \mu_{2.}) \leq 10.83 + 6.448$$

$$\mathbf{4.382} \leq (\mu_{1.} - \mu_{2.}) \leq \mathbf{17.278}$$

Conclusión: Como ambos límites del intervalo son positivos podemos decir que estadísticamente el restaurante 1 quedó mejor evaluado que el restaurante 2 por un mínimo de 4.382 y un máximo de 17.278 puntos.

Eficiencia relativa estimada (RE)

<u>Solución al inciso d.</u>

La eficiencia relativa estimada (RE) del diseño aleatorizado en bloques, en comparación con el diseño completamente aleatorizado, se puede calcular con la siguiente ecuación:

$$RE = \frac{(b-1)(CM_b) + b(k-1)(CME)}{(bk-1)CME}$$

Por lo tanto de la tabla de Anova se tiene:

$$RE = \frac{(b-1)(CM_b) + b(k-1)(CME)}{(bk-1)CME} = \frac{(6-1)(56.67) + 6(4-1)(14.98)}{(6x4-1)(14.98)} = \mathbf{1.605}$$

Conclusión:

La eficiencia relativa estimada nos indica que se necesitarían 1.605 veces tantas observaciones en cada grupo de tratamientos en un diseño ANOVA en un sentido o de una vía para obtener la misma precisión al comparar las medias de los grupos de tratamiento, de lo que se necesitaría para el diseño aleatorio en bloques, es decir hubiéramos tenido que utilizar 24 x 0.605= 15 observaciones más para obtener la precisión de este diseño por bloques.

ISBN 978-607-8285-62-4

1.3.2.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 1.3.2.1 DISEÑO DE BLOQUES AL AZAR

Se proporcionan los siguientes datos para una **ANOVA** de **bloques al azar:**

Tratamientos	**Bloques**		
	1	**2**	**3**
1	46	37	44
2	31	26	35

a) Con un nivel de significancia de 0.05 ¿existe diferencia en las medias de los dos tratamientos?.
b) Con un nivel de significancia de 0.05 ¿existe diferencia en las medias de los tres bloques?.
c) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿Cual tratamiento resulto mayor y por cuanto más?
d) Determine la eficiencia relativa del diseño aleatorizado en bloques en comparación con el diseño completamente aleatorizado.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Prueba de hipótesis para el Factor 1

Solución al inciso a.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución *F*

$$F_{Calculada} = \frac{CMt}{CME} = \frac{SCt/_{g.l.}}{SCE/_{g.l.}}$$

Donde:

$$SCT = SCt + SCb + SCE$$

Empiece por calcular las sumas y medias para cada tratamiento así como la suma y media total:

TRATAMIENTO	BLOQUES			TOTAL	MEDIA
	1	2	3		
1	46	37	44		
2	31	26	35		
TOTAL					

Para obtener la **SCT** obtenga entonces la desviación de cada observación de la media total, elevamos al cuadrado esas desviaciones y sumamos este resultado para las 24 observaciones. Para simplificar los cálculos podemos utilizar la siguiente fórmula abreviada:

Suma de Cuadrados Total.

$$SCT = \sum_{i=1}^{k}\sum_{j=1}^{b} X_{ij}^2 - \frac{X^2..}{bk} =$$

Posteriormente determine la **SCt** ó la suma de los cuadrados de los errores debido a los tratamientos. Ésta es la suma de las diferencias al cuadrado que existen entre cada media de tratamiento $(\bar{X}i.)$ y la media total $(\bar{X}..)$.

La fórmula abreviada para encontrar la **SCt** es:

Suma de Cuadrados de tratamientos.

$$SCtratamientos = \sum_{i=1}^{k} \frac{X_{i.}^2}{b} - \frac{X_{..}^2}{bk} =$$

Ahora determine la **SCb** ó la suma de los cuadrados de los errores debido a los bloques. Ésta es la suma de las diferencias al cuadrado que existen entre cada media de bloque $(\bar{X}.i)$ y la media total $(\bar{X}..)$.

ISBN 978-607-8285-62-4

La fórmula abreviada para encontrar la **SCb** es:

Suma de Cuadrados de bloques.

$$SCbloques = \sum_{i=1}^{b} \frac{X_{.j}^2}{k} - \frac{X_{..}^2}{bk} =$$

Suma de Cuadrados del Error.

Por último determine la **SCE** a través de la resta:

$$SCE = SCT - SC_{tratamientos} - SC_{bloques}$$

$$SCE =$$

Una manera de obtener la **SCE** en forma directa y como comprobación sería:

$$SCE = \sum_{I=1}^{K} \sum_{J=1}^{b} (X_{ij} - \bar{X}i. - \bar{X}_{.j} + \bar{X}_{..})^2 =$$

Para encontrar el valor calculado de ***F,*** trabaje con la tabla de ANOVA. El término de cuadrado de la media es otra expresión que se utiliza para un cálculo de la varianza. El cuadrado de la media para los tratamientos es **SCt** dividida entre sus grados de libertad. El resultado es el cuadrado de la media para los tratamientos y se escribe **CMt.**

Tabla de ANOVA.

TABLA DE ANOVA

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada}$***
Tratamientos	$v_1 = k - 1 =$	$SCt =$	$CMt = {SCt}/{g.l.} =$	$F_{calc.} = \frac{CM_t}{CME} =$
Bloques	$v_1 = b - 1 =$	$SCb =$	$CMt = {SCb}/{g.l.} =$	$F_{calc.} = \frac{CM_b}{CME} =$
Error	$v_2 = (k-1)(b-1) =$	$SCE =$	$CME = {SCE}/{g.l.} =$	
Total	$(k)(b) - 1 =$	$SCT =$		

ISBN 978-607-8285-62-4

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

Prueba de hipótesis para bloques

<u>Solución al inciso b.</u>

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución *F*

$$F_{Calculada} = \frac{CMb}{CME} = \frac{SCb/_{g.l.}}{SCE/_{g.l.}} =$$

ISBN 978-607-8285-62-4

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

Comparaciones múltiples para los tratamientos. Método T de Tukey

<u>Solución al inciso c.</u>

El método ***T* de Tukey** permite examinar, en forma simultánea, comparaciones entre todos los pares de grupos mediante los siguientes pasos:

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$$\boldsymbol{rango\ ó\ alcance\ crítico = q_{\alpha,k,(k-1)(b-1)}\sqrt{\frac{CME}{b}} =}$$

ISBN 978-607-8285-62-4

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

Paso 5. Construir la gráfica de medias.

Paso 5.- Construir la gráfica de medias:

Paso 6. Construir los intervalos de confianza de cada par de medias.

Paso 6.- Establecer el conjunto de intervalos de confianza:

$$(\bar{X}_{i.} - \bar{X}_{i'}) - q_{\alpha,k,(k-1)(b-1)}\sqrt{\frac{CME}{b}} \leq (\mu_{i.} - \mu_{i.)}^{'} \leq (\bar{X}_{i.} - \bar{X}_{i'}) + q_{\alpha,k,(k-1)(b-1)}\sqrt{\frac{CME}{b}}$$

Conclusiones.

Conclusiones:

Eficiencia relativa RE

<u>Solución al inciso d.</u>

$$RE = \frac{(b-1)(CM_b) + b(k-1)(CME)}{(bk-1)CME} =$$

Conclusión:

ISBN 978-607-8285-62-4

1.3.2.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. Las respuestas a este ejercicio de autoevaluación se encuentran al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 1.3.2.1
DISEÑO DE BLOQUES AL AZAR

Se proporcionan los siguientes datos para una ANOVA de bloques al azar:

Tratamientos	Bloques		
	1	**2**	**3**
1	11	9	8
2	17	14	12
3	8	9	8

a) Con un nivel de significancia de 0.05 ¿existe diferencia en las medias de los dos tratamientos?.
b) Con un nivel de significancia de 0.05 ¿existe diferencia en las medias de los tres bloques?.
c) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas ($1-\alpha=0.95$) de los tratamientos ó niveles del factor. ¿Cual o cuales tratamientos resultaron más grandes y por cuanto más?
d) Determine la eficiencia relativa del diseño aleatorizado en bloques en comparación con el diseño completamente aleatorizado

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ISBN 978-607-8285-62-4

Prueba de hipótesis para los tratamientos. Diseño de bloques al azar.

Solución al inciso a.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

Empiece por calcular las sumas y medias para cada tratamiento así como la suma y media total:

TRATAMIENTO	BLOQUES			TOTAL	MEDIAS
	1	2	3		
1	46	37	44		
2	31	26	35		
TOTAL					

Suma de Cuadrados Total.

$$\boldsymbol{SCT} =$$

Suam de cuadrados de tratamientos.

$$\boldsymbol{SCtratamientos} =$$

Suam de Cuadrados de bloques.

$$\boldsymbol{SCbloques} =$$

ISBN 978-607-8285-62-4

Suma de Cuadrados del Error.

Por último determine la **SCE** a través de la resta:

$SCE =$

Una manera de obtener la **SCE** en forma directa y como comprobación sería:

$SCE =$

TABLA DE ANOVA

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	$F_{calculada}$
Tratamientos				
Bloques				
Error				
Total				

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

ISBN 978-607-8285-62-4

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

Prueba de hipótesis para los bloques. Diseño de bloques al azar.

<u>Solución al inciso b.</u>

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución *F*

$$\boldsymbol{F_{Calculada}} =$$

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

ISBN 978-607-8285-62-4

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

Prueba *T* de Tukey de comparaciones múltiples para los tratamientos.

Solución al inciso c.

El método ***T* de Tukey** permite examinar, en forma simultánea, comparaciones entre todos los pares de grupos mediante los siguientes pasos:

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$$rango\ \acute{o}\ alcance\ crítico =$$

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

ISBN 978-607-8285-62-4

Paso 5. Construir la gráfica de medias.	**Paso 5.- Construir la gráfica de medias:**
Paso 6. Construir los intervalos de confianza de cada par de medias.	**Paso 6.- Establecer el conjunto de intervalos de confianza:**
Conclusiones.	**Conclusiones:**
Eficiencia relativa: RE	**<u>Solución al inciso d.</u>** $\boldsymbol{RE} =$
Conclusión.	**Conclusión:**

ISBN 978-607-8285-62-4

1.3.2 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 1.3.2 DISEÑO DE BLOQUES AL AZAR

NOTA:
El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

1.3.2.1 Se desea resaltar una campaña publicitaria, tratando de resaltar el menor desgaste que sufre una llanta de automóvil. Para esto, el anuncio pretende hacer la comparación de esta marca en particular contra otras tres que se encuentran en el mercado, por lo que el experimento utilizó cuatro autos de una misma marca y en condiciones más o menos similares. Un factor que se debería considerar en el desgaste vendría siendo la forma como el conductor maneja el automóvil por lo que este factor se registra y se obtienen los siguientes resultados:

MARCA	CONDUCTOR				TOTAL	MEDIA
	1	2	3	4		
A	17	14	12	13	56	14
B	14	14	12	11	51	12.75
C	13	13	10	11	47	11.75
D	13	8	9	9	39	9.75

a) Con un nivel de significancia de 0.05 ¿Existe evidencia suficiente que indique que cualquier marca sufre igual desgaste en promedio?.
b) Con un nivel de significancia de 0.05 ¿Existe evidencia suficiente que indique que la forma de manejo del conductor del automóvil no afecta al desgaste de las llantas?.
c) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas ($1-\alpha=0.95$) de los tratamientos ó niveles del factor. ¿Cual o cuales marcas tuvieron mayor desgaste y cuanto más?
d) Determine la eficiencia relativa del diseño aleatorizado en bloques en comparación con el diseño completamente aleatorizado.

1.3.2.2 La compañía Colorama, S.A. se dedica a la producción y comercialización de pinturas. Pretenden lanzar al mercado cuatro tipos de marcas diferentes y quieren saber cuanto tiempo aproximado se puede dar de garantía con respecto a la duración de la pintura según su desgaste. Para esto, se realizó un experimento para probar cuantas lavadas aguanta cada marca en varias paredes. Un factor a considerar en las lavadas que aguantan las pinturas es el tipo de pared en donde se probó cada pintura; se le pide a usted comprobar si este factor afecta el número de lavadas que aguanta cada pintura o si no hay que preocuparse por ello. Los datos son los siguientes:

ISBN 978-607-8285-62-4

MARCA	SUPERFICIE			
	Ladrillo	**Cemento**	**Tirol**	**Block**
Pigmento 34	10	9	7	11
Pigmento 201	12	12	11	9
Pigmento 01	15	22	18	12
Pigmento 83	8	10	9	7

a) Con un nivel de significancia de 0.05 ¿Existe evidencia suficiente que indique que las diferentes marcas de pintura aguantan el mismo número promedio de lavadas?

b) Con un nivel de significancia de 0.05 ¿Existe evidencia suficiente que indique que el tipo de pared no afecta el número de lavadas que aguanta cada pintura?

c) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿Cual o cuáles marcas aguantaron más lavadas y cuántas más?

d) Determine la eficiencia relativa del diseño aleatorizado en bloques en comparación con el diseño completamente aleatorizado.

1.3.2.1 EJEMPLO ILUSTRATIVO EN EXCEL

EJEMPLO ILUSTRATIVO INTEGRAL EN EXCEL 1.3.2.1 DISEÑO DE BLOQUES AL AZAR

Una cadena de restaurantes de comida rápida tiene cuatro sucursales en determinada zona geográfica y desea evaluar su servicio. El director de investigación de mercados contrata 24 investigadores (clasificadores) con diversas experiencias en evaluación de servicios alimentarios.

Después de las consultas preliminares, los 24 investigadores se dividen en seis bloques de cuatro. Los cuatro investigadores mas experimentados quedan ubicados en el bloque 1, los cuatro siguientes en el bloque 2 y así sucesivamente.

Dentro de cada bloque homogéneo se asignan al azar los cuatro clasificadores para evaluar el servicio de determinado restaurante bajo el siguiente criterio 0(baja) a 100(alta). Los resultados se presentan a continuación.

ISBN 978-607-8285-62-4

REST	CLASIFICADORES					
	1	**2**	**3**	**4**	**5**	**6**
1	70	77	76	80	84	78
2	61	75	67	63	66	68
3	82	88	90	96	92	98
4	74	76	80	76	84	86

a) Con un nivel de significancia de 0.05 ¿existe diferencia en la evaluación entre los cuatro restaurantes?.
b) Con un nivel de significancia de 0.05 ¿existe diferencia en la evaluación entre los seis clasificadores?.

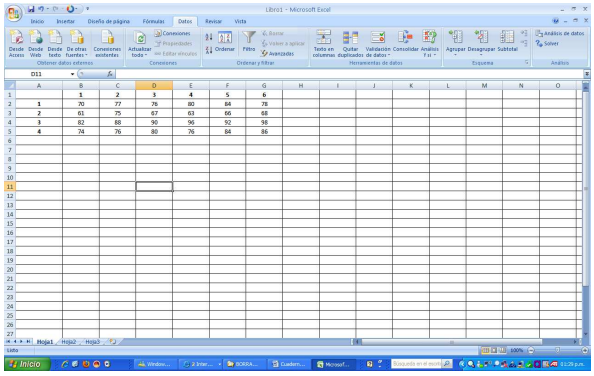
Pruebas de hipótesis para el Factor 1 y los bloques

Solución al inciso a y b.

Cuando el número de observaciones en cada tratamiento es extenso y/o existen muchos tratamientos, los cálculos manuales son tediosos. Existen muchos paquetes de software que pueden mostrar los resultados.

Comenzamos introduciendo los datos en la hoja de Excel, tal y como se muestra a continuación:

Hoja de Excel.

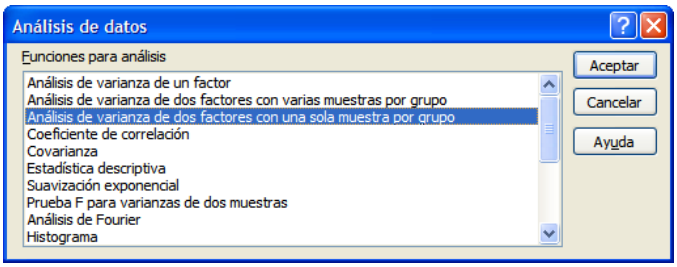


Tenemos entonces un ***diseño en bloques aleatorizados con una unidad por casilla***. Como este diseño equivale a un diseño de dos factores (restaurantes y clasificadores) sin interacción, seleccionamos la opción ***Análisis de datos*** del menú ***Datos***, utilizaremos la opción ***Análisis de la varianza de dos factores con una muestra por grupo***, del cuadro ***Análisis de datos*** de la figura siguiente:

Excel dispone de varias herramientas de análisis útiles para realizar análisis de varianza. La opción ***Análisis de datos*** del menú ***Datos (Microsoft Office Excel 2007)*** nos lleva al cuadro de diálogo ó ventana de la figura siguiente:

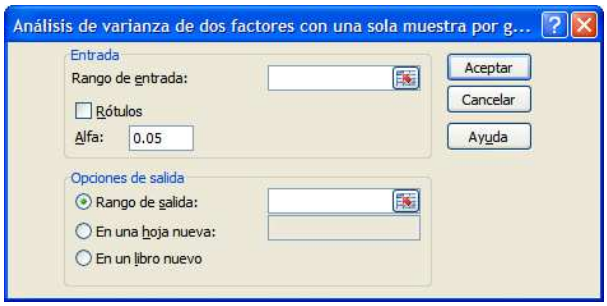
ISBN 978-607-8285-62-4

Cuadro de diálogo: Análisis de datos.



En la lista ***Funciones para análisis,*** elija la modalidad de ***Análisis de varianza de dos factores con una sola muestra por grupo*** y oprima el botón Aceptar para obtener el siguiente cuadro de dialogo rellenando su pantalla de entrada:

Cuadro de diálogo: Análisis de varianza de dos factores con una sola muestra o réplica por grupo (Diseño de bloques al azar)



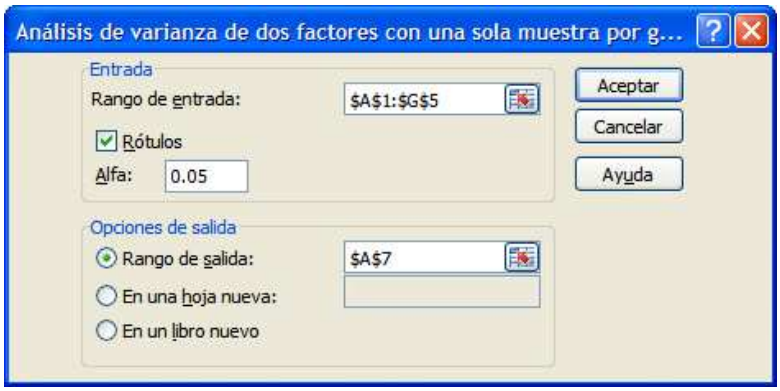
En el cuadro ***Rango de entrada*** introduzca, (seleccionando con el cursor las celdas donde están los datos incluyendo los rótulos), la referencia de celda correspondiente al rango de datos que está analizando. La referencia deberá contener dos o más rangos adyacentes organizados en columnas o filas.

Si la primera fila y primera columna del rango de entrada contiene rótulos, active la casilla de verificación Rótulos . Esta casilla de verificación debe quedar desactivada si el rango de entrada carece de rótulos; Microsoft Office Excel 2007 generará los rótulos de datos correspondientes para la tabla de resultados. **Deje sin cambio el campo Alfa con el valor de 0.05** (nivel con el que desee evaluar los valores críticos de la función estadística ***F).*** El nivel ***alfa*** es un nivel de importancia relacionado con la probabilidad de que haya un error de tipo I (rechazar una hipótesis verdadera).

En cuanto a las ***opciones de salida***, en el campo ***Rango de salida*** introduzca la referencia, (dando un clic), correspondiente a la celda superior izquierda de la tabla de resultados, en este caso la celda A7 y oprima el botón Aceptar.

ISBN 978-607-8285-62-4

Cuadro de diálogo: Análisis de varianza de dos factores con una sola muestra o réplica por grupo (Diseño de bloques al azar



A continuación se muestra la salida del análisis de la varianza de dos factores con una sola muestra por grupo:

Salida del análisis de varianza de bloques al azar.

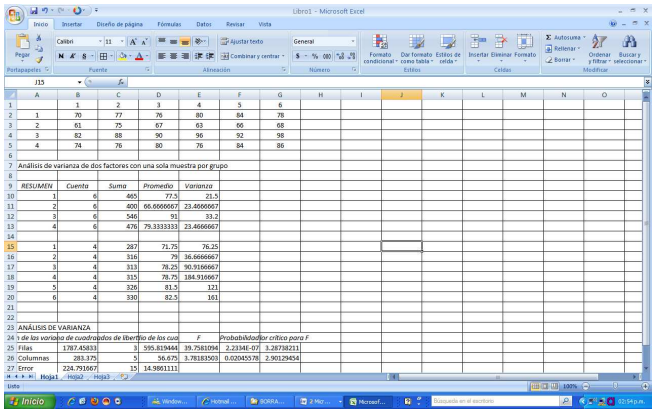
-Si el *valor p* es menor que o igual al nivel Alfa (a), rechace la hipótesis nula en favor de la hipótesis alternativa.

-Si $0.01 \leq p \leq 0.05$ se dice que la prueba es significativa (S).

-Si $p < 0.01$ se dice que la prueba es altamente significativa (AS).

-Si el *valor p* es mayor que el nivel Alfa (a), no rechace la hipótesis nula. Si $p > 0.05$ se dice que la prueba es no significativa (NS).

Los niveles de significancia más utilizado son 0.05 y 0.01



	A	B	C	D	E	F	G
1		1	2	3	4	5	6
2	1	70	77	76	80	84	78
3	2	61	75	67	63	66	68
4	3	82	88	90	96	92	98
5	4	74	76	80	76	84	86
6							
7	Análisis de varianza de dos factores con una sola muestra por grupo						
8							
9	*RESUMEN*	*Cuenta*	*Suma*	*Promedio*	*Varianza*		
10	1	6	465	77.5	21.5		
11	2	6	400	66.6666667	23.4666667		
12	3	6	546	91	33.2		
13	4	6	476	79.3333333	23.4666667		
14							
15	1	4	287	71.75	76.25		
16	2	4	316	79	36.6666667		
17	3	4	313	78.25	90.9166667		
18	4	4	315	78.75	184.916667		
19	5	4	326	81.5	121		
20	6	4	330	82.5	161		
21							
22							
23	ANÁLISIS DE VARIANZA						
24	*n de las variaba de cuadrados de libertad de los cuad*				*F*	*Probabilidad*	*or crítico para F*
25	Filas	1787.45833	3	595.819444	39.7581094	2.2134E-07	3.2873211
26	Columnas	283.375	5	56.675	3.78183503	0.02045578	2.90129454
27	Error	224.791667	15	14.9861111			

Observe que Excel utiliza el término **"Filas"** para **" Tratamientos"** , **"Columnas"** para **"bloques"** y **"Error"** para **" Error".** Sin embargo, tienen los mismos significados.

Conclusiones: A la vista de los ***p-valores***, se concluye que es **altamente significativa (AS)** la diferencia entre los **restaurantes** al 95% (p-valor menor que 0.05), y que es **significativa (S)** la diferencia entre los **Clasificadores** (p-valor entre 0.01 y 0.05).

NOTA: Excel en este caso no tiene opción para realizar pruebas Pos-hoc ó Aposteriori como la prueba de Tukey

ISBN 978-607-8285-62-4

1.3.2.1 EJEMPLO ILUSTRATIVO EN MINITAB 15

EJEMPLO ILUSTRATIVO INTEGRAL EN MINITAB 1.3.2.1 DISEÑO DE BLOQUES AL AZAR

Pruebas de hipótesis para el Factor 1 y los bloques

Una cadena de restaurantes de comida rápida tiene cuatro sucursales en determinada zona geográfica y desea evaluar su servicio. El director de investigación de mercados contrata 24 investigadores (clasificadores) con diversas experiencias en evaluación de servicios alimentarios. Después de las consultas preliminares, los 24 investigadores se dividen en seis bloques de cuatro. Los cuatro investigadores mas experimentados quedan ubicados en el bloque 1, los cuatro siguientes en el bloque 2 y así sucesivamente.

Dentro de cada bloque homogéneo se asignan al azar los cuatro clasificadores para evaluar el servicio de determinado restaurante bajo el siguiente criterio 0(baja) a 100(alta). Los resultados se presentan a continuación.

REST	**CLASIFICADORES**					
	1	**2**	**3**	**4**	**5**	**6**
1	70	77	76	80	84	78
2	61	75	67	63	66	68
3	82	88	90	96	92	98
4	74	76	80	76	84	86

a) Con un nivel de significancia de 0.05 ¿existe diferencia en la evaluación entre los cuatro restaurantes?.
b) Con un nivel de significancia de 0.05 ¿existe diferencia en la evaluación entre los seis clasificadores?.
c) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿Cuál o cuáles restaurantes obtuvieron mayor puntaje y cuánto más?.

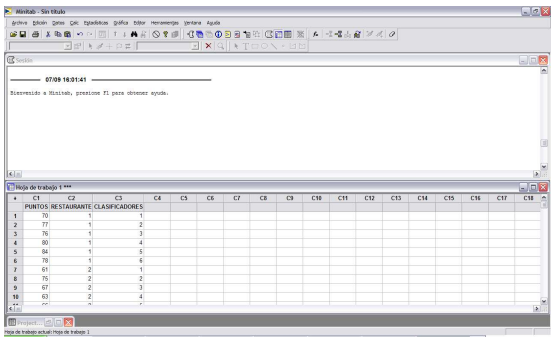
Solución al inciso a y b.

Cuando el número de observaciones en cada tratamiento es extenso y/o existen muchos tratamientos, los cálculos manuales son tediosos. Existen muchos paquetes de software que pueden mostrar los resultados, entre ellos Minitab (Versión 15)

ISBN 978-607-8285-62-4

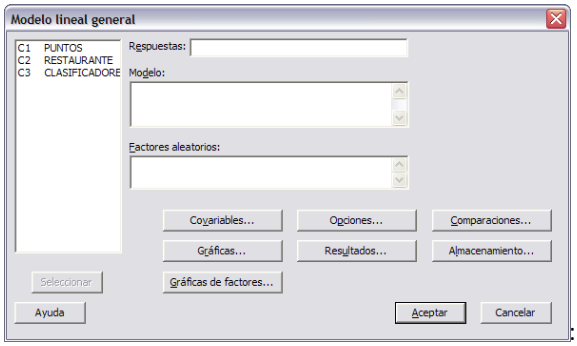
Comenzamos introduciendo los datos en la hoja de Trabajo 1 de Minitab, tal y como se muestra a continuación:

Hoja de trabajo Minitab.



Tenemos entonces un ***diseño en bloques aleatorizados con una unidad por casilla.*** Como este diseño equivale a un diseño de dos factores (restaurantes y clasificadores) sin interacción, seleccionamos la opción ***Modelo Lineal General*** del menú ***Datos Estadísticas>Anova*** que nos proporciona el siguiente cuadro de dialogo.

Cuadro de diálogo: Modelo lineal general.



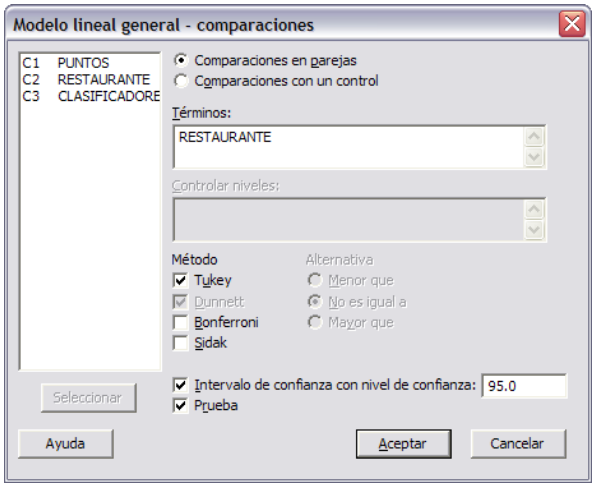
En ***Respuesta:***, ingrese **PUNTOS**. En ***Modelo:***, ingrese **RESTAURANTES CLASIFICADORES**

Cuadro de diálogo: Modelo lineal general.

Haga clic en el botón ***Comparaciones.*** En ***Términos*** ingrese **RESTAURANTE.**

ISBN 978-607-8285-62-4

Cuadro de diálogo: Modelo lineal general-comparaciones.



Haga clic en ***Aceptar*** en cada cuadro de dialogo.

Salida de la ventana Sesión.

Salida de la ventana Sesión

Modelo lineal general: PUNTOS vs. RESTAURANTE, CLASIFICADORES

```
Factor          Tipo  Niveles  Valores
RESTAURANTE     fijo        4  1, 2, 3, 4
CLASIFICADORES  fijo        6  1, 2, 3, 4, 5, 6


Análisis de varianza para PUNTOS, utilizando SC ajustada para
pruebas

Fuente          GL  SC sec.  SC ajust.  MC ajust.      F      P
RESTAURANTE      3  1787.46    1787.46     595.82  39.76  0.000
CLASIFICADORES   5   283.37     283.37      56.67   3.78  0.020
Error           15   224.79     224.79      14.99
Total           23  2295.62

S = 3.87119   R-cuad. = 90.21%   R-cuad.(ajustado) = 84.99%
```

- Si el *valor p* es menor que o igual al nivel Alfa (a), rechace la hipótesis nula en favor de la hipótesis alternativa.

-Si $0.01 \leq p \leq 0.05$ se dice que la prueba es significativa (S).

-Si $p < 0.01$ se dice que la prueba es altamente significativa (AS).

-Si el *valor p* es mayor que el nivel Alfa (a), no rechace la hipótesis nula. Si $p > 0.05$ se dice que la prueba es no significativa (NS).

Los niveles de significancia más utilizado son 0.05 y 0.01

Conclusiones: Minitab muestra una tabla de niveles, una tabla de **ANOVA**. Las pruebas ***F*** del **ANOVA** indican que **hay evidencia significativa de los efectos de los restaurantes**. A la vista de los ***p-valores***, se concluye que es **altamente significativa (AS) la diferencia entre los restaurantes al 95%** (p-valor menor que 0.05), y que es **significativa (S) la diferencia entre los Clasificadores** (p-valor entre 0.01 y 0.05).

Prueba de comparaciones múltiples. Método T de Tukey

<u>Solución al inciso c.</u>

Conclusiones: Minitab muestra los resultados de **Tukey de comparaciones múltiples** e **intervalos de confianza para las diferencias en parejas entre restaurantes** y las pruebas de hipótesis de comparaciones múltiples correspondientes.

ISBN 978-607-8285-62-4

Examine los intervalos de confianza de las comparaciones múltiples de **Tukey**. Hay tres conjuntos: **1)** uno en el que a las medias de los restaurantes 2,3 y 4 se les resta la media del restaurante 1; **2)** otro en el que a las medias de los restaurantes 3 y 4 se les resta la media del restaurante 2; y **3)** otro en el que a la media del restaurante 4 se le resta la media del restaurante 3.

Salida de la ventana Sesión.

```
Intervalos de confianza simultáneos de Tukey del 95.0%

Variable de respuesta PUNTOS

Todas las comparaciones de dos a dos entre los niveles de
RESTAURANTE

RESTAURANTE = 1  restado a:

RESTAURANTE  Inferior  Centrada  Superior
2              -17.28    -10.83    -4.385
3                7.05     13.50    19.948
4               -4.61      1.83     8.281


RESTAURANTE  --+---------+---------+---------+----
2            (----*---)
3                                (---*---)
4                      (---*----)
             --+---------+---------+---------+----
             -15         0        15        30
```

Interpretación.

INTERPRETACION:

- El primer intervalo, del primer conjunto, correspondiente a la media del restaurante 2 menos la media del restaurante 1, ambos signos del intervalo de confianza son negativos, por lo tanto existe evidencia de que ambos pares de medias son diferentes y que el restaurante 1 quedo mejor evaluado que el restaurante 2 por un mínimo de 4.385 y un máximo de 17.28 puntos. Una ventaja de los intervalos de confianza es que se puede ver la magnitud de las diferencias entre las medias.

- El segundo intervalo, del primer conjunto, correspondiente a la media del restaurante 3 menos la media del restaurante 1, ambos signos del intervalo de confianza son positivos, por lo tanto existe evidencia de que ambos pares de medias son diferentes y que el restaurante 3 quedo mejor evaluado que el restaurante 1 por un mínimo de 7.05 y un máximo de 19.948 puntos.

- El tercer intervalo, del primer conjunto, correspondiente a la media del restaurante 4 menos la media del restaurante 1, contiene cero en el intervalo de confianza, por lo tanto, no hay una evidencia significativa en alfa 0.05 en la diferencia de las medias.

ISBN 978-607-8285-62-4

Salida de la ventana Sesión.

```
RESTAURANTE = 2  restado a:

RESTAURANTE  Inferior  Centrada  Superior
3              17.885     24.33     30.78
4               6.219     12.67     19.11

RESTAURANTE  --+---------+---------+---------+----
3                                   (---*----)
4                              (---*----)
             --+---------+---------+---------+----
             -15         0        15        30
```

Interpretación.

INTERPRETACION:

- El primer intervalo, del segundo conjunto, correspondiente a la media del restaurante 3 menos la media del restaurante 2, ambos signos del intervalo de confianza son positivos, por lo tanto existe evidencia de que ambos pares de medias son diferentes y que el restaurante 3 quedo mejor evaluado que el restaurante 2 por un mínimo de 17.885 y un máximo de 30.78 puntos.
- El segundo intervalo, del segundo conjunto, correspondiente a la media del restaurante 4 menos la media del restaurante 2, ambos signos del intervalo de confianza son positivos, por lo tanto existe evidencia de que ambos pares de medias son diferentes y que el restaurante 4 quedo mejor evaluado que el restaurante 2 por un mínimo de 6.219 y un máximo de 19.11 puntos.

Salida de la ventana Sesión.

```
RESTAURANTE = 3  restado a:

RESTAURANTE  Inferior  Centrada  Superior
4              -18.11    -11.67    -5.219

RESTAURANTE  --+---------+---------+---------+----
4            (---*----)
             --+---------+---------+---------+----
             -15         0        15        30
```

Interpretación

INTERPRETACION:

- El primer intervalo, del tercer conjunto, correspondiente a la media del restaurante 4 menos la media del restaurante 3, ambos signos del intervalo de confianza son negativos, por lo tanto existe evidencia de que ambos pares de medias son diferentes y que el restaurante 3 quedo mejor evaluado que el restaurante 4 por un mínimo de 5.219 y un máximo de 18.11 puntos.

ISBN 978-607-8285-62-4

OBJETIVO 1.4. El alumno podrá realizar una prueba de hipótesis para un diseño de dos factores con repetición, utilizar el método *T* de Tukey de comparaciones múltiples y construir un gráfico de interacciones.

ANTECEDENTES

CONCEPTOS DE:

Experimento. Unidad experimental. Variable de respuesta. Ensayos ó réplicas. Aleatorización. Agrupamiento. Bloqueo. Balanceo. Factores controlados. Factores no controlados. Tratamientos ó niveles de un factor. Error experimental. Efectos del tratamiento. Variación total. Variación entre tratamientos. Variación dentro de tratamientos. Análisis de varianza (ANOVA).Modelo lineal general.

1.4.1 ELEMENTOS Y SUPUESTOS DE LOS EXPERIMENTOS FACTORIALES

CONCEPTOS BÁSICOS DISEÑO DE DOS FACTORES

Elementos y supuestos básicos del diseño de dos factores.

En la **estructura de un diseño de experimentos** se deben considerar los siguientes **elementos**:

1.- El conjunto de **tratamientos** incluidos en el estudio.
2.- El conjunto de **unidades experimentales** utilizadas en el estudio.
3.- Las reglas y procedimientos por los cuales los **tratamientos son asignados a las unidades experimentales** (o viceversa).
4.- Las **medidas** o evaluaciones que se hacen a las **unidades experimentales** luego de aplicar los **tratamientos**.

Las pruebas usuales de hipótesis **ANOVA de dos sentidos** son válidas bajo las siguientes **condicione ó supuestos básicos**:

1.- El diseño debe estar completo
2.- El diseño debe estar balanceado
3.- El número de réplicas por tratamiento, *K*, debe ser al menos 2.
4.-Dentro de cualquier tratamiento, las observaciones $X_{ij1}, \dots, X_{ijk}$ constituyen una muestra aleatoria simple de una población normal.
5.- La varianza poblacional es igual para todos los tratamientos. Esta varianza se denota mediante σ^2.

ISBN 978-607-8285-62-4

1.4.2 ANÁLISIS DE VARIANZA DE DOS FACTORES

CONCEPTOS BÁSICOS DISEÑO DE DOS FACTORES

Los tratamientos ó niveles de un factor son los tipos o grados específicos del factor que se tendrán en cuenta en la realización del experimento.

La unidad experimental, es el objeto o espacio al cual se aplica el tratamiento y donde se mide y analiza la variable que se investiga

Modelo estadístico.

El análisis de varianza de dos factores es conocida con el nombre de ***diseño factorial,*** técnica en la que se manejan los efectos de dos factores sobre una variable dependiente donde cada uno de ellos puede manejar un número de niveles diferente, dispuestos en un diseño factorial, esto es cada repetición ò replica del experimento contiene todas las combinaciones de niveles de los factores por lo que en general hay n repeticiones.

La siguiente tabla representa la **matriz de las observaciones** al efectuar el experimento:

Factor 1 (A)	**Factor 2 (B)**					**Total**	**Media**
	1	**2**	**j**	**…**	**b**		
1	X_{111}, X_{112}, …, X_{11n}	X_{121}, X_{122}, …, X_{12n}.			X_{1b1}, X_{1b2}, …, X_{1bn}	$X_{1..}$	$\overline{X}_{1..}$
2	X_{211}, X_{212}, …, X_{21n}	X_{221}, X_{222}, …, X_{22n}				$X_{2..}$	$\overline{X}_{2..}$
…							
i			X_{ijk}			$X_{i..}$	$\overline{X}_{i..}$
…							
a	X_{a11}, X_{a12}, …, X_{a1n}	X_{a21}, X_{a22}, …, X_{a2n}			X_{ab1}, X_{ab2}, …, X_{abn}		$\overline{X}_{a..}$
Total	$X_{.1.}$	$X_{.2.}$	$X_{.j.}$		$X_{.b.}$	$X_{...}$	
Media	$\overline{X}_{.1.}$	$\overline{X}_{.2.}$	$\overline{X}_{.j.}$		$\overline{X}_{.b.}$		$\overline{X}_{...}$

El modelo estadístico para este diseño es:

$$X_{ijk} = \mu + \tau_{i..} + \beta_{.j.} + (\tau\beta)_{ij.} + \varepsilon_{ijk} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, b \\ k = 1, 2, \dots n \end{cases}$$

ISBN 978-607-8285-62-4

Donde:

X_{ijk}: observación *k* de la variable dependiente medida bajo los efectos del nivel *i* del factor *a* y el nivel *j* del factor *b*, manejado en el experimento.

μ: promedio general de la variable dependiente.

$\tau_{i..}$: es el efecto del nivel *i* del factor a manejado en el experimento.

$\beta_{.j.}$:es el efecto del nivel *j* del factor b manejado en el experimento.

$(\tau\beta)_{ij.}$: representa el efecto interacción de los niveles *i* y *j* del factor *a* y *b* respectivamente, manejados en el experimento.

En el **diseño factorial de dos factores**, tanto los **factores de renglón como de columna tienen la misma importancia.**

Como se supone que ambos factores son fijos y que los efectos de tratamiento se definen como desviaciones de la media general entonces:

$$\sum_{i=1}^{a} \tau_i = 0 \quad \text{y} \quad \sum_{j=1}^{b} \beta_j = 0$$

Asimismo **los efectos de interacción** son fijos:

$$\sum_{i=1}^{a} (\tau\beta)_{ij} = 0$$

Hay un total de ***abn*** observaciones porque se realizan ***n*** replicas.

El **análisis de varianza** consiste en descomponer o subdividir la suma total de cuadrados en componentes del **factor a**, **factor b**, **interacción ab y error**, de la siguiente manera:

Suma de cuadrados.

$$\mathbf{SCT = SC_A + SC_B + SC_{AB} + SCE}$$

Los efectos directos de cada uno de los factores se llaman ***efectos principales***. Además, esta técnica considera un efecto más que se le conoce con el nombre de ***efecto interacción***.

Debido a que esta técnica analiza tres fuentes de variación (dos efectos principales y una interacción) sobre la variable dependiente, será necesario realizar tres pruebas de significancia, dos de efectos principales y una de interacción.

Prueba de hipótesis para el Factor 1.

Se desea probar la igualdad de las medias de los niveles o **tratamientos del factor 1**. El juego de hipótesis es:

$$\boldsymbol{H_0: \mu_{1..} = \mu_{2..} = \cdots = \mu_{k..}}$$

$$\boldsymbol{H_1: al\ menos\ una\ \mu_{k..}\ es\ diferente}$$

ISBN 978-607-8285-62-4

La suma de cuadrados es la cantidad calculada en el análisis de varianza y usada para obtener cuadrados medios para la prueba ***F***.

Con el fin de determinar si las medias de los diversos niveles del **Factor 1** son todos iguales, se pueden examinar dos estimadores diferentes de la varianza de la población. Uno de los estimadores se basa en **la suma de los cuadrados dentro de los niveles del Factor 1** ***(SC_A);*** el otro se basa en **la suma de los cuadrados entre los niveles del Factor 1 y el Factor 2** ***(SCE***). Si la hipótesis nula es cierta, estos estimadores deben ser aproximadamente iguales; si es falsa, el estimador basado en la suma de los cuadrados entre grupos debe ser mayor.

Los cuadrados medios son el cociente entre la suma de cuadrados y los grados de libertad.
El cuadrado medio del tratamiento es la estimación de la variación en el análisis de varianza. Se usa en el numerador de la prueba estadística ***F***.
El cuadrado medio del error es la estimación de la variación en el análisis de varianza. Se usa en el denominador de la estadística ***F***

En el **Análisis de Varianza**, el estimador de la varianza entre los niveles del **Factor 1** ***(CM_A)*** se calcula dividiendo la suma de los cuadrados de los niveles del **Factor 1** entre los grados de libertad entre los niveles del **Factor 1** ***(a-1).***

La varianza dentro de los niveles de los **Factores 1 y 2**, ***(CME),*** se estima dividiendo la suma de los cuadrados dentro de los niveles de los **Factores 1 y 2** entre los grados de libertad dentro de los niveles de los **Factores 1 y 2** ***(ab)(n-1).*** Si en realidad hay una diferencia entre los niveles del **Factor 1** el ***(CM_A),*** será significativamente ***mayor*** que el ***(CME).*** La prueba estadística se basa en la razón de las dos varianzas, ***CM_A/CME***. La distribución de esta razón se conoce como la **distribución *F,*** por lo que el estadístico de prueba es:

$$F_{CALC.} = \frac{CM_A}{CME} = \frac{SC_A/_{g.l.}}{SCE/_{g.l.}}$$

El estadístico F se utiliza en el análisis de varianza, entre otras pruebas, para comparar la magnitud de dos estimaciones de la varianza de la población y determina si ambas estimaciones son aproximadamente iguales; en el análisis de varianza, se emplea la razón de la varianza entre los niveles del factor con la varianza dentro de los niveles del factor.

La regla de decisión es rechazar la hipótesis nula de que no hay diferencia entre los niveles del Factor 1 si al nivel de significancia α

$$\boldsymbol{F_{calc} \geq F_{\alpha,(a-1),(ab)(n-1)}}$$

Prueba de hipótesis para el Factor 2.

Se desea probar la igualdad de las medias de los niveles o **tratamientos del factor 2**. El juego de hipótesis es:

$$\boldsymbol{H_0: \mu_{.1.} = \mu_{.2.} = \cdots = \mu_{.j.}}$$

$$\boldsymbol{H_1: al\ menos\ una\ \mu_{.j.}\ es\ diferente}$$

La suma de cuadrados es la cantidad calculada en el análisis de varianza y usada para obtener cuadrados medios para la prueba ***F***.

Con el fin de determinar si las medias de los diversos niveles del **Factor 2** son todos iguales, se pueden examinar dos estimadores diferentes de la varianza de la población. Uno de los estimadores se basa en **la suma de los cuadrados dentro de los niveles del Factor 2** ***(SC_B);*** el otro se basa en **la suma de los cuadrados entre los niveles de los Factores 1 y 2** ***(SCE***). Si la hipótesis nula es cierta, estos estimadores deben ser aproximadamente iguales; si es falsa, el estimador basado en la suma de los cuadrados entre grupos debe ser mayor.

Los cuadrados medios son el cociente entre la suma de cuadrados y los grados de libertad.

En el Análisis de Varianza, el estimador de la varianza entre los niveles del **Factor 2** ***(CM_B)*** se calcula dividiendo la suma de los cuadrados de los niveles del **Factor 2** entre los grados de libertad entre los niveles del **Factor 2** ***(b-1).***
La varianza dentro de los niveles de los **Factores 1 y 2**, ***(CME),*** se estima

ISBN 978-607-8285-62-4

El cuadrado medio del tratamiento es la estimación de la variación en el análisis de varianza. Se usa en el numerador de la prueba estadística ***F***.
El cuadrado medio del error es la estimación de la variación en el análisis de varianza. Se usa en el denominador de la estadística ***F***

dividiendo la suma de los cuadrados dentro de los niveles de los **Factores 1 y 2** entre los grados de libertad dentro de los niveles de los **Factores 1 y 2** ***(ab)(n-1).*** Si en realidad hay una diferencia entre los niveles del **Factor 2** el ***(CM_B),*** será significativamente ***mayor*** que el ***(CME).*** La prueba estadística se basa en la razón de las dos varianzas, ***CM_B/CME***. La distribución de esta razón se conoce como la **distribución *F,*** por lo que el estadístico de prueba es:

$$F_{CALC.} = \frac{CM_B}{CME} = \frac{SC_B/_{g.l.}}{SCE/_{g.l.}}$$

El estadístico F se utiliza en el análisis de varianza, entre otras pruebas, para comparar la magnitud de dos estimaciones de la varianza de la población y determina si ambas estimaciones son aproximadamente iguales; en el análisis de varianza, se emplea la razón de la varianza entre los niveles del factor con la varianza dentro de los niveles del factor.

La regla de decisión es rechazar la hipótesis nula de que no hay diferencia entre los niveles del **Factor 2** si al nivel de significancia α

$$\boldsymbol{F_{calc} \geq F_{\alpha,(b-1),(ab)(n-1)}}$$

También es interesante determinar si **los niveles del Factor 1 y los niveles del Factor 2 *interaccionan,*** es decir resulta conveniente probar:

$$\boldsymbol{H_0: \mu_{11.} = \mu_{12.} = \cdots = \mu_{ij.}}$$

$$\boldsymbol{H_1: al\ menos\ una\ \mu_{ij.}\ es\ diferente}$$

Prueba de hipótesis para la interacción de los niveles de los Factores 1 y 2

Con el fin de determinar si las medias de los diversos niveles del los **Factores 1 y 2** son todos iguales, se pueden examinar dos estimadores diferentes de la varianza de la población. Uno de los estimadores se basa en **la suma de los cuadrados dentro de los niveles de los Factores 1 y 2 *(SC_{AB});*** el otro se basa en **la suma de los cuadrados entre los niveles de los Factores 1 y 2 *(SCE*)**. Si la hipótesis nula es cierta, estos estimadores deben ser aproximadamente iguales; si es falsa, el estimador basado en la suma de los cuadrados entre grupos debe ser mayor.

La suma de cuadrados es la cantidad calculada en el análisis de varianza y usada para obtener cuadrados medios para la prueba ***F***.

En el Análisis de Varianza, el **estimador de la varianza** entre los niveles de los **Factores 1 y 2 *(CM_{AB})*** se calcula dividiendo la suma de los cuadrados de los niveles de **los Factores 1 y 2** entre los grados de libertad entre los niveles de **los Factores 1 y 2 *(a-1)(b-1).*** La varianza dentro de los niveles de **los Factores 1y 2**, ***(CME),*** se estima dividiendo la suma de los cuadrados dentro de los niveles de los **Factores 1 y 2** entre los grados de libertad dentro de los niveles de los **Factores 1 y 2 *(ab)(n-1).*** Si en realidad hay una diferencia entre los niveles de **los factores 1 y 2,** el ***(CM_{AB}),*** será significativamente ***mayor*** que el ***(CME).*** La prueba estadística se basa en la razón de las dos varianzas, ***CM_{AB}/CME***. La distribución de esta razón se conoce como la **distribución *F,*** por lo que el estadístico de prueba es:

Los cuadrados medios son el cociente entre la suma de cuadrados y los grados de libertad.
El cuadrado medio del tratamiento es la estimación de la variación en el análisis de varianza. Se usa en el numerador de la prueba estadística ***F***.
El cuadrado medio del error es la estimación de la variación en el análisis de varianza. Se usa en el denominador de la estadística

$$F_{CALC.} = \frac{CM_{AB}}{CME} = \frac{SC_{AB}/_{g.l.}}{SCE/_{g.l.}}$$

ISBN 978-607-8285-62-4

La regla de decisión es rechazar la hipótesis nula de que no hay diferencia entre los niveles de **los factores 1 y 2** si al nivel de significancia α

El estadístico F se utiliza en el análisis de varianza, entre otras pruebas, para comparar la magnitud de dos estimaciones de la varianza de la población y determina si ambas estimaciones son aproximadamente iguales; en el análisis de varianza, se emplea la razón de la varianza entre los niveles del factor con la varianza dentro de los niveles del factor.

$$\boldsymbol{F_{calc} \geq F_{\alpha,(a-1)(b-1),(ab)(n-1)}}$$

Para obtener la **Suma de Cuadrados** en un **diseño de dos factores** se usan las siguientes fórmulas:

Suma de Cuadrados Total.

$$SCT = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} X_{ijk}^2 - \frac{X_{...}^2}{abn}$$

Suma de Cuadrados para el Factor 1 ó A.

$$SC_A = \sum_{i=1}^{a} \frac{X_{i..}^2}{bn} - \frac{X_{...}^2}{abn}$$

Suma de Cuadrados para el Factor 2 ó B.

$$SC_B = \sum_{j=1}^{b} \frac{X_{.j.}^2}{an} - \frac{X_{...}^2}{abn}$$

La interacción:

Suma de cuadrados para interacción del Factor 1 y 2 ó A y B.

$SC_{AB} = SC_{SUBTOTALES} - SC_A - SC_B$

$$SC_{SUBTOTALES} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{X_{ij.}^2}{n} - \frac{X_{...}^2}{abn}$$

Suma de Cuadrados del Error.

$SCE = SCT - SC_A - SC_B - SC_{AB}$

ò

$SCE = SCT - SC_{SUBTOTALES}$

Los cuadrados medios son el cociente entre la suma de cuadrados y los grados de libertad.

$$\boldsymbol{CM_A} = \frac{SC_A}{a-1} \qquad \boldsymbol{CM_B} = \frac{SC_B}{b-1} \qquad \boldsymbol{CM_{AB}} = \frac{SC_{AB}}{(a-1)(b-1)} \qquad \boldsymbol{CME} = \frac{SCE}{ab(n-1)}$$

ISBN 978-607-8285-62-4

Tabla de ANOVA para el diseño de dos factores.

TABLA DE ANOVA PARA LA TÈCNICA DE ANOVA DE DOS FACTORES

Todo el conjunto de pasos para el diseño de dos factores se puede resumir en una tabla de análisis de varianza:

Fuente de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	**$F_{calculada}$**
Factor 1 (A)	$a-1$	SC_A	$CM_A = {SC_A}/{g.l.}$	${CM_A}/{CME}$
Factor 2 (B)	$b-1$	SC_B	$CM_B = {SC_B}/{g.l.}$	${CM_B}/{CME}$
Interacción AB	$(a-1)(b-1)$	SC_{AB}	$CM_{AB} = {SC_{AB}}/{g.l.}$	${CM_{AB}}/{CME}$
Error	$ab(n-1)$	SCE	$CME = {SCE}/{g.l.}$	
Total	$abn-1$	SCT		

El método T de Tukey se utiliza en el análisis ANOVA para construir intervalos de confianza para todas las diferencias en parejas entre medias de los niveles de factor mientras controla el nivel de significancia por familia en un nivel que usted especifique. Es importante considerar el nivel de significancia por familia al realizar múltiples comparaciones, porque las posibilidades de cometer un error de tipo I para una serie de comparaciones son mayores que el nivel de significancia para cualquier comparación individual. Para contrarrestar este nivel de significancia más alta, el método de Tukey ajusta el intervalo de confianza para cada intervalo individual (sobre todo en el caso de diseños desbalanceados), de manera que el nivel de confianza simultáneo resultante sea igual al valor que usted especifique.

COMPARACIONES MÚLTIPLES: EL MÉTODO T DE TUKEY

Con la finalidad de determinar cuáles de las *k* medias son significativamente diferentes de las otras podemos utilizar el procedimiento de Tukey. Este método es un ejemplo de un procedimiento de comparación *post hoc* (o ***a posteriori***), pues las hipótesis de interés son formuladas *después* de que los datos han sido inspeccionados.

Para usar el procedimiento de Tukey, simplemente se ordenan en forma descendente **las medias de los tratamientos** y se comparan las diferencias observadas entre cada par de promedios con el valor correspondiente al **rango ó alcance crítico**. Si $|\overline{X}_{i.} - \overline{X}_{i.'}| \geq$ **rango ó alcance crítico**, se concluye que las medias poblacionales $\mu_{i.}$ y $\mu_{j.}$ son diferentes.

El **rango ó alcance crítico para el Factor 1 (A)** se obtiene entonces de la cantidad dada en la ecuación siguiente:

$$\boldsymbol{rango\ ó\ alcance\ crítico = q_{\alpha,a,(ab)(n-1)}\sqrt{\frac{CME}{bn}}}$$

Para el modelo de diseño de dos factores, ***a* será el número de niveles del Factor 1 (A), *b* será el número de niveles del Factor 2 (B) y *n* el número de repeticiones .**

ISBN 978-607-8285-62-4

Con el **método de Tukey** se puede establecer también un conjunto de intervalos de confianza estimados simultáneamente para las verdaderas diferencias entre cada par de medias. Lo anterior se logra sumando y restando el alcance o rango crítico a las diferencias en cada par de medias muestrales.

Intervalo de confianza de Tukey.

$$(\bar{X}_{i..} - \bar{X}_{i'..}) - q_{\alpha,a,(ab)(n-1)}\sqrt{\frac{CME}{bn}} \leq (\mu_{i..} - \mu_{i'..}) \leq (\bar{X}_{i..} - \bar{X}_{i'..}) + q_{\alpha,a,(ab)(n-1)}\sqrt{\frac{CME}{bn}}$$

El valor $q_{\alpha,a,(ab)(n-1)}$ se obtiene de la tabla X del apéndice buscando en $\alpha = 0.05\ ó\ 0.01\ según\ se\ indique\ en\ el\ problema$, *a*= Número de niveles del Factor A y (ab)(n-1) =(Número de niveles del Factor A por el número de niveles del factor B por número de repeticiones menos 1).

NOTA: Si en la tabla no hay ninguna entrada que corresponda exactamente a los grados de libertad especificados se puede tomar el más cercano al especificado o hacer una interpolación con los valores que se encuentren con los grados de libertad entre los cuales se encuentre el especificado.

El **rango ó alcance crítico para el Factor 2 (B)** se obtiene entonces de la cantidad dada en la ecuación siguiente:

$$rango\ ó\ alcance\ crítico = q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}}$$

Para el modelo de diseño de dos factores, ***a* será el número de niveles del Factor 1 (A), *b* será el número de niveles del Factor 2 (B) y *n* el número de repeticiones.**

Con el **método de Tukey** se puede establecer también un conjunto de intervalos de confianza estimados simultáneamente para las verdaderas diferencias entre cada par de medias. Lo anterior se logra sumando y restando el alcance o rango crítico a las diferencias en cada par de medias muestrales.

$$(\bar{X}_{i..} - \bar{X}_{i'..}) - q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}} \leq (\mu_{i..} - \mu_{i'..}) \leq (\bar{X}_{i..} - \bar{X}_{i'..}) + q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}}$$

El valor $q_{\alpha,b,(ab)(n-1)}$ se obtiene de la tabla de puntos porcentuales del rango studentizado del apéndice buscando en $\alpha = 0.05\ ó\ 0.01\ según\ se\ indique\ en\ el\ problema$
, *b*= Número de niveles del Factor B y (ab)(n-1) =(Número de niveles del Factor A por el número de niveles del factor B por número de repeticiones menos 1). Si en la tabla no hay ninguna entrada que corresponda exactamente a los grados de libertad especificados se puede tomar el más cercano al especificado o hacer una interpolación con los valores que se encuentren con los grados de libertad entre los cuales se encuentre el especificado.

ISBN 978-607-8285-62-4

1.4.2.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 1.4.2.1 DISEÑO DE DOS FACTORES

Un experimento se lleva acabo para investigar la posible influencia de la altura en que se muestra un producto y la posición del estante y su efecto sobre las ventas. Para este experimento fueron manejadas 3 niveles de altura: 1)Inferior, 2)Media y 3)Superior del estante y para la posición que podría llegar a tener el mostrador en la tienda se consideraron dos situaciones: 1) A la entrada y 2) En las Cajas. La información se presenta a continuación.

UBICACIÓN DEL ESTANTE Factor 1	**ALTURA DEL ESTANTE Factor 2**		
	Inferior	***Media***	***Superior***
A la entrada	70 75 79	85 88 93	71 81 78
En las cajas	90 91 87	94 97 93	87 90 90

a) ¿Hay algún efecto debido a la ubicación del estante?.
b) ¿Afecto en algo la altura del estante?.
c) ¿Hay interacción entre la ubicación del estante y la altura del mismo?.
d) Construya una gráfica de interacción. Explique cómo la gráfica ilustra el grado con el que las interacciones están presentes.
e) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de las posiciones del estante. ¿Cuál o cuáles posiciones obtuvieron mayores ventas y por cuánto más?.
f) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de las alturas del estante. ¿Cuál o cuáles alturas obtuvieron mayores ventas y por cuánto más?.

Prueba de hipótesis para el Factor 1

<u>Solución al inciso a.</u>

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

La hipótesis nula es que las ventas promedio en las diferentes posiciones del estante es el mismo:

$$H_0:\ \mu_{1.} = \mu_{2.}$$

ISBN 978-607-8285-62-4

La hipótesis alternativa es que las ventas promedio en las diferentes posiciones del estante no es el mismo:

$$H_1\text{: No todas las ventas promedio son iguales}$$

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución F

$$F_{Calculada} = \frac{CMA}{CME} = \frac{SCA/_{g.l.}}{SCE/_{g.l.}}$$

Donde:

$$\mathbf{SCT = SCA + SCB + SC_{AB} + SCE}$$

Empiece por calcular las sumas y medias para cada tratamiento así como la suma y media total:

POSICIÓN DEL ESTANTE	**ALTURA DEL ESTANTE**			***Totales***	***Medias***
	Inferior	***Media***	***Superior***		
Posición a la entrada	$X_{111} = 70$ $X_{112} = 75$ $\underline{X_{113} = 79}$ $X_{11.} = 224$	$X_{121} = 85$ $X_{122} = 88$ $\underline{X_{123} = 93}$ $X_{12.} = 266$	$X_{131} = 71$ $X_{132} = 81$ $\underline{X_{133} = 78}$ $X_{13.} = 230$	$X_{1..} = 720$	$\bar{X}_{1..} = 80$
En las cajas	$X_{211} = 90$ $X_{212} = 91$ $\underline{X_{213} = 87}$ $X_{21.} = 268$	$X_{221} = 94$ $X_{222} = 97$ $\underline{X_{223} = 93}$ $X_{22.} = 284$	$X_{231} = 87$ $X_{232} = 90$ $\underline{X_{233} = 90}$ $X_{23.} = 267$	$X_{2..} = 819$	$\bar{X}_{2..} = 91$
Total	$X_{..1} = 492$	$X_{..2} = 550$	$X_{..3} = 497$	$X_{...} = 1539$	
Medias	$\bar{X}_{..1} = 82$	$\bar{X}_{..2} = 91.66$	$\bar{X}_{..3} = 82.83$		$\bar{X}_{...} = 85.50$

Para obtener la **SCT** obtenga entonces la desviación de cada observación de la media total, elevamos al cuadrado esas desviaciones y sumamos este resultado para las 18 observaciones. Para simplificar los cálculos podemos utilizar la siguiente fórmula abreviada:

Suma de Cuadrados Total.

$$SCT = \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{3} X_{ijk}^2 - \frac{X^2..}{abn} = (70^2 + 75^2 + \cdots + 90^2) - \frac{1539^2}{2(3)(3)}$$
$$= 132{,}683 - 131{,}584.50 = \mathbf{1{,}098.50}$$

La fórmula abreviada para encontrar la SCA es:

Suma de Cuadrados del Factor 1 ó A.

$$SCA = \sum_{i=1}^{2} \frac{X_{i..}^2}{bn} - \frac{X_{...}^2}{abn} = \left(\frac{720^2}{3(3)} + \frac{819^2}{3(3)}\right) - \frac{1539^2}{2(3)(3)}$$
$$= (57{,}600 + 74{,}529) - 131{,}584.50 = \mathbf{544.50}$$

ISBN 978-607-8285-62-4

La fórmula abreviada para encontrar la SCB es:

Suma de Cuadrados del Factor 2 ó B.

$$SCB = \sum_{j=1}^{3} \frac{X_{.j.}^2}{an} - \frac{X_{...}^2}{abn} = \left(\frac{492^2}{2(3)} + \frac{550^2}{2(3)} + \frac{497^2}{2(3)}\right) - \frac{1539^2}{2(3)(3)}$$
$$= (40{,}344 + 50{,}416.67 + 41{,}168.17) - 131{,}584.50 = \mathbf{344.33}$$

La fórmula abreviada para encontrar la SC_{AB} es:

$$SC_{AB} = SC_{Subtotal} - SCA - SCB$$

Donde:

$$SC_{Subtotal} = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{X_{ij.}^2}{3} - \frac{X_{...}^2}{2(3)(3)} = \left(\frac{224^2}{3} + \frac{266^2}{3} + \frac{230^2}{3} + \frac{268^2}{3} + \frac{284^2}{3} + \frac{267^2}{3}\right) - \frac{1539^2}{2(3)(3)} = (16{,}725.33 + 23{,}585.33 + 17{,}633.33 + 23{,}941.33 + 26{,}885.33 + 23{,}763.00) - 131{,}584.50 = 132{,}533.67 - 131{,}584.50 = \mathbf{949.166}$$

Suma de cuadrados de la interacción del Factor 1 con el Factor 2 ó del Factor A con el Factor B.

$$SC_{AB} = SC_{Subtotal} - SCA - SCB = 949.166 - 544.50 - 344.33 = \mathbf{60.34}$$

Por último determine la SCE ó la suma de los cuadrados de los errores debido a las diferencias dentro de cada celda a través de la resta:

Suma de Cuadrados del Error.

$$SCE = SCT - SCA - SCB - SC_{AB} = 1{,}098.50 - 544.50 - 344.33 - 60.34 = \mathbf{149.33}$$

Una manera de obtener la SCE en forma directa y como comprobación sería:

$$SCE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(X_{ijk} - \bar{X}_{ij.}\right)^2$$

O bien en forma abreviada:

$$SCE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} X_{ijk}^2 - \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{X_{ij.}^2}{n}$$

$$SCE = \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{3} X_{ijk}^2 - \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{X_{ij.}^2}{n}$$
$$= (70^2 + 75^2 + \cdots + 90^2) - \left(\frac{224^2}{3} + \frac{266^2}{3} + \frac{230^2}{3} + \frac{268^2}{3} + \frac{284^2}{3} + \frac{267^2}{3}\right)$$
$$= 132{,}683 - 132{,}533.67 = \mathbf{149.33}$$

ISBN 978-607-8285-62-4

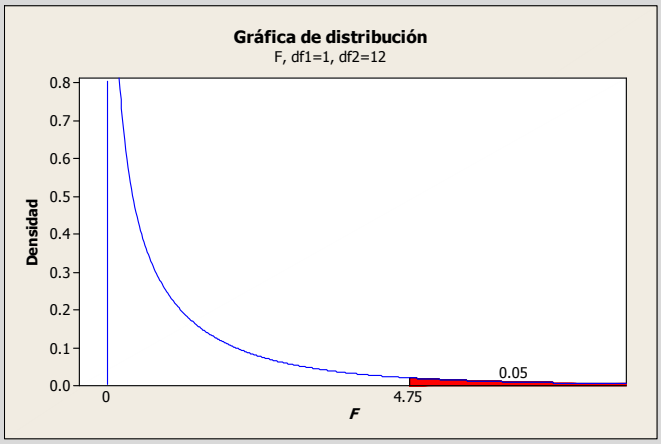
Tabla de ANOVA para dos factores.

Tabla de análisis de varianza para el modelo de dos factores con repetición

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrado Medio*	*$F_{calculada}$*
Factor A	a-1= 2-1=**1**	SCA=**544.50**	$CMA = SCA/_{g.l.} = 544.5/_{1} = \mathbf{544.50}$	$F_A = CMA/_{CME} = 544.5/_{12.44} = \mathbf{43.77}$
Factor B	b-1= 2=3-1=**2**	SCB=**344.33**	$CMB = SCB/_{g.l.} = 344.33/_{2} = \mathbf{172.16}$	$F_B = CMB/_{CME} = 172.16/_{12.44} = \mathbf{13.83}$
Interacción AB	(a-1)(b-1)= (2-1)(3-1)=**2**	SC_{AB} =**60.34**	$CM_{AB} = SC_{AB}/_{g.l.} = 60.34/_{2} = \mathbf{30.17}$	$F_{AB} = CMA/_{CME} = 30.17/_{12.44} = \mathbf{2.42}$
Error	(ab)(n-1)= (2x3)(3-1)=**12**	SCE=**149.33**	$CME = SCE/_{g.l.} = 149.33/_{12} = \mathbf{12.44}$	
Total	(abn)-1= (2x3x3)=**17**	SCT=**1,098.50**		

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).



ISBN 978-607-8285-62-4

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

La regla de decisión es rechazar H_0 si el valor calculado de F es mayor a 4.75.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como el valor calculado de F es de 43.77 que es mayor al valor crítico de 4.75; por lo tanto la hipótesis nula se rechaza y llegamos a la conclusión de que no todas las ventas promedio en las diferentes posiciones del estante son iguales, es decir al menos una de ellas es diferente.

Administrativa: Existe evidencia suficiente para concluir que estadísticamente al menos una de las ventas promedio en las diferentes posiciones del estante no es el mismo.

Prueba de hipótesis para el Factor 2

<u>Solución al inciso b.</u>

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

La hipótesis nula es que las ventas promedio en las diferentes alturas del estante es el mismo:

$$H_0:\ \mu_{.1.} = \mu_{.2.} = \mu_{.3.}$$

La hipótesis alternativa es que las ventas promedio en las diferentes posiciones del estante no es el mismo:

$$H_1: \text{No todas las ventas promedio son iguales}$$

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

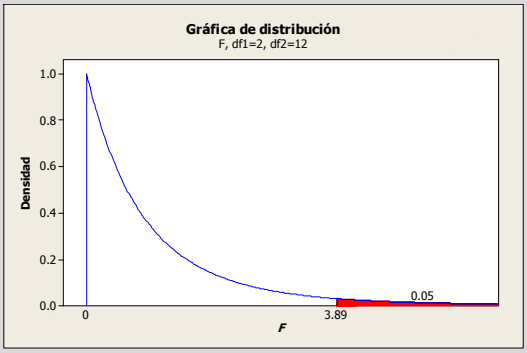
El estadístico de prueba sigue una distribución ***F***

$$F_{Calculada} = \frac{CMB}{CME} = \frac{SCB/_{g.l.}}{SCE/_{g.l.}} = \frac{172.16}{12.44} = \mathbf{13.83}$$

ISBN 978-607-8285-62-4

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).



Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

La regla de decisión es rechazar H_0 si el valor calculado de F es mayor a 3.89.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como el valor calculado de F es de 13.83 que es mayor al valor crítico de 3.89; por lo tanto la hipótesis nula se rechaza y llegamos a la conclusión de que no todas las ventas promedio en las diferentes altura del estante son iguales, es decir al menos una de ellas es diferente.

Administrativa: Existe evidencia suficiente para concluir que estadísticamente al menos una de las ventas promedio en las diferentes alturas del estante no es el mismo.

Prueba de hipótesis para la interacción.

<u>Solución al inciso c.</u>

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

La hipótesis nula es que las ventas promedio en las diferentes combinaciones de la posición y de las alturas del estante es el mismo:

$$H_0:\ \mu_{11.} = \mu_{12.} = \mu_{13.} = \mu_{21.} = \mu_{22.} = \mu_{23.}$$

La hipótesis alternativa es que las ventas promedio en las diferentes combinaciones de la posición y de las alturas del estante no es el mismo:

$$H_1: \text{No todas las ventas promedio son iguales}$$

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

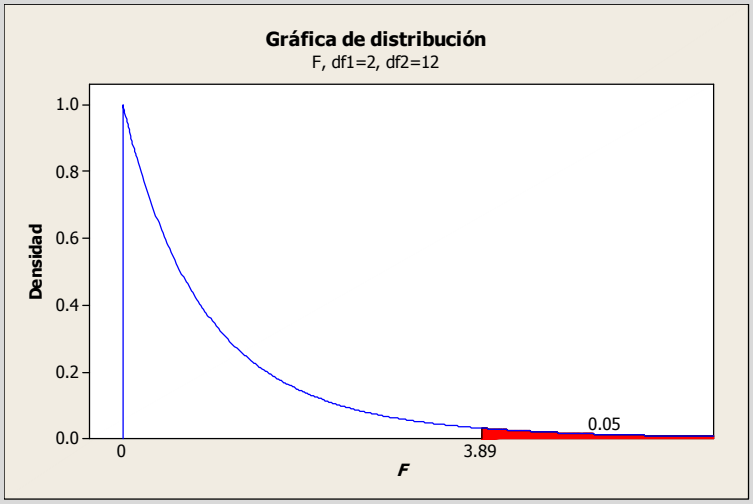
Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución F

$$F_{Calculada} = \frac{CM_{AB}}{CME} = \frac{SC_{AB}/_{g.l.}}{SCE/_{g.l.}} = \frac{30.17}{12.44} = \mathbf{2.42}$$

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).



Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

La regla de decisión es rechazar H_0 si el valor calculado de F es mayor a 3.89.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como el valor calculado de F es de 2.42 que es menor al valor crítico de 3.89; por lo tanto la hipótesis nula no se rechaza y llegamos a la conclusión de que todas las ventas promedio en las diferentes combinaciones de la posición y altura del estante son iguales.

Administrativa: Existe evidencia suficiente para concluir que estadísticamente las ventas promedio en las diferentes combinaciones de la posición y altura del estante son las mismas y no se presenta el efecto de sinergia en las ventas.

ISBN 978-607-8285-62-4

Gráfica de interacciones

Cuando el efecto de un factor depende del nivel del otro factor. Usted puede utilizar una gráfica de interacción para visualizar posibles interacciones. Usted puede utilizar una gráfica de interacción para visualizar posibles interacciones.

Las líneas paralelas en una gráfica de interacción indican que no hay interacción. Mientras mayor sea la diferencia en la pendiente entre las líneas, mayor será el grado de interacción. Sin embargo, la gráfica de interacción no dice si la interacción es estadísticamente significativa[4].

Prueba post-hoc ó aposteriori. Método T de Tukey de comparaciones múltiples para el Factor 1.

Paso 1. Ordenar las medias en forma descendente.

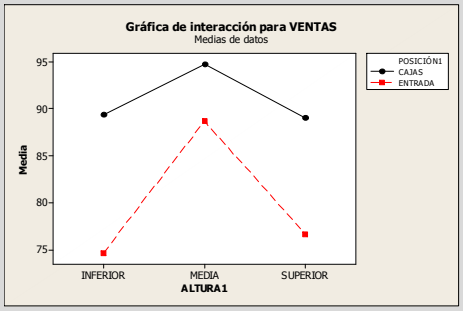
Solución al inciso d.

Una vez realizadas las pruebas para ver la importancia del factor A, del B y su combinación o interacción, se puede obtener una mejor comprensión de la interpretación del concepto de interacción, dibujando las medias de las celdas, las cuales son:

$$\bar{X}_{11.} = \frac{224}{3} = 74.67 \ ; \ \bar{X}_{12.} = \frac{266}{3} = 88.67 \ ; \ \bar{X}_{13.} = \frac{230}{3} = 76.67 \ ENTRADA$$

$$\bar{X}_{21} = \frac{268}{3} = 89.33 \ ; \ \bar{X}_{22} = \frac{284}{3} = 94.67 \ ; \ \bar{X}_{23} = \frac{267}{3} = 89 \ CAJAS$$

En la figura siguiente se han dibujado las ventas promedio para cada nivel de altura en el estante donde se muestra el producto y la posición de éste en la tienda.



Interpretación: Las dos líneas (que representan la posición del estante) parecen ser aproximadamente paralelas. Este fenómeno se puede interpretar como que la *diferencia* en ventas en las dos posiciones es prácticamente la misma para los tres niveles de altura donde se exhibe el producto. En otras palabras, no hay interacción (sinergia) entre estos dos factores-como ya se determino para la prueba *F* de interacción.

Solución al inciso e.

El método ***T* de Tukey** permite examinar, en forma simultánea, comparaciones entre todos los pares de grupos mediante los siguientes pasos:

Paso 1. Ordenar las medias en forma descendente:

$$\bar{X}_{2..} = 91; \qquad \bar{X}_{1..} = 80$$

Construida con el software estadístico Minitab 15

ISBN 978-607-8285-62-4

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

$$\overline{X}_{2..} - \overline{X}_{1..} = 91 - 80 = 11$$

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$$rango\ ó\ alcance\ crítico = q_{\alpha,a,(ab)(n-1)}\sqrt{\frac{CME}{bn}}$$

$$rango\ ó\ alcance\ crítico = q_{0.05,2,(2x3)(3-1)}\sqrt{\frac{12.44}{(3)(3)}}$$

$$rango\ ó\ alcance\ crítico = q_{0.05,2,12)}\sqrt{\frac{12.44}{9}}$$

Para determinar el rango ó alcance crítico se usa la tabla de puntos porcentuales del rango studentizado para:

$$\alpha = 0.05, a = 2\ y\ (ab)(n-1) = (2x3)(3-1) = 12, el\ valor\ crítico\ superior\ de\ q_{0.05,2,12} es\ \mathbf{3.08.}$$

Por lo tanto, al utilizar la formula para el rango ó alcance crítico se tiene,

$$rango\ ó\ alcance\ crítico = 3.08\sqrt{\frac{12.44}{9}} = 3.6$$

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

$$\overline{X}_{2..} - \overline{X}_{1..} = 91 - 80 = 11 > 3.62\ \therefore la\ prueba\ es\ (S)\ y\ \mu_{2..} > \mu_{1.}$$

Paso 5. Construir la gráfica de medias.

Paso 5.- Construir la gráfica[5] de medias:

```
                 ICs de 95% individuales para la media
                 basados en Desv.Est. agrupada
POSICIÓN  Media  -----+---------+---------+---------+----
1            80  (----*----)
2            91                       (----*----)
                 -----+---------+---------+---------+----
                   80.0      85.0      90.0      95.0
```

[5] Contruida con el software estadístico Minitab 15.

ISBN 978-607-8285-62-4

Paso 6. Construir los intervalos de confianza de cada par de medias.

Paso 6.- Establecer el conjunto de intervalos de confianza:

1. $(\bar{X}_{i..} - \bar{X}_{i'..}) - q_{\alpha,a,(ab)(n-1)}\sqrt{\frac{CME}{bn}} \leq (\mu_{i..} - \mu_{i'..}) \leq (\bar{X}_{i..} - \bar{X}_{i'..}) + q_{\alpha,a,(ab)(n-1)}\sqrt{\frac{CME}{bn}}$

$$(\bar{X}_{2..} - \bar{X}_{1..}') - q_{0.05,2,(2x3)(3-1)}\sqrt{\frac{12.44}{(3)(3)}} \leq (\mu_{2..} - \mu_{1..}') \leq (\bar{X}_{2..} - \bar{X}_{1..}') + q_{0.05,2,(2x3)(3-1)}\sqrt{\frac{12.44}{(3)(3)}}$$

$$(91 - 80) - q_{0.05,2,12}\sqrt{\frac{12.44}{9}} \leq (\mu_{2..} - \mu_{1..}') \leq (91 - 80) + q_{0.05,2,12)}\sqrt{\frac{12.44}{9}}$$

$$11 - 3.08\sqrt{\frac{12.44}{9}} \leq (\mu_{2..} - \mu_{1..}') \leq (91 - 80) + 3.08\sqrt{\frac{12.44}{9}}$$

$$11 - 3.62 \leq (\mu_{2..} - \mu_{1..}') \leq 11 + 3.62$$

$$\mathbf{7.38} \leq (\mu_{2..} - \mu_{1..}') \leq \mathbf{14.62}$$

Conclusiones.

Conclusión: Como ambos límites del intervalo son positivos podemos decir que estadísticamente la posición del estante en las cajas produce mayores ventas promedio que la posición del estante en la entrada por un mínimo de 7.38 y un máximo de 14.62 unidades venta.

Prueba post-hoc ó aposteriori. Método T de Tukey de comparaciones múltiples para el Factor 2

<u>Solución al inciso f.</u>

El método ***T* de Tukey** permite examinar, en forma simultánea, comparaciones entre todos los pares de grupos mediante los siguientes pasos:

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

$$\bar{X}_{.2.} = 91.66; \quad \bar{X}_{.3.} = 82.83; \quad \bar{X}_{.1.} = 82$$

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

$$\bar{X}_{.2.} - \bar{X}_{.3.} = 91.66 - 82.83 = 8.83$$

$$\bar{X}_{.2.} - \bar{X}_{.1.} = 91.66 - 82 = 9.66$$

$$\bar{X}_{.3.} - \bar{X}_{.1.} = 82.83 - 82 = 0.83$$

ISBN 978-607-8285-62-4

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método T:

$$rango\ ó\ alcance\ crítico = q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}}$$

$$rango\ ó\ alcance\ crítico = q_{0.05,3,(2x3)(3-1)}\sqrt{\frac{12.44}{(2)(3)}}$$

$$rango\ ó\ alcance\ crítico = q_{0.05,3,12)}\sqrt{\frac{12.44}{6}}$$

Para determinar el rango ó alcance crítico se usa la tabla de puntos porcentuales del rango studentizado para:

$$\alpha = 0.05, b = 3\ y\ (ab)(n-1) = (2x3)(3-1) = 12, el\ valor\ crítico\ superior\ de\ q_{0.05,3,12} es\ 3.77.$$

Por lo tanto, al utilizar la formula para el rango ó alcance crítico se tiene,

$$rango\ ó\ alcance\ crítico = 3.77\sqrt{\frac{12.44}{6}} = 5.43$$

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

$$\bar{X}_{.2.} - \bar{X}_{.3.} = 91.66 - 82.83 = 8.83\ > 5.43\ \therefore la\ prueba\ es\ (S) y\ \mu_{.2.} > \mu_{.3.}$$

$$\bar{X}_{.2.} - \bar{X}_{.1.} = 91.66 - 82 = 9.66\ > 5.43\ \therefore la\ prueba\ es\ (S) y\ \mu_{.2.} > \mu_{.1.}$$

$$\bar{X}_{.3.} - \bar{X}_{.1.} = 82.83 - 82 = 0.83\ < 5.43\ \therefore la\ prueba\ es\ (NS) y\ \mu_{.3.} = \mu_{.1.}$$

Paso 5. Construir la gráfica de medias.

Paso 5.- Construir la gráfica[6] de medias:

```
ICs de 95% individuales para la media
                basados en Desv.Est. agrupada
ALTURA    Media ---+---------+---------+---------+------
1       82.0000 (-------*-------)
2       91.6667                        (-------*-------)
3       82.8333   (-------*-------)
                ---+---------+---------+---------+------
                80.0      84.0      88.0      92.0
```

[6] Construida con el software estadístico Minitab 15

ISBN 978-607-8285-62-4

Paso 6. Construir los intervalos de confianza de cada par de medias.

Paso 6.- Establecer el conjunto de intervalos de confianza:

$$1.\ (\bar{X}_{.i.} - \bar{X}_{.i'}) - q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}} \leq (\mu_{.i.} - \mu_{.i.}') \leq (\bar{X}_{.i.} - \bar{X}_{.i'}) + q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}}$$

$$(\bar{X}_{.2.} - \bar{X}_{.3.}') - q_{0.05,3,(2x3)(3-1)}\sqrt{\frac{12.44}{(2)(3)}} \leq (\mu_{.2.} - \mu_{.3.}') \leq (\bar{X}_{.2.} - \bar{X}_{.3.}') + q_{0.05,3,(2x3)(3-1)}\sqrt{\frac{12.44}{(2)(3)}}$$

$$(91.66 - 82.83) - q_{0.05,3,12}\sqrt{\frac{12.44}{6}} \leq (\mu_{.2.} - \mu_{.3.}') \leq (91.66 - 82.83) + q_{0.05,3,12)}\sqrt{\frac{12.44}{6}}$$

$$8.83 - 3.77\sqrt{\frac{12.44}{6}} \leq (\mu_{.2.} - \mu_{.3.}') \leq 8.83 + 3.77\sqrt{\frac{12.44}{6}}$$

$$8.83 - 5.43 \leq (\mu_{.2.} - \mu_{.3.}') \leq 8.83 + 5.43$$

$$\mathbf{3.40} \leq (\mu_{.2.} - \mu_{.3.}') \leq \mathbf{14.26}$$

Conclusiones.

Conclusión: Como ambos límites del intervalo son positivos podemos decir que estadísticamente la altura media del estante genera mayores ventas promedio que la altura superior del estante por un mínimo de 3.40 y un máximo de 14.26 unidades venta.

$$2.\ (\bar{X}_{.i.} - \bar{X}_{.i'}) - q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}} \leq (\mu_{.i.} - \mu_{.i.}') \leq (\bar{X}_{.i.} - \bar{X}_{.i'}) + q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}}$$

$$(\bar{X}_{.2.} - \bar{X}_{.1.}') - q_{0.05,3,(2x3)(3-1)}\sqrt{\frac{12.44}{(2)(3)}} \leq (\mu_{.2.} - \mu_{.1.}') \leq (\bar{X}_{.2.} - \bar{X}_{.1.}') + q_{0.05,3,(2x3)(3-1)}\sqrt{\frac{12.44}{(2)(3)}}$$

$$(91.66 - 82) - q_{0.05,3,12}\sqrt{\frac{12.44}{6}} \leq (\mu_{.2.} - \mu_{.1.}') \leq (91.66 - 82) + q_{0.05,3,12)}\sqrt{\frac{12.44}{6}}$$

$$9.66 - 3.77\sqrt{\frac{12.44}{6}} \leq (\mu_{.2.} - \mu_{.1.}') \leq 9.66 + 3.77\sqrt{\frac{12.44}{6}}$$

$$9.66 - 5.43 \leq (\mu_{.2.} - \mu_{.1.}') \leq 9.66 + 5.43$$

$$\mathbf{4.23} \leq (\mu_{.2.} - \mu_{.1.}') \leq \mathbf{15.09}$$

Conclusión: Como ambos límites del intervalo son positivos podemos decir que estadísticamente la altura media del estante genera mayores ventas promedio que la altura inferior del estante por un mínimo de 4.23 y un máximo de 15.09 unidades venta.

ISBN 978-607-8285-62-4

$$3.\ (\bar{X}_{.i.} - \bar{X}_{.i'}) - q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}} \leq (\mu_{.i.} - \mu_{.i.}') \leq (\bar{X}_{.i.} - \bar{X}_{.i'}) + q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}}$$

$$(\bar{X}_{.3.} - \bar{X}_{.1.}') - q_{0.05,3,(2x3)(3-1)}\sqrt{\frac{12.44}{(2)(3)}} \leq (\mu_{.3.} - \mu_{.1.}') \leq (\bar{X}_{.3.} - \bar{X}_{.1.}') + q_{0.05,3,(2x3)(3-1)}\sqrt{\frac{12.44}{(2)(3)}}$$

$$(82.83 - 82) - q_{0.05,3,12}\sqrt{\frac{12.44}{6}} \leq (\mu_{.3.} - \mu_{.1.}') \leq (82.83 - 82) + q_{0.05,3,12)}\sqrt{\frac{12.44}{6}}$$

$$0.83 - 3.77\sqrt{\frac{12.44}{6}} \leq (\mu_{.3.} - \mu_{.1.}') \leq 0.83 + 3.77\sqrt{\frac{12.44}{6}}$$

$$0.83 - 5.43 \leq (\mu_{.3.} - \mu_{.1.}') \leq 0.83 + 5.43$$

$$\mathbf{-4.60} \leq (\mu_{.3.} - \mu_{.1.}') \leq \mathbf{6.26}$$

Conclusión: Como el intervalo de confianza pasa por cero podemos decir que estadísticamente la altura superior del estante y la altura inferior del estante generan las mismas ventas promedio.

1.4.2.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 1.4.2.1 DISEÑO DE DOS FACTORES

Se proporcionan los siguientes datos para una ANOVA en dos sentidos:

Factor 1	Factor 2		
	1	**2**	**3**
1	52 57	48 39	59 67
2	51 43	61 52	58 64
3	37 46	44 50	65 69

a) ¿Hay algún efecto debido al Factor 1?.
b) ¿Afecto en algo el Factor 2?.

ISBN 978-607-8285-62-4

c) ¿Hay interacción entre el Factor 1 y el Factor 2?.

d) Construya una gráfica de interacción. Explique cómo la gráfica ilustra el grado con el que las interacciones están presentes.

e) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los niveles del Factor 1. ¿Cuál o cuáles niveles resultaron mayores y por cuánto más?.

f) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los niveles del Factor 2. ¿Cuál o cuáles niveles resultaron mayores y por cuánto más?.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Prueba de hipótesis para el Factor 1

Solución al inciso a.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución *F*

$$F_{Calculada} = \frac{CMA}{CME} = \frac{SCA/_{g.l.}}{SCE/_{g.l.}} =$$

Donde:

$$SCT = SCA + SCB + SC_{AB} + SCE$$

ISBN 978-607-8285-62-4

Empiece por calcular las sumas y medias para cada tratamiento así como la suma y media total:

FACTOR 1	FACTOR 2			Totales	Medias
	1	*2*	*3*		
1	$X_{111} =$ $X_{112} =$ $X_{113} =$ $X_{11.} =$	$X_{121} =$ $X_{122} =$ $X_{123} =$ $X_{12.} =$	$X_{131} =$ $X_{132} =$ $X_{133} =$ $X_{13.} =$	$X_{1..} =$	$\bar{X}_{1..} =$
2	$X_{211} =$ $X_{212} =$ $X_{213} =$ $X_{21.} =$	$X_{221} =$ $X_{222} =$ $X_{223} =$ $X_{22.} =$	$X_{231} =$ $X_{232} =$ $X_{233} =$ $X_{23.} =$	$X_{2..} =$	$\bar{X}_{2..} =$
3	$X_{311} =$ $X_{312} =$ $X_{213} =$ $X_{31.} =$	$X_{321} =$ $X_{322} =$ $X_{223} =$ $X_{32.} =$	$X_{331} =$ $X_{332} =$ $X_{333} =$ $X_{23.} =$	$X_{3..} =$	$\bar{X}_{3..} =$
Total	$X_{..1} =$	$X_{..2} =$	$X_{..3} =$	$X_{...} =$	
Medias	$\bar{X}_{..1} =$	$\bar{X}_{..2} =$	$\bar{X}_{..3} =$		$\bar{X}_{...} =$

Suma de Cuadrados Total.

$$\boldsymbol{SCT = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} X_{ijk}^2 - \frac{X^2_{..}}{abn} =}$$

Suma de Cuadrados de los tratamientos ó niveles del Factor 1

$$\boldsymbol{SCA = \sum_{i=1}^{a} \frac{X_{i..}^2}{bn} - \frac{X_{...}^2}{abn} =}$$

Suma de Cuadrados de los tratamientos ó niveles del factor

$$\boldsymbol{SCB = \sum_{J=1}^{b} \frac{X_{.j.}^2}{an} - \frac{X_{...}^2}{abn} =}$$

$$\boldsymbol{SC_{Subtotal} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{X_{ij.}^2}{n} - \frac{X_{...}^2}{abn} =}$$

Entonces,

Suma de Cuadrados de la interacción de los niveles de los Factores 1 y 2.

$$\boldsymbol{SC_{AB} = SC_{Subtotal} - SCA - SCB =}$$

Por último determine la SCE ó la suma de los cuadrados de los errores debido a las diferencias dentro de cada celda a través de la resta:

Suma de Cuadrados del Error.

$$\boldsymbol{SCE = SCT - SCA - SCB - SC_{AB} =}$$

ISBN 978-607-8285-62-4

Tabla de ANOVA para un modelo de dos factores con repetición.

Tabla de análisis de varianza para el modelo de dos factores con repetición

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada}$***
Factor A	a-1=	SCA=	$CMA = SCA/_{g.l.} =$	$F_A = CMA/_{CME} =$
Factor B	b-1=	SCB=	$CMB = SCB/_{g.l.} =$	$F_B = CMB/_{CME} =$
Interacción AB	(a-1)(b-1)=	$SC_{AB} =$	$CM_{AB} = SC_{AB}/_{g.l.} =$	$F_{AB} = CMA/_{CME} =$
Error	(ab)(n-1)=	SCE=	$CME = SCE/_{g.l.} =$	
Total	(abn)-1=	SCT=		

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

ISBN 978-607-8285-62-4

Administrativa:

Prueba de hipótesis para el Factor 2

Solución al inciso b.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución F

$$F_{Calculada} = \frac{CMB}{CME} = \frac{SCB/_{g.l.}}{SCE/_{g.l.}} =$$

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

ISBN 978-607-8285-62-4

Administrativa:

Prueba de hipótesis para la interacción de los niveles del Factor 1 y 2.

Solución al inciso c.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución ***F***

$$F_{Calculada} = \frac{CM_{AB}}{CME} = \frac{SC_{AB}/_{g.l.}}{SCE/_{g.l.}} =$$

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

ISBN 978-607-8285-62-4

Administrativa:

Gráfica de interacciones.

Solución al inciso d.

Una vez realizadas las pruebas para ver la importancia del factor A, del B y su combinación o interacción, se puede obtener una mejor comprensión de la interpretación del concepto de interacción, dibujando las medias de las celdas, las cuales son:

$\bar{X}_{11.} =$; $\bar{X}_{12.} =$; $\bar{X}_{13.} =$ $NIVEL\ 1$

$\bar{X}_{21.} =$; $\bar{X}_{22.} =$; $\bar{X}_{23.} =$ $NIVEL\ 2$

$\bar{X}_{31.} =$; $\bar{X}_{32.} =$; $\bar{X}_{33.} =$ $NIVEL\ 3$

A continuación elabore la gráfica de interacciones o combinaciones de ambos factores:

Interpretación:

Prueba *T* de Tukey de comparaciones múltiples para el Factor 1.

Solución al inciso e.

El método ***T* de Tukey** permite examinar, en forma simultánea, comparaciones entre todos los pares de grupos mediante los siguientes pasos:

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

ISBN 978-607-8285-62-4

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$$\boldsymbol{rango\ ó\ alcance\ crítico = q_{\alpha,a,(ab)(n-1)}\sqrt{\frac{CME}{bn}} =}$$

Nota: Para determinar el rango ó alcance crítico se usa la tabla de puntos porcentuales del rango studentizado.

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

Paso 5. Construir la gráfica de medias.

Paso 5.- Construir la gráfica de medias:

Paso 6. Construir los intervalos de confianza de cada par de medias.

Paso 6.- Establecer el conjunto de intervalos de confianza:

$$(\bar{X}_{i..} - \bar{X}_{i'..}) - q_{\alpha,a,(ab)(n-1)}\sqrt{\frac{CME}{bn}} \leq (\mu_{i..} - \mu_{i'..}) \leq (\bar{X}_{i..} - \bar{X}_{i'..}) + q_{\alpha,a,(ab)(n-1)}\sqrt{\frac{CME}{bn}}$$

ISBN 978-607-8285-62-4

Conclusiones.

Conclusiones:

Prueba *T* de Tukey de comparaciones múltiples para el Factor 2.

Solución al inciso f.

El método ***T* de Tukey** permite examinar, en forma simultánea, comparaciones entre todos los pares de grupos mediante los siguientes pasos:

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$$\boldsymbol{rango\ ó\ alcance\ crítico = q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}}}$$

ISBN 978-607-8285-62-4

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

Paso 5. Construir la gráfica de medias.

Paso 5.- Construir la gráfica de medias:

Paso 6. Construir los intervalos de confianza de cada par de medias.

Paso 6.- Establecer el conjunto de intervalos de confianza:

$$(\bar{X}_{.i.} - \bar{X}'_{.i.}) - q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}} \leq (\mu_{.i.} - \mu'_{.i.}) \leq (\bar{X}_{.i.} - \bar{X}'_{.i.}) + q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}}$$

Conclusiones.

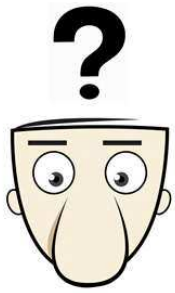
Conclusiones:

ISBN 978-607-8285-62-4

1.4.2.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. Las respuestas a este ejercicio de autoevaluación se encuentran al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 1.4.2.1
DISEÑO DE DOS FACTORES



Se proporcionan los siguientes datos para una ANOVA en dos sentidos:

FACTOR 1	FACTOR 2		
	1	**2**	**3**
1	1	4	3
	3	3	5
	2	2	2
	6	3	3
	2	2	3
2	3	4	4
	10	5	8
	6	11	12
	7	5	10
	8	6	3

a) ¿Hay algún efecto debido al Factor 1?.
b) ¿Afecto en algo el Factor 2?.
c) ¿Hay interacción entre el Factor 1 y el Factor 2?.
d) Construya una gráfica de interacción. Explique cómo la gráfica ilustra el grado con el que las interacciones están presentes.
e) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas ($1-\alpha=0.95$) de los niveles del Factor 1. ¿Cuál o cuáles niveles resultaron mayores y por cuánto más?.
f) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas ($1-\alpha=0.95$) de los niveles del Factor 2. ¿Cuál o cuáles niveles resultaron mayores y por cuánto más?.

ISBN 978-607-8285-62-4

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Prueba de hipótesis para el Factor 1

Solución al inciso a.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

Empiece por calcular las sumas y medias para cada tratamiento así como la suma y media total:

FACTOR 1	**FACTOR 2**			***Totales***	***Medias***
	1	***2***	***3***		
1	$X_{111} =$ $X_{112} =$ $X_{113} =$ $X_{114} =$ $X_{115} =$ $X_{11.} =$	$X_{121} =$ $X_{122} =$ $X_{123} =$ $X_{124} =$ $X_{125} =$ $X_{12.} =$	$X_{131} =$ $X_{132} =$ $X_{133} =$ $X_{134} =$ $X_{135} =$ $X_{13.} =$	$X_{1..} =$	$\bar{X}_{1..} =$
2	$X_{211} =$ $X_{212} =$ $X_{213} =$ $X_{214} =$ $X_{215} =$ $X_{21.} =$	$X_{221} =$ $X_{222} =$ $X_{223} =$ $X_{224} =$ $X_{225} =$ $X_{22.} =$	$X_{231} =$ $X_{232} =$ $X_{233} =$ $X_{234} =$ $X_{235} =$ $X_{23.} =$	$X_{2..} =$	$\bar{X}_{2..} =$
3	$X_{311} =$ $X_{312} =$ $X_{313} =$ $X_{314} =$ $X_{215} =$ $X_{31.} =$	$X_{321} =$ $X_{322} =$ $X_{323} =$ $X_{324} =$ $X_{225} =$ $X_{32.} =$	$X_{331} =$ $X_{332} =$ $X_{333} =$ $X_{334} =$ $X_{335} =$ $X_{23.} =$	$X_{3..} =$	$\bar{X}_{3..} =$
Total	$X_{..1} =$	$X_{..2} =$	$X_{..3} =$	$X_{...} =$	
Medias	$\bar{X}_{..1} =$	$\bar{X}_{..2} =$	$\bar{X}_{..3} =$		$\bar{X}_{...} =$

ISBN 978-607-8285-62-4

Suma de Cuadrados Total.

$$\boldsymbol{SCT} =$$

Suma de Cuadrados de los tratamientos ó niveles del Factor 1

$$\boldsymbol{SCA} =$$

Suma de Cuadrados de los tratamientos ó niveles del Factor 2

$$\boldsymbol{SCB} =$$

$$\boldsymbol{SC_{Subtotal}} =$$

Entonces,

Suma de Cuadrados de la interacción de los niveles de los Factores 1 y 2.

$$\boldsymbol{SC_{AB}} =$$

Por último determine la SCE ó la suma de los cuadrados de los errores debido a las diferencias dentro de cada celda a través de la resta:

Suma de Cuadrados del Error.

$$\boldsymbol{SCE = SCT - SCA - SCB - SC_{AB}} =$$

ISBN 978-607-8285-62-4

Tabla de ANOVA para un modelo de dos factores con repetición.

Tabla de análisis de varianza para el modelo de dos factores con repetición

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrado***	***Cuadrado Medio***	***$F_{calculada}$***
Factor A				
Factor B				
Interacción AB				
Error				
Total				

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

ISBN 978-607-8285-62-4

Prueba de hipótesis para el Factor 2

Solución al inciso b.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución F

$\boldsymbol{F_{Calculada}} =$

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

ISBN 978-607-8285-62-4

Prueba de hipótesis para la interacción de los niveles del Factor 1 y 2.

<u>Solución al inciso c.</u>

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

El estadístico de prueba sigue una distribución ***F***

$\boldsymbol{F_{Calculada}} =$

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

ISBN 978-607-8285-62-4

Gráfica de interacciones.

Solución al inciso d.

Una vez realizadas las pruebas para ver la importancia del factor A, del B y su combinación o interacción, se puede obtener una mejor comprensión de la interpretación del concepto de interacción, dibujando las medias de las celdas, las cuales son:

A continuación elabore la gráfica de interacciones o combinaciones de ambos factores:

Interpretación:

Prueba *T* de Tukey de comparaciones múltiples para el Factor 1.

Solución al inciso e.

El método ***T* de Tukey** permite examinar, en forma simultánea, comparaciones entre todos los pares de grupos mediante los siguientes pasos:

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

ISBN 978-607-8285-62-4

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.	**Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :**
Paso 3. Obtener el alcance crítico para todas las diferencias de medias.	**Paso 3. Obtener el rango crítico para el método *T*:** $rango\ ó\ alcance\ crítico =$ **Nota:** Para determinar el rango ó alcance crítico se usa la tabla de puntos porcentuales del rango studentizado.
Paso 4. Comparar el alcance crítico con las diferencias del paso 2.	**Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:**
Paso 5. Construir la gráfica de medias.	**Paso 5.- Construir la gráfica de medias:**
Paso 6. Construir los intervalos de confianza de cada par de medias.	**Paso 6.- Establecer el conjunto de intervalos de confianza:**
Conclusiones.	**Conclusiones:**

ISBN 978-607-8285-62-4

Prueba *T* de Tukey de comparaciones múltiples para el Factor 2.

Solución al inciso f.

El método ***T* de Tukey** permite examinar, en forma simultánea, comparaciones entre todos los pares de grupos mediante los siguientes pasos:

Paso 1. Ordenar las medias en forma descendente.

Paso 1. Ordenar las medias en forma descendente:

Paso 2. Formar todas las combinaciones posibles de medias de dos en dos y su diferencia.

Paso 2.- Formar todas las combinaciones de medias de dos en dos y calcular las diferencias :

Paso 3. Obtener el alcance crítico para todas las diferencias de medias.

Paso 3. Obtener el rango crítico para el método *T*:

$$rango\ \acute{o}\ alcance\ crítico =$$

Paso 4. Comparar el alcance crítico con las diferencias del paso 2.

Paso 4. Comparar el rango crítico con las diferencias de las medias del paso 2:

Paso 5. Construir la gráfica de medias.

Paso 5.- Construir la gráfica de medias:

ISBN 978-607-8285-62-4

Paso 6. Construir los intervalos de confianza de cada par de medias.

Paso 6.- Establecer el conjunto de intervalos de confianza:

Conclusiones.

Conclusiones:

1.4.2 EJERCICIOS DE REFUERZO

EJERCICIOS DE REFUERZO 1.4.2. DISEÑO DE DOS FACTORES

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de

1.4.2.1 En una investigación se evaluaron diferentes campañas publicitarias sobre las ventas con técnicas diversas de promoción: descuentos, precios bajos y regalos así como con distintos personajes para un mismo producto.

Los datos obtenidos fueron los siguientes:

PROMOCIÓN	PERSONAJE A	PERSONAJE B	PERSONAJE C
DESCUENTOS	486	398	514
	474	389	521
	467	396	507
	473	400	508
PRECIOS BAJOS	580	469	402
	568	485	387
	583	457	409
	567	423	415
REGALOS	407	561	482
	402	537	477
	401	531	479
	401	539	470

a) ¿Hay algún efecto debido a la campaña de publicidad?.
b) ¿Afecto en algo los personajes?.

ISBN 978-607-8285-62-4

cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

c) ¿Hay interacción entre la campaña de publicidad y los personajes?.
d) Construya una gráfica de interacción. Explique cómo la gráfica ilustra el grado con el que las interacciones están presentes.
e) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de las distintas promociones. ¿Cuál o cuáles promociones obtuvieron mayores ventas y por cuánto más?.
f) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los personajes. ¿Cuál o cuáles personajes obtuvieron mayores ventas y por cuánto más?.

1.4.2.2 En una investigación se analizo el efecto de los descuentos sobre las ventas, se tomó en cuenta que la ubicación de los anuncios de los descuentos en la tienda podrían afectar las ventas. Se consideraron posibles puntos de ubicación:

1. Anunciar las promociones en los aparadores.
2. Poner *displays* en puntos clave de la tienda.

UBICACION	**DESCUENTO**		
	10%	**20%**	**30%**
Aparadores	30 34 39	43 46 51	36 42 37
Displays	50 52 47	53 56 52	46 50 50

a) ¿Hay algún efecto debido a la ubicación de los anuncios?.
b) ¿Afecto en algo el porcentaje de los descuentos?.
c) ¿Hay interacción entre la ubicación de los anuncios y los porcentajes de descuento?.
d) Construya una gráfica de interacción. Explique cómo la gráfica ilustra el grado con el que las interacciones están presentes.
e) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de las distintas ubicaciones de los anuncios. ¿Cuál o cuáles ubicaciones obtuvieron mayores ventas promedio y por cuánto más?.
f) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) del porcentaje de descuento. ¿Cual o cuales porcentajes obtuvieron mayores ventas promedio y por cuanto más?

ISBN 978-607-8285-62-4

1.4.2.1 EJEMPLO ILUSTRATIVO EN EXCEL

EJEMPLO ILUSTRATIVO INTEGRAL EN EXCEL 1.4.2.1 DISEÑO DE DOS FACTORES

Un experimento se lleva acabo para investigar la posible influencia de la altura en que se muestra un producto y la posición del estante y su efecto sobre las ventas. Para este experimento fueron manejadas 3 niveles de altura: 1)Inferior, 2)Media y 3)Superior del estante y para la posición que podría llegar a tener el mostrador en la tienda se consideraron dos situaciones: 1) A la entrada y 2) En las Cajas.

La información se presenta a continuación.

UBICACIÓN DEL ESTANTE Factor 1	**ALTURA DEL ESTANTE Factor 2**		
	Inferior	***Media***	***Superior***
A la entrada	70 75 79	85 88 93	71 81 78
En las cajas	90 91 87	94 97 93	87 90 90

a) ¿Hay algún efecto debido a la ubicación del estante?.

b) ¿Afecto en algo la altura del estante?.

c) ¿Hay interacción entre la ubicación del estante y la altura del mismo?.

Pruebas de hipótesis para el Factor 1, para el Factor 2 y para la interacción del Factor 1 y Factor 2.

<u>Solución al inciso a, b y c.</u>

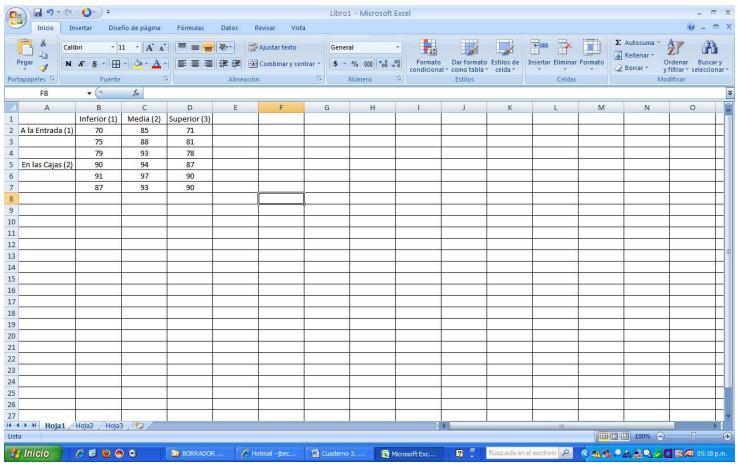
Estamos ante un ***diseño de dos factores en medidas repetidas o replicado 3 veces***, ya que se realiza el mismo test tres veces en cada posición del estante bajo tres alturas diferentes del estante donde se exhibe el producto.

La variable de respuesta es el volumen de ventas y los dos factores son la posición y la altura.

Para resolver el problema con **Excel**, introducimos los datos tal y como se muestra a continuación:

ISBN 978-607-8285-62-4

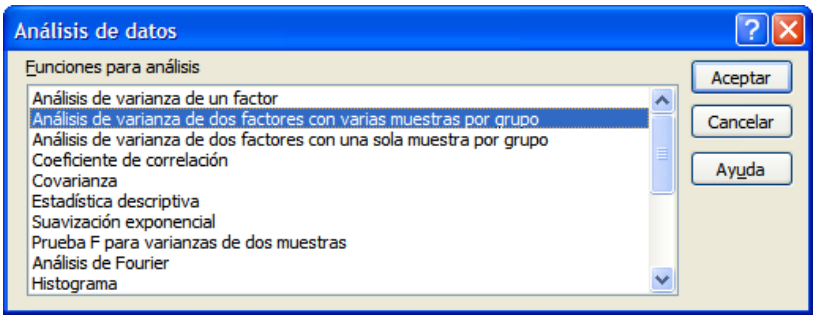
Hoja de trabajo Excel



	Inferior (1)	Media (2)	Superior (3)
A la Entrada (1)	70	85	71
	75	88	81
	79	93	78
En las Cajas (2)	90	94	87
	91	97	90
	87	93	90

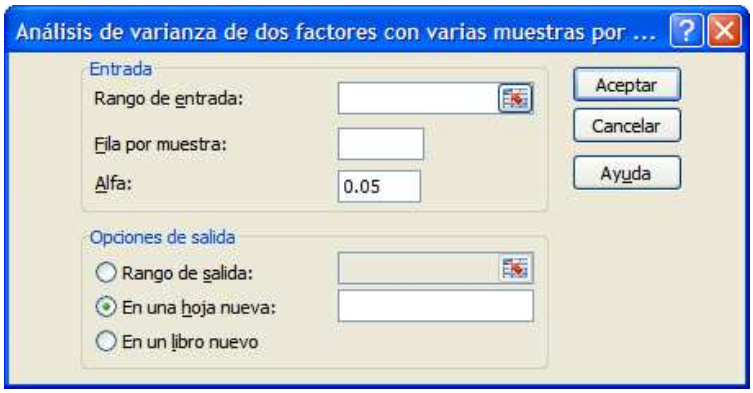
Seleccionamos la opción ***Análisis de la varianza de dos factores con muestras repetidas***, del cuadro ***Análisis de datos*** del menú **Datos** ***(Microsoft Office Excel 2007)*** que nos lleva al cuadro de diálogo ó ventana de la figura siguiente:

Cuadro de diálogo: Análisis de datos.



En la lista ***Funciones para análisis***, elija la modalidad de ***Análisis de varianza de dos factores con varias muestras por grupo*** y oprima el botón ***Aceptar*** para obtener el siguiente cuadro de dialogo:

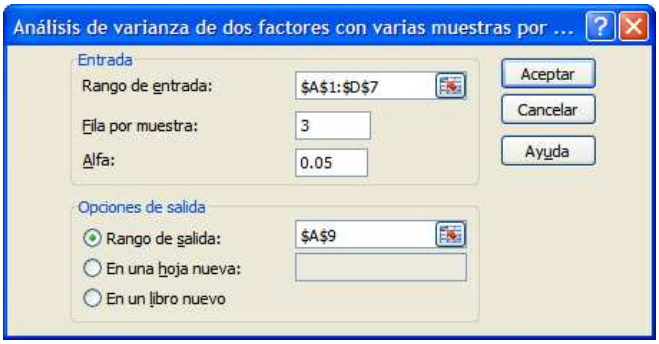
Cuadro de diálogo: Análisis de varianza de dos factores con repeticiones.



ISBN 978-607-8285-62-4

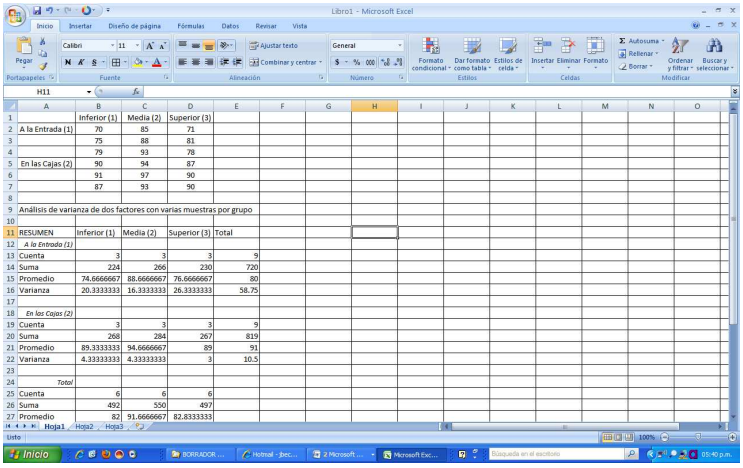
En el cuadro ***Rango de entrada*** introduzca, (seleccionando con el cursor las celdas donde están los datos incluyendo los rótulos), la referencia de celda correspondiente al rango de datos que está analizando. La referencia deberá contener dos o más rangos adyacentes organizados en columnas o filas. En el cuadro ***Fila por muestra*** introduzca el número de filas que contiene cada muestra (réplicas). Todas las muestras deben contener el mismo número de filas, ya que cada fila representa una réplica de los datos. Deje sin cambio el campo ***Alfa*** con el valor de ***0.05*** (nivel con el que desee evaluar los valores críticos de la función estadística ***F).*** El nivel ***alfa*** es un nivel de importancia relacionado con la probabilidad de que haya un error de tipo I (rechazar una hipótesis verdadera).

En cuanto a las ***opciones de salida***, en el campo ***Rango de salida*** introduzca la referencia , (dando un clic), correspondiente a la celda superior izquierda de la tabla de resultados, en este caso la **celda A9**



Cuadro de diálogo: Análisis de varianza de dos factores con repeticiones.

Oprima el botón ***Aceptar.*** A continuación se muestra la salida del análisis de la varianza de dos factores con muestras repetidas:



	Inferior (1)	Media (2)	Superior (3)	
A la Entrada (1)	70	85	71	
	75	88	81	
	79	93	78	
En las Cajas (2)	90	94	87	
	91	97	90	
	87	93	90	

Análisis de varianza de dos factores con varias muestras por grupo

RESUMEN	Inferior (1)	Media (2)	Superior (3)	Total
A la Entrada (1)				
Cuenta	3	3	3	9
Suma	224	266	230	720
Promedio	74.6666667	88.6666667	76.6666667	80
Varianza	20.3333333	16.3333333	26.3333333	58.75
En las Cajas (2)				
Cuenta	3	3	3	9
Suma	268	284	267	819
Promedio	89.3333333	94.6666667	89	91
Varianza	4.33333333	4.33333333	3	10.5
Total				
Cuenta	6	6	6	
Suma	492	550	497	
Promedio	82	91.6666667	82.8333333	

Salida del análisis de varianza de dos factores con medidas repetidas

ISBN 978-607-8285-62-4

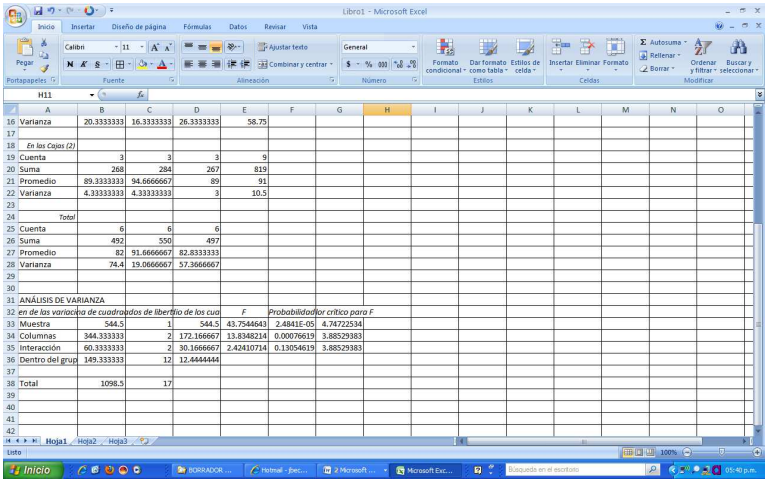
- Si el *valor p* es menor que o igual al nivel Alfa (a), rechace la hipótesis nula en favor de la hipótesis alternativa.

-Si $0.01 \leq p \leq 0.05$ se dice que la prueba es significativa (S).

-Si $p < 0.01$ se dice que la prueba es altamente significativa (AS).

- Si el *valor p* es mayor que el nivel Alfa (a), no rechace la hipótesis nula. Si $p > 0.05$ se dice que la prueba es no significativa (NS).

Los niveles de significancia más utilizado son 0.05 y 0.01



	A	B	C	D	E	F	G
16	Varianza	20.3333333	16.3333333	26.3333333	58.75		
17							
18	*En los Cajas (2)*						
19	Cuenta	3	3	3	9		
20	Suma	268	284	267	819		
21	Promedio	89.3333333	94.6666667	89	91		
22	Varianza	4.33333333	4.33333333	3	10.5		
23							
24	*Total*						
25	Cuenta	6	6	6			
26	Suma	492	550	497			
27	Promedio	82	91.6666667	82.8333333			
28	Varianza	74.4	19.0666667	57.3666667			
29							
30							
31	ANÁLISIS DE VARIANZA						
32	*en de las variacia de cuadrados de libertio de los cua*				*F*	*Probabilidad*	*lor critico para F*
33	Muestra	544.5	1	544.5	43.7544643	2.4841E-05	4.74722534
34	Columnas	344.333333	2	172.166667	13.8348214	0.00076619	3.88529383
35	Interacción	60.3333333	2	30.1666667	2.42410714	0.13054619	3.88529383
36	Dentro del grup	149.333333	12	12.4444444			
37							
38	Total	1098.5	17				
39							
40							
41							
42							

Conclusión: A la vista de los ***p-valores*** obtenidos, se concluye que es **altamente significativa (AS) para la posición del estante y para la altura del mismo en la que se exhiben los productos** (p-valores menores que 0.05), **pero no es significativa (NS)** la **interacción entre posición y altura** (p-valor mayor que 0.05, es decir **no hay sinergia al unir los dos factores.**

NOTA: Excel en este caso no tiene opción para construir la gráfica de interacciones ni realizar pruebas Pos-hoc ó Aposteriori como la prueba de Tukey.

ISBN 978-607-8285-62-4

1.4.2.1 EJEMPLO ILUSTRATIVO EN MINITAB 15

EJEMPLO ILUSTRATIVO INTEGRAL EN MINITAB 1.4.2.1 DISEÑO DE DOS FACTORES

Un experimento se lleva acabo para investigar la posible influencia de la altura en que se muestra un producto y la posición del estante y su efecto sobre las ventas. Para este experimento fueron manejadas 3 niveles de altura: 1)Inferior, 2)Media y 3)Superior del estante y para la posición que podría llegar a tener el mostrador en la tienda se consideraron dos situaciones: 1) A la entrada y 2) En las Cajas.

La información se presenta a continuación.

UBICACIÓN DEL ESTANTE Factor 1	**ALTURA DEL ESTANTE Factor 2**		
	Inferior	***Media***	***Superior***
A la entrada	70 75 79	85 88 93	71 81 78
En las cajas	90 91 87	94 97 93	87 90 90

a) ¿Hay algún efecto debido a la ubicación del estante?.
b) ¿Afecto en algo la altura del estante?.
c) ¿Hay interacción entre la ubicación del estante y la altura del mismo?.
d) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de las posiciones del estante. ¿Cuál o cuáles posiciones obtuvieron mayores ventas y por cuánto más?.
e) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de las alturas del estante. ¿Cuál o cuáles alturas obtuvieron mayores ventas y por cuánto más?.
f) Construya una gráfica de interacción. Explique cómo la gráfica ilustra el grado con el que las interacciones están presentes.

Pruebas de hipótesis para el Factor 1, el Factor 2 y la interacción.

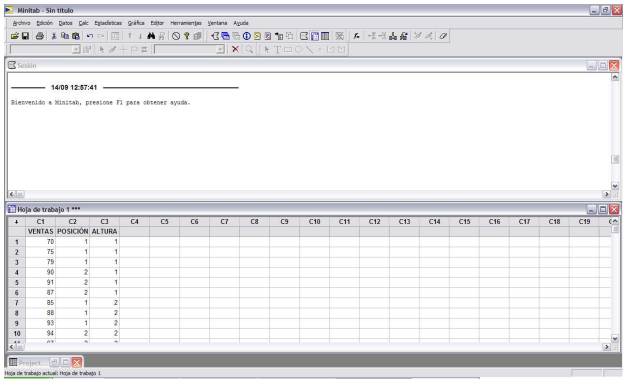
Solución al inciso a, b y c.

Cuando el número de observaciones en cada tratamiento es extenso y/o existen muchos tratamientos, los cálculos manuales son tediosos. Existen muchos paquetes de software que pueden mostrar los resultados, entre ellos **Minitab**.

ISBN 978-607-8285-62-4

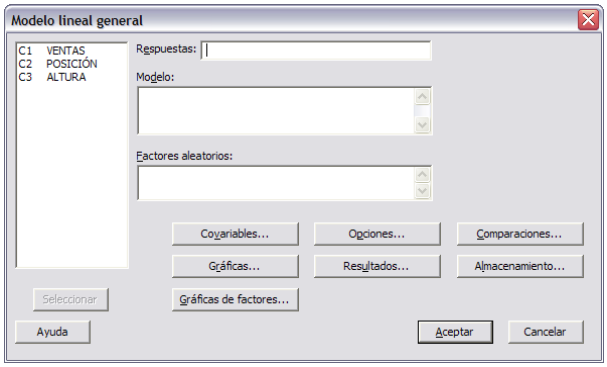
Comenzamos introduciendo los datos en **la hoja de Trabajo 1** de Minitab, tal y como se muestra a continuación:

Hoja de trabajo de Minitab.



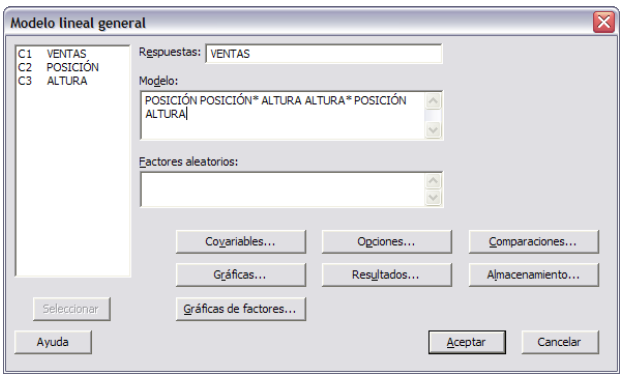
Seleccionamos la opción ***Anova>Modelo lineal general,*** del menú ***Estadísticas*** que nos lleva al cuadro de diálogo ó ventana de la figura siguiente:

Cuadro de diálogo: Modelo lineal general.



En ***Respuesta***, ingrese **VENTAS**. En ***MODELO***, ingrese **POSICIÓN POSICIÓN*ALTURA ALTURA*POSICIÓN ALTURA**.

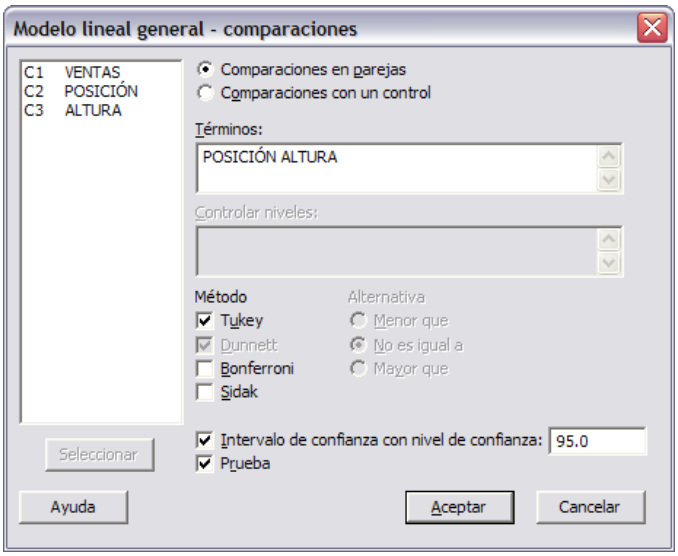
Cuadro de diálogo: Modelo lineal general.



Haga clic en el botón ***Comparaciones***. En ***Términos*** ingrese **POSICIÓN ALTURA**

ISBN 978-607-8285-62-4

Cuadro de diálogo: Modelo lineal general-comparaciones.



Haga clic en ***Aceptar*** en cada cuadro de dialogo.

Salida de la ventana Sesión

Modelo lineal general: VENTAS vs. POSICIÓN, ALTURA

```
Factor     Tipo  Niveles  Valores
POSICIÓN   fijo        2  1, 2
ALTURA     fijo        3  1, 2, 3


Análisis de varianza para VENTAS, utilizando SC ajustada
para pruebas

Fuente            GL  SC sec.  SC ajust.  MC ajust.      F      P
POSICIÓN           1   544.50     544.50     544.50  43.75  0.000
POSICIÓN*ALTURA    2    60.33      60.33      30.17   2.42  0.131
ALTURA             2   344.33     344.33     172.17  13.83  0.001
Error             12   149.33     149.33      12.44
Total             17  1098.50
```

-Si el *valor p* es menor que o igual al nivel Alfa (α), rechace la hipótesis nula en favor de la hipótesis alternativa.

-Si $0.01 \leq p \leq 0.05$ se dice que la prueba es significativa (S).

-Si $p < 0.01$ se dice que la prueba es altamente significativa (AS).

-Si el *valor p* es mayor que el nivel Alfa (α), no rechace la hipótesis nula. Si $p > 0.05$ se dice que la prueba es no significativa (NS).

Los niveles de significancia más utilizado son 0.05 y 0.01.

El método T de Tukey es una prueba *a posteriori o post-hoc* para hacer comparaciones apareadas múltiples entre medias después de obtener una prueba *F* significativa en el análisis de varianza. Es el método más recomendado por los estadígrafos

Interpretación de resultados: Minitab muestra una tabla de los niveles de cada uno de los factores y una tabla de **ANOVA**. Las pruebas ***F*** del ANOVA indican que **hay evidencias significativas de los efectos de la POSICIÓN y de la ALTURA con p-level menor a 0.05 , no así de la combinación de ambos factores o interacción con un p-level mayor a 0.05.**

<u>Solución al inciso d y e.</u>

Minitab muestra los intervalos de confianza simultáneos de Tukey para las diferencias en parejas entre POSICIÓN y ALTURA.

Examine los intervalos de confianza de la comparación múltiple. Hay tres conjuntos, uno para POSICIÓN, en el que a la media de la posición en las cajas se le resta la media de la posición en la entrada y dos para

ISBN 978-607-8285-62-4

ALTURA: 1) uno en el que a las medias de las ventas de las alturas media y superior (2 y 3) se les resta la media de la altura inferior (1); 2) otro en el que a la media de la altura superior (3) se le resta la media de la altura media (2).

Intervalos de confianza simultáneos de Tukey del 95%.

```
Intervalos de confianza simultáneos de Tukey del 95.0%
Variable de respuesta VENTAS
Todas las comparaciones de dos a dos entre los niveles de POSICIÓN
POSICIÓN = 1  restado a:

POSICIÓNInferior Centrada Superior  ---+---------+---------+---+---
2        7.377     11.00     14.62  (-----------------*-----------)
                                    ---+---------+---------+----+---
                                       8.0    10.0      12.0      14.0
```

Interpretación para posición:

- En el intervalo correspondiente al primer conjunto, a la media de la posición 2 menos la media de la posición 1, ambos signos del intervalo de confianza son positivos, por lo tanto existe evidencia de que ambos pares de medias son diferentes y que las ventas promedio cuando el estante se encuentra en la posición de las cajas son mayores que cuando se encuentra en la entrada por un mínimo de 7.377 y un máximo de 14.62 unidades venta.

```
Intervalos de confianza simultáneos de Tukey del 95.0%
Variable de respuesta VENTAS
Todas las comparaciones de dos a dos entre los niveles de ALTURA
ALTURA = 1  restado a:

ALTURA Inferior  Centrada  Superior --------+---------+---------+--------
2       4.237    9.6667    15.096                        (------*------)
3      -4.596    0.8333     6.263             (------*------)
                                     --------+---------+---------+--------
                                          -8.0       0.0       8.0
```

Interpretación para altura:

- El primer intervalo del segundo conjunto correspondiente a la media de la altura 2 menos la media de la altura 1, ambos signos del intervalo de confianza son positivos, por lo tanto existe evidencia de que ambos pares de medias son diferentes y que las ventas promedio cuando el producto se exhibe en la altura media del estante son mayores que cuando se exhibe en la parte inferior del estante por un mínimo de 4.237 y un máximo de 15.096.
- El segundo intervalo del segundo conjunto correspondiente a la media de la altura 3 menos la media de la altura 1 contiene cero en el intervalo de confianza, por lo tanto, no hay una evidencia significativa en alfa 0.05 en la diferencia de las medias.

```
ALTURA = 2  restado a:

ALTURA Inferior  Centrada  Superior  --------+---------+---------+--------
3        -14.26    -8.833    -3.404  (------*------)
                                     --------+---------+---------+--------
                                           -8.0       0.0       8.0
```


ISBN 978-607-8285-62-4

- El primer intervalo del tercer conjunto correspondiente a la media de la altura 3 menos la media de la altura 2, ambos signos del intervalo de confianza son negativos, por lo tanto existe evidencia de que ambos pares de medias son diferentes y que las ventas promedio cuando el producto se exhibe en la altura media del estante son mayores que cuando se exhibe en la parte superior del estante por un mínimo de 3.404 y un máximo de 14.26 unidades venta.

Efecto de la interacción entre los niveles del Factor 1 y los niveles del Factor 2.

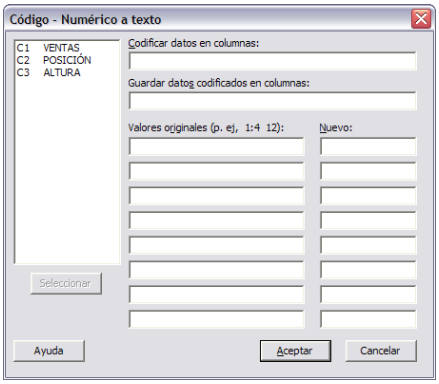
Solución al inciso f.

Efectos de la Interacción

Para construir la gráfica de interacción primero hay que descodificar los niveles de los dos factores **POSICIÓN Y ALTURA** de la siguiente manera:

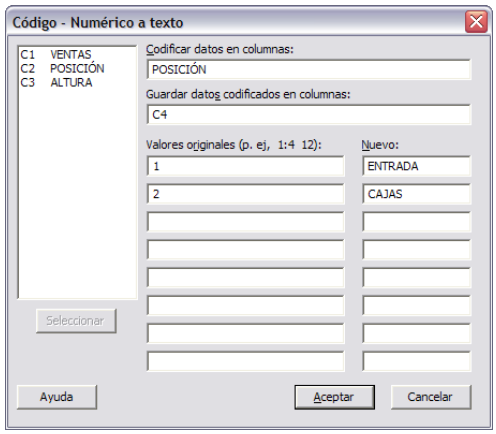
Seleccione las opciones ***Codificar>Numérico a texto*** de la barra menú ***Datos*** para que aparezca la siguiente ventana:

Cuadro de diálogo: Código-Numérico a texto.



En ***Codificar datos en columnas*** ingrese **C2 POSICIÓN**. En ***Guardar datos codificados en columnas***: ingrese **C4**. En el primer renglón de ***Valores originales*** ingrese **1** y en ***Nuevo***: ingrese **ENTRADA**; en el segundo renglón de ***Valores originales*** ingrese **2** y en ***Nuevo***: ingrese **CAJAS**. Haga clic en ***Aceptar***.

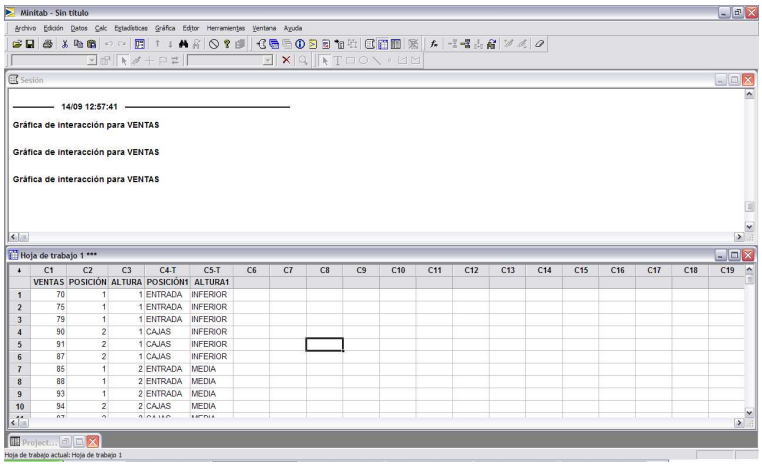
Cuadro de diálogo: Código-Numérico a texto.



ISBN 978-607-8285-62-4

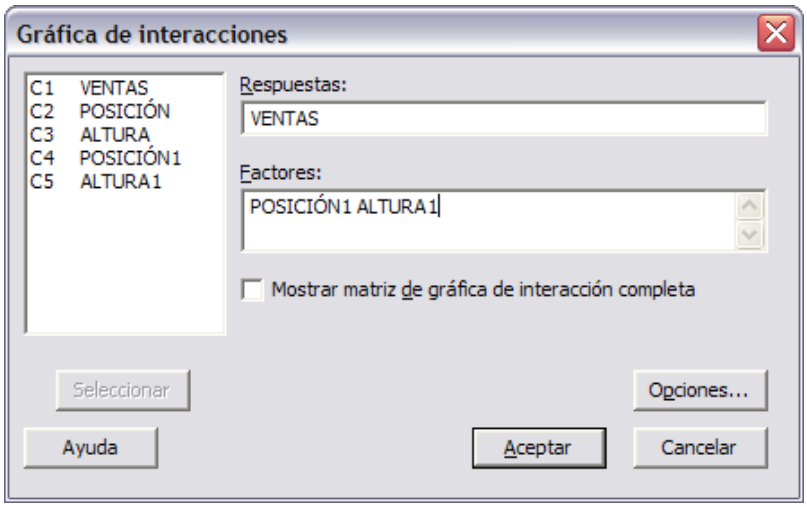
Nuevamente seleccione las opciones ***Codificar>Numérico a texto*** de la barra menú ***Datos.*** En ***Codificar datos en columnas*** ingrese **C3 ALTURA**. En ***Guardar datos codificados en columnas***: ingrese **C5**. En el primer renglón de ***Valores originales*** ingrese **1** y en ***Nuevo***: ingrese **INFERIOR**; en el segundo renglón de ***Valores originales*** ingrese **2** y en ***Nuevo***: ingrese **MEDIA**. En el tercer renglón de ***Valores originales*** ingrese 3 y en ***Nuevo***: ingrese **SUPERIOR** .Haga clic en ***Aceptar***. Nombre la ***columna C4*** como **POSICIÓN 1** y la ***columna C5*** como **ALTURA 1.**

Hoja de trabajo de Minitap



Seleccione las opciones ***ANOVA>Gráfica de interacciones*** de la barra menú ***Estadísticas*** para que aparezca la siguiente ventana:

Cuadro de diálogo: Gráfica de interacciones.

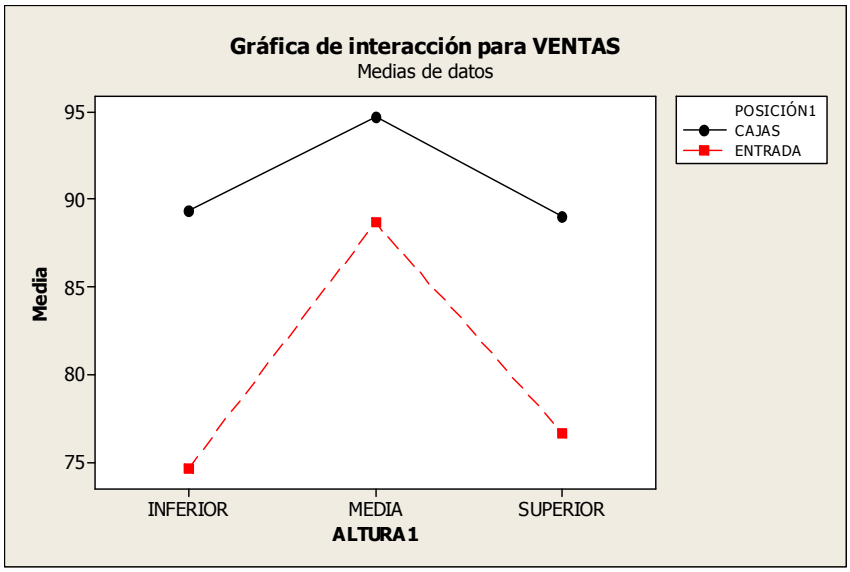


En ***Respuesta***, ingrese **VENTAS**. En ***Factores***, ingrese **POSICIÓN1ALTURA1 .** Haga clic en ***Aceptar***

ISBN 978-607-8285-62-4

La interacción (Sinergia) se presenta cuando el efecto de un factor depende del nivel del otro factor. Se puede utilizar una gráfica de interacción para visualizar posibles interacciones.

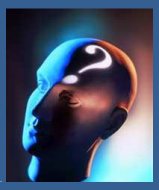
Las líneas paralelas en una gráfica de interacción indican que no hay interacción. Mientras mayor sea la diferencia en la pendiente entre las líneas cuando éstas se cruzan, mayor será el grado de interacción (sinergia). Sin embargo, la gráfica de interacción no dice si la interacción es estadísticamente significativa.



Interpretación de resultados:

En la figura anterior se han graficado las ventas promedio para cada nivel de altura en el estante donde se muestra el producto y la posición de éste en la tienda. **Las dos líneas** (que representan la posición del estante) **parecen ser aproximadamente paralelas.** Este fenómeno se puede interpretar como que la ***diferencia*** en ventas en las **dos posiciones es prácticamente la misma** para los tres niveles de altura donde se exhibe el producto. En otras palabras, **no hay interacción (sinergia**) entre estos dos factores como ya se determinó para la prueba *F* de interacción.

ISBN 978-607-8285-62-4



EJERCICIOS COMPLEMENTARIOS

1 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 1**

El gerente de una compañía de software desea estudiar, a través del tipo de industria, el número de horas que los directivos pasan frente a sus computadoras de escritorio. El gerente selecciono una muestra de cinco ejecutivos de cada una de las tres industrias.

Industria	**Observaciones o repeticiones**				
	1	**2**	**3**	**4**	**5**
Detallista	8	8	6	8	9
Bancaria	12	10	10	12	11
De seguros	10	8	6	8	9

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre el número promedio de horas que los directivos pasan frente a sus computadoras de escritorio en las tres industrias.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿En cuál o cuáles industrias los directivos pasan más número de horas promedio frente a sus computadoras y cuántas más?.

ISBN 978-607-8285-62-4

2 EJERCICIO COMPLEMENTARIO

EJERCICIO COMPLEMENTARIO 2

Una organización de consumidores tiene interés en determinar si hay alguna diferencia en la vida promedio de cuatro marcas diferentes de pilas para radios de transistores. Se probó una muestra aleatoria de cuatro pilas de cada marca con los siguientes resultados (en horas):

Marca	**Observaciones o repeticiones**			
	1	**2**	**3**	**4**
1	12	14	15	14
2	18	17	19	21
3	18	22	20	25
4	20	19	23	20

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre el número promedio de horas que duran las pilas de cuatro marcas diferentes.

b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿En cuál o cuáles marcas de pilas el número de horas promedio de duración es más grande y por cuánto más?.

3 EJERCICIO COMPLEMENTARIO

EJERCICIO COMPLEMENTARIO 3

El director de personal de una prestigiada marca, desea investigar el "perfeccionamiento" en el trabajo. A una muestra aleatoria de 18 empleados les aplicó un examen diseñado para medir el perfeccionamiento. Las puntuaciones van desde 20 hasta casi 40. Una de las facetas del estudio incluyó los antecedentes de cada empleado. . Las puntuaciones son:

Antecedentes	**Observaciones o repeticiones**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Región rural	35	30	36	38	39	34	31
Ciudad pequeña	28	24	25	30	32	28	
Metrópoli	24	28	26	30	32		

ISBN 978-607-8285-62-4

a) Utilice el nivel de significancia 0.05 para probar la hipótesis nula de que no existen diferencias significativas entre la puntuación promedio obtenida por los empleados y sus antecedentes.
b) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿En cuál o cuáles antecedentes los empleados obtuvieron un puntaje promedio más alto y por cuánto más alto?.

4 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 4**

Se tomaron muestras aleatorias de tamaño cinco, en cada una de tres poblaciones. La SCT fue de 100. La SCt fue de 40.

a) Establezca la hipótesis nula (H_0) y la hipótesis alternativa (H_1).
b) Seleccione y calcule el valor del estadístico de prueba apropiado.
c) Establezca la región de rechazo de (H_0). Use el nivel de significancia 0.05.
d) Formule una regla de decisión basada en los pasos 1,2 y 3 anteriores.
e) Tome una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interprete los resultados de la prueba (conclusión administrativa).

5 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 5**

Una compañía de publicidad a nivel nacional quiere saber si el tamaño de un anuncio y su colorido producen diferencia en la respuesta de los lectores de revistas. A una muestra aleatoria de lectores les fue presentada una serie de anuncios con tres colores distintos y tres tamaños diferentes. A cada lector se le pide que asigne una calificación, de 1 a 10, a cada combinación de color y tamaño. Supóngase que las calificaciones se distribuyen en forma aproximadamente normal. Las puntuaciones de cada combinación se muestran en la siguiente tabla.

Tamaño del anuncio	**Color del anuncio**		
	Rojo	**Azul**	**Naranja**
Pequeño	2	3	3
Mediano	3	5	6
Grande	6	7	8

ISBN 978-607-8285-62-4

a) Con un nivel de significancia de 0.05 ¿existe diferencia en la calificación promedio asignada por los lectores de acuerdo al tamaño del anuncio?.
b) Con un nivel de significancia de 0.05 ¿existe diferencia en la calificación promedio asignado por los lectores de acuerdo al color del anuncio?.
c) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿Cuál o cuáles tamaños de anuncio obtuvieron mayor calificación promedio y cuánto más?.
d) Determine la eficiencia relativa del diseño aleatorizado en bloques en comparación con el diseño completamente aleatorizado.

6 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 6**

Hay cuatro restaurantes de una cadena de comida rápida. Las cantidades de hamburguesas vendidas en cada uno de los establecimientos de 14 a 15 horas durante las últimas 6 semanas, se muestra a continuación.

Restaurantes	**Semana**					
	1	**2**	**3**	**4**	**5**	**6**
1	12	23	43	10	24	31
2	16	22	29	25	21	26
3	32	34	41	31	37	35
4	19	23	34	17	18	21

a) Con un nivel de significancia de 0.05 ¿existe diferencia en la cantidad de hamburguesas promedio vendidas de 14 a 14 horas en los cuatro restaurantes?.

b) Con un nivel de significancia de 0.05 ¿existe diferencia en la cantidad de hamburguesas promedio vendidas de 14 a 15 horas en las seis semanas?.

c) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿En cuál o cuáles restaurantes se vendieron en promedio más hamburguesas de 14 a 15 horas y cuántas más?.

d) Determine la eficiencia relativa del diseño aleatorizado en bloques en comparación con el diseño completamente aleatorizado.

ISBN 978-607-8285-62-4

7 EJERCICIO COMPLEMENTARIO

EJERCICIO COMPLEMENTARIO 7

En la ciudad de Aguascalientes, se emplea personal para estimar el valor de las casas con el propósito de establecer un impuesto sobre bienes raíces. El alcalde de la ciudad envía regularmente a cuatro asesores a cinco inmuebles y después compara los resultados. A continuación se proporciona la información, en miles de pesos. El resultado de un paquete de software estadístico es:

RESUMEN		**TABLA DE ANOVA**					
Inmueble	**Media**	**Fuente**	**GL**	**SC**	**MC**	$F_{Calc.}$	$F_{Crítica.}$
1	505.0	Inmueble	4				
2	877.5	Asesor			5,813		
3	515.0	Error			6,293		
4	500.0	Total		506,120			
5	592.5						

a) Con un nivel de significancia de 0.05 ¿existe diferencia en el valor promedio de las casas en los cinco inmuebles?.

b) Con un nivel de significancia de 0.05 ¿existe diferencia en el valor promedio de las casas de acuerdo con los cuatro asesores?.

c) Según el método *T* de Tukey de comparaciones múltiples, obtenga el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas (1-α=0.95) de los tratamientos ó niveles del factor. ¿Cuál o cuáles avalúos promedio de los inmuebles resultaron más grandes y por cuánto más?.

d) Determine la eficiencia relativa del diseño aleatorizado en bloques en comparación con el diseño completamente aleatorizado.

ISBN 978-607-8285-62-4

8 **EJERCICIO COMPLEMENTARIO**

EJERCICIO COMPLEMENTARIO 8

Se llevó a cabo un experimento para determinar el efecto del tamaño de la habitación y el color de la pared sobre un nivel medido de ansiedad en los sujetos. Para cada tamaño y color de habitación se probaron dos personas. Un valor grande significa un alto nivel de ansiedad. Los resultados fueron los siguientes:

Tamaño	**Color de la habitación**		
	Rojo (1)	**Amarillo (2)**	**Azul (3)**
Pequeña (1)	50 65	31 26	12 16
Mediana (2)	63 49	48 54	26 19
Grande (3)	68 59	63 57	20 16

a) ¿Hay algún efecto debido al tamaño de la habitación?.

b) ¿Afecto en algo el color de la habitación?.

c) ¿Hay interacción entre el tamaño y el color de la habitación?.

d) Construya una gráfica de interacción. Explique cómo la gráfica ilustra el grado con el que las interacciones están presentes.

e) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas ($1-\alpha=0.95$) de los distintos tamaños de habitación. ¿Cuál o cuáles tamaños de habitación obtuvieron mayores niveles de ansiedad promedio y por cuánto más?.

f) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas ($1-\alpha=0.95$) del color de la habitación. ¿Cuál o cuáles colores de habitación obtuvieron mayores niveles de ansiedad promedio y por cuánto más?.

ISBN 978-607-8285-62-4

9 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 9**

La junta de educación de un estado desea estudiar las diferencias en el número de alumnos de las clases entre las escuelas primaria, secundaria y preparatoria, en varias ciudades. Se seleccionó una muestra aleatoria de tres ciudades. Se eligieron dos escuelas al mismo nivel dentro de cada ciudad y se registró el número de alumnos promedio de clase para la escuela con los resultados siguientes:

Nivel educativo	**Ciudad**		
	A (1)	**B (2)**	**C (3)**
Primaria (1)	32 34	26 30	20 23
Secundaria (2)	35 39	33 30	24 27
Preparatoria (3)	43 38	37 34	31 28

a) ¿Hay algún efecto debido al nivel educativo de las escuelas?.

b) ¿Afecto en algo la ciudad donde se encuentran las escuelas?.

c) ¿Hay interacción entre el nivel educativo y la ciudad donde se encuentran las escuelas?.

d) Construya una gráfica de interacción. Explique cómo la gráfica ilustra el grado con el que las interacciones están presentes.

e) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas ($1-\alpha=0.95$) de los distintos niveles educativos de las escuelas. ¿Cuál o cuáles niveles educativos obtuvieron mayor número de alumnos promedio y por cuánto más?.

f) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas ($1-\alpha=0.95$) de las distintas ciudades. ¿En cuál o cuáles ciudades hubo mayor número de alumnos promedio y cuántos más?.

ISBN 978-607-8285-62-4

10 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 10**

La tabla adjunta se presentan los tiempos, en minutos, de conexión con una dirección de internet desde cuatro puntos geográficos de una región y en tres horas determinadas. El experimento se repetía cuatro veces y era diseñado para estudiar la influencia del factor "*hora de conexión*" y el factor "*lugar de la conexión*" en la variable de interés "*tiempo de conexión*". Analizar estos datos y estudiar la influencia de los dos factores.

Hora de conexión	**Lugar de la conexión**			
	1	**2**	**3**	**4**
1	31 45 46 43	82 110 88 72	43 45 63 76	45 71 66 62
2	36 29 40 23	92 61 49 124	44 35 31 40	56 102 71 58
3	22 21 18 23	30 37 38 29	23 25 24 22	50 56 51 53

a) ¿Hay algún efecto debido a la hora de conexión?.

b) ¿Afecto en algo el lugar de conexión?.

c) ¿Hay interacción entre la hora y el lugar de conexión?.

d) Construya una gráfica de interacción. Explique cómo la gráfica ilustra el grado con el que las interacciones están presentes.

e) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas ($1-\alpha=0.95$) de las diferentes horas de conexión. ¿En cuál o cuáles horas de conexión se obtuvo mayor número de minutos promedio y por cuánto más?.

f) Utilice el método *T* de Tukey de comparaciones múltiples, para obtener el rango crítico y compárelo con todos los contrastes y estime intervalos para todas las comparaciones apareadas ($1-\alpha=0.95$) de los lugares de conexión. ¿En cuál o cuáles lugares de conexión se obtuvo mayor número de minutos promedio y cuánto más?.

ISBN 978-607-8285-62-4

AUTOEVALUACIÓN CON REACTIVOS DE **FALSO Ó VERDADERO**

EN CADA UNO DE LOS REACTIVOS, CONTESTE CON UNA F SI CONSIDERA QUE LA AFIRMACIÓN ES FALSA Y CON UNA V SI CONSIDERA QUE LA AFIRMACIÓN ES VERDADERA.

1. La forma específica de una distribución F depende del número de grados de libertad en el numerador. **( )**

2. En el análisis de varianza, debemos suponer que las muestras se seleccionan de una población normal sin importar si cada una de estas poblaciones tiene la misma varianza. **( )**

3. En el diseño aleatorizado en bloques además de las suposiciones del análisis de varianza en un sentido se necesita suponer que no existe efecto de interacción entre los tratamientos y los bloques. **( )**

4. La distribución F es discreta, esto significa que puede tocar una cantidad infinita de valores entre cero y más infinito. **( )**

5. En una distribución F, conforme el número de grados de libertad aumenta, tanto en el numerador como en el denominador, la distribución se aproxima a una distribución normal. **( )**

6. Para utilizar ANOVA se debe suponer que la mayoría de las poblaciones deben seguir la distribución normal. **( )**

7. Para utilizar ANOVA se debe suponer que las poblaciones deben tener varianzas iguales (homoscedasticidad). **( )**

8. Para probar la homoscedasticidad en el análisis de varianza se utiliza la prueba T de Tukey. **( )**

9. En el diseño aleatorizado en bloques puede haber mas de una réplica para cada combinación de tratamientos y bloques. **( )**

10. El método de Tukey está diseñado específicamente para comparar dos ó más medias cualesquiera. **( )**

11. Para probar la igualdad de varianzas de dos ó más muestras se utiliza la prueba de Levene. **( )**

12. La forma específica de una distribución F depende del numero de grados de libertad en el numerador y denominador de la razón F. **( )**

ISBN 978-607-8285-62-4

( )

13. Una prueba de hipótesis donde se utiliza el estadístico F debe ser siempre de extremo derecho.

( )

14. Los tamaños de las muestras ó tratamientos en el análisis de varianza siempre deben ser iguales.

( )

15. La forma específica de una distribución F depende del número de grados de libertad en el denominador.

( )

16. El análisis de varianza se aplica para probar si las medias de dos o más poblaciones pueden considerarse iguales.

( )

17. En el análisis de varianza si la razón de las varianzas entre tratamientos y la varianza dentro de los tratamientos se acerca a 1 se rechaza la hipótesis nula de igualdad de medias.

( )

18. Cuando hay una gran interacción entre los niveles de dos factores, el diagrama de interacción la exhibe por medio de líneas no paralelas.

( )

19. La distribución F es sesgada positivamente, es decir la cola larga de la distribución se encuentra a la derecha.

( )

20. Usamos el análisis de varianza para decidir si las muestras fueron extraídas de poblaciones que tiene la misma media.

( )

21. Una segunda variable de tratamiento ó factor que cuando se incluye en el análisis de varianza tiene el efecto de reducir el error experimental se llama variable de bloqueo.

( )

22. A la suma de las diferencias elevadas al cuadrado entre las observaciones y sus medias de tratamiento se le denomina variación de tratamiento.

( )

23. A la suma de las diferencias elevadas al cuadrado entre cada observación y la media total se le denomina variación total.

( )

24. Si el valor del estadístico calculado de F en un análisis de varianza es pequeño, la tendencia a pensar que exista una diferencia entre las medias de los tratamientos es mayor.

( )

25. Si el valor del estadístico calculado de F en un análisis de varianza es pequeño, la tendencia a pensar que no exista una diferencia entre las medias de los tratamientos es mayor.

( )

26. Una técnica estadística con la cual se puede probar la igualdad de dos medias se denomina análisis de varianza.

( )

27. A la suma de las diferencias elevadas al cuadrado entre la media de cada tratamiento y la media total o global se le denomina variación de tratamiento.

( )

28. Una familia de distribuciones diferenciadas por dos parámetros y utilizada frecuentemente en el análisis de varianza se denomina distribución t de Student.

( )

ISBN 978-607-8285-62-4

29. Una familia de distribuciones diferenciadas por un solo parámetro y utilizada frecuentemente para probar la igualdad de dos medias se denomina distribución t de Student. **( )**

30. El análisis de varianza se basa en comparar la varianza dentro de los tratamientos con la varianza entre tratamientos. F **( )**

31. En el análisis de varianza si la razón de las varianzas entre tratamientos y la varianza dentro de los tratamientos se aleja de 1 se rechaza la hipótesis nula de igualdad de medias. **( )**

32. En el análisis de varianza, una vez calculada la razón F, el numerador y el denominador deberán ser aproximadamente iguales si la hipótesis nula de igualdad de medias no es verdadera. F **( )**

33. En un diseño de bloques aleatorizados debe haber por lo menos dos réplicas por celda. **( )**

34. Para utilizar ANOVA se debe suponer que las muestras se deben seleccionar en forma aleatoria. **( )**

35. A la suma de las diferencias elevadas al cuadrado entre la media de cada tratamiento y la media total o global se le denomina variación total. **( )**

36. El número máximo de réplicas que puede tener un diseño de dos factores es tres. **( )**

37. En el diseño aleatorizado en bloques además de las suposiciones del análisis de varianza en un sentido se necesita suponer que existe efecto de interacción entre los tratamientos y los bloques. **( )**

38. El análisis de varianza se basa en comparar la varianza entre tratamientos con la varianza dentro de los tratamientos. **( )**

39. A la suma de las diferencias elevadas al cuadrado entre las observaciones y sus medias de tratamiento se le denomina variación aleatoria (error). **( )**

40. Cuando no hay interacción, las diferencias entre las medias para los niveles de un factor son iguales para todos los niveles del otro factor; esta es la razón por la que los factores que no interactúan muestran líneas cruzadas en un diagrama de interacción. **( )**

41. En el diseño aleatorizado en bloques no todos los tratamientos ni todos los bloques deben tener el mismo tamaño **( )**

42. En el análisis de varianza en dos sentidos o dos factores debe haber por lo menos dos replicas para cada combinación de factores. **( )**

43. A la suma de las diferencias elevadas al cuadrado entre cada observación y la media total se le denomina variación de tratamiento. **( )**

44. El numero máximo de réplicas que puede tener el diseño aleatorizado en bloques es una. **( )**

45. En el diseño de dos factores además de las suposiciones del análisis de varianza en un sentido se necesita suponer que existe efecto de interacción entre los niveles de ambos factores. **( )**

ISBN 978-607-8285-62-4

AUTOEVALUACIÓN CON REACTIVOS DE OPCIÓN MÚLTIPLE

EN CADA UNO DE LOS REACTIVOS SIGUIENTES, SELECCIONE LA OPCIÓN QUE CONSIDERE CORRECTA.

1. La distribución *F*:
- a) No puede ser negativa.
- b) No puede ser positiva.
- c) Es la misma que la distribución t.
- d) Es la misma que la distribución X^2.

2. Los valores de *t*:
- a) Todos son negativos.
- b) Todos son positivos.
- c) Dependen de los grados de libertad.
- d) Todos los anteriores.

3. Conforme se aumenta el tamaño de la muestra, la distribución *t* se aproxima a:
- a) La distribución binomial.
- b) La distribución normal estándar o distribución z.
- c) La distribución de Poisson.
- d) La distribución *F.*

4. El estadístico de prueba para probar una hipótesis con muestras pequeñas, cuando no se conoce la desviación estándar poblacional es:

a)z b)t c)*F* d)X^2 .

5. Una muestra tiene 20 observaciones. Si se realiza una prueba de extremo derecho en la que se usa la distribución t como estadístico de prueba y el nivel de significación es 0.01, el valor crítico es:
- a) 2.5280.
- b) -2.8609.
- c) -2.0930.
- d) 2.8609.

6. Una muestra tiene 31 observaciones. Si se realiza una prueba bilateral en la que se usa la distribución t como estadístico de prueba y el nivel de significación es 0.05, el valor crítico es:
- a) $\mp$2.4573.
- b) -2.4573.
- c) $\mp$2.0423.
- d) +2.7500.

ISBN 978-607-8285-62-4

7. Cometemos un error de tipo I cuando:
- a) Se rechaza una hipótesis nula que es verdadera.
- b) No se rechaza una hipótesis alternativa que es verdadera.
- c) Se rechaza una hipótesis alternativa que es verdadera.
- d) No se rechaza una hipótesis nula que es falsa.

8. Cometemos un error de tipo II cuando:
- a) Se rechaza una hipótesis nula que es verdadera.
- b) No se rechaza una hipótesis alternativa que es verdadera.
- c) Se rechaza una hipótesis alternativa que es verdadera.
- d) No se rechaza una hipótesis nula que es falsa.

9. Una muestra tiene 21 observaciones y otra 30. Si se realiza una prueba de ANOVA en la que se usa la distribución F como estadístico de prueba y el nivel de significación es 0.05, el valor crítico es:
- a) 2.21.
- b) 1.94.
- c) 2.57.
- d) 2.86.

10. Si planteamos el siguiente juego de hipótesis $H_o: \mu \geq 240$ y $H_{1:}\mu < 240$
- a) Se trata de una prueba de extremo derecho.
- b) Se trata de una prueba bilateral o de dos colas.
- c) Se trata de una prueba de extremo izquierdo.
- d) Todas las anteriores.

11. Si planteamos el siguiente juego de hipótesis $H_o: \mu = 240$ y $H_{1:}\mu \neq 240$
- a) Se trata de una prueba de extremo derecho.
- b) Se trata de una prueba bilateral o de dos colas.
- c) Se trata de una prueba de extremo izquierdo.
- d) Todas las anteriores.

12. Si planteamos el siguiente juego de hipótesis $H_o: \mu \leq 240$ y $H_{1:}\mu > 240$
- a) Se trata de una prueba de extremo derecho.
- b) Se trata de una prueba bilateral o de dos colas.
- c) Se trata de una prueba de extremo izquierdo.
- d) Todas las anteriores.

13. Se obtiene una muestra aleatoria de 21 observaciones. Se realiza una prueba de hipótesis de extremo izquierdo con un nivel de significancia de 0.05. El valor del estadístico de t calculada es -2.8453, esto indica que: -2.086
- a) No se debe rechazar la H_0.
- b) Se debe rechazar H_0 .
- c) No se debe rechazar H_1 .
- d) b y c solamente.

ISBN 978-607-8285-62-4

14. Se quiere probar la hipótesis de que μ_2 es mayor que μ_1

a) Se debe emplear una prueba de extremo izquierdo.
b) Se debe emplear una prueba bilateral.
c) Se debe emplear una prueba de extremo derecho.
d) Con la información que se tiene no se puede determinar si se debe utilizar una prueba de extremo izquierdo, de extremo derecho, o una prueba bilateral.

15. Se quiere probar la hipótesis nula de que μ_2 es igual que μ_1

a) Se debe emplear una prueba de extremo izquierdo.
b) Se debe emplear una prueba bilateral.
c) Se debe emplear una prueba de extremo derecho.
d) Con la información que se tiene no se puede determinar si se debe utilizar una prueba de extremo izquierdo, de extremo derecho, o una prueba bilateral.

16. Se quiere probar la hipótesis nula de que μ_1 es mayor que μ_2

a) Se debe emplear una prueba de extremo izquierdo.
b) Se debe emplear una prueba bilateral.
c) Se debe emplear una prueba de extremo derecho.
d) Con la información que se tiene no se puede determinar si se debe utilizar una prueba de extremo izquierdo, de extremo derecho, o una prueba bilateral.

17. Se quiere probar la hipótesis nula de que μ_1 es igual que μ_2 y μ_3

a) Se debe emplear una prueba de extremo izquierdo.
b) Se debe emplear una prueba bilateral.
c) Se debe emplear una prueba de extremo derecho
d) Con la información que se tiene no se puede determinar si se debe utilizar una prueba de extremo izquierdo, de extremo derecho, o una prueba bilateral.

18. Se realizó una prueba ANOVA de dos factores y se rechazó la hipótesis nula para el Factor 1. Esto indica que:

a) Hay demasiados grados de libertad.
b) No hay diferencia entre las medias de los tratamientos del Factor 1.
c) Hay diferencia, en por lo menos uno de los tratamientos del Factor 1.
d) a y c solamente.

19. Se realizó una prueba ANOVA de dos factores y no se rechazó la hipótesis nula para el Factor 1. Esto indica que:

a) Hay demasiados grados de libertad.
b) No hay diferencia entre las medias de los tratamientos del Factor 1.
c) Hay diferencia, en por lo menos uno de los tratamientos del Factor 1.
d) a y b solamente.

20. ¿Cúal de los siguientes enunciados es un paso en la realización del análisis de variancia?

a) Determinar una estimación de la variancia de la población dentro de los tratamientos.
b) Determinar una estimación de la variancia de la población entre los tratamientos.
c) Determinar la diferencia entre la frecuencia observada y la esperada en cada clase.
d) a y b pero no c.

ISBN 978-607-8285-62-4

21. Una prueba bilateral con dos variancias va a realizarse en las muestra 1 y 2 con n_1 = 15 y n_2 = 12. Si a = .10, ¿Cuál de los siguientes enunciados representa el valor superior con que debería compararse s_1^2 / s_2^2 ?

a) $\frac{1}{F_{0.05,11,14}}$
b) $\frac{1}{F_{0.05,14,11}}$
$c)$ $F_{0.05,11,14}$
$d)$ $F_{0.05,14,11}$

22. Se realizó una prueba ANOVA de dos factores y se rechazó la hipótesis nula para el Factor 1. Esto indica que:

a) Hay demasiados grados de libertad.
b) No hay diferencia entre las medias de los tratamientos del Factor 1.
c) Hay diferencia, en por lo menos uno de los tratamientos del Factor 1
d) a y c solamente.

23. Una prueba bilateral con dos variancias va a realizarse en las muestra 1 y 2 con n_1 = 12 y n_2 = 15. Si a = .10, ¿Cuál de los siguientes enunciados representa el valor inferior con que debería compararse s_1^2 / s_2^2 ?

a) $\frac{1}{F_{0.05,11,14}}$
b) $\frac{1}{F_{0.05,14,11}}$
c) $F_{0.05,11,14}$
d) $F_{0.05,14,11}$

24. Suponga que está comparando 5 grupos sometidos a diferentes métodos de tratamiento y que ha seleccionado una muestra de tamaño 10 en cada uno y se ha calculado $\bar{X}$ en cada tratamiento ¿Cómo debe calcular ahora la gran media?

a) Multiplicando cada media muestral por $1/5$ y sumando estos valores. Dividiendo luego la suma entre 50.
b) Sumando las 5 medias muestrales y dividiendo entre 50 el resultado.
c) Sumando las 5 medias muestrales y multiplicando por $1/5$ el resultado.
d) Sumando las 5 medias muestrales.

25. ¿Cúal de las siguientes distribuciones tiene un par de grados de libertad?

a) De Poisson.
b) Normal.
c) *F.*
d) Binominal.

ISBN 978-607-8285-62-4

GLOSARIO DE ANOVA. PARTE 1

ALEATORIA, ASIGNACIÓN. Uso de métodos aleatorios para designar pacientes a tratamientos distintos o viceversa**.**

ALFA La probabilidad de cometer un error de tipo I. Se representa por la letra griega α.

ANCOVA. Análisis de la covarianza, del inglés Analysis of Covariance. Es una combinación del análisis de la varianza y de la regresión. Corrige el ANOVA del efecto que pueda tener una variable no controlada, covariable, sobre la variable respuesta mediante la función de regresión de la variable respuesta sobre la covariable.

ANOVA. Análisis de la varianza, del inglés Analysis of Variance. Se utiliza también el término ADEVA, de su traducción al castellano. Es un método para contrastar la homogeneidad de más de dos variables poblacionales mediante la descomposición de la suma total de los cuadrados de las diferencias entre valores observados de esas variables poblacionales y su media muestral, como medida de la heterogeneidad de esas poblaciones. Las variables poblacionales para cada tratamiento son concreciones de una variable general llamada variable respuesta o variable dependiente.

ANALISIS DE VARIANZA (ANOVA). Técnica estadística con que se prueba la igualdad de 3 o más medias muéstrales y que, por tanto, permite hacer inferencias sobre si las muestras provienen de poblaciones que tienen la misma media.

BETA Probabilidad de cometer un error de tipo II. Se representa por la letra griega β.

BLOQUE Es una segunda fuente de variación, además de los tratamientos.

BLOQUE, DISEÑO EN. En el análisis de varianza, un protocolo donde los sujetos en cada bloque o grupo están asignados a tratamiento diferente.

CONFIANZA, INTERVALO DE (IC). Espacio calculado a partir de los datos de una muestra, que tiene una probabilidad dada de comprender el parámetro desconocido.

CONFIANZA, LIMITES DE. Delimitan a un intervalo de confianza. Se calculan de los datos de la muestra y tienen una probabilidad dada de que el parámetro desconocido se ubique entre éstos.

CUADRADOS MEDIOS. Cociente entre la suma de cuadrados y los grados de libertad.

CUADRADO MEDIO DENTRO DEGRUPO. Estimación de la variación en el análisis de varianza. Se usa en el denominador de la prueba estadística ***F***.

CUADRADO MEDIO ENTRE GRUPOS. Estimación de la variación en el análisis de varianza. Se usa en el numerador de la estadística ***F***.

CUADRADOS SUMAS DE. Cantidad calculada en el análisis de varianza y usada para obtener cuadrados medios para la prueba ***F***.

ISBN 978-607-8285-62-4

GLOSARIO DE ANOVA. PARTE 2

DISTRIBUCIÓN *F* Se emplea como el estadístico de prueba en el ANOVA, así como en otras pruebas. Sus características principales son las siguientes:

1. El valor *F* nunca es negativo.
2. Es una distribución continua que se aproxima indefinidamente al eje X, pero nunca lo toca.
3. Tiene sesgo positivo.
4. Se basa en dos conjuntos de grados de libertad.
5. Como en el caso de la distribución *t*, existe una "familia" de distribuciones *F*.

DISTRIBUCIÓN *t*. Fue investigada y dada a conocer por William S. Gosset, en 1908, bajo el seudónimo de *Student.* Es similar a la distribución normal. Sus principales características son:

1. Es una distribución continua.
2. Puede tomar valores comprendidos entre menos infinito y más infinito.
3. Es simétrica respecto a su media de cero. Es más extendida y plana en su ápice que la distribución normal estándar.
4. Se aproxima a una distribución normal estándar conforme aumenta *n*.

Hay una "familia" de distribuciones *t*. Hay una distribución *t* para una muestra de 15 observaciones, otra distribución para una muestra de 16, y así sucesivamente.

EFECTOS. Medida de la influencia de los tratamientos en la heterogeneidad de las poblaciones.

EFECTOS FIJOS. Los tratamientos son fijos cuando los tratamientos de la experimentación coinciden con los tratamientos totales sobre los que tenemos que sacar conclusiones

EFECTOS ALEATORIOS. Los tratamientos son aleatorios cuando los tratamientos de la experimentación son una parte aleatoria del colectivo total de tratamientos sobre el que tenemos que inducir las conclusiones del contraste.

ERROR MUESTRAL. Error debido a la aleatoriedad muestral, a que sea un elemento y no otro el observado.

ESTADÍSTICO F. Función de datos muestrales que sigue la ley F de Snedecor al definirse como cociente de dos sumas de cuadrados divididas por sus grados de libertad, es decir como cociente de dos cuadrados medios.

ERROR DE TIPO I Se presenta cuando se rechaza una H_0 verdadera.

ERROR DE TIPO II Se presenta cundo no se rechaza H_0 falsa.

EXPERIMENTO. Proceso planeado de obtención de datos.

EXPERIMENTAL, ESTUDIO. Estudio comparativo en donde hay una intervención o manipulación. Se designa como una prueba o serie clínica cuando se aplica a seres humanos.

ISBN 978-607-8285-62-4

GLOSARIO DE ANOVA. PARTE 3

EXPERIMENTAL, ESTUDIO. Estudio comparativo en donde hay una intervención o manipulación. Se designa como una prueba o serie clínica cuando se aplica a seres humanos.

FACTOR. Característica que es el centro de indagación en un estudio**.** Cada una de las causas que influyen en la heterogeneidad de las poblaciones en estudio. Si hay un solo factor lo representamos por A, si hay dos o más lo representamos por A, B, C.

FACTOR DE BLOQUEO. Factor que puede afectar a la variable dependiente, o dicho de otra forma a la homogeneidad de las poblaciones, y, por ello, lo introducimos como factor secundario para eliminar la influencia que pudiera tener sobre las conclusiones del contraste. Véase explicación detallada en el modelo III del ANOVA.

GRADO DE LIBERTAD. Número de términos independientes de una suma.

GRADO DE LIBERTAD Es el número de elementos en una muestra que pueden variar libremente. Un parámetro en algunas distribuciones de probabilidad de uso común, por ejemplo, distribución ***t***, ji cuadrada y ***F***

GRAN MEDIA. Media del grupo completo de sujetos de todas las muestras del experimento.

GRAN MEDIA. Media del grupo completo sujeto a las muestras del experimento.

HIPÓTESIS Suposición, o conjetura, que hacemos sobre un parámetro de la población.

HIPÓTESIS ALTERNATIVA La conclusión que se acepta cuando se demuestra que la hipótesis nula es falsa. También se conoce como hipótesis de investigación.

HIPÓTESIS NULA Hipótesis, o suposición, acerca de un parámetro de la población que deseamos probar, generalmente una suposición del status quo (situación actual).

HOMOCEDASTICIDAD. Supuesto de igualdad de as varianzas poblacionales.

IGUALDAD FUNDAMENTAL DEL ANÁLISIS DE LA VARIANZA. Descomposición de la suma total de cuadrados de los elementos muestrales respecto de su media global en dos términos, el que mide la influencia del factor en la heterogeneidad y el que mide la influencia del error muestral o aleatoriedad muestral en dicha heterogeneidad.

INFORMACIÓN COMPLETAMENTE ALEATORIZADA. Esto ocurre cuando la información de la experimentación se reparte de forma totalmente aleatoria entre los tratamientos.

INFORMACIÓN ALEATORIZADA EN BLOQUES. Esto sucede cuando la información reparte por bloques de manera aleatoria entre los tratamientos.

MUESTRA ALEATORIA. Muestra de n sujetos (u objetos) seleccionada de una población de modo que cada uno tenga una probabilidad conocida de estar en la muestra**.**

MUESTRAS DEPENDIENTES. Las muestras dependientes se caracterizan porque se hace una medición, después una intervención, y de nuevo se realiza una medición. Las muestras en pares también son dependientes, ya que un mismo individuo u objeto es miembro de ambas muestras. Ejemplo: diez participantes en un maratón se pesaron antes y después de la carrera. Se quiere estudiar la cantidad media de peso corporal que pierden los participantes.

ISBN 978-607-8285-62-4

GLOSARIO DE ANOVA. PARTE 4

MUESTRAS INDEPENDIENTES. Las muestras tomadas aleatoriamente no están relacionadas una con otra. Se quiere estudiar la edad promedio de los internos en las prisiones A y B. Se toma una muestra aleatoria de 28 internos en la prisión A, y una muestra de 19 internos en la prisión B. Una persona no puede ser interno de ambas prisiones. Las muestras son independientes, no están relacionadas.

NIVELES DEL FACTOR O TRATAMIENTOS. Cada uno de los valores posibles del factor. Los representamos por A, si el factor se representa por A.

NIVEL DE SIGNIFICANCIA Valor que indica el porcentaje de los valores muestrales que se halla fuera de ciertos límites, suponiendo que la hipótesis nula sea correcta, esto es, la probabilidad de rechazarla cuando es verdadera.

Post-hoc, COMPARACIONES. Métodos para comparar medias después del Análisis de varianza.

PRUEBA DE DOS COLAS O BILATERAL. Método donde la hipótesis alternativa especifica una desviación a partir de la hipótesis nula en las dos direcciones. La región crítica o de rechazo se localiza en ambos extremos de la distribución de la estadística de prueba.

PRUEBA DE UNA COLA O DE UN EXTREMO. Prueba donde la hipótesis alternativa especifica una desviación de la hipótesis nula sólo en una dirección. La región crítica o de rechazo se localiza en un extremo de la distribución de la prueba estadística.

PRUEBA ESTADÍSTICA. Procedimiento empleado para probar una hipótesis nula (por ejemplo, prueba *t*, prueba ji cuadrada, prueba *F*).

PRUEBA DE HIPÓTESIS. Un enfoque para la inferencia estadística que conduce a una decisión para rechazar o no la hipótesis nula.

PRUEBA T DE TUKEY. Una prueba *a posteriori* para hacer comparaciones apareadas múltiples entre medias y después de obtener una prueba ***F*** significativa en el análisis de varianza. Es el método más recomendado por los estadígrafos.

RAZON F. aquella que se utiliza en el análisis de varianza, entre otras pruebas para comparar la magnitud de dos estimaciones de la variancia de la población y determina si ambas estimaciones son aproximadamente iguales; en el análisis de variancia, se emplea la razón de la variancia ente dos columnas con la variancia dentro de las columnas.

REGIÓN CRÍTICA. Región (o conjunto de valores) donde debe ocurrir una prueba estadística para rechazar la hipótesis nula.

SUMA DE CUADRADOS DENTRO DE LAS MUESTRAS. Suma de cuadrados debida al error muestral.

SUMA DE CUADRADOS ENTRE LAS MUESTRAS. Suma de cuadrados debida al factor.

TABLA DEL ANÁLISIS DE LA VARIANZA. Tabla donde se recogen todos los datos necesarios para realizar el contraste con el ANOVA. En el mismo sentido hablamos de la tabla del ANCOVA o del MANOVA.

TABLA DE CONTINGENCIA. La que tiene renglones R y columnas C. cada renglón corresponde a un nivel de una variable cada columna, a un nivel de otra variable. Las partes del cuerpo de las tablas son las frecuencias con que ocurre cada combinación de variables.

ISBN 978-607-8285-62-4

GLOSARIO DE ANOVA. PARTE 5

VALOR CRÍTICO. Cantidad que una prueba estadística debe exceder (en un sentido de valor absoluto) para poder rechazar la hipótesis nula.

VALOR CRÍTICO. Valor que nos delimita la región crítica, es decir, la región de rechazo de la hipótesis.

VALOR CRÍTICO Un valor que es el punto divisorio entre la región en la que se rechaza la hipótesis nula, y la región en la si no se rechaza. En una prueba de una cola hay sólo un valor crítico. En una prueba de dos colas hay dos valores críticos, uno en cada cola.

VALOR *p* Probabilidad de encontrar, para el estadístico de prueba, un valor tan extremo o más que el obtenido con los datos muestrales, dado que la hipótesis nula sea verdadera.

VARIABLE. En un estudio, característica de interés que tiene valores diferentes para distintos sujetos u objetos.

VARIABLE DEPENDIENTE. Variable en estudio o variable respuesta, es decir a variable que nos interesa medir al aplicar los tratamientos como niveles del factor. Dicho factor sería independiente y el objeto del ANOVA es medir la influencia del factor (independiente) sobre la variable dependiente. Las variables que se asocian a los tratamientos son concreciones, para esos tratamientos, de la variable dependiente general. De tal manera que decir que esas variables son homogéneas equivale a decir que no existe influencia del factor sobre la variable dependiente.

VARIABLE INDEPENDIENTE. Variable explicatoria o "predictora" en un estudio. Se conoce en ocasiones como un factor en **ANOVA**.

VARIABLES POBLACIONALES. Las representamos por ξ_i y están asociadas a los conjuntos o grupos de valores que se obtienen aplicando los tratamientos A_i. Por ello podemos hablar de variables de tratamiento o grupos de tratamiento.

VARIANCIA DENTRO DE TRATAMIENTOS. Estimación de la variancia de la población que se basa en las variancias dentro de las muestras o tratamientos *k*, que usa un promedio ponderado de k variancias muéstrales.

VARIANCIA ENTRE TRATAMIENTOS. Estimación de la variancia de la población derivada de la variancia entre las medias muéstrales.

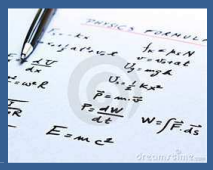
ISBN 978-607-8285-62-4

αA

SIMBOLOGÍA

$=$	Igual	ε	Letra griega épsilon; usada para simbolizar el error experimental
$\neq$	Desigual	t	Símbolo para la razón t (la razón crítica que sigue a una distribución t)
$<$	Menor que	X	Variable independiente (explicatoria, predictora) en regresión
$\leq$	Menor que o igual a	$\bar{X}$	Media de la muestra; X con barra
$>$	Mayor que	Y	Variable dependiente (resultado, respuesta, criterio) en regresión
$\geq$	Mayor que o igual que	SCT	Suma de Cuadrados Total
H_o	Hipótesis nula	SC_t	Suma de cuadrados de Tratamientos
H_1	Hipótesis alterna	SCE	Suma de Cuadrados del Error
α	Letra griega alfa; probabilidad de un error tipo I	SC_b	Suma de Cuadrados de Bloque
β	Letra griega beta; probabilidad de un error tipo II	SC_A	Suma de cuadrados del Factor A
μ	Letra griega mu; media de la población	SC_B	Suma de Cuadrados del Factor B
σ	Legra griega minúscula sigma; desviación estándar de población	SC_{AB}	Suma de Cuadrados de la interacción de A con B
τ	Legra griega tau; usada para simbolizar términos en el modelo ANOVA	CMt	Cuadrado Medio de Tratamientos
Σ	Letra griega mayúscula signa; símbolo que indica una suma	CME	Cuadrado Medio del Error
$g.l.$	Grados de libertad	CMb	Cuadrado Medio de Bloques
F	Símbolo para la prueba y la distribución F	CM_A	Cuadrado Medio del Factor A
S	Desviación estándar de la muestra	CM_B	Cuadrado Medio del Factor B
SE	Error estándar de la muestra	CM_{AB}	Cuadrado Medio de la Interacción de A con B
n	Tamaño de la muestra		

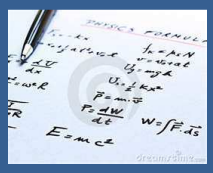
ISBN 978-607-8285-62-4



FÓRMULAS CLAVE. PARTE 1

• **Suma de Cuadrados, Total** **Anova de una vía. Diseño balanceado** $SCT \sum_{I=1}^{k}\sum_{j=1}^{n} X_{ij} - \bar{X}..)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij}^2 - \frac{X_{..}^2}{N}$	**(1)**	• **Suma de Cuadrados, Tratamientos** **Anova de una vía. Diseño balanceado** $SC_{Tratamientos} = \sum_{i=1}^{k} \frac{X_{i.}^2}{n} - \frac{X_{..}^2}{N}$	**(2)**
• **Suma de cuadrados, Error** **Anova de una vía. Diseño balanceado** $SCE = \sum_{i=1}^{k}\sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2$	**(3)**	• **Suma de Cuadrados, Total** **Anova de una vía. Diseño desbalanceado** $SCT = \sum_{I=1}^{k}\sum_{j=1}^{n_i} X_{ij} - \bar{X}..)^2$ $= \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{X_{..}^2}{N}$	**(4)**
• **Cuadrado Medio, Tratamientos** **Anova de una vía.** $CM_{trat} = \frac{SC_{trat}}{k-1}$	**(5)**	• **Cuadrado Medio, Error** **Anova de una vía** $CME = \frac{SCE}{N-k}$	**(6)**
• **Prueba *F* para tratamientos** **Anova de una vía** $F_{calc} = \frac{CMt}{CME}$	**(7)**	• **Rango crítico. Prueba T de Tukey** **Anova de una vía. Diseño balanceado** $Rango\ ó alcance\ crítico = q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}}$	**(8)**
• **Intervalo de confianza para comparaciones múltiples de Tukey.** **Anova de una vía. Diseño balanceado** $(\mu_{i.} - \mu_{i.}') = (\bar{X}_{i.} - \bar{X}_{i'}') \mp q_{\alpha(k,N-k)}\sqrt{\frac{CME}{n}}$	**(9)**	• **Intervalo de confianza para comparaciones múltiples de Tukey. Ecuación alternativa** **Anova de una vía. Diseño balanceado.** $(\mu_{i.} - \mu_{i.}') = (\bar{X}_{i.} - \bar{X}_{i'}') \mp q_{\alpha(k,N-k)}\sqrt{\frac{CME}{2}\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}$	**(10)**

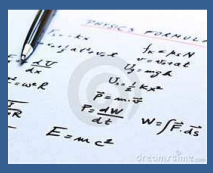
ISBN 978-607-8285-62-4



FÓRMULAS CLAVE. PARTE 2

• **Suma de Cuadrados, Tratamientos Anova de una vía. Diseño desbalanceado** $SC_{Tratamientos} = \sum_{i=1}^{k} \frac{X_{i.}^2}{n_i} - \frac{X_{..}^2}{N}$	**(11)**	• **Suma de Cuadrados, Error Anova de una vía. Diseño desbalanceado** $SCE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$	**(12)**
• **Rango crítico. Prueba T de Tukey Anova de una vía. Diseño desbalanceado** $Rango\ ó\ alcance\ crítico = q_{\alpha(k,N-k)} \sqrt{\frac{CME}{n_i}}$	**(13)**	• **Media armónica Anova de una vía. Diseño desbalanceado** $n_h(media\ armónica) = \frac{k}{\sum_{i=1}^{k} \frac{1}{n_i}}$	**(14)**
• **Intervalo de confianza para comparaciones múltiples de Tukey. Anova de una vía. Diseño desbalanceado** $(\mu_{i.} - \mu_{i.}') = (\bar{X}_{i.} - \bar{X}_{i.}') \mp q_{\alpha(k,N-k)} \sqrt{\frac{CME}{n_i}}$	**(15)**	• **Media armónica para intervalos Anova de una vía. Diseño desbalanceado** $n_h(media\ armónica) = \frac{k(grupos\ comparados)}{\sum_{i=1}^{k} \frac{1}{n_{i(grupos\ comparados)}}}$	**(16)**
• **Suma de Cuadrados, Total Diseño de bloques al azar** $SCT = \sum_{i=1}^{k} \sum_{j=1}^{b} X_{ij}^2 - \frac{X_{..}^2}{bk}$	**(17)**	• **Suma de cuadrados, Tratamientos Diseño de bloques al azar** $SC_{tratamientos} = \sum_{i=1}^{k} \frac{X_{i.}^2}{b} - \frac{X_{..}^2}{bk}$	**(18)**
• **Suma de cuadrados, Bloque Diseño de bloques al azar** $SC_{bloques} = \sum_{j=1}^{b} \frac{X_{.j}^2}{k} - \frac{X_{..}^2}{bk}$	**(19)**	• **Suma de cuadrados, Error Diseño de bloques al azar** $SCE = \sum_{I=1}^{K} \sum_{J=1}^{b} \left(X_{ij} - \bar{X}i. - \bar{X}_{.j} + \bar{X}_{..}\right)^2$	**(20)**

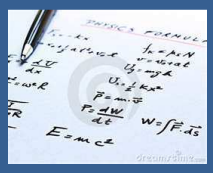
ISBN 978-607-8285-62-4



FÓRMULAS CLAVE. PARTE 3

• **Cuadrado Medio, Tratamientos** **Diseño de bloques al azar** $CM_{trat} = \frac{SC_{trat}}{k-1}$	**(21)**	• **Cuadrado Medio, Bloques** **Diseño de bloques al azar** $CM_{bloque} = \frac{SC_{bloque}}{b-1}$	**(22)**
• **Cuadrado Medio, Error** **Diseño de bloques al azar** $CME = \frac{SCE}{(k-1)(b-1)}$	**(23)**	• **Eficiencia Relativa** **Diseño de Bloques al azar** RE = (b-1) CM_{bloque} + b(k-1)CME / (bk-1)CME	**(24)**
• **Prueba F para Bloques** $F_{CALC.} = \frac{CM_b}{CME}$	**(25)**	• **Rango crítico. Prueba T de Tukey** **Anova de bloques al azar** $rango\ ó\ alcance\ crítico = q_{\alpha,k,(k-1)(b-1)}\sqrt{\frac{CME}{b}}$	**(26)**
• **Intervalo de confianza para comparaciones múltiples de Tukey. Anova de bloques al azar** $(\mu_{i.} - \mu_{i.}') = (\bar{X}_{i.} - \bar{X}_{i.}') \mp q_{\alpha,k,(k-1)(b-1)}\sqrt{\frac{CME}{b}}$	**(27)**	• **Suma de Cuadrados, Total** **Anova de dos factores** $SCT = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} X_{ijk}^2 - \frac{X_{...}^2}{abn}$	**(28)**
• **Suma de Cuadrados, Factor A** **Anova de dos factores** $SC_A = \sum_{i=1}^{a}\frac{X_{i..}^2}{bn} - \frac{X_{...}^2}{abn}$	**(29)**	• **Suma de Cuadrados, Factor B** **Anova de dos factores** $SC_B = \sum_{j=1}^{b}\frac{X_{.j.}^2}{an} - \frac{X_{...}^2}{abn}$	**(30)**

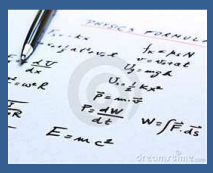
ISBN 978-607-8285-62-4



FÓRMULAS CLAVE. PARTE 4

• **Suma de cuadrados, Subtotales** **Anova de dos Factores** $SC_{SUBTOTALES} = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{X_{ij.}^2}{n} - \frac{X_{...}^2}{abn}$	**(31)**	• **Suma de Cuadrados, Interacción de AB** **Anova de dos Factores** $SC_{AB} = SC_{SUBTOTALES} - SC_A - SC_B$	**(32)**
• **Cuadrado Medio, Factor A** **Anova de dos Factores** $CM_A = \frac{SC_A}{a-1}$	**(33)**	• **Cuadrado Medio, Factor B** **Anova de dos Factores** $CM_B = \frac{SC_B}{b-1}$	**(34)**
• **Cuadrado Medio, Interacción de AB** **Anova de dos Factores** $CM_{AB} = \frac{SC_{AB}}{(a-1)(b-1)}$	**(35)**	• **Cuadrado Medio, Error** **Anova de dos Factores** $CME = \frac{SCE}{ab(n-1)}$	**(36)**
• **Prueba F para Factor A** **Anova de dos Factores** $F_{CALC.} = \frac{CM_A}{CME}$	**(37)**	• **Prueba F para Factor B** **Anova de dos factores** $F_{CALC.} = \frac{CM_B}{CME}$	**(38)**
• **Prueba F para Interacción de AB** **Anova de dos Factores** $F_{CALC.} = \frac{CM_{AB}}{CME}$	**(39)**	• **Rango crítico. Prueba T de Tukey para Factor A** **Anova de dos Factores** $rango\ ó\ alcance\ crítico = q_{\alpha,a,(ab)(n-1)}\sqrt{\frac{CME}{bn}}$	**(40)**

ISBN 978-607-8285-62-4



FÓRMULAS CLAVE. PARTE 5

• **Rango crítico. Prueba T de Tukey para Factor B** **Anova de dos Factores** $rango\ ó\ alcance\ crítico$ $= q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}}$	**(41)**	• **Intervalo de confianza para comparaciones múltiples de Tukey para Factor A** **Anova de dos Factores** $(\mu_{i..} - \mu_{i..}^{'})$ $= (\bar{X}_{i..} - \bar{X}_{i..}^{'})$ $\mp q_{\alpha,a,(ab)(n-1)}\sqrt{\frac{CME}{bn}}$	**(42)**
• **Intervalo de confianza para comparaciones múltiples de Tukey para Factor B** **Anova de dos Factores** $(\mu_{i..} - \mu_{i..}^{'})$ $= (\bar{X}_{i..} - \bar{X}_{i..}^{'})$ $\mp q_{\alpha,b,(ab)(n-1)}\sqrt{\frac{CME}{an}}$	**(43)**		

ISBN 978-607-8285-62-4

ESTADÍSTICA II

CUADERNO DE TRABAJO

CAPÍTULO 2
REGRESIÓN LÍNEAL SIMPLE

Javier Bech Vertti

ISBN 978-607-8285-62-4

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

Hecho en México / *Made in Mexico*

ISBN 978-607-8285-62-4

CONTENIDO

ISBN 978-607-8285-62-4

ISBN 978-607-8285-62-4

ISBN 978-607-8285-62-4

CAPÍTULO 2. ANÁLISIS DE REGRESIÓN LINEAL SIMPLE

OBJETIVO 2.1 El alumno podrá identificar las propiedades y la estructura de la covarianza y la correlación entre variables.

ANTECEDENTES

CONCEPTOS DE:

Población. Muestra. Variable. Tipos de variable. Escalas de medición de las variables. La media de la población. Tamaño de la muestra. Ejes cartesianos. Varianza de la población. Desviación estándar de la población. Varianza muestral. Desviación estándar muestral.

2.1.1 CONCEPTO DE PARÁMETRO, DIAGRAMAS DE DISPERSIÓN Y COEFICIENTE DE CORRELACIÓN MUESTRAL.

CONCEPTOS BÁSICOS REGRESIÓN LINEAL SIMPLE

En estadística, un **parámetro** es un número que resume una enorme cantidad de datos que pueden derivarse del estudio de una **variable estadística**. El cálculo de este número está **bien definido**, usualmente mediante una **fórmula aritmética** obtenida a partir de **datos de la población.**

ISBN 978-607-8285-62-4

Los **parámetros** estadísticos son una consecuencia inevitable del propósito esencial de la estadística: **crear un modelo** de la realidad. Los **parámetros** $\boldsymbol{\beta_0}$ **y** $\boldsymbol{\beta_1}$ del modelo de regresión se estiman mediante los valores $\widehat{\boldsymbol{\beta}}_0$ y $\widehat{\boldsymbol{\beta}}_1$ con base en los datos muestrales.

Transferencia de información tabular a una gráfica

El **diagrama de dispersión** es un tipo de diagrama matemático que utiliza las **coordenadas cartesianas** para mostrar los valores de **dos variables** para un conjunto de datos. Los datos se muestran como **un conjunto de puntos**, cada uno con el valor de la variable independiente ***X*** que determina la posición en el eje horizontal y el valor de la variable dependiente ***Y*** determinado por la posición en el eje vertical. Un diagrama de dispersión se llama también ***gráfico de dispersión***.

Relación entre variables
Relación directa entra X y Y
Relación inversa entre X y Y

La **covarianza** mide la medida en que dos variables **"varían juntas"**. Un **signo positivo** indica una **relación directa**, en tanto que un **signo negativo** indica **relación inversa**. La fórmula para las covarianzas muestrales es

$$\boldsymbol{cov(X,Y) = \frac{\sum[(X-\bar{X})(Y-\bar{Y})]}{n-1}}$$

Coeficiente de correlación muestral

Mientras que el **coeficiente de correlación** puede variar solamente entre **-1.00 y +1.00** y, por lo general, se le considera como **una medida de la relación**, la covarianza no tiene esos límites y no es una medida generalizada. La fórmula que permite transformar la covarianza en el **coeficiente de correlación muestral** es :

$$\boldsymbol{r = \frac{cov(X,Y)}{S_X S_Y}}$$

Desviación estándar de X **Desviación estándar de Y**

$$\boldsymbol{S_X = \sqrt{\frac{\sum_{i=1}^{n} X^2 - n\bar{X}^2}{n-1}}} \qquad \boldsymbol{S_Y = \sqrt{\frac{\sum_{i=1}^{n} Y^2 - n\bar{Y}^2}{n-1}}}$$

2.1.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 2.1.1.1 CÁLCULO DE LA COVARIANZA Y CORRELACIÓN

El director general de una cadena de tiendas de autoservicio en expansión desea conocer el comportamiento de las ventas en los diferentes establecimientos con base en la superficie de piso en la que se exhiben los diferentes productos con el fin de contar con un modelo que le permita llevar un control adecuado de la eficiencia con la que trabaja cada establecimiento. Para ello utiliza el volumen de ventas mensuales (en millones de pesos) y la superficie de piso (en miles de metros cuadrados). En forma aleatoria recopila el volumen de ventas del último mes en diez tiendas de la cadena que correspondan más o menos entre 2,000 y 12,000 metros cuadrados de superficie de piso.

ISBN 978-607-8285-62-4

Tienda	1	2	3	4	5	6	7	8	9	10
Superficie *(X)* en miles de m^2	2.15	9.20	6.70	13.50	5.50	12.15	4.80	10.70	3.25	8.25
Ventas *(Y)* en millones de pesos	1.0	3.0	3.0	4.5	2.0	5.0	1.0	4.0	1.5	3.5

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación

Cálculo de la covarianza

Solución al inciso a.

Tienda	Superficie de piso (X) en miles de metros cuadrados	Volumen Ventas (Y) en millones de pesos	X^2	Y^2	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
1	2.15	1.0	4.6225	1.0	-5.47	-1.85	10.12
2	9.20	3.0	84.64	9.0	1.58	0.15	0.237
3	6.70	3.0	44.89	9.0	-0.920	0.15	-0.138
4	13.50	4.5	182.25	20.3	5.88	1.65	9.7□2
5	5.50	2.0	30.25	4.0	-2.12	-0.850	1.802
6	12.15	5.0	147.6225	25.0	4.53	2.15	9.74
7	4.80	1.0	23.04	1.0	-2.82	-1.85	5.217
8	10.70	4.0	114.49	16.0	3.08	1.15	3.542
9	3.25	1.5	10.5625	2.3	-4.37	-1.35	5.90
10	8.25	3.5	68.0625	12.3	0.63	0.62	0.410
SUMAS	**76.2**	**28.5**	**710.43**	**99.75**			**46.532**
	PROMEDIOS						
	7.62	2.85					

$$cov(X,Y)=\frac{\sum[(X-\bar{X})(Y-\bar{Y})]}{n-1}=\frac{46.532}{9}=5.17$$

Cálculo del coeficiente de correlación muestral

Solución al inciso b.

$$r=\frac{cov(X,Y)}{S_X S_Y}$$

ISBN 978-607-8285-62-4

$$S_X = \sqrt{\frac{\sum_{i=1}^{n} X^2 - n\bar{X}^2}{n-1}} = \sqrt{\frac{710.43 - 10(7.62)^2}{9}} = 3.797$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n} Y^2 - n\bar{Y}^2}{n-1}} = \sqrt{\frac{99.75 - 10(2.85)^2}{9}} = 1.435$$

$$r = \frac{cov(X,Y)}{S_X S_Y} = \frac{5.17}{(3.797)(1.435)} = 0.9489$$

2.1.1.1 ACTIVIDAD DE APRENDIZAJE

ACTIVIDAD DE **APRENDIZAJE 2.1.1.1**
CÁLCULO DE LA **COVARIANZA Y CORRELACIÓN**

Una compañía refresquera está estudiando el efecto de su última campaña publicitaria. A un grupo de personas a quienes escogió al azar se les pregunto por teléfono cuantas latas del nuevo refresco habían comprado en la semana anterior y cuantos anuncios de él habían leído o visto esa semana.

Persona	1	2	3	4	5	6	7	8	9	10
No. de anuncios (X)	4	10	3	0	1	4	2	5	6	8
Latas compradas (Y)	12	19	7	6	3	5	6	10	7	11

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ISBN 978-607-8285-62-4

Cálculo de la covarianza

Solución al inciso a.

Obs.	X	Y	X^2	Y^2	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
SUMAS							
	PROMEDIOS						

$$cov(X,Y) = \frac{\sum[(X-\bar{X})(Y-\bar{Y})]}{n-1} =$$

Cálculo del coeficiente de correlación muestral

Solución al inciso b.

$$r = \frac{cov(X,Y)}{S_X S_Y}$$

$$S_X = \sqrt{\frac{\sum_{i=1}^{n} X^2 - n\bar{X}^2}{n-1}} =$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n} Y^2 - n\bar{Y}^2}{n-1}} =$$

$$r = \frac{cov(X,Y)}{S_X S_Y} =$$

ISBN 978-607-8285-62-4

2.1.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 2.1.1.1
CÁLCULO DE LA COVARIANZA Y CORRELACIÓN

El propietario de una cadena de heladerías desea estudiar el efecto de la temperatura atmosférica sobre las ventas durante la temporada de verano. Seleccionó una muestra aleatoria de 12 días con los resultados siguientes:

Día	1	2	3	4	5	6	7	8	9	10	11	12
Temperatura en º C. (X)	22.8	23.9	24.1	25.3	26.7	27.8	29.5	31.1	32.2	33.4	36.7	37.8
Ventas en miles de pesos (Y)	18.0	20.5	21.8	22.9	23.6	22.5	26.8	29.0	31.4	32.4	34.0	32.8

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ISBN 978-607-8285-62-4

Cálculo de la covarianza

Solución al inciso a.

Obs.	X	Y	X^2	Y^2	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
SUMAS							
	PROMEDIOS						

$$cov(X,Y) =$$

Cálculo del coeficiente de correlación muestral

Solución al inciso b.

$$r =$$

$$S_X =$$

$$S_Y =$$

$$r =$$

ISBN 978-607-8285-62-4

2.1.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 2.1.1. CÁLCULO DE LA COVARIANZA Y CORRELACIÓN

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

2.1.1.1 El presidente de una fábrica de computadoras desea estudiar la relación que hay entre el tamaño del incremento anual de sueldos y rendimientos de un representante de ventas en el año siguiente. Muestreo a 10 representantes y determino los tamaños de sus respectivos incrementos (dados en porcentaje de sus sueldos individuales) y el numero de ventas realizadas por cada uno durante los 12 meses después del incremento.

Representante	1	2	3	4	5	6	7	8	9	10
Incremento en porcentaje (X)	7.2	6.5	7.4	5.7	6.8	8.9	7.7	10.6	11.3	8.2
No. de ventas (Y)	55	39	58	36	41	70	53	74	66	73

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación

2.1.1.2 Una cadena de tiendas de repostería ha tenido grandes fluctuaciones en sus ingresos durante los últimos años. Abundantes baratas, nuevos productos y técnicas de publicidad se han utilizado durante este tiempo, por lo cual es difícil determinar cuáles de estos factores tienen la influencia más profunda en las ventas (Y). El departamento de mercadotecnia ha estudiado varias relaciones y piensa que los gastos mensuales destinados a carteles (X) pueden ser significativos. Muestreó 10 meses y descubrió lo siguiente:

Mes	1	2	3	4	5	6	7	8	9	10
(X)	25	16	43	34	10	21	19	13	23	38
(Y)	34	14	55	32	26	29	20	20	30	39

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación

ISBN 978-607-8285-62-4

OBJETIVO 2.2 El alumno podrá identificar las características, principios y propósitos del Análisis de Regresión Lineal Simple.

ANTECEDENTES

CONCEPTOS DE:

Variable aleatoria. Tipos de variable. Escala de medición de las variables. Ecuación de tendencia lineal. Ordenada al origen. Pendiente de la recta. Relación lineal. Relación curvilínea.

2.2.1 NATURALEZA DE LA REGRESIÓN: LINEAL Y NO LINEAL. LA ECUACIÓN DE PREDICCIÓN. VARIABLES DE RESPUESTA Y PREDICTORA

CONCEPTOS BÁSICOS NATURALEZA DE LA REGRESIÓN LINEAL SIMPLE

Variables independientes y dependiente

La finalidad del **análisis de regresión lineal simple** es conocer **la forma y función matemática** en que están relacionadas **las variables**. En este análisis los elementos presentan **dos valores**, uno para cada variable que se considera.

Una vez conocida tal función es posible determinar el comportamiento de la variable objeto de estudio, denominada variable **dependiente o predictiva (Y)** , en términos de las variaciones de otra variable **denominada independiente o predictor (X).**

Cuando la función de regresión esta conformada por **dos variables**, se **llama modelo de regresión simple,** en el caso **de dos o más variables independientes**, este modelo es conocido como **de regresión múltiple**. Dependiendo de la **forma de relación** entre las dos o más variables tenemos: **Regresión lineal si la relación se expresa mediante una línea recta; y regresión curvilínea cuando la relación es del tipo exponencial, parabólico, potencial, etc.**

ISBN 978-607-8285-62-4

Una vez identificado el modelo de regresión, es posible determinar los **parámetros** de la función elegida.

Los **supuestos** sobre los que descansa el **análisis de regresión** son los siguientes:

Relación entre variables

1.- La **variable dependiente** debe ser una **variable aleatoria**.
2.- La **relación** entre **ambas variables** debe ser **lineal.**
3.- La **distribución de los valores** de la **variable dependiente** para cada uno de los valores de la variable independiente **debe ser normal**.
4.- La **variancia** de las **distribuciones de la variable dependiente** para cada valor de la variable independiente **debe ser la misma** (Homoscedasticidad).
5.- El **error** (diferencia "residual" entre un valor observado y uno predicho de Y) **debe ser independiente** para **cada valor de *X***.

El **análisis** de **regresión lineal simple** y el de **correlación** son un **proceso** que consta de los siguientes pasos en general:

Pasos para el análisis de regresión y correlación lineal simple

1.- Definir la **ecuación de regresión lineal simple**.
2.- Examinar el **error estándar de estimación** para la regresión lineal simple.
3.- Probar la **significación de la relación** entre la variable dependiente y la variable explicativa.
4.- Construir **intervalos de confianza** para $\mu_{Y.X} = Y$.
5.- Calcular el **coeficiente de determinación** para medir la proporción de la variación que se explica por la variable independiente en el modelo de regresión y aplicar el análisis de **correlación lineal** simple para medir la fuerza de la asociación en el modelo de regresión lineal simple.
6.- Realizar un **diagnóstico de la regresión** mediante el **análisis de los residuales** estandarizados para estudiar **posibles violaciones a las suposiciones** del modelo de regresión.
7.- Realizar un **diagnóstico de la regresión** mediante el **análisis de influencias** para evaluar lo apropiado de un modelo en particular en relación con el **efecto potencial o la "influencia"** de cada punto sobre ese modelo ajustado.

ISBN 978-607-8285-62-4

OBJETIVO 2.3 El alumno podrá calcular e interpretar la recta de regresión por el método de mínimos cuadrados y elaborar un diagrama de dispersión.

ANTECEDENTES

CONCEPTOS DE:

Variable aleatoria. Tipos de variable. Variable dependiente. Variable independiente. Ecuación de tendencia lineal. Ordenada al origen. Pendiente de la recta. Relación directa de dos variables. Relación inversa de dos variables. Curva normal. Normal estándar. Estimador de punto. Distribución de probabilidad. Diagrama de dispersión. Coordenadas cartesianas.

2.3.1 EL MÉTODO DE MÍNIMOS CUADRADOS. COEFICIENTES DE REGRESIÓN Y DIAGRAMA DE DISPERSIÓN

CONCEPTOS BÁSICOS
COEFICIENTES DE REGRESIÓN

Desarrollo de una ecuación de estimación

El modelo de regresión lineal simple está dado por la **función:**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Donde:

Y_i = Variable dependiente
X_i = Variable independiente.
β_0 = Primer parámetro de la regresión (ordenada al origen).
β_1 = Segundo parámetro de la regresión (pendiente de la recta).
ε_i = Error aleatorio de muestreo.

ISBN 978-607-8285-62-4

Para estimar los parámetros de la regresión se utiliza el **método de mínimos cuadrados**.

Cálculo matemático de la línea de mínimos cuadrado del mejor ajuste

El **método de mínimos cuadrados** determina la ecuación de la recta de regresión minimizando la suma de los cuadrados de las distancias verticales entre los valores reales de *Y* y los valores pronosticados para *Y.*

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Pendiente de la línea de regresión de los mínimos cuadrados

La **pendiente estimada de la recta de regresión** es la pendiente estimada de la recta, o el cambio promedio en la variable dependiente $\widehat{\boldsymbol{Y}}_{\boldsymbol{i}}$ para cada cambio de una unidad (ya sea aumento o reducción) en la variable independiente $\boldsymbol{X_i}$ y la podemos calcular mediante la ecuación normal

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}} = \frac{\sum_{i=1}^{n} \boldsymbol{XY} - \boldsymbol{n}\overline{\boldsymbol{X}}\overline{\boldsymbol{Y}}}{\sum_{i=1}^{n} \boldsymbol{X}^{\boldsymbol{2}} - \boldsymbol{n}\overline{\boldsymbol{X}}^{\boldsymbol{2}}}$$

Intersección de la línea de regresión de mínimos cuadrados

La **ordenada en el origen o intersección con el eje Y** es la intersección ***Y***. Es el valor estimado de la variable dependiente $\widehat{\boldsymbol{Y}}_{\boldsymbol{i}}$ cuando $\boldsymbol{X_i}$ **=0.** En otras palabras, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}$ es el valor estimado de $\widehat{\boldsymbol{Y}}_{\boldsymbol{i}}$ cuando la línea de regresión cruza el eje ***Y*** cuando ***X*** es cero y la podemos calcular mediante la ecuación normal

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} = \overline{\boldsymbol{Y}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\overline{\boldsymbol{X}}$$

Dibujo, o "ajuste", de una recta a través de un diagrama de dispersión

El **diagrama de dispersión e**s un tipo de diagrama matemático que utiliza las **coordenadas cartesianas** para mostrar los valores de dos variables para un conjunto de datos. Los datos se muestran como un conjunto de puntos, cada uno con el valor de la variable independiente ***X*** que determina la posición en el eje horizontal y el valor de la variable dependiente ***Y*** determinado por la posición en el eje vertical. Un diagrama de dispersión se llama también **gráfico o diagrama de dispersión**.

ISBN 978-607-8285-62-4

2.3.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 2.3.1.1 COEFICIENTES DE REGRESIÓN

El director general de una cadena de tiendas de autoservicio en expansión desea conocer el comportamiento de las ventas en los diferentes establecimientos con base en la superficie de piso en la que se exhiben los diferentes productos con el fin de contar con un modelo que le permita llevar un control adecuado de la eficiencia con la que trabaja cada establecimiento. Para ello utiliza el volumen de ventas mensuales (en millones de pesos) y la superficie de piso (en miles de metros cuadrados). En forma aleatoria recopila el volumen de ventas del último mes en diez tiendas de la cadena que correspondan más o menos entre 2,000 y 12,000 metros cuadrados de superficie de piso.

Tienda	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Superficie *(X)* en miles de m^2	**2.15**	**9.2**	**6.70**	**13.5**	**5.5**	**12.15**	**4.8**	**10.7**	**3.25**	**8.25**
Ventas *(Y)* en millones de pesos	**1.0**	**3.0**	**3.0**	**4.5**	**2.0**	**5.0**	**1.0**	**4.0**	**1.5**	**3.5**

c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el volumen de ventas cuando la superficie de piso donde se exhiben los productos $\hat{Y}$ es de 10,000 metros cuadrados $X_0 = 10$ (por estar en miles de m^2).

Cálculo de la línea de regresión mediante una ecuación

Solución al inciso c.

Las estimaciones mínimas cuadráticas se pueden obtener de los cálculos realizados en la siguiente tabla:

Tienda	**Superficie de piso (X) en miles de metros cuadrados**	**Volumen Ventas (Y) en millones de pesos**	**XY**	**X*X**	**Y*Y**
1	2.15	1.0	2.2	4.6225	1.0
2	9.20	3.0	27.6	84.64	9.0
3	6.70	3.0	20.1	44.89	9.0
4	13.50	4.5	60.8	182.25	20.3
5	5.50	2.0	11.0	30.25	4.0

ISBN 978-607-8285-62-4

6	12.15	5.0	60.8	147.6225	25.0
7	4.80	1.0	4.8	23.04	1.0
8	10.70	4.0	42.8	114.49	16.0
9	3.25	1.5	4.9	10.5625	2.3
10	8.25	3.5	28.9	68.0625	12.3
SUMAS	76.2	28.5	263.7	710.43	99.75
	PROMEDIOS				
	7.62	2.85			

Los cálculos necesarios para determinar la ecuación de la recta de regresión son:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} XY - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X^2 - n\bar{X}^2} = \frac{263.7 - 10(7.62)(2.85)}{710.43 - 10\,(7.62)^2} = \frac{263.7 - 217.17}{710.43 - 580.64} = \frac{46.53}{129.786}$$
$$\cong \mathbf{0.35851}^{1}$$

$$\hat{\beta}_0 = 2.85 - (0.358513)(7.62) \cong \mathbf{0.11813}$$

El modelo ajustado se puede expresar como:

$$\hat{Y}_i = \mathbf{0.11813} + \mathbf{0.35851}X_i$$

Usando la calculadora (opcional):

En el modo **REG**

1 (Lin)

SHIFT **CLR** **1**(**Scl**) **=** (para borra la memoria estadística)

2.15 ▫ **1.0** **M+** | REG n = 1. |

Cada vez que presiona **M+** para registrar un ingreso (par ordenado) , el número de dato ingresado (par ordenado) hasta este punto se indica sobre la presentación (valor *n*) .

9.20 ▫ **3.0** **M+** **6.70** ▫ **3.0** **M+** … … … **8.25** ▫ **3.5** **M+** | REG n = 10. |

Coeficiente de regresión **A= 0.118129074 ≅ 0.11813**

SHIFT **S − VAR** **REPLAY →** **REPLAY →** **1** **=** **0.11813**

(Especifica cinco lugares decimales) **MODE** **MODE** **MODE** **1**(**Fix**) **5** | FIX 0.11813 |

Coeficiente de regresión **B= 0.35851**

SHIFT **S − VAR** **REPLAY →** **REPLAY →** **2** **=** **0.35851**

[1] Cálculos efectuados con el módulo de Regresión Lineal Simple de la calculadora de mano Casio fx.82 MS

ISBN 978-607-8285-62-4

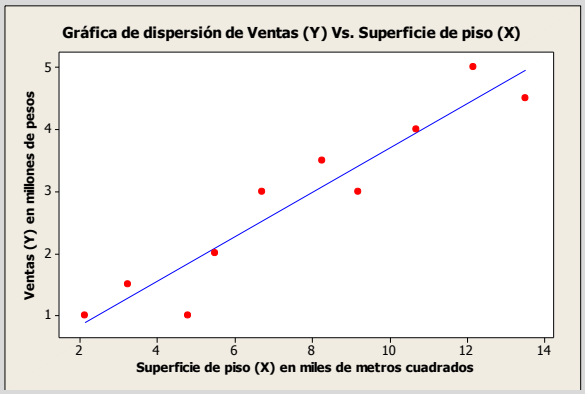
Interpretación de la ecuación

Solución al inciso d.

Con este modelo se podría llegar a la conclusión en cuanto a la ordenada al origen $\hat{\beta}_o$ de que cuando la superficie donde se exhibe la mercancía es cero, el volumen de ventas es de **0.11813** millones de pesos ó **$ 118,130.00** pesos. Puesto que el resultado de la variable independiente ***(X)*** raramente puede ser cero, la ordenada al origen se puede considerar como expresión del volumen de ventas ***(Y)*** que varía con factores ajenos al resultado de la superficie de piso de la tienda ***(X).*** Asimismo en cuanto a la pendiente de la recta $\hat{\beta}_1$,por cada mil metros cuadrado que se incrementan a la superficie de piso donde se exhibe la mercancía de la tienda ***(X)*** , el volumen de ventas (***Y)*** se incrementa en **0.358513** millones de pesos ó por cada metro cuadrado que se incrementa a la superficie de piso donde se exhibe la mercancía de la tienda, el volumen de ventas se incrementa (la pendiente es positiva) en **$ 358.51 pesos**. Esta pendiente también se puede contemplar como representante del volumen de ventas ***(Y)*** que varía de acuerdo a la superficie de piso de la tienda (***X)***.

Diagrama de dispersión

Solución al inciso e.



Cálculo de *Y* a partir de *X* aplicando la ecuación de una recta

Solución al inciso f.

$$\hat{Y}_{10} = \hat{\beta}_0 + \hat{\beta}_1 X_0 = 0.11813 + 0.35851(10) = \mathbf{3.70326}$$

Volumen de ventas de $\hat{\boldsymbol{Y}}$ cuando $\boldsymbol{X_0\ es\ 10 = 3.70326}$

Usando la calculadora:

$\mathbf{10}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-VAR}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{2}}$ $\boxed{=}$ $\mathbf{3.70326}$

ISBN 978-607-8285-62-4

2.3.1.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 2.3.1.1 COEFICIENTES DE REGRESIÓN

Una compañía refresquera está estudiando el efecto de su última campaña publicitaria. A un grupo de personas a quienes escogió al azar se les pregunto por teléfono cuantas latas del nuevo refresco habían comprado en la semana anterior y cuantos anuncios de él habían leído o visto esa semana.

Persona	1	2	3	4	5	6	7	8	9	10
No. de anuncios (X)	4	10	3	0	1	4	2	5	6	8
Latas compradas (Y)	12	19	7	6	3	5	6	10	7	11

c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 7$

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Cálculo de la línea de regresión mediante una ecuación

<u>Solución al inciso c.</u>

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} XY - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X^2 - n\bar{X}^2} =$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} =$$

El modelo ajustado se puede expresar como:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i =$$

ISBN 978-607-8285-62-4

Usando la calculadora:

En el modo **REG**

$\boxed{\mathbf{1\ (Lin)}}$

$\boxed{\mathbf{SHIFT}}\ \boxed{\mathbf{CLR}}\ \boxed{\mathbf{1}}\mathbf{(Scl)}\ \boxed{=}$ (para borra la memoria estadística)

$\mathbf{X_1}\ \boxed{,}\ \mathbf{Y_1}\ \boxed{\mathbf{M+}}$ $\boxed{\begin{array}{c}\mathbf{REG}\\ \mathbf{n = \qquad\qquad 1.}\end{array}}$

Cada vez que presiona $\boxed{\mathbf{M+}}$ para registrar un ingreso (par ordenado) , el número de dato ingresado (par ordenado) hasta este punto se indica sobre la presentación (valor *n*) .

$X_2\ \boxed{,}\ Y_2\ \boxed{\mathbf{M+}}\ X_3\ \boxed{,}\ Y_3\ \boxed{\mathbf{M+}} \ldots\ldots\ldots X_n\ \boxed{,}\ Y_n\ \boxed{\mathbf{M+}}$ $\boxed{\begin{array}{c}\mathbf{REG}\\ \mathbf{n = \qquad\qquad n.}\end{array}}$

Coeficiente de regresión **A=**

$\boxed{\mathbf{SHIFT}}\ \boxed{\mathbf{S-VAR}}\ \boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{1}}\ \boxed{=}$

(Especifica cinco lugares decimales) $\boxed{\mathbf{MODE}}\ \boxed{\mathbf{MODE}}\ \boxed{\mathbf{MODE}}\ \boxed{\mathbf{1}}\mathbf{(Fix)}\ \boxed{\mathbf{5}}\ \boxed{\mathbf{FIX}}$

Coeficiente de regresión **B=** $\boxed{\mathbf{SHIFT}}\ \boxed{\mathbf{S-VAR}}\ \boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{2}}\ \boxed{=}$

Interpretación de la ecuación

<u>Solución al inciso d.</u>

Dibujo, o "ajuste", de una recta a través de un diagrama de dispersión

<u>Solución al inciso e.</u>

Cálculo de *Y* a partir de *X* aplicando la ecuación de una recta

<u>Solución al inciso f.</u>

Valor de $\widehat{Y}$ cuando $X_0\ es\ 7 =$

Usando la calculadora:

$\mathbf{7}\ \boxed{\mathbf{SHIFT}}\ \boxed{\mathbf{S-VAR}}\ \boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{2}}\ \boxed{=}$

ISBN 978-607-8285-62-4

2.3.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 2.3.1.1
COEFICIENTES DE REGRESIÓN

El propietario de una cadena de heladerías desea estudiar el efecto de la temperatura atmosférica (X) sobre las ventas (Y) durante la temporada de verano. Seleccionó una muestra aleatoria de 12 días con los resultados siguientes:

Día	1	2	3	4	5	6	7	8	9	10	11	12
(X)	22.8	23.9	24.1	25.3	26.7	27.8	29.5	31.1	32.2	33.4	36.7	37.8
(Y)	18.0	20.5	21.8	22.9	23.6	22.5	26.8	29.0	31.4	32.4	34.0	32.8

c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 30$

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Cálculo de la línea de regresión mediante una ecuación

Solución al inciso c.

$\boldsymbol{\hat{\beta}_1} =$

$\boldsymbol{\hat{\beta}_0} =$

ISBN 978-607-8285-62-4

El modelo ajustado se puede expresar como:

$\hat{Y}_i =$

Usando la calculadora:

En el modo **REG**

$\boxed{\mathbf{1\ (Lin)}}$

$\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{CLR}}$ $\boxed{\mathbf{1}}$(**Scl**) $\boxed{=}$ (para borra la memoria estadística)

$\mathbf{X_1}$ $\boxed{,}$ $\mathbf{Y_1}$ $\boxed{\mathbf{M+}}$ $\boxed{\begin{array}{l} \mathbf{REG} \\ \mathbf{n =\ \ \ \ \ \ \ \ \ \ \ \ \ 1.} \end{array}}$

Cada vez que presiona $\boxed{\mathbf{M+}}$ para registrar un ingreso (par ordenado) , el número de dato ingresado (par ordenado) hasta este punto se indica sobre la presentación (valor *n*) .

X_2 $\boxed{,}$ Y_2 $\boxed{\mathbf{M+}}$ X_3 $\boxed{,}$ Y_3 $\boxed{\mathbf{M+}}$ … … … X_n $\boxed{,}$ Y_n $\boxed{\mathbf{M+}}$ $\boxed{\begin{array}{l} \mathbf{REG} \\ \mathbf{n =\ \ \ \ \ \ \ \ \ \ \ \ \ n.} \end{array}}$

Coeficiente de regresión **A=**

$\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-VAR}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{1}}$ $\boxed{=}$

(Especifica cinco lugares decimales) $\boxed{\mathbf{MODE}}$ $\boxed{\mathbf{MODE}}$ $\boxed{\mathbf{MODE}}$ $\boxed{\mathbf{1}}$(**Fix**)$\boxed{\mathbf{5}}$ $\boxed{\mathbf{FIX}}$

Coeficiente de regresión **B=** $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-VAR}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{2}}$ $\boxed{=}$

Interpretación de la recta de regresión

Solución al inciso d.

Dibujo, o "ajuste", de una recta a través de un diagrama de dispersión

Solución al inciso e.

Cálculo de Y a partir de X aplicando la ecuación de una recta

Solución al inciso f.

Valor de $\hat{Y}$ cuando $X_0\ es\ 30 =$

Usando la calculadora:

7 $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-VAR}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{2}}$ $\boxed{=}$

ISBN 978-607-8285-62-4

2.3.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 2.3.1 COEFICIENTES DE REGRESIÓN

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

2.3.1.1 El presidente de una fábrica de computadoras desea estudiar la relación que hay entre el tamaño del incremento anual de sueldos y rendimientos de un representante de ventas en el año siguiente. Muestreo a 10 representantes y determino los tamaños de sus respectivos incrementos (dados en porcentaje de sus sueldos individuales) y el numero de ventas realizadas por cada uno durante los 12 meses después del incremento.

Representante	1	2	3	4	5	6	7	8	9	10
Incremento en porcentaje (X)	7.2	6.5	7.4	5.7	6.8	8.9	7.7	10.6	11.3	8.2
No. de ventas (Y)	55	39	58	36	41	70	53	74	66	73

c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0 \; y \; \hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 9.6\%$

2.3.1.2 Una cadena de tiendas de repostería ha tenido grandes fluctuaciones en sus ingresos durante los últimos años. Abundantes baratas, nuevos productos y técnicas de publicidad se han utilizado durante este tiempo, por lo cual es difícil determinar cuáles de estos factores tienen la influencia más profunda en las ventas (Y). El departamento de mercadotecnia ha estudiado varias relaciones y piensa que los gastos mensuales destinados a carteles (X) pueden ser significativos. Muestreó 10 meses y descubrió lo siguiente:

Mes	1	2	3	4	5	6	7	8	9	10
(X)	25	16	43	34	10	21	19	13	23	38
(Y)	34	14	55	32	26	29	20	20	30	39

c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0 \; y \; \hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 28$

ISBN 978-607-8285-62-4

OBJETIVO 2.4. El alumno podrá calcular e interpretar el error estándar del estimador, probar la significancia entre la variable dependiente e independiente y elaborar e interpretar intervalos de confianza para el verdadero valor de la variable dependiente *Y*.

ANTECEDENTES

CONCEPTOS DE:

Varianza poblacional. Desviación estándar poblacional. Varianza muestral. Desviación estándar de la muestra. Error estándar de la muestra. Tamaño de la muestra.

2.4.1 ERROR ESTÁNDAR DE ESTIMACIÓN

CONCEPTOS BÁSICOS
ERROR ESTÁNDAR DEL ESTIMADOR

Definición y uso del error estándar de estimación
Ecuación con que se calcula el error estándar de estimación

El error estándar de estimación es una medida de la dispersión, o extensión, de los valores observados alrededor de la recta de regresión.

El error estándar del estimador, proporcionado por el símbolo $S_{Y.X}$, se define como:

$$S_{Y:X} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i}{n-2}}$$

Donde:

Y= valores de la variable dependiente
$\hat{Y}$ = valores estimados obtenidos de la ecuación de estimación que corresponden a cada valor de Y
n=número de observaciones usadas para ajustar la línea de regresión

Observe que la ecuación en su estructura es muy parecida a la que utilizamos para la **desviación estándar de una muestra:**

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

ISBN 978-607-8285-62-4

2.4.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 2.4.1.1
ERROR ESTÁNDAR DEL ESTIMADOR

El director general de una cadena de tiendas de autoservicio en expansión desea conocer el comportamiento de las ventas (Y) en los diferentes establecimientos con base en la superficie de piso (X) en la que se exhiben los diferentes productos con el fin de contar con un modelo que le permita llevar un control adecuado de la eficiencia con la que trabaja cada establecimiento. Para ello utiliza el volumen de ventas mensuales (en millones de pesos) y la superficie de piso (en miles de metros cuadrados). En forma aleatoria recopila el volumen de ventas del último mes en diez tiendas de la cadena que correspondan más o menos entre 2,000 y 12,000 metros cuadrados de superficie de piso.

Tienda	1	2	3	4	5	6	7	8	9	10
(X)	2.15	9.2	6.70	13.5	5.5	12.15	4.8	10.7	3.25	8.25
(Y)	1.0	3.0	3.0	4.5	2.0	5.0	1.0	4.0	1.5	3.5

g) Determine el error estándar de estimación.

Cálculo del error estándar de la estimación

Solución al inciso g.

El error estándar del estimador, proporcionado por el símbolo $S_{Y.X}$, se define como

$$S_{Y:X} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i}{n-2}}$$

$$= \sqrt{\frac{99.75 - 0.11813(28.5) - 0.35851(263.7)}{10-2}} = \sqrt{\frac{1.84338}{8}}$$

$$= \sqrt{0.23042} = \mathbf{0.48002}$$

Usando la calculadora:

$\boxed{\sqrt{\ }}$ $\boxed{(}$ $\boxed{(}$ $\boxed{SHIFT}$ $\boxed{S-SUM}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{1}$ $\boxed{-}$ $\boxed{SHIFT}$ $\boxed{S-VAR}$ $\boxed{\text{REPLAY}\rightarrow}$
$\boxed{\text{REPLAY}\rightarrow}$ $\boxed{1}$ $\boxed{X}$ $\boxed{SHIFT}$
$\boxed{S-SUM}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{2}$ $\boxed{-}$ $\boxed{SHIFT}$ $\boxed{S-VAR}$ $\boxed{\text{REPLAY}\rightarrow}$
$\boxed{\text{REPLAY}\rightarrow}$ $\boxed{2}$ $\boxed{X}$ $\boxed{SHIFT}$ $\boxed{S-SUM}$ $\boxed{\text{REPLAY}\rightarrow}$
$\boxed{3}$ $\boxed{)}$ $\boxed{\div}$ $\boxed{8}$ $\boxed{)}$ $\boxed{=}$ $\mathbf{0.48002}$

ISBN 978-607-8285-62-4

2.4.1.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 2.4.1.1
ERROR ESTÁNDAR DEL ESTIMADOR

Una compañía refresquera está estudiando el efecto de su última campaña publicitaria. A un grupo de personas a quienes escogió al azar se les pregunto por teléfono cuantas latas del nuevo refresco habían comprado en la semana anterior y cuantos anuncios de él habían leído o visto esa semana.

Persona	1	2	3	4	5	6	7	8	9	10
No. de anuncios (X)	4	10	3	0	1	4	2	5	6	8
Latas compradas (Y)	12	19	7	6	3	5	6	10	7	11

g) Determine el error estándar de estimación.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Solución al inciso g.

Cálculo del error estándar de la estimación

$$S_{Y:X} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i}{n-2}} =$$

Usando la calculadora:

$\boxed{\sqrt{\ }}$ $\boxed{(}$ $\boxed{(}$ $\boxed{SHIFT}$ $\boxed{S-SUM}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{1}$ $\boxed{-}$ $\boxed{SHIFT}$ $\boxed{S-VAR}$ $\boxed{\text{REPLAY}\rightarrow}$
$\boxed{\text{REPLAY}\rightarrow}$ $\boxed{1}$ $\boxed{X}$ $\boxed{SHIFT}$
$\boxed{S-SUM}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{2}$ $\boxed{-}$ $\boxed{SHIFT}$ $\boxed{S-VAR}$ $\boxed{\text{REPLAY}\rightarrow}$
$\boxed{\text{REPLAY}\rightarrow}$ $\boxed{2}$ $\boxed{X}$ $\boxed{SHIFT}$ $\boxed{S-SUM}$ $\boxed{\text{REPLAY}\rightarrow}$
$\boxed{3}$ $\boxed{)}$ $\boxed{\div}$ $\boxed{8}$ $\boxed{)}$ $\boxed{=}$

ISBN 978-607-8285-62-4

2.4.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 2.4.1.1
ERROR ESTÁNDAR DEL ESTIMADOR

Cálculo del error estándar de la estimación

El propietario de una cadena de heladerías desea estudiar el efecto de la temperatura atmosférica sobre las ventas durante la temporada de verano. Seleccionó una muestra aleatoria de 12 días con los resultados siguientes:

Día	1	2	3	4	5	6	7	8	9	10	11	12
Temperatura en °C. (X)	22.8	23.9	24.1	25.3	26.7	27.8	29.5	31.1	32.2	33.4	36.7	37.8
Ventas en miles de pesos (Y)	18.0	20.5	21.8	22.9	23.6	22.5	26.8	29.0	31.4	32.4	34.0	32.8

g) Determine el error estándar de estimación.

<u>Solución al inciso g.</u>

$S_{Y:X} =$

Usando la calculadora:

$\boxed{\sqrt{\ }}$ $\boxed{(}$ $\boxed{(}$ $\boxed{SHIFT}$ $\boxed{S-SUM}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{1}$ $\boxed{-}$ $\boxed{SHIFT}$ $\boxed{S-VAR}$ $\boxed{\text{REPLAY}\rightarrow}$
$\boxed{\text{REPLAY}\rightarrow}$ $\boxed{1}$ $\boxed{X}$ $\boxed{SHIFT}$
$\boxed{S-SUM}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{2}$ $\boxed{-}$ $\boxed{SHIFT}$ $\boxed{S-VAR}$ $\boxed{\text{REPLAY}\rightarrow}$
$\boxed{\text{REPLAY}\rightarrow}$ $\boxed{2}$ $\boxed{X}$ $\boxed{SHIFT}$ $\boxed{S-SUM}$ $\boxed{\text{REPLAY}\rightarrow}$
$\boxed{3}$ $\boxed{)}$ $\boxed{\div}$ $\boxed{8}$ $\boxed{)}$ $\boxed{=}$

ISBN 978-607-8285-62-4

2.4.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 2.4.1

ERROR ESTÁNDAR DEL ESTIMADOR

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

2.4.1.1 El presidente de una fábrica de computadoras desea estudiar la relación que hay entre el tamaño del incremento anual de sueldos y rendimientos de un representante de ventas en el año siguiente. Muestreo a 10 representantes y determino los tamaños de sus respectivos incrementos (dados en porcentaje de sus sueldos individuales) y el numero de ventas realizadas por cada uno durante los 12 meses después del incremento.

Representante	1	2	3	4	5	6	7	8	9	10
Incremento en porcentaje (X)	7.2	6.5	7.4	5.7	6.8	8.9	7.7	10.6	11.3	8.2
No. de ventas (Y)	55	39	58	36	41	70	53	74	66	73

g) Determine el error estándar de estimación.

2.4.1.2 Una cadena de tiendas de repostería ha tenido grandes fluctuaciones en sus ingresos durante los últimos años. Abundantes baratas, nuevos productos y técnicas de publicidad se han utilizado durante este tiempo, por lo cual es difícil determinar cuáles de estos factores tienen la influencia más profunda en las ventas (Y). El departamento de mercadotecnia ha estudiado varias relaciones y piensa que los gastos mensuales (X) destinados a carteles pueden ser significativos. Muestreó 10 meses y descubrió lo siguiente:

Mes	1	2	3	4	5	6	7	8	9	10
(X)	25	16	43	34	10	21	19	13	23	38
(Y)	34	14	55	32	26	29	20	20	30	39

g) Determine el error estándar de estimación.

ISBN 978-607-8285-62-4

ANTECEDENTES

CONCEPTOS DE:

Variables aleatorias. Variable dependiente. Variable independiente. Población, marco y muestra. Parámetro. La función de probabilidad, Las distribuciones de probabilidad, Características de la forma de una distribución de probabilidad. Prueba de hipótesis. Estructura de las hipótesis nula y alternativa, Error tipo I y tipo II. Distribución t de Student. Prueba t. Nivel de significancia. Distribución *F*. Prueba *F* para la razón de varianzas. Estadístico de prueba. Análisis de Varianza. La significancia observada (valor p). Estimador puntual. Varianza poblacional. Desviación estándar poblacional. Varianza muestral. Desviación estándar de la muestra. Error estándar de la muestra. Estructura de un intervalo de confianza.

2.4.2 PRUEBAS DE SIGNIFICANCIA. RELACIÓN LINEAL ENTRE LAS VARIABLES

CONCEPTOS BÁSICOS PRUEBAS DE SIGNIFICANCIA

Prueba de una hipótesis con respecto a β_1

El error estándar del coeficiente de regresión de β_1 es decir S_{β_1}

Se puede determinar si hay **relación significativa** entre las variables ***X*** y ***Y*** al probar si $\boldsymbol{\beta_1}$ (la pendiente real) es igual a cero. Si se rechaza esta hipótesis se concluiría que hay relación lineal.

Las **hipótesis nula y alternativa** se expresarían de la manera siguiente:

$$H_0: \beta_1 = 0 \ (no\ existe\ relación)$$

$$H_1: \beta_1 \neq 0 \ (\ existe\ relación)$$

La **distribución *t*** se puede utilizar para realizar pruebas de significancia y construir intervalos de confianza para la verdadera pendiente.

Entonces el **estadístico de prueba** sería:

$$t_{calculada(n-2)} = \frac{\hat{\beta}_1}{S_{\beta_1}}$$

Donde:

$$S_{\beta_1} = \frac{S_{y.x}}{\sqrt{\sum_{i=1}^{n} X^2 - n\bar{X}^2}}$$

NOTA: En caso de no rechazarse Ho; esto indicaría que el modelo no sería apropiado para predecir los valores de la variable dependiente con base a los valores de la variable independiente.

ISBN 978-607-8285-62-4

Uso de la distribución *t* para los intervalos de confianza

Un **segundo y equivalente método** para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de β_1 y determinar si el valor hipotético ($\beta_1 = 0$) está incluido en el intervalo. El estimado del intervalo de confianza de β_1 se obtendría de la siguiente manera:

$$\beta_1 = \hat{\beta}_1 \pm t_{n-2} S_{\hat{\beta}_1}$$

Prueba *F* de la regresión como un todo

Hay una **prueba alternativa**, una prueba ***F*** para la hipótesis nula de un valor predictivo nulo. Esta prueba proporciona el mismo resultado que una prueba ***t*** bilateral de $H_0: \beta_1 = 0$ en la regresión lineal simple. A continuación se presenta un resumen de la prueba ***F***:

Las **hipótesis nula y alternativa** se expresarían de la manera siguiente:

$$H_0: \beta_1 = 0 \ (no\ existe\ relación)$$

$$H_1: \beta_1 \neq 0 \ (\ existe\ relación)$$

Análisis de varianza para la regresión

Entonces el **estadístico de prueba** sería

$$F_{calculada} = \frac{CMR}{CME} = \frac{SCR/_{G.L.}}{SCE/_{G.L.}}$$

Donde:

SCT= SCR+SCE

$$SCR = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} XY - n\bar{Y}^2$$

$$SCE = \sum_{i=1}^{n} Y^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y - \widehat{\beta}_1 \sum_{i=1}^{n} XY$$

$$SCT = SCR + SCE$$

Como Comprobación,

$$SCT = \sum_{i=1}^{n} Y^2 - n\bar{Y}^2$$

ISBN 978-607-8285-62-4

Tabla de ANOVA

La suma de cuadrados y sus grados de libertad

Tabla de Anova

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	$F_{calculada}$
Regresión	*k*	*SCR*	*SCR/G.L.*	$\frac{CMR}{CME}$
Error	*n-k-1*	*SCE*	*SCE/G.L.*	
Total	*n-1*	*SCT*		

Donde:

n= número de observaciones.
k= número de variables independientes
La regla de decisión es de rechazar Ho si *F* calculada es mayor o igual a un valor crítico determinado para alfa de 0.05 y 0.01 y para v_1= k g.l. y v_2 = n-k-1 g.l.

NOTA: En caso de no rechazarse Ho; esto indicaría que el modelo no sería apropiado para predecir los valores de la variable dependiente con base a los valores de la variable independiente.

ISBN 978-607-8285-62-4

2.4.2.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 2.4.2.1 PRUEBAS DE SIGNIFICANCIA

El director general de una cadena de tiendas de autoservicio en expansión desea conocer el comportamiento de las ventas en los diferentes establecimientos con base en la superficie de piso en la que se exhiben los diferentes productos con el fin de contar con un modelo que le permita llevar un control adecuado de la eficiencia con la que trabaja cada establecimiento. Para ello utiliza el volumen de ventas mensuales (en millones de pesos) y la superficie de piso (en miles de metros cuadrados). En forma aleatoria recopila el volumen de ventas del último mes en diez tiendas de la cadena que correspondan más o menos entre 2,000 y 12,000 metros cuadrados de superficie de piso.

Tienda	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Superficie *(X)* en miles de m^2	**2.15**	**9.2**	**6.70**	**13.5**	**5.5**	**12.15**	**4.8**	**10.7**	**3.25**	**8.25**
Ventas *(Y)* en millones de pesos	**1.0**	**3.0**	**3.0**	**4.5**	**2.0**	**5.0**	**1.0**	**4.0**	**1.5**	**3.5**

h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba *t.*

i) Determine un intervalo de confianza de 95% para la pendiente para el volumen de ventas

j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba *F.*

Prueba de una hipótesis con respecto a β_1

Solución al inciso h.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

$$H_0: \beta_1 = 0 \ (no\ existe\ relación)$$

$$H_1: \beta_1 \neq 0 \ (\ existe\ relación)$$

ISBN 978-607-8285-62-4

Paso 2.Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$t_{calculada(n-2)} = \frac{\hat{\beta}_1}{S_{\beta_1}}$$

Donde:

$$S_{\beta_1} = \frac{S_{y.x}}{\sqrt{\sum_{i=1}^{n} X^2 - n\bar{X}^2}}$$

Entonces,

$$S_{\beta_1} = \frac{S_{y.x}}{\sqrt{\sum_{i=1}^{n} X^2 - n\bar{X}^2}} = \frac{0.48002}{\sqrt{710.43 - 10(7.62)^2}} = \frac{0.48002}{11.39237} = \mathbf{0.04214}$$

El error estándar del coeficiente de regresión de β_1 es decir S_{β_1}

Usando la calculadora:

$\mathbf{0.48002}$ $\boxed{\div}$ $\boxed{\sqrt{\ }}$ $\boxed{(}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-SUM}}$ $\boxed{\mathbf{1}}$ $\boxed{-}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-SUM}}$ $\boxed{\mathbf{3}}$ $\boxed{\mathbf{x}}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-VAR}}$ $\boxed{\mathbf{1}}$ $\boxed{\mathbf{x^2}}$ $\boxed{)}$ $\boxed{=}$ $\mathbf{0.04214}$

$$t_{calculada(n-2)} = \frac{\hat{\beta}_1}{S_{\beta_1}} = \frac{0.35851}{0.04214} \cong \mathbf{8.50767} \cong \mathbf{8.51}$$

Usando la calculadora:

$\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-VAR}}$ $\boxed{\mathbf{REPLAY\rightarrow}}$ $\boxed{\mathbf{REPLAY\rightarrow}}$ $\boxed{\mathbf{2}}$ $\boxed{\div}$ $\mathbf{0.04214}$ $\boxed{=}$ **8.51**

Paso 3.- Establecer la región de rechazo de (H_0).

La hipótesis alternativa no indica una dirección por lo que esta es una prueba de dos colas. Hay 8 grados de libertad, obtenidos de n-2 = 10-2=8. El valor de ***t*** es **2.306 ó 2.31** que se obtiene buscando en la tabla de valores críticos de ***t*** bajo prueba de dos colas, usando .05 como nivel de significancia y por tanto 0.025 como área de la cola superior, con 8 grados de libertad de la siguiente manera:

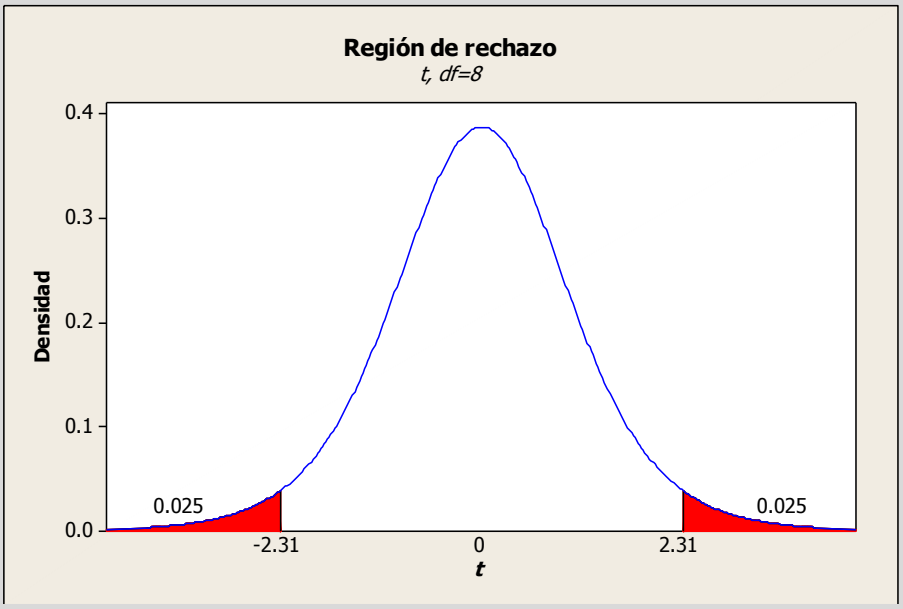
Paso 3. Región de rechazo.

ISBN 978-607-8285-62-4

Grados de liber-tad	Áreas de la cola superior					
	.25	.10	.05	.025	.01	.005
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	**2.3060**	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1993

Esta información se presenta en el siguiente diagrama

Gráfica con valores críticos.



Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $t_{calculada} \leq -2.31$ *ó* ≥ 2.31

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: como 8.51 > 2.31, se rechaza H_0.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación lineal significativa entre el volumen de ventas y la superficie de piso de las tiendas.

ISBN 978-607-8285-62-4

Solución al inciso i.

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de β_1 y determinar si el valor hipotético $(\beta_1 = 0)$ está incluido en el intervalo. El estimado del intervalo de confianza de β_1 se obtendría de la siguiente manera:

Obtención de los limites superior e inferior de la región de no rechazo de H_0

$$\boldsymbol{\beta_1} = \boldsymbol{\widehat{\beta}_1 \pm t_{n-2} S_{\widehat{\beta}_1}}$$

$$\beta_1 = 0.35851 \pm 2.31(0.04213)$$

$$\boldsymbol{\beta_1} = 0.35851 \pm 0.09732 \begin{cases} \boldsymbol{LIC} = 0.35851 - 0.09732 = \mathbf{0.26119} \\ \boldsymbol{LSC} = 0.35851 + 0.09732 = \mathbf{0.45583} \end{cases}$$

Otra forma de expresar el intervalo de confianza es:

$$\boldsymbol{\widehat{\beta}_1 - t_{n-2} S_{\widehat{\beta}_1} \leq \beta_1 \leq \widehat{\beta}_{1+} t_{n-2} S_{\widehat{\beta}_1}}$$

$$0.35851 - 2.31(0.04213) \leq \beta_1 \leq 0.35851 + 2.31(0.04213)$$

$$0.35851 - 0.09732 \leq \beta_1 \leq 0.35851 + 0.09732$$

$$\mathbf{0.26119} \leq \boldsymbol{\beta_1} \leq \mathbf{0.45583}$$

Interpretación: Se estima, con una confianza del 95%, que la pendiente real se encuentra entre 0.26119 y 0.45583. Puesto que estos valores son superiores a cero se puede concluir que hay relación lineal significativa entre el volumen de ventas y la superficie de piso de las tiendas.

Solución al inciso j.

Prueba *F* de la regresión como un todo

Hay una prueba alternativa, una prueba ***F*** para la hipótesis nula de un valor predictivo nulo. Esta prueba proporciona el mismo resultado que una prueba ***t*** bilateral de $H_0: \beta_1 = 0$ en la regresión lineal simple. A continuación se presenta un resumen de la prueba ***F***:

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 1. Juego de hipótesis.

$$H_0: \beta_1 = 0 \ (no\ existe\ relación)$$

$$H_1: \beta_1 \neq 0 \ (\ existe\ relación)$$

ISBN 978-607-8285-62-4

Paso 2.Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$F_{calculada} = \frac{CMR}{CME} = \frac{SCR/_{G.L.}}{SCE/_{G.L.}} = \mathbf{72.40}$$

Análisis de varianza para la regresión

SCT= SCR+SCE

La Suma de Cuadrados de la Regresión.

$$SCR = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} XY - n\bar{Y}^2$$
$$= 0.11813(28.5) + 0.35851(263.7) - 10(2.85)^2$$
$$\cong \mathbf{16.68162}$$

Usando la calculadora:

SHIFT **S – VAR** **REPLAY →** **REPLAY →** **1** **X** **SHIFT** **S – SUM** **REPLAY →** **2** **+** **SHIFT** **S – VAR** **REPLAY →** **REPLAY →** **2** **X** **SHIFT** **S – SUM** **REPLAY →** **3** **−** **SHIFT** **S – SUM** **3** **X** **SHIFT** **S – VAR** **REPLAY →** **1** **X²** **=** **16.68162**

La Suma deCcuadrados del Error.

$$SCE = \sum_{i=1}^{n} Y^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y - \hat{\beta}_1 \sum_{i=1}^{n} XY$$
$$= 99.75 - 0.11813(28.5) - 0.35851(263.7) \cong \mathbf{1.84338}$$

Usando la calculadora:

SHIFT **S – SUM** **REPLAY →** **1** **−** **SHIFT** **S – VAR** **REPLAY →** **REPLAY →** **1** **X** **SHIFT** **S – SUM** **REPLAY →** **2** **−** **SHIFT** **S – VAR** **REPLAY →** **REPLAY →** **2** **X** **SHIFT** **S – SUM** **REPLAY →** **3** **=** **1.84338**

La Suma de Cuadrados Total.

$$SCT = SCR + SCE = 16.68162 + 1.84338 = \mathbf{18.525}$$

Como Comprobación,

$$SCT = \sum_{i=1}^{n} Y^2 - n\bar{Y}^2 = 99.75 - 10(2.85)^2 = 99.75 - 81.225 = \mathbf{18.52500}$$

ISBN 978-607-8285-62-4

Usando la calculadora:

$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{\text{REPLAY}\rightarrow}\boxed{1}\boxed{=}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{3}\boxed{\text{X}}$
$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{\text{REPLAY}\rightarrow}\boxed{1}\boxed{\text{X}^2}\boxed{=}\ 18.52500$

Tabla de Anova

Tabla de ANOVA

La suma de cuadrados y sus grados de libertad

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculad}$***
Regresión	*k*=1	*SCR*=16.68162	*SCR/G.L.*=16.68162	$\frac{CMR}{CME} = \frac{16.682}{0.230} \cong 72.40$
Error	*n-k-1*= 8	*SCE*=1.84338	*SCE/G.L.*=0.23042	
Total	*n-1*=9	*SCT*=18.525		

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Para determinar la región de rechazo, se necesita el valor crítico. El valor crítico en el estadístico ***F*** se encuentra en las tablas de valores críticos de ***F*** . Para utilizar esta tabla se necesita conocer los grados de libertad en el numerador y en el denominador. Los grados de libertad en el numerador son iguales al número de variables independientes, designados como "k". Los grados de libertad en el denominador son el número de observaciones menos el número de variables independientes menos 1. Para este problema existe una solo variable independiente, por lo tanto los grados de libertad en el numerador son*: k= 1 g.l.* y los grados de libertad del denominador para 10 observaciones y una sola variable independiente son: *n-k-1=10-1-1=8 g.l.*

Como existen tablas para niveles de Alfa diferentes, busque la que corresponda al nivel de significancia solicitada, en este caso 0.05, y desplácese horizontalmente sobre la parte superior de la página hasta llegar a los 1 grados de libertad del numerador. Luego descienda en esa columna hasta llegar a la fila que presenta 8 grados de libertad. El valor en esta intersección es 5.32 que en este caso es el valor crítico.

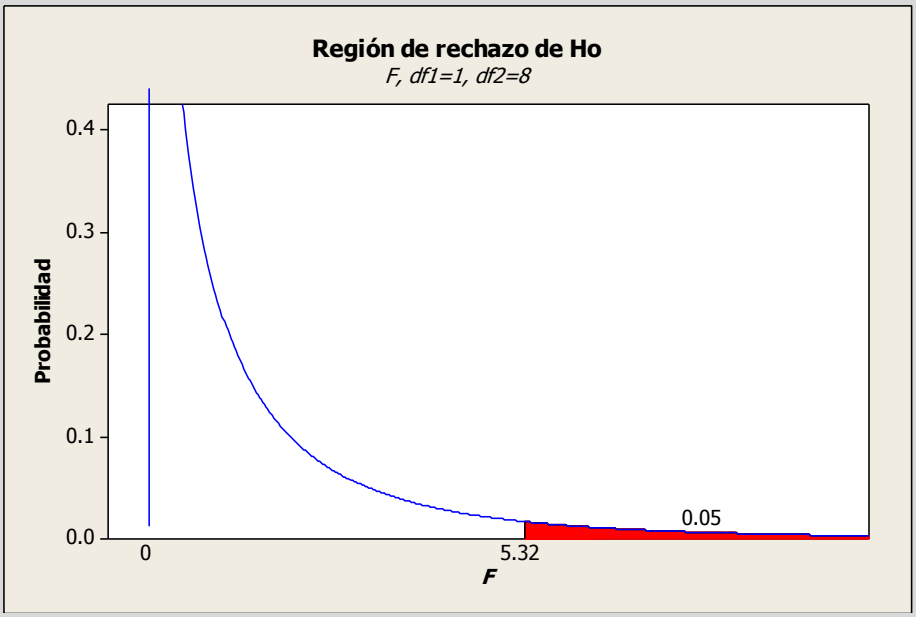
ISBN 978-607-8285-62-4

Tabla de valores crítico de *F.*

Grados de libertad en el denominador	Grados de libertad en el numerador			
	1	2	3	4
8	**5.32**	4.46	4.07	3.84
9	5.12	4.26	3.86	3.63
10	4.96	4.10	3.71	3.48
11	4.84	3.98	3.59	3.36
12	4.75	3.89	3.49	3.26
13	4.67	3.81	3.41	3.18
14	4.60	3.74	3.34	3.11
15	4.54	3.68	3.29	3.06

Esta información se presenta en el siguiente diagrama

Gráfico con valores críticos de F.



Si se desea aplicar el criterio *p-level* en la conclusión busque en las tablas de valores críticos de ***F*** la que corresponda al nivel de significancia de 0.01, y desplácese horizontalmente sobre la parte superior de la página hasta llegar a los 1 grados de libertad del numerador. Luego descienda en esa columna hasta llegar a la fila que presenta 8 grados de libertad. El valor en esta intersección es 11.26 que en este caso es el valor crítico.

ISBN 978-607-8285-62-4

Tabla de valores crítico de *F.*

Grados de libertad en el denominador	Grados de libertad en el numerador			
	1	2	3	4
8	**11.26**	8.65	7.59	7.01
9	10.6	8.02	6.99	6.42
10	10.0	7.56	6.55	5.99
11	9.65	7.21	6.22	5.67
12	9.33	6.93	5.95	5.41
13	9.07	6.70	5.74	5.21
14	8.86	6.51	5.56	5.04
15	8.68	6.36	5.42	4.89

Esta información se presenta en el siguiente diagrama

Gráfico con valores críticos de F.

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si fcalc ≥ 5.32

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).
Estadística: como 72.40
$> 5.32 > 11.26 \therefore la\ prueba\ es\ (AS)\ y\ se\ rechaza H_0$.
Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación entre el volumen de ventas y la superficie se piso donde se exhiben los productos.
Nota importante:
Observe que $t^2 = 8.51^2 = 72.4 = F = 72.4$ por eso se dice que son equivalentes.

ISBN 978-607-8285-62-4

2.4.2.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 2.4.2.1 PRUEBAS DE SIGNIFICANCIA

Una compañía refresquera está estudiando el efecto de su última campaña publicitaria. A un grupo de personas a quienes escogió al azar se les pregunto por teléfono cuantas latas del nuevo refresco habían comprado en la semana anterior y cuantos anuncios de él habían leído o visto esa semana.

Persona	1	2	3	4	5	6	7	8	9	10
No. de anuncios (X)	4	10	3	0	1	4	2	5	6	8
Latas compradas (Y)	12	19	7	6	3	5	6	10	7	11

h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba *t.*

i) Determine un intervalo de confianza de 95% para la pendiente para el volumen de ventas.

j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba *F.*

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Prueba de una hipótesis con respecto a β_1

<u>Solución al inciso h.</u>

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0:
H_1:

ISBN 978-607-8285-62-4

Paso 2.Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$S_{\beta_1}=\frac{S_{y.x}}{\sqrt{\sum_{i=1}^{n}X^2-n\bar{X}^2}}=$$

Usando la calculadora:

$S_{Y.X}$ $\boxed{\div}$ $\boxed{\sqrt{\ }}$ $\boxed{(}$ $\boxed{\text{SHIFT}}$ $\boxed{\text{S}-\text{SUM}}$ $\boxed{1}$ $\boxed{-}$ $\boxed{\text{SHIFT}}$ $\boxed{\text{S}-\text{SUM}}$ $\boxed{3}$ $\boxed{\text{x}}$ $\boxed{\text{SHIFT}}$ $\boxed{\text{S}-\text{VAR}}$ $\boxed{1}$ $\boxed{\text{X}^2}$ $\boxed{)}$ $\boxed{=}$

$$t_{calculada(n-2)}=\frac{\hat{\beta}_1}{S_{\beta_1}}=$$

Usando la calculadora:

$\boxed{\text{SHIFT}}$ $\boxed{\text{S}-\text{VAR}}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{2}$ $\boxed{\div}$ S_{β_1} $\boxed{=}$

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

ISBN 978-607-8285-62-4

Estadística:

Administrativa:

Obtención de los limites superior e inferior de la región de no rechazo de H_0

Solución al inciso i.

$$\boldsymbol{\beta_1 = \widehat{\beta}_1 \pm t_{n-2} S_{\widehat{\beta}_1}}$$

$$\boldsymbol{\beta_1 = \qquad \begin{cases} LIC = \\ LSC = \end{cases}}$$

Otra forma de expresar el intervalo de confianza es:

$$\boldsymbol{\widehat{\beta}_1 - t_{n-2} S_{\widehat{\beta}_1} \leq \beta_1 \leq \widehat{\beta}_{1+} t_{n-2} S_{\widehat{\beta}_1}}$$

$$\boldsymbol{LIC = \qquad \leq \beta_1 \leq LSC =}$$

Interpretación:

Prueba *F* de la regresión como un todo

Solución al inciso j.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0:
H_1:

Paso 2.Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

Análisis de varianza para la regresión

$$\boldsymbol{F_{calculada} = \frac{CMR}{CME} = \frac{SCR/_{G.L.}}{SCE/_{G.L.}} =}$$

ISBN 978-607-8285-62-4

SCT= SCR+SCE

Suma de Cuadrados de la Regresión.

$$SCR = \widehat{\beta}_0 \sum_{i=1}^{n} Y + \widehat{\beta}_1 \sum_{i=1}^{n} XY - n\overline{Y}^2 =$$

Usando la calculadora:

La suma de cuadrados

[SHIFT] [S – VAR] [REPLAY →] [REPLAY →] [1] [X] [SHIFT] [S – SUM] [REPLAY →] [2] [+] [SHIFT] [S – VAR] [REPLAY →] [REPLAY →] [2] [X] [SHIFT] [S – SUM] [REPLAY →] [3] [−] [SHIFT] [S – SUM] [3] [X] [SHIFT] [S – VAR] [REPLAY →] [1] [X^2] [=]

Suma de Cuadrados del Error.

$$SCE = \sum_{i=1}^{n} Y^2 - \widehat{\beta}_0 \sum_{i=1}^{n} Y - \widehat{\beta}_1 \sum_{i=1}^{n} XY =$$

Usando la calculadora:

[SHIFT] [S – SUM] [REPLAY →] [1] [−] [SHIFT] [S – VAR] [REPLAY →] [REPLAY →] [1] [X] [SHIFT] [S – SUM] [REPLAY →] [2] [−] [SHIFT] [S – VAR] [REPLAY →] [REPLAY →] [2] [X] [SHIFT] [S – SUM] [REPLAY →] [3] [=]

$$SCT = SCR + SCE =$$

Como Comprobación,

Suma de Cuadrados Total.

$$SCT = \sum_{i=1}^{n} Y^2 - n\overline{Y}^2 =$$

Usando la calculadora:

[SHIFT] [S – SUM] [REPLAY →] [1] [−] [SHIFT] [S – SUM] [3] [X] [SHIFT] [S – VAR] [REPLAY →] [1] [X^2] [=]

ISBN 978-607-8285-62-4

Tabla de ANOVA

Tabla de Anova

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada}$***
Regresión	*k=*	*SCR=*	*SCR/G.L.=*	*CMR/CME=*
Error	*n-k-1=*	*SCE=*	*SCE/G.L.=*	
Total	*n-1=*	*SCT=*		

La suma de cuadrados y sus grados de libertad

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

ISBN 978-607-8285-62-4

2.4.2.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 2.4.2.1
PRUEBAS DE SIGNIFICANCIA

El propietario de una cadena de heladerías desea estudiar el efecto de la temperatura atmosférica (X) sobre las ventas (Y) durante la temporada de verano. Seleccionó una muestra aleatoria de 12 días con los resultados siguientes:

Día	1	2	3	4	5	6	7	8	9	10	11	12
X	22.8	23.9	24.1	25.3	26.7	27.8	29.5	31.1	32.2	33.4	36.7	37.8
Y	18.0	20.5	21.8	22.9	23.6	22.5	26.8	29.0	31.4	32.4	34.0	32.8

h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Prueba de una hipótesis con respecto a β_1

Solución al inciso h.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.-

ISBN 978-607-8285-62-4

Paso 2.Estadístico de prueba.

Paso 2.-

Paso 3. Región de rechazo.

Paso 3.-

Paso 4. Regla de decisión.

Paso 4.-

Paso 5. Conclusiones.

Paso 5.-

Intervalo de confianza para β_1

<u>Solución al inciso i.</u>

ISBN 978-607-8285-62-4

Prueba *F* de la regresión como un todo

<u>Solución al inciso j.</u>

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.-

Paso 2.Estadístico de prueba.

Paso 2.-

Paso 3. Región de rechazo.

Paso 3.-

Paso 4. Regla de decisión.

Paso 4.-

Paso 5. Conclusiones.

Paso 5.-

ISBN 978-607-8285-62-4

2.4.2 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 2.4.2 PRUEBAS DE SIGNIFICANCIA

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

2.4.2.1 El presidente de una fábrica de computadoras desea estudiar la relación que hay entre el tamaño del incremento anual de sueldos y rendimientos de un representante de ventas en el año siguiente. Muestreo a 10 representantes y determino los tamaños de sus respectivos incrementos (dados en porcentaje de sus sueldos individuales) y el numero de ventas realizadas por cada uno durante los 12 meses después del incremento.

Representante	1	2	3	4	5	6	7	8	9	10
Incremento en porcentaje (X)	7.2	6.5	7.4	5.7	6.8	8.9	7.7	10.6	11.3	8.2
No. de ventas (Y)	55	39	58	36	41	70	53	74	66	73

h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***

2.4.2.2 Una cadena de tiendas de repostería ha tenido grandes fluctuaciones en sus ingresos (Y) durante los últimos años. Abundantes baratas, nuevos productos y técnicas de publicidad se han utilizado durante este tiempo, por lo cual es difícil determinar cuáles de estos factores tienen la influencia más profunda en las ventas. El departamento de mercadotecnia ha estudiado varias relaciones y piensa que los gastos mensuales (X) destinados a carteles pueden ser significativos. Muestreó 10 meses y descubrió lo siguiente:

Mes	1	2	3	4	5	6	7	8	9	10
(X)	25	16	43	34	10	21	19	13	23	38
(Y)	34	14	55	32	26	29	20	20	30	39

h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***

ISBN 978-607-8285-62-4

ANTECEDENTES

CONCEPTOS DE:

La función de probabilidad, Las distribuciones de probabilidad, Características de la forma de una distribución de probabilidad. Distribución t de Student. Nivel de significancia. La significancia observada (valor p). Estimador puntual. Varianza poblacional. Desviación estándar poblacional. Varianza muestral. Desviación estándar de la muestra. Error estándar de la estimación. Estructura de un intervalo de confianza.

2.4.3 INTERVALOS DE CONFIANZA PARA LA MEDIA *Y*, DADO X_0

CONCEPTOS BÁSICOS INTERVALO DE CONFIANZA DE LA MEDIA "Y"

Uso de $S_{Y.X}$ para construir límites alrededor de la línea de regresión

El error estándar del estimador se utiliza también para establecer **intervalos de confianza** para reportar el valor **medio** de **Y** para una $\mathbf{X_0}$ determinada, si el tamaño de la muestra es suficientemente grande y la dispersión alrededor de la recta de regresión se aproxima a la distribución normal, se puede desarrollar una estimación por intervalo de confianza para hacer inferencia sobre el valor predicho de ***Y;*** la fórmula es:

$$\boldsymbol{\mu_{Y:X} = \hat{Y}_i \pm t_{\alpha/2.n-2} S_{Y:X}\sqrt{h_i}}$$

Donde

$$\boldsymbol{h_i = \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2}}$$

Al pronosticar ***Y*** para los valores de X_0 cercanos a $\bar{X}$, el intervalo es mucho más estrecho que para las predicciones de valores de ***X*** más distantes de la media. Este efecto se puede observar por la sección de la raíz cuadrada en la ecuación donde h_i son **los "elementos diagonales de la matriz sombrero"**, que reflejan **la influencia de cada** X_0 **en el modelo de regresión lineal simple**. La **estimación por intervalo de la media real de *Y*** varía **hiperbólicamente** como una función de la cercanía de la X_0 dada a la $\bar{X}$. Cuando es necesario hacer predicciones de valores de X_0 que **están distantes del valor promedio de *X***, el intervalo mucho más amplio es la compensación por predecir esos valores de ***X***. Por tanto se observa un **efecto de banda de confianza** para las predicciones.

ISBN 978-607-8285-62-4

2.4.3.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 2.4.3.1
INTERVALO DE CONFIANZA DE LA MEDIA "Y"

El director general de una cadena de tiendas de autoservicio en expansión desea conocer el comportamiento de las ventas en los diferentes establecimientos con base en la superficie de piso en la que se exhiben los diferentes productos con el fin de contar con un modelo que le permita llevar un control adecuado de la eficiencia con la que trabaja cada establecimiento. Para ello utiliza el volumen de ventas mensuales (en millones de pesos) y la superficie de piso (en miles de metros cuadrados). En forma aleatoria recopila el volumen de ventas del último mes en diez tiendas de la cadena que correspondan más o menos entre 2,000 y 12,000 metros cuadrados de superficie de piso.

Tienda	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Superficie *(X)* en miles de m^2	**2.15**	**9.2**	**6.70**	**13.5**	**5.5**	**12.15**	**4.8**	**10.7**	**3.25**	**8.25**
Ventas *(Y)* en millones de pesos	**1.0**	**3.0**	**3.0**	**4.5**	**2.0**	**5.0**	**1.0**	**4.0**	**1.5**	**3.5**

k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del volumen de ventas cuando se tenga una superficie de piso de 10,000 metros cuadrados o sea $X_0 = 10$.

Intervalo de confianza del 95% para el verdadero valor de Y

<u>Solución al inciso k.</u>

$$\mu_{Y:X} = \hat{Y}_i \pm t_{\alpha/2 \cdot n-2} S_{Y:X}\sqrt{h_i}$$

Donde

$$h_i = \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2}$$

Entonces

$$\hat{Y}_{10} = \hat{\beta}_0 + \hat{\beta}_1 X_0 = 0.11813 + 0.35851(10) = \mathbf{3.70326}$$

ISBN 978-607-8285-62-4

Usando la calculadora:

10 [SHIFT] [S − VAR] [REPLAY →] [REPLAY →] [REPLAY →] [2] [=] **3.70326**

$$h_i = \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2} = \frac{1}{10} + \frac{(10 - 7.62)^2}{710.43 - 10(7.62)^2} =$$

$$\frac{1}{10} + \frac{5.6644}{129.786} = \mathbf{0.14364}$$

Usando la calculadora:

1 [÷] [SHIFT] [S − SUM] [3] [=] [+] [(] [(] **10** [−] [SHIFT] [S − VAR] [1] [)] [X²] [÷] [(] [SHIFT] [S − SUM] [1] [−] [SHIFT] [S − SUM] [3] [X] [SHIFT] [S − VAR] [1] [X²] [)] [=] **0.14364**

Por lo tanto

$$\mu_{Y:X} = \hat{Y}_i \pm t_{0.05.8} S_{Y:X}\sqrt{h_i} = 3.70326 \pm 2.306(0.48)\sqrt{0.14364}$$
$$= 3.70326 \pm 0.41951 \begin{cases} LIC = 3.70326 - 0.41951 = \mathbf{3.28375} \\ LSC = 3.70326 + 0.41951 = \mathbf{4.12277} \end{cases}$$

Otra forma de expresar el intervalo de confianza es:

$$\mathbf{3.28375} \leq \mu_{Y.X} \leq \mathbf{4.12277}$$

Interpretación: En **95** de cada **100** muestras (95% de confianza) de tamaño **10**, el verdadero volumen de ventas promedio mensuales de una tienda que tiene una superficie de piso de 10,000 metros cuadrados oscilará entre 3'283,750 y 4'122,770 millones de pesos.

ISBN 978-607-8285-62-4

2.4.3.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 2.4.3.1
INTERVALO DE CONFIANZA DE LA MEDIA "Y"

Una compañía refresquera está estudiando el efecto de su última campaña publicitaria. A un grupo de personas a quienes escogió al azar se les pregunto por teléfono cuantas latas del nuevo refresco habían comprado en la semana anterior y cuantos anuncios de él habían leído o visto esa semana.

Persona	1	2	3	4	5	6	7	8	9	10
No. de anuncios (X)	4	10	3	0	1	4	2	5	6	8
Latas compradas (Y)	12	19	7	6	3	5	6	10	7	11

k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 7 unidades o sea $X_0 = 7$

Intervalo de confianza del 95% para el verdadero valor de Y

<u>Solución al inciso k.</u>

$$\hat{Y}_7 = \hat{\beta}_0 + \hat{\beta}_1 X_0 =$$

Usando la calculadora:

7 [SHIFT] [S – VAR] [REPLAY →] [REPLAY →] [REPLAY →] [2] [=]

$$h_i = \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2} =$$

ISBN 978-607-8285-62-4

Usando la calculadora:

$$1 \boxed{\div} \boxed{\text{SHIFT}} \boxed{\text{S}-\text{SUM}} \boxed{3} \boxed{=} \boxed{+} \boxed{(} \boxed{(} \boxed{X_0} \boxed{-} \boxed{\text{SHIFT}} \boxed{\text{S}-\text{VAR}} \boxed{1} \boxed{)} \boxed{X^2} \boxed{\div} \boxed{(} \boxed{\text{SHIFT}} \boxed{\text{S}-\text{SUM}} \boxed{1} \boxed{-} \boxed{\text{SHIFT}} \boxed{\text{S}-\text{SUM}} \boxed{3} \boxed{X} \boxed{\text{SHIFT}} \boxed{\text{S}-\text{VAR}} \boxed{1} \boxed{X^2} \boxed{)} \boxed{=}$$

$$\mu_{Y:X} = \hat{Y}_i \pm t_{0.05.n-2} S_{Y:X} \sqrt{h_i} = \qquad \begin{cases} LIC = \\ LSC = \end{cases}$$

Interpretación:

ISBN 978-607-8285-62-4

2.4.3.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 2.4.3.1
INTERVALO DE CONFIANZA DE LA MEDIA "Y"

Intervalo de confianza del 95% para el verdadero valor de Y

El propietario de una cadena de heladerías desea estudiar el efecto de la temperatura atmosférica (X) en ºC. sobre las ventas (Y) en miles de pesos, durante la temporada de verano. Seleccionó una muestra aleatoria de 12 días con los resultados siguientes:

Día	1	2	3	4	5	6	7	8	9	10	11	12
(X)	22.8	23.9	24.1	25.3	26.7	27.8	29.5	31.1	32.2	33.4	36.7	37.8
(Y)	18.0	20.5	21.8	22.9	23.6	22.5	26.8	29.0	31.4	32.4	34.0	32.8

k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 30 unidades o sea $X_0 = 30$

Solución al inciso k.

$$\hat{Y}_{30} = \hat{\beta}_0 + \hat{\beta}_1 X_0 =$$

Usando la calculadora:

$$h_i =$$

ISBN 978-607-8285-62-4

Usando la calculadora:

$\mu_{Y:X} =$

$\begin{cases} LIC = \\ LSC = \end{cases}$

Interpretación:

ISBN 978-607-8285-62-4

2.4.3 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 2.4.3

INTERVALO DE CONFIANZA DE LA MEDIA "Y"

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

2.4.3.1 **E**l presidente de una fábrica de computadoras desea estudiar la relación que hay entre el tamaño del incremento anual de sueldos y rendimientos de un representante de ventas en el año siguiente. Muestreo a 10 representantes y determino los tamaños de sus respectivos incrementos (dados en porcentaje de sus sueldos individuales) y el numero de ventas realizadas por cada uno durante los 12 meses después del incremento.

Representante	1	2	3	4	5	6	7	8	9	10
Incremento en porcentaje (X)	7.2	6.5	7.4	5.7	6.8	8.9	7.7	10.6	11.3	8.2
No. de ventas (Y)	55	39	58	36	41	70	53	74	66	73

k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 9.6% unidades o sea $X_0 = 9.6\%$.

2.4.3.2 **U**na cadena de tiendas de repostería ha tenido grandes fluctuaciones en sus ingresos durante los últimos años. Abundantes baratas, nuevos productos y técnicas de publicidad se han utilizado durante este tiempo, por lo cual es difícil determinar cuáles de estos factores tienen la influencia más profunda en las ventas. El departamento de mercadotecnia ha estudiado varias relaciones y piensa que los gastos mensuales destinados a carteles pueden ser significativos. Muestreó 10 meses y descubrió lo siguiente:

Mes	1	2	3	4	5	6	7	8	9	10
Gasto mensual en carteles en miles de pesos (X)	25	16	43	34	10	21	19	13	23	38
Ingreso mensual por ventas en miles de pesos (Y)	34	14	55	32	26	29	20	20	30	39

k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 28 unidades o sea $X_0 = 28$.

ISBN 978-607-8285-62-4

OBJETIVO 2.5. El alumno podrá calcular e interpretar el coeficiente de determinación y el coeficiente de correlación

ANTECEDENTES

CONCEPTO DE:

Población. Muestra. Variable. Tipos de variable. Escalas de medición de las variables. La media de la población. Tamaño de la muestra. Ejes cartesianos. Diagrama de dispersión. Covarianza de la muestra. Desviación estándar de la muestra para X. Desviación estándar de la muestra para Y. Sumas de cuadrados de la Regresión. Suma de Cuadrados Total.

2.5.1 COEFICIENTE DE DETERMINACIÓN Y DE CORRELACIÓN

CONCEPTOS BÁSICOS COEFICIENTES DE DETERMINACIÓN Y CORRELACIÓN

Coeficiente muestral de determinación

Covarianza de la muestra

El coeficiente de determinación mide la proporción que se explica por la variable independiente en el modelo de regresión y se puede expresar como el cociente de la suma explicada de cuadrados o suma del cuadrado de la regresión ***SCR (Variación explicada)*** entre la suma de cuadrados tota ***SCT(Variación Total).***

$$r^2_{Y:X} = \frac{SCR}{SCT} = \frac{\hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} XY - n\bar{Y}^2}{\sum_{i=1}^{n} Y^2 - n\bar{Y}^2}$$

Por lo general la fuerza de una relación entre dos variables en una población se mide mediante el **coeficiente de correlación**, cuyos valores oscilan entre **-1** para la **correlación negativa perfecta** hasta **+1** para la **correlación positiva perfecta.**

En el caso de los problemas orientados hacia la **regresión**, el **coeficiente de correlación** se define, a partir de las ***n*** pares de observaciones, mediante

$$r_{y.x} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

ISBN 978-607-8285-62-4

Desviación estándar de X

Desviación estándar de Y

$$S_{xy} = \sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X)(\sum_{i=1}^{n} Y)}{n} \qquad S_{xx} = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} \qquad S_{yy} = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$$

se puede obtener con facilidad el **coeficiente de correlación** mediante la fórmula:

Coeficiente de correlación muestral

$$r_{y.x} = \sqrt{r_{y.x}^2}$$

2.5.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 2.5.1.1 COEFICIENTES DE DETERMINACIÓN Y CORRELACIÓN

El director general de una cadena de tiendas de autoservicio en expansión desea conocer el comportamiento de las ventas en los diferentes establecimientos con base en la superficie de piso en la que se exhiben los diferentes productos con el fin de contar con un modelo que le permita llevar un control adecuado de la eficiencia con la que trabaja cada establecimiento. Para ello utiliza el volumen de ventas mensuales (en millones de pesos) y la superficie de piso (en miles de metros cuadrados). En forma aleatoria recopila el volumen de ventas del último mes en diez tiendas de la cadena que correspondan más o menos entre 2,000 y 12,000 metros cuadrados de superficie de piso.

Tienda	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Superficie *(X)* en miles de m^2	**2.15**	**9.2**	**6.70**	**13.5**	**5.5**	**12.15**	**4.8**	**10.7**	**3.25**	**8.25**
Ventas *(Y)* en millones de pesos	**1.0**	**3.0**	**3.0**	**4.5**	**2.0**	**5.0**	**1.0**	**4.0**	**1.5**	**3.5**

l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.

ISBN 978-607-8285-62-4

Coeficiente muestral de determinación

Solución al inciso l.

$$r^2_{Y:X} = \frac{SCR}{SCT} = \frac{\hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} XY - n\bar{Y}^2}{\sum_{i=1}^{n} Y^2 - n\bar{Y}^2}$$

$$= \frac{0.11813(28.5) + 0.35851(263.7) - 10(2.85)^2}{99.75 - 10(2.85)^2}$$

$$= \frac{16.68162}{18.525} = 0.90049 \text{ ó } 0.90049x\,100 = 90.05\%$$

Usando la calculadora:

$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{\text{REPLAY}\rightarrow}\boxed{\text{REPLAY}\rightarrow}\boxed{3}\boxed{\text{X}^2}\boxed{=}\ 0.90049$

Interpretación: El 90.05% de la variación del volumen de ventas (en millones de pesos) se puede explicar por la superficie de piso (en miles de metros cuadrados). Este es un ejemplo donde hay una fuerte relación lineal entre dos variables, dado que el uso de un modelo de regresión ha reducido la variabilidad en la predicción del volumen de ventas en 90%. Solo el 10% de la variabilidad en el volumen de ventas se puede explicar por factores distintos que los explicados por el modelo de regresión lineal simple.

Coeficiente de correlación muestral

Solución al inciso m.

$$r_{Y.X} = \sqrt{r^2_{Y.X}} = \sqrt{0.90049} = 0.94894 \text{ ó } 0.94894\ x\ 100 = 94.89\%$$

Usando la calculadora:

$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{\text{REPLAY}\rightarrow}\boxed{\text{REPLAY}\rightarrow}\boxed{3} = 0.94894$

Interpretación: En este problema del volumen de ventas, puesto que $r^2 = 0.90049$ y la pendiente $\hat{\beta}_1$ es positiva, el coeficiente de correlación se interpreta como ***+0.94894***. La cercanía del coeficiente de correlación con +1.0 implica una fuerte asociación entre el volumen de ventas y la superficie de piso en que se exhiben las mercancías.

ISBN 978-607-8285-62-4

2.5.1.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 2.5.1.1 COEFICIENTES DE DETERMINACIÓN Y CORRELACIÓN

Una compañía refresquera está estudiando el efecto de su última campaña publicitaria. A un grupo de personas a quienes escogió al azar se les pregunto por teléfono cuantas latas del nuevo refresco habían comprado en la semana anterior y cuantos anuncios de él habían leído o visto esa semana.

Persona	1	2	3	4	5	6	7	8	9	10
No. de anuncios (X)	4	10	3	0	1	4	2	5	6	8
Latas compradas (Y)	12	19	7	6	3	5	6	10	7	11

l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Desarrollo del coeficiente muestral de determinación

<u>Solución al inciso l.</u>

$$r^2_{Y:X} = \frac{SCR}{SCT} = \frac{\widehat{\beta}_0 \sum_{i=1}^{n} Y + \widehat{\beta}_1 \sum_{i=1}^{n} XY - n\overline{Y}^2}{\sum_{i=1}^{n} Y^2 - n\overline{Y}^2} =$$

Usando la calculadora:

$\boxed{\text{SHIFT}}\ \boxed{\text{S} - \text{VAR}}\ \boxed{\text{REPLAY} \rightarrow}\ \boxed{\text{REPLAY} \rightarrow}\ \boxed{3}\ \boxed{\text{X}^2}\ \boxed{=}$

ISBN 978-607-8285-62-4

Interpretación:

Cálculo del coeficiente de correlación muestral

<u>Solución al inciso m.</u>

$$r_{Y.X} = \sqrt{r_{Y.X}^2} =$$

Usando la calculadora:

$\boxed{\text{SHIFT}}\boxed{\text{S} - \text{VAR}}\boxed{\text{REPLAY} \rightarrow}\boxed{\text{REPLAY} \rightarrow}\boxed{3} =$

Interpretación:

ISBN 978-607-8285-62-4

2.5.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 2.5.1.1
COEFICIENTES DE DETERMINACIÓN Y CORRELACIÓN

El propietario de una cadena de heladerías desea estudiar el efecto de la temperatura atmosférica sobre las ventas durante la temporada de verano. Seleccionó una muestra aleatoria de 12 días con los resultados siguientes:

Día	1	2	3	4	5	6	7	8	9	10	11	12
Temperatura en º C. (X)	22.8	23.9	24.1	25.3	26.7	27.8	29.5	31.1	32.2	33.4	36.7	37.8
Ventas en miles de pesos (Y)	18.0	20.5	21.8	22.9	23.6	22.5	26.8	29.0	31.4	32.4	34.0	32.8

l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Cálculo del coeficiente de determinación

Solución al inciso l.

$r^2_{Y:X} =$

Usando la calculadora:

ISBN 978-607-8285-62-4

Interpretación:

Cálculo del coeficiente de correlación muestral

Solución al inciso m.

$r_{Y.X} =$

Usando la calculadora:

Interpretación:

ISBN 978-607-8285-62-4

2.5.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 2.5.1 COEFICIENTES DE DETERMINACIÓN Y CORRELACIÓN

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

2.5.1.1 El presidente de una fábrica de computadoras desea estudiar la relación que hay entre el tamaño del incremento anual de sueldos y rendimientos de un representante de ventas en el año siguiente. Muestreo a 10 representantes y determino los tamaños de sus respectivos incrementos (dados en porcentaje de sus sueldos individuales) y el numero de ventas realizadas por cada uno durante los 12 meses después del incremento.

Representante	1	2	3	4	5	6	7	8	9	10
Incremento en porcentaje (X)	7.2	6.5	7.4	5.7	6.8	8.9	7.7	10.6	11.3	8.2
No. de ventas (Y)	55	39	58	36	41	70	53	74	66	73

l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.

2.5.1.2 Una cadena de tiendas de repostería ha tenido grandes fluctuaciones en sus ingresos (Y) durante los últimos años. Abundantes baratas, nuevos productos y técnicas de publicidad se han utilizado durante este tiempo, por lo cual es difícil determinar cuáles de estos factores tienen la influencia más profunda en las ventas. El departamento de mercadotecnia ha estudiado varias relaciones y piensa que los gastos mensuales (X) destinados a carteles pueden ser significativos. Muestreó 10 meses y descubrió lo siguiente:

2.5.1.3

Mes	1	2	3	4	5	6	7	8	9	10
(X)	25	16	43	34	10	21	19	13	23	38
(Y)	34	14	55	32	26	29	20	20	30	39

l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.

ISBN 978-607-8285-62-4

OBJETIVO 2.6. El alumno podrá calcular los residuales estandarizados y determinará lo apropiado del ajuste del modelo.

ANTECEDENTES

CONCEPTOS DE:

Variable dependiente. Variable independiente. Coeficientes de regresión. Valor estimado de Y. Error aleatorio. Tipos de relación. Relación lineal. Correlación. Tablas de frecuencia. Histograma. Diagrama de tallo y hojas. Diagrama de caja y brazos. Valores atípicos. Distribuciones de probabilidad. Características de la distribución normal. Varianza muestral. Desviación estándar muestral. Elementos de la matriz sombrero.

2.6.1 ANÁLISIS DE RESIDUALES. DIAGNÓSTICO DE LA REGRESIÓN

CONCEPTOS BÁSICOS
ANÁLISIS DE RESIDUALES

El análisis de residuales es un **enfoque gráfico** para evaluar lo adecuado del modelo de regresión ajustado a los datos. Este enfoque también permitirá estudiar posibles violaciones a las suposiciones del modelo de regresión. Las suposiciones generales en las que se basa el modelo de regresión son: **(1)** las variables dependiente e independiente tienen una relación lineal **(linealidad); (2)** las varianzas de las distribuciones condicionales de la variable dependiente, para diversos valores de la variable independiente, son iguales **(homoscedasticidad).** La primera suposición indica que, aunque puedan controlarse los valores de la variable independiente, los valores de la variable dependiente se deben obtener a través del proceso de muestreo.

Adicionalmente si se van a utilizar intervalos de confianza en el análisis de regresión, se requiere una suposición adicional **(3)** que las distribuciones condicionales de la variable dependiente, para valores diferentes de la variable independiente, sean todas distribuciones normales para la población de valores **(normalidad).**

ISBN 978-607-8285-62-4

El estadístico de Durbin-Watson se utiliza para detectar la presencia de autocorrelación en los residuos. La autocorrelación significa que las observaciones adyacentes están correlacionadas

Por otro lado una de las hipótesis más importantes del análisis de regresión es que los términos de error $(\boldsymbol{\varepsilon_i})$, que se podrían llamar los **"residuos verdaderos"** , **son independientes**. Gran parte de la teoría estadística de la regresión depende de esta hipótesis. Los datos de series temporales, medidos en periodos sucesivos, a menudo muestran un comportamiento más o menos cíclico. Este problema restringido principalmente a los **datos de series temporales**, se llama **autocorrelación.** Una prueba formal para la autocorrelación se apoya en el estadístico de **Durbin-Watson**. El estadístico de Durbin-Watson es:

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2}$$

donde: e_i = residual del periodo *i*.

Un residuo es la diferencia entre un valor observado (y) y su valor ajustado correspondiente (ŷ). Los valores residuales son útiles especialmente en procedimientos de regresión y ANOVA porque ellos indican el grado hasta el cual un modelo representa la variación en los datos observados

Si los **verdaderos errores** son en realidad **independientes**, el valor esperado de ***d*** es **alrededor de 2.0**. Cualquier valor de ***d*** menor que 1.5 o 1.6 nos lleva a sospechar que hay autocorrelación.

Los **valores residuales o de error** $(\boldsymbol{\varepsilon_i})$ se pueden definir como **la diferencia entre los valores observados** $(\boldsymbol{Y_i})$ **y los predichos** $(\boldsymbol{\hat{Y}})_i$ de la variable dependiente para los valores $\boldsymbol{X_i}$ determinados. Lo anterior se puede representar como:

$$\boldsymbol{\varepsilon_i = Y_i - \hat{Y}_i}$$

Para poder **evaluar las suposiciones** en que se basa la regresión lineal simple se requiere considerar la **magnitud de los residuales en unidades que reflejan la variación estandarizada** en torno a la línea de regresión. El residual estandarizado se presenta como la ecuación:

Los residuos estandarizados son útiles porque los residuos sin procesar pueden ser escasos indicadores de valores atípicos debido a su varianza no constante: los residuos con valores x correspondientes que se encuentran lejos de $\bar{X}$ presentan una varianza mayor que los residuos con valores x correspondientes más cercanos a $\bar{X}$. Los controles de estandarización para esta varianza no constante y todos los residuos estandarizados tienen la misma desviación estándar.

$$\boldsymbol{SR_i = \frac{\varepsilon_i}{S_{Y:X}\sqrt{1 - h_i}}}$$

Donde,

$$\boldsymbol{h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2}}$$

Estos **valores estandarizados** permiten considerar **la magnitud de los residuales en unidades que reflejan la variación estandarizada** en torno a la línea de regresión. Los residuales estandarizados se trazan contra la variable independiente.

ISBN 978-607-8285-62-4

Evaluación de las suposiciones

Linealidad

Linealidad

Se puede evaluar lo apropiado del modelo de regresión, **trazando los "residuales estandarizados" sobre el eje vertical contra los valores** X_i correspondientes a la variable independiente en el eje horizontal. Si el **modelo ajustado es apropiado** para los datos **no habrá un patrón aparente** en esta gráfica de los residuales contra X_i. Sin embargo, si el modelo ajustado no es apropiado, habrá relación entre los valores X_i y los residuales ε_i.

Homoscedasticidad

Homoscedasticidad

La suposición de **homoscedasticidad** se puede evaluar también de la gráfica de residuales estandarizados con X_i. Si parece haber un ***"efecto de abanico"*** en el cual aumenta ó disminuye la variabilidad de los residuales al aumentar ***X*** se demuestra la falta de homogeneidad en las varianzas de Y_i a cada nivel de ***X.***

Normalidad

Normalidad

El histograma de residuos es una herramienta exploratoria que muestra las características generales de los datos, incluyendo:

- Valores atípicos, dispersión o variación y forma
- Valores inusuales en los datos

La presencia de largas colas en la gráfica podrían indicar sesgo en los datos. Si una o dos barras están lejos de las demás, esos puntos pueden ser valores atípicos. Debido a que el aspecto del histograma cambia según el número de intervalos utilizados para agrupar los datos, utilice la gráfica de probabilidad normal y las pruebas de bondad de ajuste para evaluar la normalidad de los residuos.

El supuesto de **normalidad** en la regresión es posible evaluarlo de un análisis residual colocando los **residuales estandarizados en una distribución de frecuencias** y mostrando los resultados en un **histograma**. Si el **histograma de frecuencias** de los residuales **no se ajusta al de una normal pueden existir valores atípicos**. Eliminando los pares (X_iY_i) que producen los valores atípicos, se puede conseguir **normalidad** en los residuos.

Si contamos con papel normal o acceso a la computadora, podemos construir una gráfica de probabilidad normal de residuos: **Los puntos de esta gráfica deben generalmente formar una línea recta si los residuos se están normalmente distribuidos.** Si los puntos en la gráfica **salen de una línea recta,** el supuesto de normalidad **puede ser inválido**.

Si sus datos tienen menos de 50 observaciones, la gráfica podría mostrar una curvatura en las colas, aun si los residuos están normalmente distribuidos. A medida que el número de observaciones disminuye, la gráfica de probabilidad podría mostrar una variación sustancial no linealidad, aun si los residuos están normalmente distribuidos. Utilice la gráfica de probabilidad y las pruebas de bondad de ajuste, tales como el **estadístico de Anderson-Darling**, para evaluar si los residuos están normalmente distribuidos.

ISBN 978-607-8285-62-4

Estadístico de Anderson-Darling

El **estadístico de Anderson-Darling** mide si los datos siguen una distribución particular. Mientras mejor se ajuste la distribución a los datos, menor será este estadístico. Utilice el **estadístico de Anderson-Darling** para comparar el ajuste de varias distribuciones para ver cual es la mejor o probar si una **muestra de datos proviene de una población con una distribución específica.** Por ejemplo, puede utilizar el **estadístico de Anderson-Darling** para probar si los datos cumplen con el **supuesto de normalidad de una prueba t.**

Las hipótesis para la **prueba de Anderson-Darling** son:

H_0: Los datos siguen una distribución normal

H_1: Los datos no siguen una distribución normal

Si el valor ***p*** (al estar disponible) para la **prueba de Anderson-Darling** es **inferior al nivel de significación seleccionado** (generalmente 0.05 ó 0.10), concluya que **los datos no siguen la distribución especificada**.

Independencia

Independencia

La suposición de **independencia** requiere que el **error (diferencia "residual" entre un valor observado y uno predicho de Y**) sea independiente para cada valor de X. Con frecuencia esta suposición se refiere a **datos que se recopilan a lo largo de un periodo**. Estos tipos de modelos caen bajo la denominación general de series de tiempo. **La suposición de independencia se puede evaluar trazando los residuales en el orden o la sucesión en que se obtuvieron los datos observados.**

El estadístico de Durbin-Watson se utiliza para detectar la presencia de autocorrelación en los residuos. La autocorrelación significa que las observaciones adyacentes están correlacionadas. Si están correlacionadas, la regresión de los cuadrados mínimos subestima el error estándar de los coeficientes; sus predictores podrían parecer significativos, cuando en realidad es posible que no lo sean.

Una **prueba formal para la autocorrelación** se apoya en el **estadístico de Durbin-Watson**. El estadístico de Durbin-Watson es:

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2}$$

donde: e_i = residual del periodo *i*.

Si los **verdaderos errores son en realidad independientes**, el valor esperado de ***d*** **es alrededor de 2.0**. Cualquier valor de ***d*** **menor que 1.5 o 1.6** nos lleva a sospechar que **hay autocorrelación.**

ISBN 978-607-8285-62-4

2.6.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 2.6.1.1 ANÁLISIS DE RESIDUALES

El director general de una cadena de tiendas de autoservicio en expansión desea conocer el comportamiento de las ventas en los diferentes establecimientos con base en la superficie de piso en la que se exhiben los diferentes productos con el fin de contar con un modelo que le permita llevar un control adecuado de la eficiencia con la que trabaja cada establecimiento. Para ello utiliza el volumen de ventas mensuales (en millones de pesos) y la superficie de piso (en miles de metros cuadrados). En forma aleatoria recopila el volumen de ventas del último mes en diez tiendas de la cadena que correspondan más o menos entre 2,000 y 12,000 metros cuadrados de superficie de piso.

Tienda	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Superficie *(X)* en miles de m^2	**2.15**	**9.2**	**6.70**	**13.5**	**5.5**	**12.15**	**4.8**	**10.7**	**3.25**	**8.25**
Ventas *(Y)* en millones de pesos	**1.0**	**3.0**	**3.0**	**4.5**	**2.0**	**5.0**	**1.0**	**4.0**	**1.5**	**3.5**

n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.

Análisis de residuales

Solución al inciso n.

$$\varepsilon_i = Y_i - \hat{Y}_i$$

El residual estandarizado se presenta como la ecuación:

$$SR_i = \frac{\varepsilon_i}{S_{Y:X}\sqrt{1-h_i}}$$

Donde,

Elementos de la matriz sombrero h_i

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2}$$

ISBN 978-607-8285-62-4

Por tanto para calcular el primer residual haremos lo siguiente:

$\varepsilon_1 = Y_1 - \hat{Y}_1$ donde $\hat{Y}_1 = 0.11813 + 0.35851(2.15) = 0.88893$ entonces,

$$\varepsilon_1 = 1.0 - 0.88893 = \mathbf{0.11107}$$

Usando la calculadora:

$\mathbf{1 - 2.15}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-VAR}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{REPLAY \rightarrow}}$ $\boxed{\mathbf{2}}$ $\boxed{=}$ $\mathbf{0.11107}$

Y así sucesivamente con los siguientes 9 datos.

Para estandarizar estos residuales haremos lo siguiente:

Elementos de la matriz sombrero h_i

$$h_1 = \frac{1}{n} + \frac{(X_1 - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2} = \frac{1}{10} + \frac{(2.15 - 7.62)^2}{710.43 - 10(7.62)^2} = \frac{1}{10} + \frac{29.92090}{129.786} = \mathbf{0.33054}$$

Usando la calculadora:

$\mathbf{1}$ $\boxed{\div}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-SUM}}$ $\boxed{\mathbf{3}}$ $\boxed{=}$ $\boxed{+}$ $\boxed{(}$ $\boxed{(}$ $\mathbf{2.15}$ $\boxed{-}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-VAR}}$ $\boxed{\mathbf{1}}$ $\boxed{)}$ $\boxed{\mathbf{X^2}}$ $\boxed{\div}$ $\boxed{(}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-SUM}}$ $\boxed{\mathbf{1}}$ $\boxed{-}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-SUM}}$ $\boxed{\mathbf{3}}$ $\boxed{\mathbf{X}}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-VAR}}$ $\boxed{\mathbf{1}}$ $\boxed{\mathbf{X^2}}$ $\boxed{)}$ $\boxed{=}$ $\mathbf{0.33054}$

Residual estandarizado para la primera observación.

Entonces

$$\boldsymbol{SR_1} = \frac{\varepsilon_1}{S_{Y:X}\sqrt{1 - h_1}} = \frac{0.11107}{0.48002\sqrt{1 - 0.33054}} = \frac{0.11107}{0.39275} \cong \mathbf{0.28279}$$

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

ISBN 978-607-8285-62-4

Tienda	Superficie de piso (*X*) En miles de m^2	Ventas *(Y)* En millones de pesos	y gorro$(\widehat{Y})$	Residual $\varepsilon_i = Y_i - \widehat{Y}_i$	h_i	$S_{Y:X}$	Residual Estandarizado SR_i
1	2.15	1.0	0.88893	0.11107	0.33054	0.48002	0.28279
2	9.2	3.0	3.41645	-0.4164	0.1192	0.48002	-0.92442
3	6.7	3.0	2.52017	0.47983	0.10652	0.48002	1.05751
4	13.5	4.5	4.95806	0.45806	0.36640	0.48002	-1.19881
5	5.5	2.0	2.08995	0.08995	0.13463	0.48002	-0.20144
6	12.15	5.0	4.47406	0.52594	0.25811	0.48002	1.27204
7	4.8	1.0	1.83899	0.83899	0.16127	0.48002	-1.90847
8	10.7	4.0	3.95422	0.04578	0.17309	0.48002	0.10488
9	3.25	1.5	1.28330	0.21670	0.24714	0.48002	0.52029
10	8.25	3.5	3.07586	0.42414	0.10306	0.48002	0.93296
SUMAS	76.2	28.5					
	PROME-DIOS						
	7.62	2.85					

El estadístico de Durbin-Watson determina si la correlación entre los términos de error adyacentes es cero

Estadístico de Durbin Watson

Volumen de ventas (Y)	Superficie de piso (X)	$\boxed{\widehat{Y}_i}$	$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$	$\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i$	$(\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i)^2$	$\hat{\varepsilon}^2$
33	3	0.88893	0.11107	-0.528	0.27828	0.01234
61	6	3.41645	-0.41645	0.89	0.80332	0.17343
70	10	2.52017	0.47983	-0.938	0.87964	0.23024
82	13	4.95806	-0.45806	0.368	0.13550	0.20982
17	9	2.08995	-0.08995	0.616	0.37932	0.00809
24	6	4.47406	0.52594	-1.365	1.86303	0.27661
75	11	1.83899	-0.83899	0.885	0.78282	0.70391
80	12	3.95422	0.04578	0.171	0.02921	0.00210
35	4	1.28330	0.21670	0.207	0.04303	0.04696
20	8	3.07586	0.42414			0.17989
					5.19415	1.84338
				d=	5.19415/ 1.84338= **2.81773**	

El estadístico de Durbin-Watson es ***d=* 2.81773.** Este valor es mayor a 1.5 por lo que no se puede pensar en que la autocorrelación sea un problema.

Diagnóstico de la regresión

<u>Solución al inciso o.</u>

Evaluación de las suposiciones

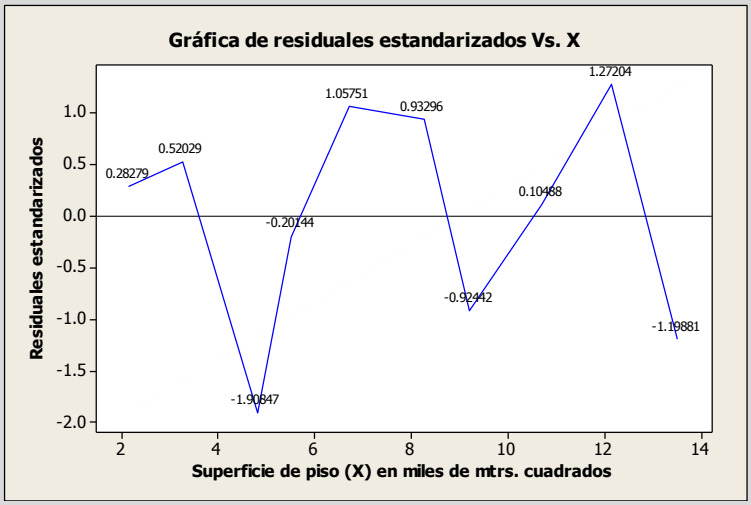
Linealidad

Linealidad

Se puede evaluar lo apropiado del modelo de regresión, trazando los "residuales estandarizados" sobre el eje vertical contra los valores X_i

ISBN 978-607-8285-62-4

correspondientes a la variable independiente en el eje horizontal. Si el modelo ajustado es apropiado para los datos no habrá un patrón aparente en esta gráfica de los residuales contra X_i. Sin embargo, si el modelo ajustado no es apropiado, habrá relación entre los valores X_i y los residuales ε_i.



Así, se puede observar que aunque haya una amplia dispersión en la gráfica residual, no hay patrón ó relación aparente entre los residuales estandarizados y X_i . Los residuales parecen estar distribuidos en forma pareja por encima y por debajo de 0 para diferentes valores de ***X***. Por lo tanto se puede concluir que el modelo ajustado parece ser el apropiado.

Homoscedasticidad

Homoscedasticidad

La suposición de homoscedasticidad se puede evaluar también de la gráfica de residuales estandarizados con X_i. Si parece haber un *"efecto de abanico"* en el cual aumenta ó disminuye la variabilidad de los residuales al aumentar ***X*** se demuestra la falta de homogeneidad en las varianzas de Y_i a cada nivel de ***X.*** Para los datos del volumen de ventas de una tienda no parece haber diferencias importantes en la variabilidad de SR_i para diferentes valores de X_i. Por lo tanto se puede concluir que para este modelo ajustado no hay violación aparente a la suposición de igual varianza en cada nivel de ***X***.

Normalidad

Normalidad

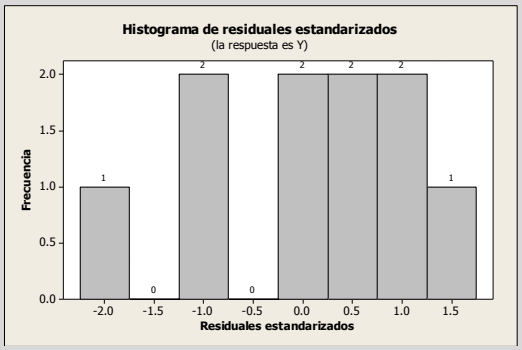
El supuesto de normalidad en la regresión es posible evaluarlo de un análisis residual colocando los residuales estandarizados en una distribución de frecuencias y mostrando los resultados en un histograma. Para los datos de las ventas en los diferentes establecimientos, los residuales estandarizados se colocaron en la siguiente distribución de frecuencias como se muestra en la siguiente tabla:

ISBN 978-607-8285-62-4

Residuales estandarizados	No.
De -2.25 a menos de -1.75	1
De -1.75 a menos de -1.25	0
De -1.25 a menos de -0.75	2
De -0.75 a menos de -0.25	0
De -0.25 a menos de 0.25	2
De 0.25 a menos de 0.75	2
De 0.75 a menos de 1.25	2
De 1.25 a menos de 1.75	1
Totales	10

Los resultados se graficaron en el siguiente histograma:

Histograma de residuales estandarizados.



Es difícil evaluar la suposición de normalidad para una muestra de tan sólo 10 observaciones y los procedimientos de pruebas disponibles quedan fuera del alcance del presente trabajo, sin embargo se puede observar que los datos aunque no parecen tener una "forma de campana" exacta, la mayor parte de los residuales están ubicados cerca del centro de la distribución por lo que parece razonable llegar a la conclusión de que no hay en modo alguno violación a la suposición de normalidad. El histograma indica que los datos podrían tener valores atípicos, lo cual se muestra mediante dos barras, en el extremo izquierdo de la gráfica.

Si contamos con papel normal o acceso a la computadora, podemos construir una gráfica de probabilidad normal de residuos: Los puntos de esta gráfica deben generalmente formar una línea recta si los residuos se están normalmente distribuidos. Si los puntos en la gráfica salen de una línea recta, el supuesto de normalidad puede ser inválido. Si sus datos tienen menos de 50 observaciones, la gráfica podría mostrar una curvatura en las colas, aun si los residuos están normalmente distribuidos. A medida que el número de observaciones disminuye, la gráfica de probabilidad podría mostrar una variación sustancial no linealidad, aun si los residuos están normalmente distribuidos.

ISBN 978-607-8285-62-4

El estadístico de Anderson-Darling mide si los residuales estandarizados siguen una distribución normal. Mientras mejor se ajuste la distribución a los residuales estandarizados, menor será este estadístico.

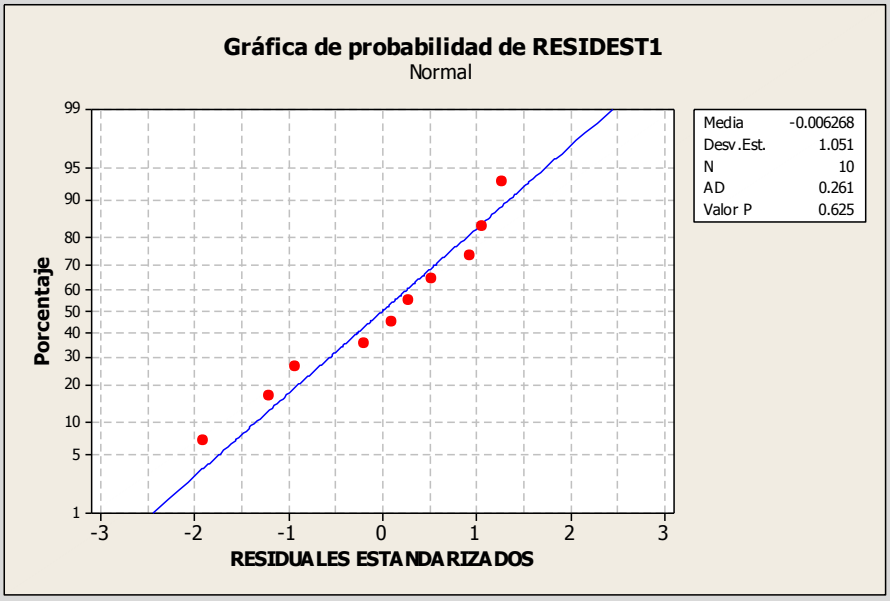
Las hipótesis para la prueba de Anderson-Darling son:

H_0: Los residuales estandarizados siguen una distribución especificada

H_1: Los residuales estandarizados no siguen una distribución especificada

Si el valor p (al estar disponible) para la prueba de Anderson-Darling es inferior al nivel de significación seleccionado (generalmente 0.05 ó 0.01), concluya que los datos no siguen la distribución especificada.

Utilice la gráfica de probabilidad y las pruebas de bondad de ajuste, tales como el **estadístico de Anderson-Darling**, para evaluar si los residuos están normalmente distribuidos.



La gráfica de probabilidad normal muestra un patrón aproximadamente lineal que concuerda con una distribución normal. El último punto de la esquina inferior izquierda de la gráfica puede ser un valor atípico. El destacado de la gráfica identifica este punto como 7, punto que deberá verificarse como observación inusual ó identificación de valores atípicos. A la derecha de la gráfica se presenta la prueba de Anderson-Darling que arroja un estadístico de prueba de 0.261 con un nivel *p* de 0.625 que al ser mayor a 0.05 nos hace no rechazar la hipótesis nula concluyendo que la distribución de los residuales estandarizados es normal.

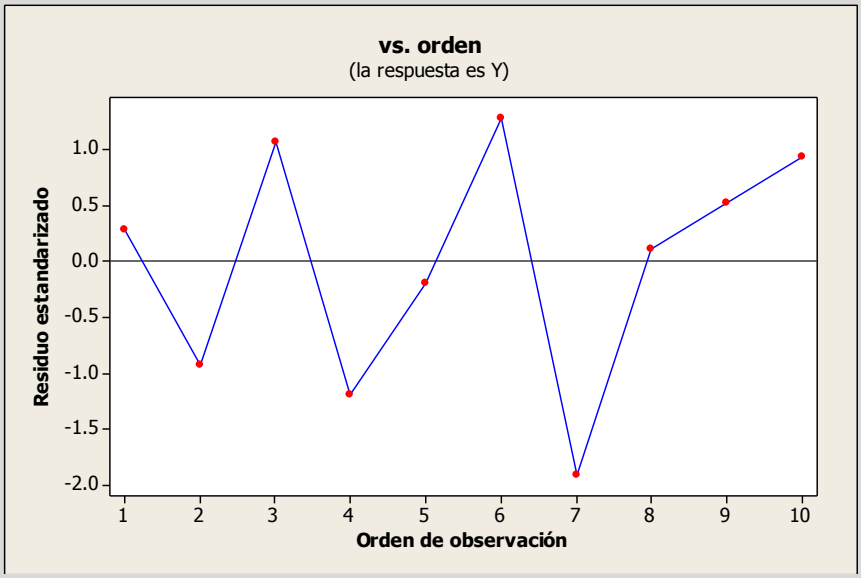
Independencia

Independencia

La suposición de independencia requiere que el error (diferencia "residual" entre un valor observado y uno predicho de Y) sea independiente para cada valor de X. Con frecuencia esta suposición se refiere a datos que se recopilan a lo largo de un periodo. Estos tipos de modelos caen bajo la denominación general de series de tiempo. La suposición de independencia se puede evaluar trazando los residuales en el orden o la sucesión en que se obtuvieron los datos observados.

ISBN 978-607-8285-62-4

Si los verdaderos errores son en realidad independientes, el valor esperado del estadístico de Durbin-Watson *d* es alrededor de 2.0. Cualquier valor de *d* menor que 1.5 o 1.6 nos lleva a sospechar que hay autocorrelación.



La gráfica de residuos versus orden no muestra un efecto de "autocorrelación" entre observaciones sucesivas, es decir no hay correlación entre una observación en particular y aquellos valores que la precedieron y la siguieron no afectando la suposición de independencia. Además el estadístico de Durbin-Watson es ***d*= 2.81773.** Este valor es mayor a 1.5 por lo que no se puede pensar en que la autocorrelación sea un problema.

ISBN 978-607-8285-62-4

2.6.1.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE **APRENDIZAJE 2.6.1.1** ANÁLISIS DE **RESIDUALES**

Una compañía refresquera está estudiando el efecto de su última campaña publicitaria. A un grupo de personas a quienes escogió al azar se les pregunto por teléfono cuantas latas del nuevo refresco habían comprado en la semana anterior y cuantos anuncios de él habían leído o visto esa semana.

Persona	1	2	3	4	5	6	7	8	9	10
No. de anuncios (X)	4	10	3	0	1	4	2	5	6	8
Latas compradas (Y)	12	19	7	6	3	5	6	10	7	11

n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Análisis de residuales

<u>Solución al inciso n.</u>

Por tanto para calcular el primer residual haremos lo siguiente:

$$\varepsilon_1 = Y_1 - \hat{Y}_1 =$$

Usando la calculadora:

$Y_1 - X_i$ $\boxed{\text{SHIFT}}$ $\boxed{\text{S} - \text{VAR}}$ $\boxed{\text{REPLAY} \rightarrow}$ $\boxed{\text{REPLAY} \rightarrow}$ $\boxed{\text{REPLAY} \rightarrow}$ $\boxed{2}$ $\boxed{=}$

Y así sucesivamente con los siguientes 9 datos.

ISBN 978-607-8285-62-4

Elementos de la matriz sombrero h_i

Para estandarizar estos residuales haremos lo siguiente:

$$h_1 = \frac{1}{n} + \frac{(X_1 - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2} =$$

Usando la calculadora:

$\mathbf{1}$ $\boxed{\div}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-SUM}}$ $\boxed{\mathbf{3}}$ $\boxed{=}$ $\boxed{+}$ $\boxed{(}$ $\boxed{(}$ $\mathbf{X_i}$ $\boxed{-}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-VAR}}$ $\boxed{\mathbf{1}}$ $\boxed{)}$ $\boxed{\mathbf{X^2}}$ $\boxed{\div}$
$\boxed{(}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-SUM}}$ $\boxed{\mathbf{1}}$ $\boxed{-}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-SUM}}$ $\boxed{\mathbf{3}}$ $\boxed{\mathbf{X}}$ $\boxed{\mathbf{SHIFT}}$ $\boxed{\mathbf{S-VAR}}$ $\boxed{\mathbf{1}}$ $\boxed{\mathbf{X^2}}$ $\boxed{)}$ $\boxed{=}$

Y así sucesivamente con los siguientes 9 datos.

ISBN 978-607-8285-62-4

Entonces

$$SR_1 = \frac{\varepsilon_1}{S_{Y:X}\sqrt{1-h_1}} =$$

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Obs.	Variable Independiente (*X*)	Variable Dependiente (*Y*)	y gorro$(\hat{Y})_i$	Residual $\varepsilon_i = Y_i - \hat{Y}_i$	h_i	$S_{Y:X}$	Residual Estandarizado SR_i
1	13	1.0					
2	16	2.0					
3	14	1.4					
4	11	0.8					
5	17	2.2					
6	8	0.5					
7	13	1.1					
8	17	2.8					
9	1	3.0					
10	12	1.2					
SUMAS	139	16					
	PROMEDIOS						
	13.9	1.6					

El estadístico de Durbin-Watson determina si la correlación entre los términos de error adyacentes es cero

Estadístico de Durbin-Watson

Variable Independiente (*X*)	Variable Dependiente (*Y*)	$\boxed{\hat{Y}_i}$	$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$	$\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i$	$(\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i)^2$	$\hat{\varepsilon}^2$
13	1.0					
16	2.0					
14	1.4					
11	0.8					
17	2.2					
8	0.5					
13	1.1					
17	2.8					
1	3.0					
12	1.2					
				d=		

ISBN 978-607-8285-62-4

Diagnóstico de la regresión

Solución al inciso o.

Evaluación de las suposiciones

Linealidad

Linealidad

Homoscedasticidad

Homoscedasticidad

ISBN 978-607-8285-62-4

Normalidad

Normalidad

Independencia

Independencia

ISBN 978-607-8285-62-4

2.6.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 2.6.1.1
ANÁLISIS DE RESIDUALES

El propietario de una cadena de heladerías desea estudiar el efecto de la temperatura atmosférica (X) sobre las ventas (Y) durante la temporada de verano. Seleccionó una muestra aleatoria de 12 días con los resultados siguientes:

Día	1	2	3	4	5	6	7	8	9	10	11	12
X	22.8	23.9	24.1	25.3	26.7	27.8	29.5	31.1	32.2	33.4	36.7	37.8
Y	18.0	20.5	21.8	22.9	23.6	22.5	26.8	29.0	31.4	32.4	34.0	32.8

n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Análisis de residuales

Solución al inciso n.

Por tanto para calcular el primer residual haremos lo siguiente:

$$\varepsilon_1 =$$

ISBN 978-607-8285-62-4

Usando la calculadora:

Y así sucesivamente con los siguientes 11 datos.

Elementos de la matriz sombrero h_i

Para estandarizar estos residuales haremos lo siguiente:

$$h_1 =$$

Usando la calculadora:

Y así sucesivamente con los siguientes 11 datos.

Entonces

$$SR_1 =$$

ISBN 978-607-8285-62-4

Y así sucesivamente con los siguientes 11 datos con lo que se construye la siguiente tabla resumen:

Obs.	Variable Independiente (*X*)	Variable Dependiente (*Y*)	y gorro$(\hat{Y})_i$	Residual $\varepsilon_i = Y_i - \hat{Y}_i$	h_i	$S_{Y:X}$	Residual Estandarizado SR_i
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
SUMAS							
	PROMEDIOS						

El estadístico de Durbin-Watson determina si la correlación entre los términos de error adyacentes es cero

Estadístico de Durbin-Watson

Variable Independiente (*X*)	Variable Dependiente (*Y*)	$\boxed{\hat{Y}_i}$	$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$	$\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i$	$(\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i)^2$	$\hat{\varepsilon}^2$
				d=		

ISBN 978-607-8285-62-4

Diagnóstico de la regresión

Solución al inciso o.

Evaluación de las suposiciones

Linealidad

Linealidad

Homoscedasticidad

Homoscedasticidad

ISBN 978-607-8285-62-4

Normalidad

Normalidad

Independencia

Independencia

ISBN 978-607-8285-62-4

2.6.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 2.6.1 ANÁLISIS DE RESIDUALES

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

2.6.1.1 El presidente de una fábrica de computadoras desea estudiar la relación que hay entre el tamaño del incremento anual de sueldos y rendimientos de un representante de ventas en el año siguiente. Muestreo a 10 representantes y determino los tamaños de sus respectivos incrementos (dados en porcentaje de sus sueldos individuales) y el numero de ventas realizadas por cada uno durante los 12 meses después del incremento.

Representante	1	2	3	4	5	6	7	8	9	10
Incremento en porcentaje (X)	7.2	6.5	7.4	5.7	6.8	8.9	7.7	10.6	11.3	8.2
No. de ventas (Y)	55	39	58	36	41	70	53	74	66	73

n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.

2.6.1.2 Una cadena de tiendas de repostería ha tenido grandes fluctuaciones en sus ingresos durante los últimos años. Abundantes baratas, nuevos productos y técnicas de publicidad se han utilizado durante este tiempo, por lo cual es difícil determinar cuáles de estos factores tienen la influencia más profunda en las ventas. El departamento de mercadotecnia ha estudiado varias relaciones y piensa que los gastos mensuales destinados a carteles pueden ser significativos. Muestreó 10 meses y descubrió lo siguiente:

Mes	1	2	3	4	5	6	7	8	9	10
Gasto mensual en carteles en miles de pesos (X)	25	16	43	34	10	21	19	13	23	38
Ingreso mensual por ventas en miles de pesos (Y)	34	14	55	32	26	29	20	20	30	39

n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.

ISBN 978-607-8285-62-4

OBJETIVO 2.7. El alumno podrá calcular e interpretar los criterios desarrollados para realizar un análisis de influencias e identificar observaciones influyentes.

ANTECEDENTES

CONCEPTOS DE:

Variable dependiente. Variable independiente. Principio de mínimos cuadrados. Forma general de la ecuación de regresión lineal simple. Punto donde se intercepta la línea de regresión con el eje Y. Pendiente de la línea de regresión. Coeficientes de regresión. Error estándar del estimador. Residual. Residual estandarizado. Distribución t de Student. Elementos de la matriz sombrero. Valores atípicos.

2.7.1 DIAGNÓSTICO DE LA REGRESIÓN: ANÁLISIS DE INFLUENCIAS

CONCEPTOS BÁSICOS ANÁLISIS DE INFLUENCIA

En modelos de regresión y ANOVA, h_i mide la distancia de un valor x de observación hasta el promedio de los valores x para todas las observaciones en un conjunto de datos. Las

En los **análisis de regresión** se relacionan tanto la **evaluación de lo apropiado de un modelo en particular** como del **efecto potencial ó la "influencia" de cada punto sobre ese modelo ajustado.**

Existen varios métodos que miden la influencia de ciertos puntos de datos. Entre diversos criterios recientemente desarrollados destacan los siguientes:

1. Los elementos de la matriz sombrero, h_i.
2. Los residuales eliminados de Student, t_i^* .
3. El estadístico de distancia de Cook, D_i

1.- Los elementos de la matriz sombrero, h_i :

En este caso cada h_i , refleja la "influencia" de cada X_i sobre el modelo de regresión ajustado. **Si existen esos puntos de influencia** quizás sea necesario **evaluar de nuevo la necesidad de mantenerlos en el modelo**. Se puede representar de la siguiente manera:

ISBN 978-607-8285-62-4

observaciones con valores apalancamiento grandes pudieran ejercer una influencia desproporcionada sobre el modelo y producir resultados desviados

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2}$$

Según algunos autores sostienen que si $\mathbf{h_i > 4/n}$, entonces X_i es **un punto de influencia** que puede **afectar de modo adverso al modelo** y se puede considerar candidato a ser retirado del modelo.

Los residuales eliminados studentizados son útiles en la detección de valores atípicos. La observación se omite para ver cómo se comporta el modelo sin este valor atípico potencial

2.- Los residuales eliminados de Student, t_i^*:

Para **medir mejor** la **repercusión adversa** sobre el modelo de cada caso individual algunos autores desarrollaron el **residual eliminado de Student** t^*_i el cual se puede representar mediante la siguiente ecuación:

$$t_i^* = \frac{\varepsilon_i}{S_{(i)}\sqrt{1 - h_i}}$$

Donde $\mathbf{S_{(i)}}$ = error estándar de la estimación para un modelo que incluye todas las observaciones excepto la observación "i" .

Entonces si $|t^*_i| > t_{0.10.\ n-3}$

Significaría que los valores Y observados y los predichos **son tan diferentes** que X_i es **un punto de influencia que afecta de modo adverso al modelo** y se puede considerar como un candidato para ser eliminado.

Si hubiera **falta de consistencia** se debe tomar en cuenta otro criterio, el D_i de Cook, que se basa tanto en h_i como en el **estadístico residual estandarizado** t^*_i.

La distancia de Cook (D_i) es una medidas de la influencia de una observación sobre el conjunto de coeficientes de regresión en un modelo de regresión o ANOVA. Las observaciones influyentes tienen un impacto desproporcionado sobre el modelo y pueden generar resultados engañosos

3.- Estadístico de distancia de Cook, D_i**:**

Para **decidir** si un punto que ha sido **destacado mediante el criterio** h_i ó t_i^* está **afectando indebidamente al modelo**, Cook sugiere el uso del estadístico D_i el cual se puede expresar mediante la siguiente ecuación,

$$D_i = \frac{SR_i^2 h_i}{2(1 - h_i)}$$

Donde SR_i es el residual estandarizado.

Si $\mathbf{D_i > F_{0.50.\ n+1.n-k-1}}$

Significaría que quizás **la observación tenga una repercusión sobre los resultados** del ajuste del modelo de regresión lineal.

ISBN 978-607-8285-62-4

2.7.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 2.7.1.1 ANÁLISIS DE INFLUENCIA

El director general de una cadena de tiendas de autoservicio en expansión desea conocer el comportamiento de las ventas en los diferentes establecimientos con base en la superficie de piso en la que se exhiben los diferentes productos con el fin de contar con un modelo que le permita llevar un control adecuado de la eficiencia con la que trabaja cada establecimiento. Para ello utiliza el volumen de ventas mensuales (en millones de pesos) y la superficie de piso (en miles de metros cuadrados). En forma aleatoria recopila el volumen de ventas del último mes en diez tiendas de la cadena que correspondan más o menos entre 2,000 y 12,000 metros cuadrados de superficie de piso.

Tienda	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Superficie *(X)* en miles de m^2	2.15	9.20	6.70	13.5	5.5	12.15	4.80	10.7	3.25	8.25
Ventas *(Y)* en millones de pesos	1.0	3.0	3.0	4.5	2.0	5.0	1.0	4.0	1.5	3.5

p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.

q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.

r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original**.**

Elementos de la matriz sombrero h_i

<u>Solución al inciso p.</u>

Cada h_i refleja la "influencia" de cada X_i sobre el modelo de regresión ajustado. Si existen esos puntos de influencia quizá sea necesario evaluar de nuevo la necesidad de mantenerlos en el modelo. En la regresión lineal simple Hoaglin y Welsch sugieren la siguiente regla de decisión: Si $h_i > {}^4/_n$ entonces X_i es un punto de influencia y se puede considerar candidato a ser eliminado del modelo.

Cuando se desarrollo una estimación por intervalo de confianza $\mu_{Y.X}$, se definieron los "elementos diagonales de la matriz sombrero" h_i como

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2}$$

ISBN 978-607-8285-62-4

Así para el primer punto u observación,

$$h_1 = \frac{1}{n} + \frac{(X_1 - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2} = \frac{1}{10} + \frac{(2.15 - 7.62)^2}{710.43 - 10(7.62)^2} = \frac{1}{10} + \frac{29.92090}{129.786} = \mathbf{0.33054}$$

Usando la calculadora:

$$\mathbf{1}\ \boxed{\div}\ \boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-SUM}}\boxed{\mathbf{3}}\ \boxed{=}\ \boxed{+}\boxed{(}\boxed{(}\mathbf{2.15}\ \boxed{-}\ \boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-VAR}}\boxed{\mathbf{1}}\boxed{)}\boxed{\mathbf{X^2}}\ \boxed{\div}\ \boxed{(}\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-SUM}}\boxed{\mathbf{1}}\ \boxed{-}\ \boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-SUM}}\boxed{\mathbf{3}}\boxed{\mathbf{X}}\ \boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-VAR}}\boxed{\mathbf{1}}\boxed{\mathbf{X^2}}\boxed{)}\ \boxed{=}\ \mathbf{0.33054}$$

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Obs	1	2	3	4	5	6	7	8	9	10
X	2.15	9.20	6.00	13.50	5.50	12.15	4.80	10.70	3.25	8.25
Y	1.0	3.0	3.0	4.5	2.0	5.0	1.0	4.0	1.5	3.5
hi	**0.33054**	**0.11923**	**0.10652**	**0.36640**	**0.13463**	**0.25811**	**0.16127**	**0.17309**	**0.24714**	**0.10306**

Interpretación: Para los datos del volumen de ventas, puesto que n=10, los criterios deben ser "destacar" cualquier valor h_i superior a 4/10=0.40. Consultando la tabla anterior se puede observar que ninguna observación es candidata potencial para ser removida del modelo del tiempo de entrega.

Residuales de Student eliminados, t_i^*

<u>Solución al inciso q.</u>

En el estudio del análisis de residuales se definieron los residuales estandarizados en la ecuación como:

$$SR_i = \frac{\varepsilon_i}{S_{Y:X}\sqrt{1 - h_i}}$$

ISBN 978-607-8285-62-4

Residual de Student eliminado t_i^*.

Para medir la repercusión adversa sobre el modelo de cada caso individual, Hoaglin y Welsch desarrollaron también el residual de Student eliminado t_i^* que se presenta en la siguiente ecuación:

$$t_i^* = \frac{\varepsilon_i}{S_{(i)}\sqrt{1-h_i}}$$

Donde $S_{(i)}$ = error estándar de la estimación para un modelo que incluye todas las observaciones excepto la observación i.

En regresión lineal simple Hoaglin y Welsch sugieren que si

$$|\boldsymbol{t_i^*}| > \boldsymbol{t_{0.10.n-3}}$$

Los valores Y observados y predichos son tan diferentes que X_i es un punto de influencia que afecta de modo adverso el modelo y se puede considerar como un candidato para ser eliminado.

Así para la observación No. 1 primero debemos calcular la ecuación de regresión considerando sólo las observaciones de la 2 a la 10,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} XY - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X^2 - n\bar{X}^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

$$\boldsymbol{\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i}$$

Tienda	Superficie de piso (X) en miles	Volumen Ventas (Y)	XY	X*X	Y*Y
	de metros	en millones			
	cuadrados	de pesos			
2	9.2	3.0	27.6	84.64	9.0
3	6.7	3.0	20.1	44.89	9.0
4	13.5	4.5	60.8	182.25	20.3
5	5.5	2.0	11.0	30.25	4.0
6	12.15	5.0	60.8	147.6225	25.0
7	4.8	1.0	4.8	23.04	1.0
8	10.7	4.0	42.8	114.49	16.0
9	3.25	1.5	4.9	10.5625	2.3
10	8.25	3.5	28.9	68.0625	12.3
SUMAS	74.1	27.5	261.55	705.8075	98.75
	PROMEDIOS				
	8.22778	3.05556			

ISBN 978-607-8285-62-4

$$\hat{\beta}_1 = \frac{261.55 - 9(8.22778)(3.05556)}{705.8075 - 9\,(3.05556)^2} = \frac{261.55 - 226.26389}{705.8075 - 609.26694}$$
$$= \frac{35.28611}{96.54056} \cong 0.36551$$

$$\hat{\beta}_0 = 3.05556 - (0.36551)(8.22778) \cong 0.04826$$

El modelo ajustado sin considerar la observación No. 1, se puede expresar como:

Modelo ajustado sin considerar la observación No. 1.

$$\boldsymbol{\hat{Y}_i \cong 0.04826 + 0.36551 X_i}$$

Usando la calculadora:

En el modo **REG**

$\boxed{\mathbf{1\ (Lin)}}$

$\boxed{\mathbf{SHIFT}}\boxed{\mathbf{CLR}}\boxed{\mathbf{1}}\mathbf{(Scl)}\ \boxed{=}$ (para borra la memoria estadística)

$\mathbf{9.2}\ \boxed{,}\ \mathbf{3.0}\ \boxed{\mathbf{M+}}$ $\boxed{\begin{array}{l}\mathbf{REG}\\ \mathbf{n =\qquad\qquad 1.}\end{array}}$

Cada vez que presiona $\boxed{M+}$ para registrar un ingreso (par ordenado) , el número de dato ingresado (par ordenado) hasta este punto se indica sobre la presentación (valor *n*) .

$\mathbf{6.7}\ \boxed{,}\ \mathbf{3.0}\ \boxed{\mathbf{M+}}\ \ \mathbf{13.5}\ \boxed{,}\ \mathbf{4.0}\ \boxed{\mathbf{M+}}\ \ldots\ldots\ldots\ \mathbf{8.25}\ \boxed{,}\ \mathbf{3.5}\ \boxed{\mathbf{M+}}$ $\boxed{\begin{array}{l}\mathbf{REG}\\ \mathbf{n =\qquad\qquad 9.}\end{array}}$

Coeficiente de regresión A= 0.04826

$\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-VAR}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{1}}\ \boxed{=}\ \mathbf{0.04826}\ldots$

(Especifica cinco lugares decimales) $\boxed{\mathbf{MODE}}\boxed{\mathbf{MODE}}\boxed{\mathbf{MODE}}\boxed{\mathbf{1}}\mathbf{(Fix)}\boxed{\mathbf{5}}$ $\boxed{\begin{array}{l}\mathbf{FIX}\\ \mathbf{0.04826}\end{array}}$

Coeficiente de regresión B= 0.36551

$\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-VAR}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{2}} = \mathbf{0.36551}$

ISBN 978-607-8285-62-4

Ahora se debe calcular el error estándar del estimador para esta ecuación,

Error estándar del estimador.

$$S_{(1)} = \sqrt{\frac{\sum_{i=1}^{n}(Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} Y^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y - \hat{\beta}_1 \sum_{i=1}^{n} XY}{n-2}}$$

$$= \sqrt{\frac{98.75 - 0.04826(27.50000) - 0.36551(261.55)}{9-2}} \cong \sqrt{\frac{1.82495}{7}}$$

$$\cong \sqrt{0.26071} \cong \mathbf{0.51060}$$

Usando la calculadora:

$\boxed{\sqrt{\ }}$ $\boxed{(}$ $\boxed{(}$ $\boxed{SHIFT}$ $\boxed{S-SUM}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{1}$ $\boxed{-}$ $\boxed{SHIFT}$ $\boxed{S-VAR}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{1}$ $\boxed{X}$ $\boxed{SHIFT}$ $\boxed{S-SUM}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{2}$ $\boxed{-}$ $\boxed{SHIFT}$ $\boxed{S-VAR}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{2}$ $\boxed{X}$ $\boxed{SHIFT}$ $\boxed{S-SUM}$ $\boxed{\text{REPLAY}\rightarrow}$ $\boxed{3}$ $\boxed{)}$ $\boxed{\div}$ $\boxed{7}$ $\boxed{)}$ $\boxed{=}$ $\mathbf{0.51060}$

Por lo tanto el residual de Student eliminado t_i^*

Residual de Student eliminado t_i^* sin la observación 1.

$$\boldsymbol{t_1^*} = \frac{\varepsilon_1}{S_{(1)}\sqrt{1-h_1}} = \frac{0.11107}{0.51060\sqrt{1-0.33054}} = \frac{0.11107}{0.41778} \cong \mathbf{0.26586}$$

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Ob	1	2	3	4	5	6	7	8	9	10
X	2.15	9.20	6.70	13.50	5.50	12.15	4.80	10.70	3.25	8.25
Y	1.0	3.0	3.0	4.5	2.0	5.0	1.0	4.0	1.5	3.5
$S_{(i)}$	**0.51060**	**0.48498**	**0.47595**	**0.46479**	**0.51186**	**0.45834**	**0.37874**	**0.51281**	**0.50441**	**0.48445**
h_i	0.33054	0.11923	0.10652	0.36640	0.13463	0.25811	0.16127	0.17309	0.24714	0.10306
t_i^*	**0.26586**	**0.91497**	**1.06656**	**1.23809**	**0.18891**	**1.33222**	**2.41882**	**0.09817**	**0.49514**	**0.92444**

ISBN 978-607-8285-62-4

Interpretación: Para los datos del volumen de ventas, puesto que n=10, los criterios deben ser "destacar" cualquier valor superior a $|t_i^*| > t_{0.10.7} = 1.89458$. Consultando la tabla anterior se puede visualizar que $t_7^* = -2.41882$. Por lo tanto la séptima tienda puede tener un efecto adverso sobre el modelo y se puede considerar candidato a ser retirado del modelo, sin embargo como de acuerdo al criterio h_i la tienda 7 no presento un efecto adverso, se debe tomar en cuenta otro criterio antes de tomar esa decisión como el criterio D_i de Cook, que se basa tanto en h_i como en el estadístico residual estandarizado t_i^*.

Estadístico de distancia de Cook, D_i

Solución al inciso r.

Para decidir si un punto que ha sido destacado mediante h_i ó t_i está afectando indebidamente al modelo, Cook y Weisberg sugieren el uso del estadístico D_i. En el modelo de regresión lineal simple se muestra $D_i$en la ecuación:

$$D_i = \frac{SR_i^2 h_i}{2(1-h_i)}$$

En la regresión lineal simple Cook y Weisberg sugieren que si

$$D_i > F_{0.50.2.n-2}$$

Lo que significaría que posiblemente la observación tenga una repercusión sobre los resultados del ajuste del modelo de regresión lineal simple.

Así para el primer punto u observación,

Tienda	Superficie de piso (X)	Volumen Ventas (Y)	*hi*	Residual Estandarizado *SRi*
1	2.15	1.0	0.33054	0.28279
2	9.2	3.0	0.11923	-0.92442
3	6.7	3.0	0.10652	1.05751
4	13.5	4.5	0.36640	-1.19881
5	5.5	2.0	0.13463	-0.20144
6	12.15	5.0	0.25811	1.27204
7	4.8	1.0	0.16127	-1.90847
8	10.7	4.0	0.17309	0.10488
9	3.25	1.5	0.24714	0.52029
10	8.25	3.5	0.10306	0.93296
SUMAS	76.2	28.5		
	PROMEDIOS			
	7.62	2.85		

ISBN 978-607-8285-62-4

$$\boldsymbol{D_1} = \frac{SR_1^2 h_1}{2(1-h_1)} = \frac{0.28279^2(0.33054)}{2(1-0.33054)} = \frac{0.02643}{1.33892} = \mathbf{0.01974}$$

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Ob	1	2	3	4	5	6	7	8	9	10
X	2.15	9.20	6.70	13.50	5.50	12.15	4.80	10.70	3.25	8.25
Y	1.0	3.0	3.0	4.5	2.0	5.0	1.0	4.0	1.5	3.5
Di	**0.01974**	**0.05784**	**0.06666**	**0.41553**	**0.00316**	**0.28148**	**0.35017**	**0.00115**	**0.04443**	**0.05000**

Interpretación: Para los datos del volumen de ventas (en millones de pesos), puesto que n=10, el criterio sería "destacar" cualquier $D_i > F_{0.50.2.8} = 0.756828$. Consultando la tabla anterior se puede observar que ninguna observación es candidata potencial para ser removida del modelo del volumen de ventas. En caso de que alguna observación una vez estudiados los tres criterios fuera necesario eliminar alguna(s) observación(es) se debería estudiar un modelo alterno en el que se hayan eliminado dichas observaciones que no fue el caso en este modelo.

ISBN 978-607-8285-62-4

2.7.1.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 2.7.1.1 ANÁLISIS DE INFLUENCIA

Una compañía refresquera está estudiando el efecto de su última campaña publicitaria. A un grupo de personas a quienes escogió al azar se les pregunto por teléfono cuantas latas del nuevo refresco habían comprado en la semana anterior y cuantos anuncios de él habían leído o visto esa semana.

Persona	1	2	3	4	5	6	7	8	9	10
No. de anuncios (X)	4	10	3	0	1	4	2	5	6	8
Latas compradas (Y)	12	19	7	6	3	5	6	10	7	11

p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.

q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.

r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Elementos de la matriz sombrero h_i

Solución al inciso p.

Así para el primer punto u observación,

$$\boldsymbol{h_1 = \frac{1}{n} + \frac{(X_1 - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2} =}$$

ISBN 978-607-8285-62-4

Usando la calculadora:

$$1 \;\boxed{\div}\; \boxed{\text{SHIFT}}\,\boxed{\text{S}-\text{SUM}}\,\boxed{3}\;\boxed{=}\;\boxed{+}\,\boxed{(}\,\boxed{(}\,\boxed{X_i}\;\boxed{-}\;\boxed{\text{SHIFT}}\,\boxed{\text{S}-\text{VAR}}\,\boxed{1}\,\boxed{)}\,\boxed{X^2}\;\boxed{\div}$$
$$\boxed{(}\,\boxed{\text{SHIFT}}\,\boxed{\text{S}-\text{SUM}}\,\boxed{1}\;\boxed{-}\;\boxed{\text{SHIFT}}\,\boxed{\text{S}-\text{SUM}}\,\boxed{3}\,\boxed{X}\;\boxed{\text{SHIFT}}\,\boxed{\text{S}-\text{VAR}}\,\boxed{1}\,\boxed{X^2}\,\boxed{)}\;\boxed{=}$$

Y así sucesivamente con los siguientes 9 datos:

con lo que se construye la siguiente tabla resumen:

Obs.	1	2	3	4	5	6	7	8	9	10
X	4	10	3	0	1	4		5	6	8
Y	12	19	7	6	3	5	6	10	7	11
hi										

ISBN 978-607-8285-62-4

Interpretación:

Residuales de Student eliminados, t_i^*

Solución al inciso q.

Así para la observación No. 1 primero debemos calcular la ecuación de regresión considerando sólo las observaciones de la 2 a la 10,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} XY - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X^2 - n\bar{X}^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

$$\boldsymbol{\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i}$$

Obs.	X	Y	XY	X*X	Y*Y
2	10	19			
3	3	7			
4	0	6			
5	1	3			
6	4	5			
7	2	6			
8	5	10			
9	6	7			
10	8	11			
SUMAS					
	PROMEDIOS				

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} XY - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X^2 - n\bar{X}^2} =$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} =$$

El modelo ajustado sin considerar la observación No. 1, se puede expresar como:

$$\boldsymbol{\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i =}$$

ISBN 978-607-8285-62-4

Usando la calculadora:

En el modo **REG**

$\boxed{\mathbf{1\ (Lin)}}$

$\boxed{\mathbf{SHIFT}}\boxed{\mathbf{CLR}}\boxed{\mathbf{1}}(\mathbf{Scl})\ \boxed{=}$ (para borra la memoria estadística)

$\mathbf{X_1}\ \boxed{,}\ \mathbf{Y_1}\ \boxed{\mathbf{M+}}$ $\boxed{\begin{array}{c}\mathbf{REG}\\ \mathbf{n =\qquad\qquad 1.}\end{array}}$

Cada vez que presiona $\boxed{\mathbf{M+}}$ para registrar un ingreso (par ordenado) , el número de dato ingresado (par ordenado) hasta este punto se indica sobre la presentación (valor *n*) .

$X_2\ \boxed{,}\ Y_2\ \boxed{\mathbf{M+}}\ X_3\ \boxed{,}\ Y_3\ \boxed{\mathbf{M+}} \ldots\ldots\ldots X_n\ \boxed{,}\ Y_n\ \boxed{\mathbf{M+}}$ $\boxed{\begin{array}{c}\mathbf{REG}\\ \mathbf{n =\qquad\qquad n.}\end{array}}$

Coeficiente de regresión **A=**

$\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-VAR}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{1}}\ \boxed{=}$

(Especifica cinco lugares decimales) $\boxed{\mathbf{MODE}}\boxed{\mathbf{MODE}}\boxed{\mathbf{MODE}}\boxed{\mathbf{1}}(\mathbf{Fix})\boxed{\mathbf{5}}\ \boxed{\mathbf{FIX}}$

Coeficiente de regresión **B=** $\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-VAR}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{2}}\ \boxed{=}$

Ahora se debe calcular el error estándar del estimador para esta ecuación,

$$S_{(1)} = \sqrt{\frac{\sum_{i=1}^{n}(Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}Y^2 - \hat{\beta}_0\sum_{i=1}^{n}Y - \hat{\beta}_1\sum_{i=1}^{n}XY}{n-2}} =$$

Usando la calculadora:

$\boxed{\sqrt{\ }}\boxed{(}\boxed{(}\boxed{SHIFT}\boxed{S-SUM}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{1}\boxed{-}\boxed{SHIFT}\boxed{S-VAR}\boxed{\mathbf{REPLAY\rightarrow}}$
$\boxed{\mathbf{REPLAY\rightarrow}}\boxed{1}\boxed{X}\boxed{SHIFT}$
$\boxed{S-SUM}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{2}\boxed{-}\boxed{SHIFT}\boxed{S-VAR}\boxed{\mathbf{REPLAY\rightarrow}}$
$\boxed{\mathbf{REPLAY\rightarrow}}\boxed{2}\boxed{X}\boxed{SHIFT}\boxed{S-SUM}\boxed{\mathbf{REPLAY\rightarrow}}$
$\boxed{3}\boxed{)}\boxed{\div}\ 7\boxed{)}\boxed{=}$

Por lo tanto el residual de Student eliminado t_i^*

$$t_i^* = \frac{\varepsilon_i}{S_{(i)}\sqrt{1-h_i}} =$$

ISBN 978-607-8285-62-4

Y así sucesivamente con los siguientes 9 datos

con lo que se construye la siguiente tabla resumen:

Obs.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X	4	10	3	0	1	4	2	5	6	8
Y	12	19	7	6	3	5	6	10	7	11
$S_{(i)}$										
h_i										
t_i^*										

Interpretación:

ISBN 978-607-8285-62-4

Estadístico de distancia de Cook, D_i

Solución al inciso r.

Así para el primer punto u observación,

$$D_1 = \frac{SR_1^2 h_1}{2(1-h_1)} =$$

Y así sucesivamente con los siguientes 9 datos

con lo que se construye la siguiente tabla resumen:

Obs.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X	4	10	3	0	1	4	2	5	6	8
Y	12	19	7	6	3	5	6	10	7	11
Di										

Interpretación:

ISBN 978-607-8285-62-4

2.7.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 2.7.1.1
ANÁLISIS DE INFLUENCIA

El propietario de una cadena de heladerías desea estudiar el efecto de la temperatura atmosférica (X) sobre las ventas (Y) durante la temporada de verano. Seleccionó una muestra aleatoria de 12 días con los resultados siguientes:

Día	1	2	3	4	5	6	7	8	9	10	11	12
X	22.8	23.9	24.1	25.3	26.7	27.8	29.5	31.1	32.2	33.4	36.7	37.8
Y	18.0	20.5	21.8	22.9	23.6	22.5	26.8	29.0	31.4	32.4	34.0	32.8

p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.

q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.

r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ISBN 978-607-8285-62-4

Elementos de la matriz sombrero h_i

<u>Solución al inciso p.</u>

Así para el primer punto u observación,

$$\boldsymbol{h_1} =$$

Usando la calculadora:

Y así sucesivamente con los siguientes 11 datos:

con lo que se construye la siguiente tabla resumen:

Obs	1	2	3	4	5	6	7	8	9	10	11	12
(X)												
(Y)												
h_i												

Interpretación:

ISBN 978-607-8285-62-4

Residuales de Student eliminados, t_i^*

Solución al inciso q.

Así para la observación No. 1 primero debemos calcular la ecuación de regresión considerando sólo las observaciones de la 2 a la 12,

Obs.	X	Y	XY	X*X	Y*Y
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
SUMAS					
	PROMEDIOS				

$\hat{\beta}_1 =$

$\hat{\beta}_0 =$

El modelo ajustado sin considerar la observación No. 1, se puede expresar como:

$\hat{Y}_i =$

Usando la calculadora:

$\hat{Y}_i =$

Ahora se debe calcular el error estándar del estimador para esta ecuación,

$S_{(1)} =$

ISBN 978-607-8285-62-4

Usando la calculadora:

$S_{(1)} =$

Por lo tanto el residual de Student eliminado t_i^*

$t_i^* =$

Y así sucesivamente con los siguientes 11 datos

con lo que se construye la siguiente tabla resumen:

Ob	1	2	3	4	5	6	7	8	9	10	11	12
(X)												
(Y)												
$s_{(i)}$												
h_i												
t_i^*												

Interpretación:

ISBN 978-607-8285-62-4

Estadístico de distancia de Cook, D_i

<u>Solución al inciso r.</u>

Así para el primer punto u observación,

$$D_1 =$$

Y así sucesivamente con los siguientes 11 datos

con lo que se construye la siguiente tabla resumen:

Ob	1	2	3	4	5	6	7	8	9	10	11	12
(X)												
(Y)												
D_i												

Interpretación:

ISBN 978-607-8285-62-4

2.7.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 2.7.1 ANÁLISIS DE INFLUENCIA

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

2.7.1.1 El presidente de una fábrica de computadoras desea estudiar la relación que hay entre el tamaño del incremento anual de sueldos y rendimientos de un representante de ventas en el año siguiente. Muestreo a 10 representantes y determino los tamaños de sus respectivos incrementos (dados en porcentaje de sus sueldos individuales) y el numero de ventas realizadas por cada uno durante los 12 meses después del incremento.

Representante	1	2	3	4	5	6	7	8	9	10
Incremento en porcentaje (X)	7.2	6.5	7.4	5.7	6.8	8.9	7.7	10.6	11.3	8.2
No. de ventas (Y)	55	39	58	36	41	70	53	74	66	73

p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.

q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.

r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

2.7.1.2 Una cadena de tiendas de repostería ha tenido grandes fluctuaciones en sus ingresos (Y) durante los últimos años. Abundantes baratas, nuevos productos y técnicas de publicidad se han utilizado durante este tiempo, por lo cual es difícil determinar cuáles de estos factores tienen la influencia más profunda en las ventas. El departamento de mercadotecnia ha estudiado varias relaciones y piensa que los gastos mensuales destinados a carteles (X) pueden ser significativos. Muestreó 10 meses y descubrió lo siguiente:

ISBN 978-607-8285-62-4

Mes	1	2	3	4	5	6	7	8	9	10
(X)	25	16	43	34	10	21	19	13	23	38
Ingreso mensual por ventas en miles de pesos (Y)	34	14	55	32	26	29	20	20	30	39

p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.

q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.

r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

R.L.S.

EJEMPLO **ILUSTRATIVO EN EXCEL**

EJEMPLO ILUSTRATIVO INTEGRAL EN EXCEL REGRESIÓN LINEAL SIMPLE

El director general de una cadena de tiendas de autoservicio en expansión desea conocer el comportamiento de las ventas en los diferentes establecimientos con base en la superficie de piso en la que se exhiben los diferentes productos con el fin de contar con un modelo que le permita llevar un control adecuado de la eficiencia con la que trabaja cada establecimiento. Para ello utiliza el volumen de ventas mensuales (en millones de pesos) y la superficie de piso (en miles de metros cuadrados). En forma aleatoria recopila el volumen de ventas del último mes en diez tiendas de la cadena que correspondan más o menos entre 2,000 y 12,000 metros cuadrados de superficie de piso.

Tienda muestriada	1	2	3	4	5	6	7	8	9	10
Superficie *(X)* en miles de m^2	2.15	9.20	6.70	13.50	5.50	12.15	4.80	10.70	3.25	8.25
Ventas *(Y)* en millones de pesos	1.0	3.0	3.0	4.5	2.0	5.0	1.0	4.0	1.5	3.5

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación
c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0 \; y \; \hat{\beta}_1$
e) Determine el error estándar de estimación.
f) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba *t.*
g) Determine un intervalo de confianza de 95% para la pendiente para el volumen de ventas
h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba *F.*
i) Determine e interprete el coeficiente de determinación.
j) Determine e interprete el coeficiente de correlación.
k) Realice un análisis de residuales sobre sus resultados
l) Determine lo adecuado del ajuste del modelo.

Cálculo de la covarianza

<u>Solución al inciso a.</u>

Cuando el número de observaciones en cada variable es extenso, los cálculos manuales son tediosos. Existen muchos paquetes de software que pueden mostrar los resultados entre ellos Excel.

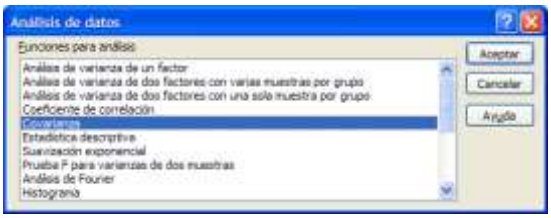
ISBN 978-607-8285-62-4

Comenzamos introduciendo los datos en la hoja de Excel, tal y como se muestra a continuación:

Hoja de trabajo.

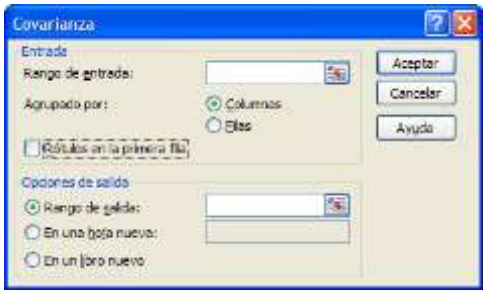
Como tenemos un ***modelo de regresión y correlación lineal simple*** seleccionamos la opción ***Análisis de datos*** del menú ***Datos,*** utilizaremos la opción ***Covarianza***, del cuadro ***Análisis de datos*** de la figura siguiente:

Cuadro de dialogo: Análisis de datos.



En la lista *Funciones para análisis,* elija la modalidad de *Covarianza* y oprima el botón Aceptar para obtener el siguiente cuadro de dialogo rellenando su pantalla de entrada:

Cuadro de diálogo: Covarianza.



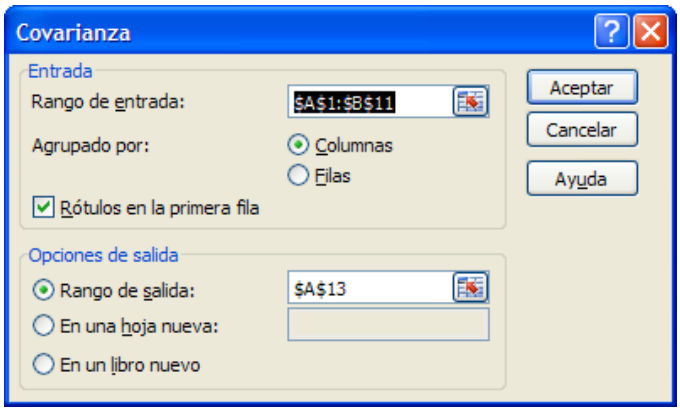
En el cuadro ***Rango de entrada*** introduzca, (seleccionando con el cursor las celdas donde están los datos incluyendo los rótulos de la primera fila), la referencia de celda correspondiente al rango de datos que está analizando. La referencia deberá contener dos o más rangos adyacentes organizados en columnas o filas.

ISBN 978-607-8285-62-4

En el campo ***Agrupado por*** haga clic en el botón ***Columnas*** para indicar que los datos del rango de entrada están organizados en columnas. Si la primera fila del rango de entrada contiene rótulos, active la casilla de verificación ***Rótulos en la primera fila*** . Esta casilla de verificación debe quedar desactivada si el rango de entrada carece de rótulos; Microsoft Office Excel 2007 generará los rótulos de datos correspondientes para la tabla de resultados. Deje sin cambio el campo Alfa con el valor de 0.05 (nivel con el que desee evaluar los valores críticos de la función estadística *F* . El nivel ***alfa*** es un nivel de importancia relacionado con la probabilidad de que haya un error de tipo I (rechazar una hipótesis verdadera).

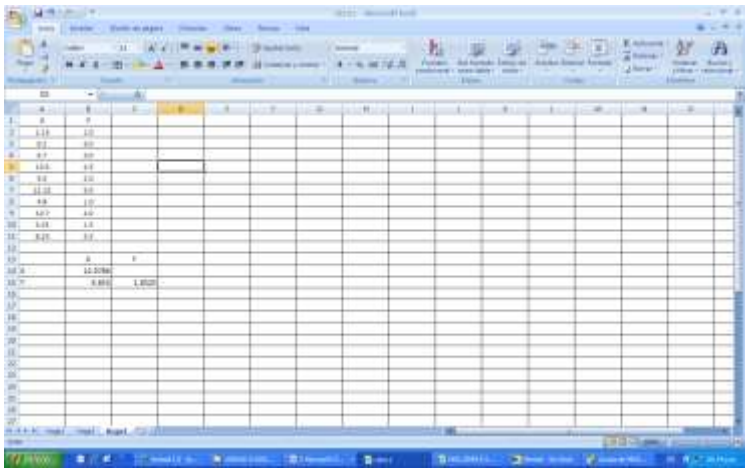
En cuanto a las ***opciones de salida***, en el campo ***Rango de salida*** introduzca la referencia , (dando un clic), correspondiente a la celda superior izquierda de la tabla de resultados, en este caso la celda A13.

Cuadro de diálogo: Covarianza.



Oprima el botón ***Aceptar***
A continuación se muestra la salida de la Covarianza:

Salida de resultados.



ISBN 978-607-8285-62-4

$$cov(X,Y) = \frac{\sum[(X-\bar{X})(Y-\bar{Y})]}{n-1} = 5.17$$

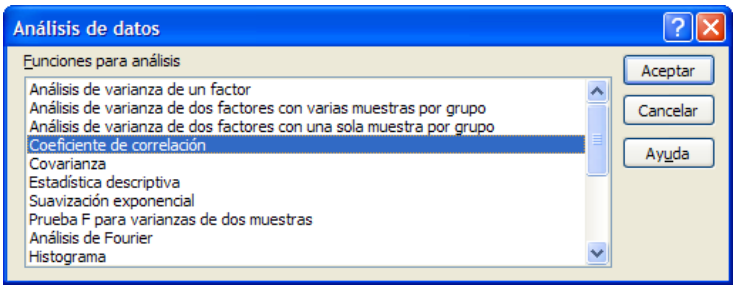
Interpretación. A la vista de los resultados, podemos decir que todas las covarianzas son positivas.

Cálculo del coeficiente de correlación muestral

Solución al inciso b.

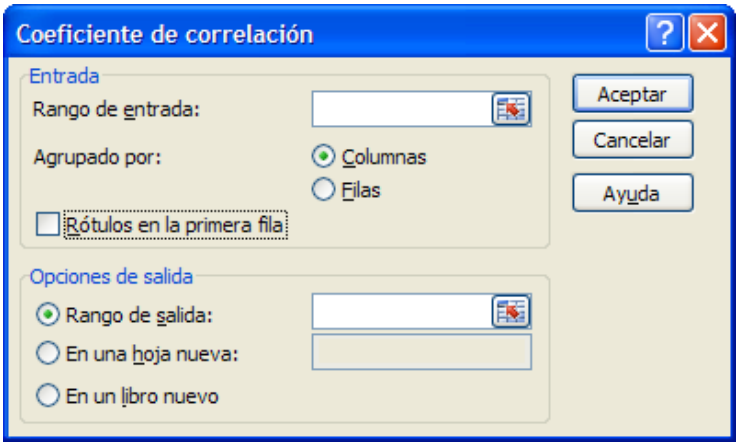
Como tenemos un ***modelo de regresión y correlación lineal simple*** seleccionamos la opción ***Análisis de datos*** del menú *Datos*, utilizaremos la opción ***Coeficiente de correlación*** del cuadro *Análisis de datos* de la figura siguiente:

Cuadro de diálogo: Análisis de datos.



En la lista ***Funciones para análisis,*** elija la modalidad de ***Coeficiente de correlación*** y oprima el botón ***Aceptar*** para obtener el siguiente cuadro de dialogo rellenando su pantalla de entrada:

Cuadro de diálogo: Coeficiente de correlación.



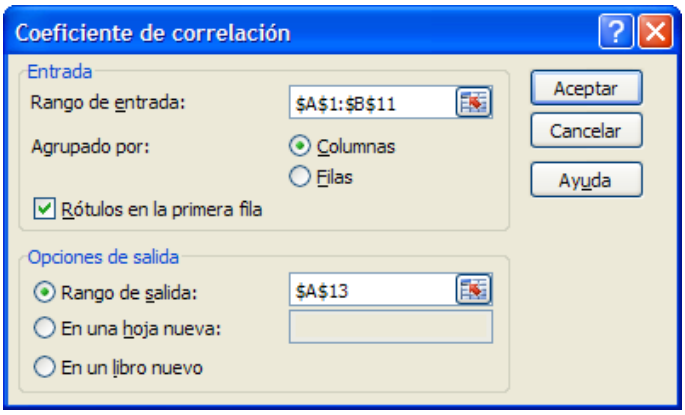
En el cuadro ***Rango de entrada*** introduzca, (seleccionando con el cursor las celdas donde están los datos incluyendo los rótulos de la primera fila), la referencia de celda correspondiente al rango de datos que está analizando. La referencia deberá contener dos o más rangos adyacentes organizados en columnas o filas.

ISBN 978-607-8285-62-4

En el campo ***Agrupado por*** haga clic en el botón ***Columnas*** para indicar que los datos del rango de entrada están organizados en columnas. Si la primera fila del rango de entrada contiene rótulos, active la casilla de verificación ***Rótulos en la primera fila*** . Esta casilla de verificación debe quedar desactivada si el rango de entrada carece de rótulos; Microsoft Office Excel 2007 generará los rótulos de datos correspondientes para la tabla de resultados. Deje sin cambio el campo Alfa con el valor de 0.05 (nivel con el que desee evaluar los valores críticos de la función estadística *F* . El nivel ***alfa*** es un nivel de importancia relacionado con la probabilidad de que haya un error de tipo I (rechazar una hipótesis verdadera).

En cuanto a las ***opciones de salida***, en el campo ***Rango de salida*** introduzca la referencia , (dando un clic), correspondiente a la celda superior izquierda de la tabla de resultados, en este caso la celda A13.

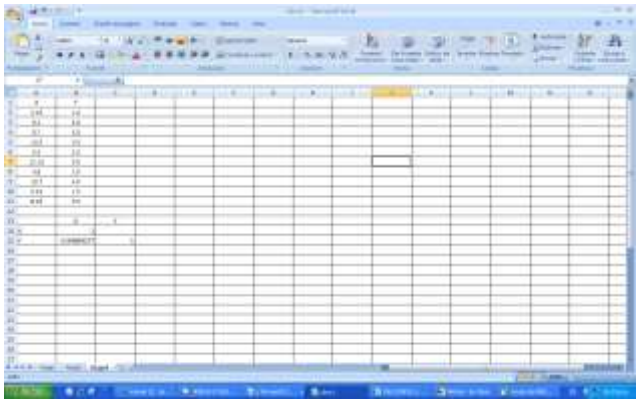
Cuadro de diálogo: Coeficiente de correlación.



Oprima el botón ***Aceptar***

A continuación se muestra la salida del coeficiente de correlación:

Salida de resultados.



$$r = \frac{cov(X,Y)}{S_X S_Y} = \frac{5.17}{(3.797)(1.435)} = 0.94894$$

ISBN 978-607-8285-62-4

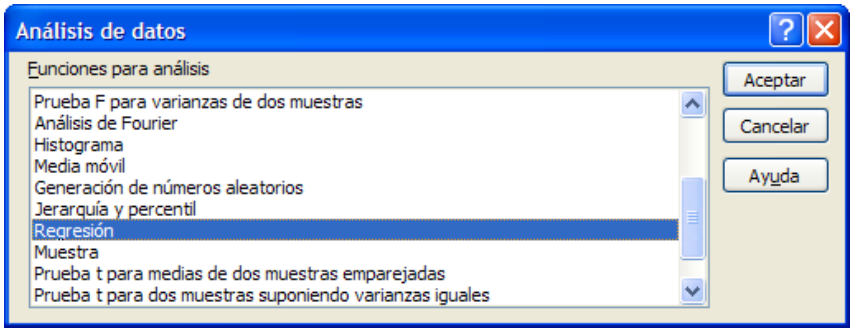
Interpretación. En la salida anterior, se observa que el coeficiente de correlación entre las variables ***X*** e ***Y*** es 0.94894, muy cercano a 1 y como las covarianzas resultaron positivas, esto indica una fuerte dependencia lineal positiva entre el par de variables.

Cálculo de la línea de regresión mediante una ecuación

Solución al inciso c.

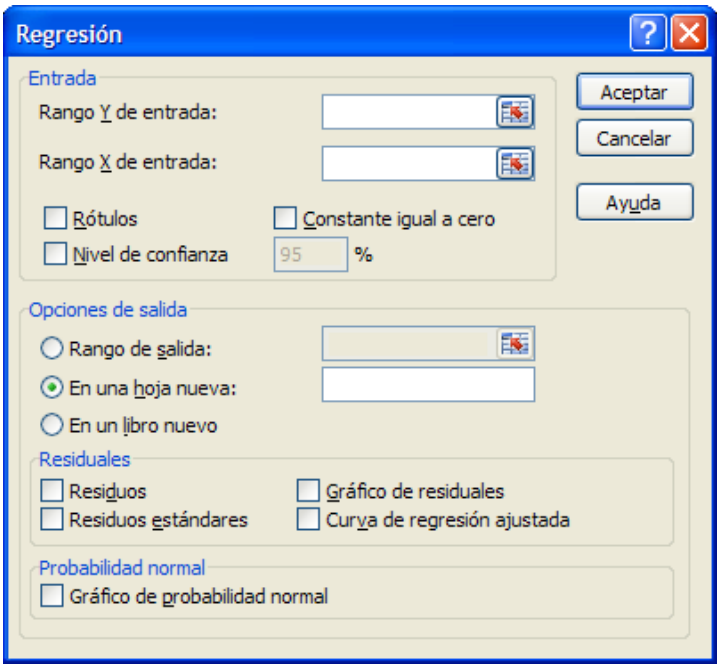
Como tenemos un ***modelo de regresión y correlación lineal simple*** seleccionamos la opción ***Análisis de datos*** del menú ***Datos***, utilizaremos la opción ***Regresión*** del cuadro ***Análisis de datos*** de la figura siguiente:

Cuadro de diálogo: Análisis de datos.



En la lista ***Funciones para análisis,*** elija la modalidad de ***Regresión*** y oprima el botón ***Aceptar*** para obtener el siguiente cuadro de dialogo rellenando su pantalla de entrada:

Cuadro de diálogo: Regresión.



ISBN 978-607-8285-62-4

Cuadros de diálogo.

Los campos del cuadro de dialogo anterior tienen las siguientes funcionalidades:

En el cuadro ***Rango Y de entrada***: Introduzca la ***referencia*** correspondiente al rango de datos dependientes. El rango debe constar de una única columna o una única columna de datos.

En el cuadro ***Rango X de entrada***: Introduzca la ***referencia*** correspondiente al rango de datos de la variable independiente.

En el cuadro ***Rótulos***: Active esta casilla si la primera fila o la primera columna del rango (o rangos) de entrada contiene rótulos. Desactívela si el rango de entrada carece de rótulos; Excel generará los rótulos de datos correspondientes para la tabla de resultados.

En el cuadro ***Nivel de confianza***: Active esta casilla para incluir más niveles en la tabla resumen de resultados. Teclee el nivel de confianza a aplicar además del nivel predeterminado del 95%.

En el cuadro ***Constante igual a cero***: Active esta casilla para que la línea o plano de regresión pase por el origen.

En el cuadro ***Rango de salida***: Introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados. Deje por lo menos siete columnas disponibles para la tabla de resultados sumarios, que incluirá una tabla de análisis de datos, coeficientes, error típico del pronóstico Y, valores de R^2, número de observaciones y error típico de coeficientes.

En el cuadro ***En una hoja nueva***: Haga clic en esta opción para insertar una hoja nueva en el libro actual y pegar los resultados, comenzando por la celda A1 de la nueva hoja de cálculo. Para darle un nombre a la nueva hoja de cálculo, escríbalo en el cuadro.

En el cuadro ***En un libro nuevo***: Haga clic en esta opción para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado.

En el cuadro ***Residuos***: Active esta casilla para incluir residuos en la tabla de resultados de residuos.

En el cuadro ***Residuos estándares***: Active esta casilla para incluir residuos estándares en la tabla de resultados de residuos.

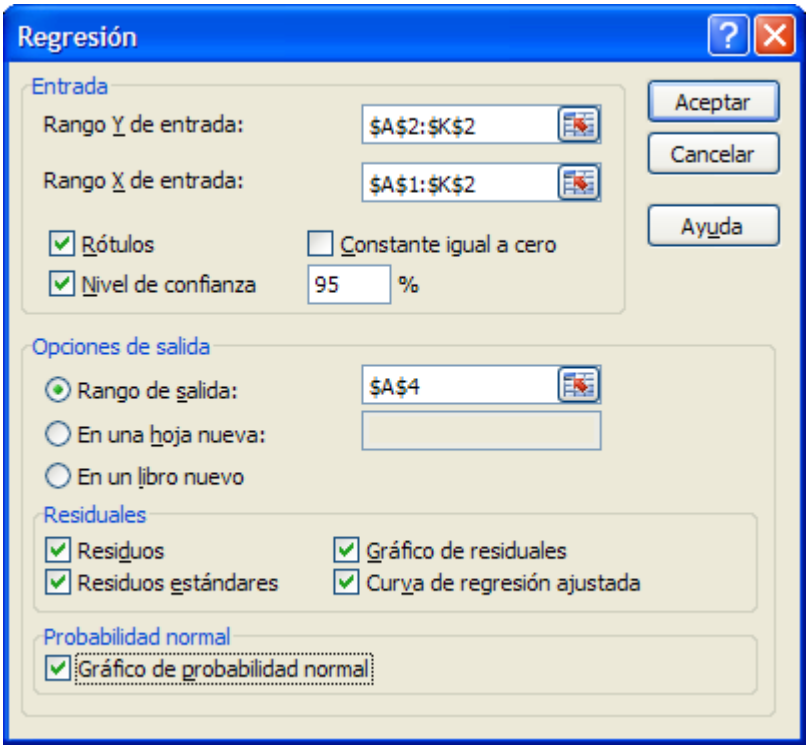
En el cuadro ***Gráficos de residuos***: Active esta casilla para generar un gráfico por cada variable independiente frente al residuo.

En el cuadro ***Curva de regresión ajustada***: Active esta casilla para generar un gráfico con los valores pronosticados frente a los observados.

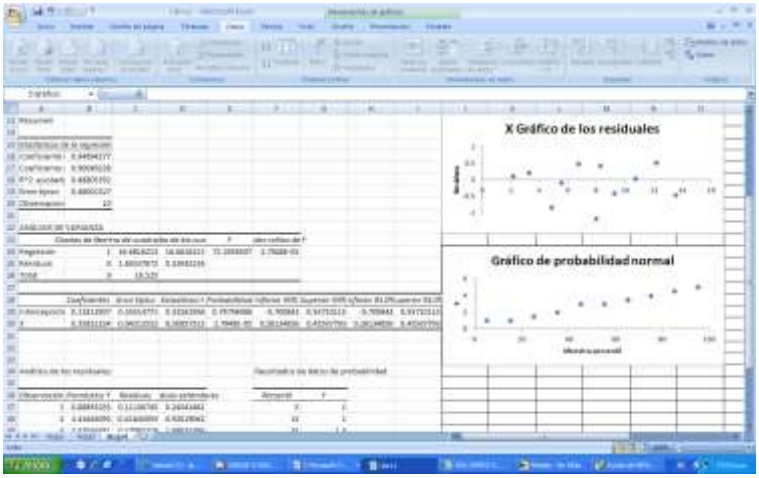
En el cuadro *Trazado de probabilidad normal*: Active esta casilla para generar un gráfico con probabilidad normal.

ISBN 978-607-8285-62-4

Cuadro de diálogo: Regresión.



Al pulsar ***Aceptar*** en el cuadro de dialogo anterior, se obtiene la siguiente salida numérica que incluye estadísticos de regresión, cuadro del análisis de varianza del modelo, estimadores, contrastes de significancia de ***F*** y ***T*** con sus *p*-valores asociados, intervalos de confianza para los parámetros y para las predicciones al 95%, y residuos



El modelo ajustado se puede expresar como:

$$\hat{Y}_i = \mathbf{0.11813} + \mathbf{0.35851}X_i$$

ISBN 978-607-8285-62-4

Interpretación de la recta de regresión

Solución al inciso d.

Con este modelo se podría llegar a la conclusión de que por cada mil metros cuadrado que se incrementan a la superficie de piso donde se exhibe la mercancía de la tienda ***(X)*** , el volumen de ventas (***Y)*** se incrementa en 0.358513 millones de pesos ó por cada metro cuadrado que se incrementa a la superficie de piso donde se exhibe la mercancía de la tienda, el volumen de ventas se incrementa (la pendiente es positiva) en **$ 358.51 pesos**. Esta pendiente también se puede contemplar como representante del volumen de ventas ***(Y)*** que varía de acuerdo a la superficie de piso de la tienda (***X)***. La ordenada al origen $\hat{\beta}_o$ se calculó como **0.11813 millones de pesos ó $ 118,130 pesos** . La ordenada al origen representa el valor del volumen de ventas ***(Y)*** cuando la superficie de piso de la tienda ***(X)*** es igual a cero (en este problema **0.11813 millones de pesos ó $ 118,130 pesos**). Puesto que el resultado de la variable independiente ***(X)*** raramente puede ser cero, esta ordenada se puede considerar como expresión del volumen de ventas ***(Y)*** que varía con factores ajenos al resultado de la superficie de piso de la tienda ***(X).***

Medición de la dispersión alrededor del plano de regresión múltiple: el error estándar de estimación

Solución al inciso e.

El error estándar del estimador, proporcionado por el símbolo $S_{Y.X}$, se define como

$$S_{Y:X} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i}{n-2}} = \mathbf{0.48002}$$

$$S_{Y:X} = \mathbf{0.48002}$$

Prueba de una hipótesis con respecto a β_1

Solución al inciso f.

Se puede determinar si hay relación significativa entre las variables *X* y *Y* al probar si $\beta_1(la\ pendiente\ real)$ es igual a cero. Si se rechaza esta hipótesis se concluiría que hay relación lineal. Si no se rechaza , un cambio en *X* no proporciona ningún cambio pronosticado en *Y*, y se sigue que *X* no tiene ningún valor para predecir *Y* .
La distribución t se puede utilizar para realizar pruebas de significancia y construir intervalos de confianza para la verdadera pendiente y ordenada en el origen.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0: β_1 =0 (no existe relación)
$H_{1:}$ $\beta_1 \neq 0$ (existe relación)

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$t_{calculada(n-2)} = \frac{\hat{\beta}_1}{S_{\beta_1}} = 8.50857$$

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

$$\text{p-level} \leq 0.05$$

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $t_{cal.}$ tiene un p-level ≤ 0.05

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como p-level de 0.000027948 es <0.05 y <0.01 se considera que la prueba es Altamente Significativa (AS) y se rechaza H_o.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación lineal significativa entre el volumen de ventas y la superficie de piso de las tiendas.

Obtención de los limites superior e inferior de la región de no rechazo de H_0

<u>Solución al inciso g.</u>

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de $\hat{\beta}_1$ y determinar si el valor hipotético $(\hat{\beta}_1 = 0)$ está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_1$ se obtendría de la siguiente manera:

$$\boldsymbol{\beta_1 = \hat{\beta}_1 \pm t_{n-2} S_{\hat{\beta}_1}}$$

$$\boldsymbol{\beta_1 = \begin{cases} LIC = 0.26134 \\ LSC = 0.45567 \end{cases}}$$

Interpretación: Se estima, con una confianza del 95%, que la pendiente real se encuentra entre 0.26134 y 0.45567. Puesto que estos valores son superiores a cero se puede concluir que hay relación lineal significativa entre el volumen de ventas y la superficie de piso de las tiendas.

ISBN 978-607-8285-62-4

Prueba *F* de la regresión como un todo

Solución al inciso h.

Hay una prueba alternativa, una prueba ***F*** para la hipótesis nula de un valor predictivo nulo. Esta prueba proporciona el mismo resultado que una prueba ***t*** bilateral de $H_0: \beta_1 = 0$ en la regresión lineal simple. A continuación se presenta un resumen de la prueba ***F***:

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0: β_1 =0 (no existe relación)
$H_{1:}$ $\beta_1 \neq 0$ (existe relación)

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$F_{calculada} = \frac{CMR}{CME} = \frac{SCR/_{G.L.}}{SCE/_{G.L.}} = 72.40$$

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

p-level ≤ 0.05

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $f_{cal.}$ tiene un p-level ≤ 0.05

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como p-level de 0.00002794 es <0.05 y <0.01 se considera que la prueba es Altamente Significativa (AS) y se rechaza H_o.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación entre el volumen de ventas y la superficie se piso donde se exhiben los productos.

Nota importante:

Observe que $t^2 = 8.51^2 = 72.4 = F = 72.4$ por eso se dice que son equivalentes.

ISBN 978-607-8285-62-4

Desarrollo del coeficiente muestral de determinación

Solución al inciso i.

El coeficiente de determinación mide la proporción que se explica por la variable independiente en el modelo de regresión y se puede expresar como el cociente de la suma explicada de cuadrados o suma del cuadrado de la regresión ***SCR (Variación explicada)*** entre la suma de cuadrados tota ***SCT(Variación Total).***

$$r^2_{Y:X} = \frac{SCR}{SCT} = 0.90049x\ 100 = 90.05\%$$

Interpretación: El 90.05% de la variación del volumen de ventas (en millones de pesos) se puede explicar por la superficie de piso (en miles de metros cuadrados). Este es un ejemplo donde hay una fuerte relación lineal entre dos variables, dado que el uso de un modelo de regresión ha reducido la variabilidad en la predicción del volumen de ventas en 90%. Solo el 10% de la variabilidad en el volumen de ventas se puede explicar por factores distintos que los explicados por el modelo de regresión lineal simple.

Cálculo del coeficiente de correlación muestral

Solución al inciso j.

Por lo general la fuerza de una relación entre dos variables en una población se mide mediante el **coeficiente de correlación**, cuyos valores oscilan entre -1 para la correlación negativa perfecta hasta +1 para la correlación positiva perfecta. Se puede obtener con facilidad el coeficiente de correlación mediante la fórmula:

$$r_{y.x} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} 0.94894\ x\ 100 = 94.89\%$$

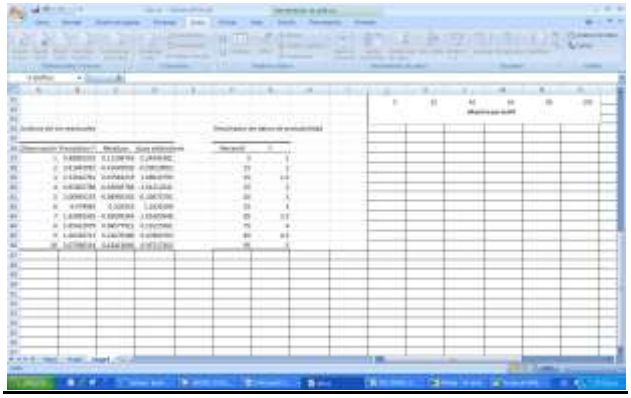
Interpretación: En este problema del volumen de ventas, puesto que $r^2 = 0.90049$ y la pendiente $\hat{\beta}_1$ es positiva, el coeficiente de correlación se interpreta como ***+0.94894***. La cercanía del coeficiente de correlación con +1.0 implica una fuerte asociación entre el volumen de ventas y la superficie de piso en que se exhiben las mercancías.

ISBN 978-607-8285-62-4

Análisis de residuales

Solución al inciso k.

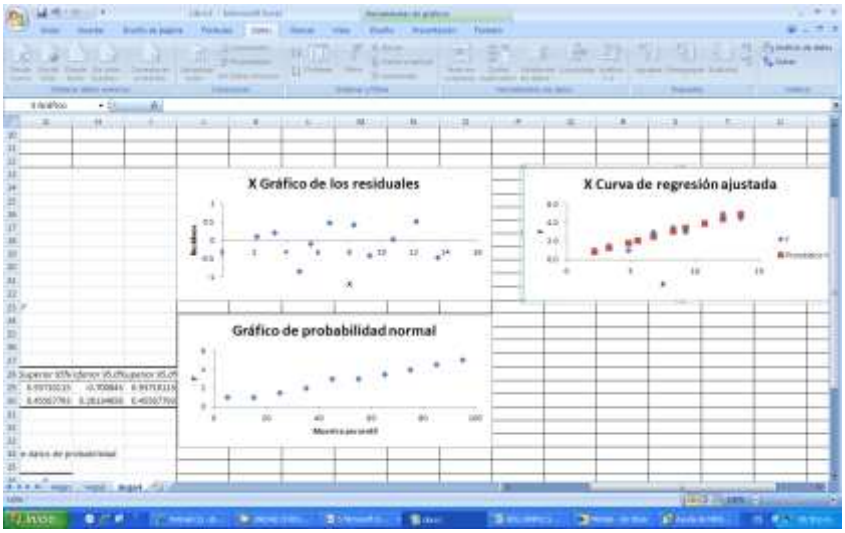
Análisis de residuales



Diagnóstico de la regresión

Solución al inciso l.

La figura siguiente presenta el gráfico de la variable independiente contra los residuales, que sirve para detectar problemas de **linealidad, normalidad, homoscedasticidad y autocorrelación** en el modelo de ajuste. En la parte derecha de la figura se presenta el gráfico de la variable independiente contra los valores predichos, que sirve para detectar problemas de **homoscedasticidad**. **Lo ideal es que todas las gráficas presenten una estructura aleatoria en sus puntos**. En la **parte inferior de la figura** se presenta el gráfico para detectar **hipótesis de normalidad en el modelo**. La gráfica ideal es la diagonal del primer cuadrante.



ISBN 978-607-8285-62-4

Interpretación:

Linealidad

Linealidad

Así, se puede observar que aunque haya una **amplia dispersión en la gráfica residual, no hay patrón** ó relación aparente entre los residuales estandarizados y X_i . Los **residuales parecen estar distribuidos en forma pareja por encima y por debajo de 0** para diferentes valores de ***X***. Por lo tanto se puede concluir que el modelo ajustado parece ser el apropiado.

Homoscedasticidad

Homoscedasticidad

La suposición de **homoscedasticidad** se puede evaluar también de la gráfica de residuales estandarizados con X_i. Si parece haber **un *"efecto de abanico"* en el cual aumenta ó disminuye la variabilidad** de los residuales al aumentar ***X*** se demuestra la falta de homogeneidad en las varianzas de Y_i a cada nivel de ***X.*** Para los datos del volumen de ventas de una tienda **no parece haber diferencias importantes en la variabilidad de SR_i para diferentes valores de X_i.** Por lo tanto se puede concluir que para este modelo ajustado **no hay violación aparente a la suposición de igual varianza en cada nivel de *X*.**

Normalidad

Normalidad

La **gráfica de probabilidad normal** muestra **un patrón aproximadamente lineal que concuerda con una distribución normal**. El **último punto de la esquina inferior izquierda** de la gráfica **puede ser un valor atípico**. El destacado de la gráfica identifica **este punto como 7**, punto que deberá verificarse como observación inusual ó identificación de valores atípicos. Excel no proporciona la prueba de normalidad de Anderson-Darling.

NOTA: Excel en este caso no tiene opción para construir intervalos de confianza para los verdaderos valores de Y, ni realizar un análisis de influencia.

ISBN 978-607-8285-62-4

R.L.S.

EJEMPLO **ILUSTRATIVO EN MINITAB 15**

EJEMPLO ILUSTRATIVO INTEGRAL EN MINITAB 15 REGRESIÓN LINEAL SIMPLE

El director general de una cadena de tiendas de autoservicio en expansión desea conocer el comportamiento de las ventas en los diferentes establecimientos con base en la superficie de piso en la que se exhiben los diferentes productos con el fin de contar con un modelo que le permita llevar un control adecuado de la eficiencia con la que trabaja cada establecimiento. Para ello utiliza el volumen de ventas mensuales (en millones de pesos) y la superficie de piso (en miles de metros cuadrados). En forma aleatoria recopila el volumen de ventas del último mes en diez tiendas de la cadena que correspondan más o menos entre 2,000 y 12,000 metros cuadrados de superficie de piso.

Tienda muestriada	1	2	3	4	5	6	7	8	9	10
Superficie *(X)* en miles de m^2	2.15	9.20	6.70	13.50	5.50	12.15	4.80	10.70	3.25	8.25
Ventas *(Y)* en millones de pesos	1.0	3.0	3.0	4.5	2.0	5.0	1.0	4.0	1.5	3.5

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación
c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0 \; y \; \hat{\beta}_1$:
e) Represente gráficamente los datos *X* y *Y* y la ecuación de predicción.
f) Determine el error estándar de estimación.
g) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba *t.*
h) Determine un intervalo de confianza de 95% para la pendiente para el volumen de ventas.
i) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba *F.*
j) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del volumen de ventas cuando se tenga una superficie de piso de 10,000 metros cuadrados o sea $X_0 = 10$.
k) Determine e interprete el coeficiente de determinación y de correlación.
l) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
m) Determine lo adecuado del ajuste del modelo.

ISBN 978-607-8285-62-4

n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

Cálculo de la covarianza

Solución al inciso a.

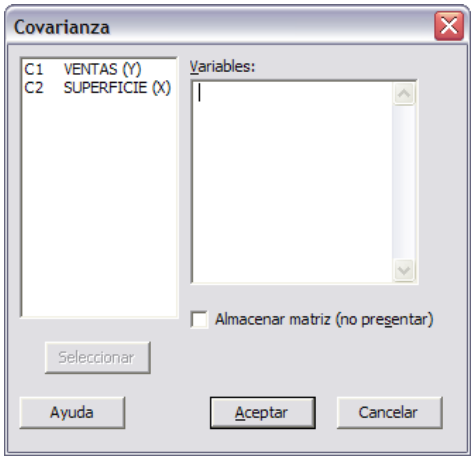
Cuando el número de observaciones en cada variable es extenso, los cálculos manuales son tediosos. Existen muchos paquetes de software que pueden mostrar los resultados entre ellos **Minitab (Versión 15)**

Comenzamos introduciendo los datos en la hoja de Trabajo 1 de Minitab, tal y como se muestra a continuación:

Hoja de trabajo.

Como tenemos un ***modelo de regresión y correlación lineal simple*** seleccionamos la opción ***Estadísticas básicas y Covarianza*** del menú ***Estadísticas,***

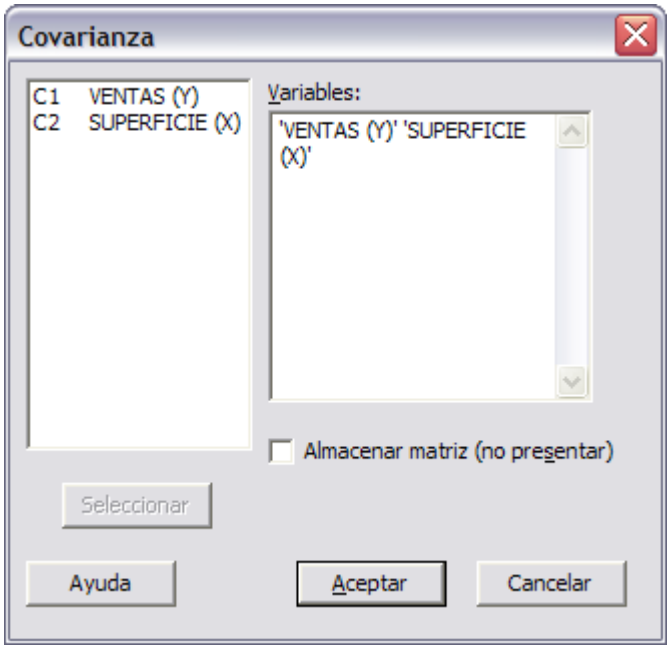
Cuadro de diálogo: Covarianza.



ISBN 978-607-8285-62-4

En ***Variables***, ingrese C1 VENTAS (Y) y C2 SUPERFICIE (X)

Cuadro de diálogo: Covarianza.



Haga clic en ***Aceptar*** en el cuadro de dialogo.

Salida del programa Minitab 15

Salida de la ventana Sesión

Covarianzas: VENTAS (Y), SUPERFICIE (X)

	VENTAS (Y)	SUPERFICIE (X)
VENTAS (Y)	2.05833	
SUPERFICIE (X)	5.17000	14.42067

$$cov(X,Y) = \frac{\sum[(X-\bar{X})(Y-\bar{Y})]}{n-1} = 5.17$$

Interpretación. A la vista de los resultados, podemos decir que todas las covarianzas son positivas.

Cálculo del coeficiente de correlación muestral

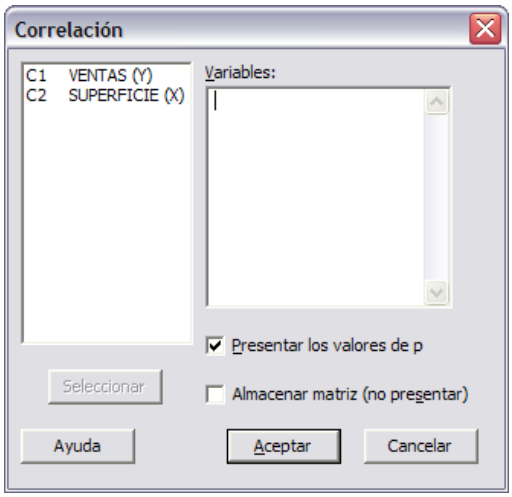
Solución al inciso b.

Comenzamos introduciendo nuevamente los datos en la hoja de Trabajo 1 de Minitab o bien borramos la información de la salida de la ventana Sesión.

Como tenemos un ***modelo de regresión y correlación lineal simple*** seleccionamos la opción ***Estadísticas básicas y Correlación*** del menú ***Estadísticas***,

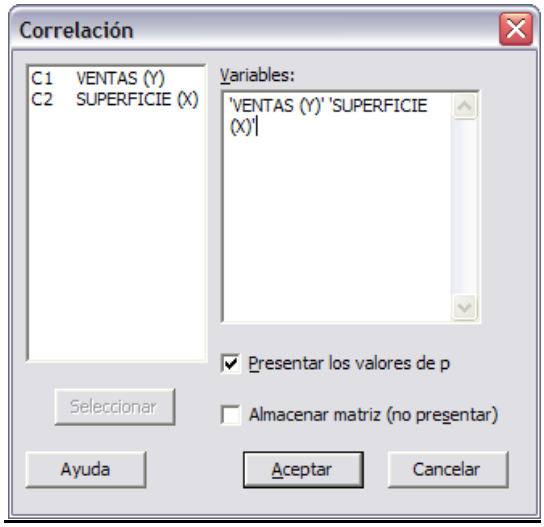
ISBN 978-607-8285-62-4

Cuadro de diálogo: Correlación.



En ***Variables***, ingrese C1 VENTAS (Y) y C2 SUPERFICIE (X)

Cuadro de diálogo: Correlación.



Haga clic en ***Aceptar*** en el cuadro de dialogo.

Salida del programa Minitab 15

Salida de la ventana Sesión

Correlaciones: VENTAS (Y), SUPERFICIE (X)

```
Correlación de Pearson de VENTAS (Y) y SUPERFICIE (X) = 0.949
Valor P = 0.000
```

$$r = \frac{cov(X,Y)}{S_X S_Y} = \frac{5.17}{(3.797)(1.435)} = 0.949$$

ISBN 978-607-8285-62-4

Interpretación. En la salida anterior, se observa que el coeficiente de correlación entre las variables ***X*** e ***Y*** es 0.94894, muy cercano a 1 y como las covarianzas resultaron positivas, esto indica una fuerte dependencia lineal positiva entre el par de variables, lo cual se corrobora con el nivel *p* el cual es mucho menor que 0.01.

Uso de la ecuación de regresión lineal múltiple para hacer la estimación

Solución al inciso c.

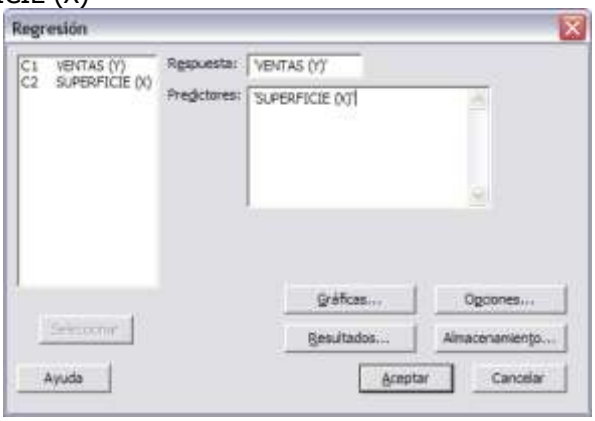
Comenzamos introduciendo nuevamente los datos en la hoja de Trabajo 1 de Minitab o bien borramos la información de la salida de la ventana Sesión.

Como tenemos un ***modelo de regresión y correlación lineal simple*** seleccionamos la opción ***Regresión y Regresión*** del menú ***Estadísticas***,

Cuadro de diálogo regresión.

En ***Respuesta,*** ingrese C1 VENTAS (Y). En ***Predictores***, ingrese C2 SUPERFICIE (X)

Cuadro de diálogo regresión.



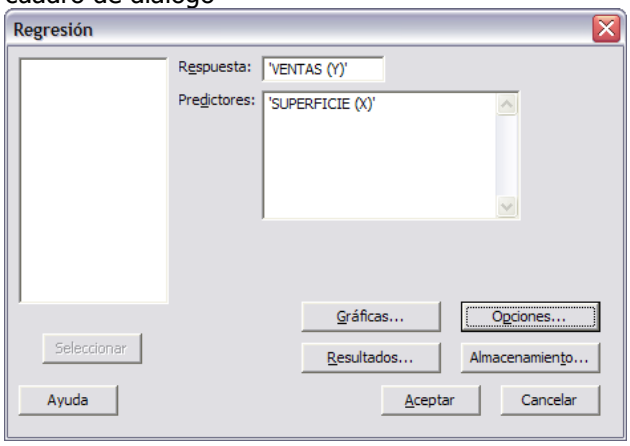
Haga clic en el botón ***Opciones***. Active la casilla que dice **Estadístico de Durbin-Watson.** En la opción ***Intervalos de confianza para nuevas observaciones*** escriba 10.

ISBN 978-607-8285-62-4

Cuadro de diálogo: Regresión-Opciones.

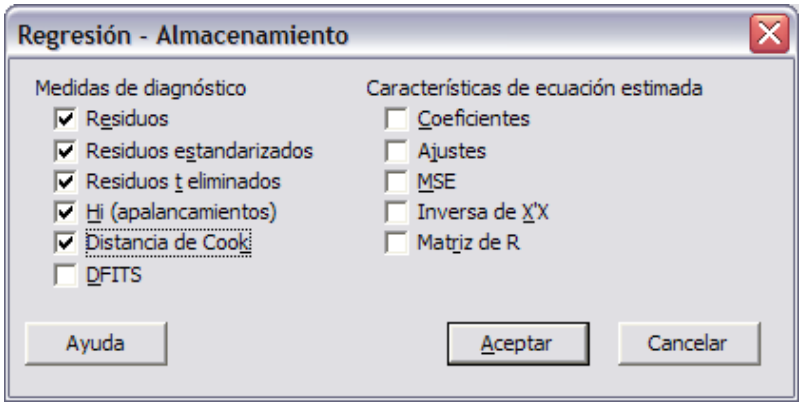
Haga clic en ***Aceptar*** en el cuadro de dialogo para que lo regrese al primer cuadro de dialogo

Cuadro de diálogo: Regresión.



Haga clic en el botón ***Almacenamiento***. Active las casillas que dicen ***Residuos, Residuos estandarizados, Residuos t eliminados, H_i (apalancamientos) y Distancia de Cook***

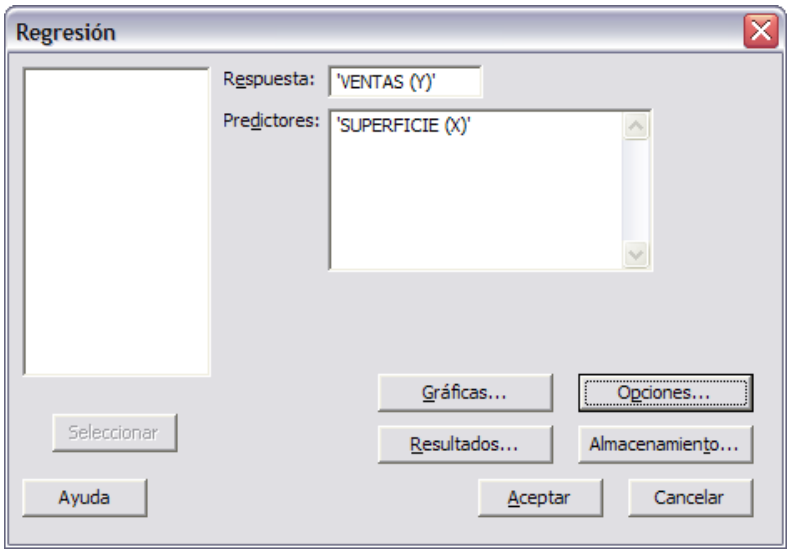
Cuadro de diálogo: Regresión-Almacenamiento.



Haga clic en ***Aceptar*** en el cuadro de dialogo para que lo regrese al primer cuadro de dialogo

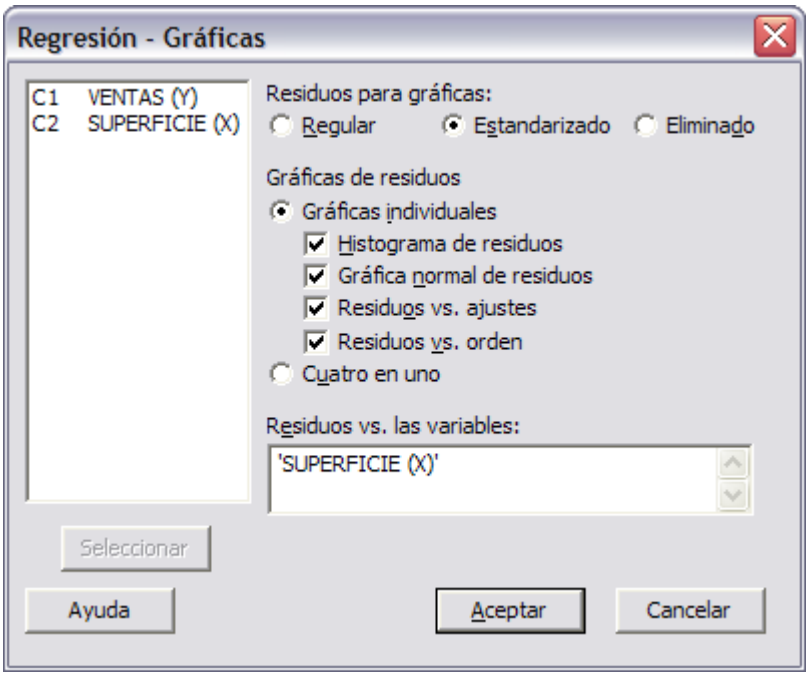
ISBN 978-607-8285-62-4

Cuadro de diálogo: Regresión.



Haga clic en el botón ***Gráficas.*** Active la casilla ***Estandarizado*** en la opción ***Residuos para gráfica.*** En la opción ***Gráficas de residuos*** active las casillas ***Histograma de residuos, Gráfica normal de residuos, Residuos vs. ajustes y Residuos vs. orden***. En la opción ***Residuos vs. las variables*** coloca el curso para que se active el cuadro izquierdo y selecciona 'SUPERFICIE (X)'

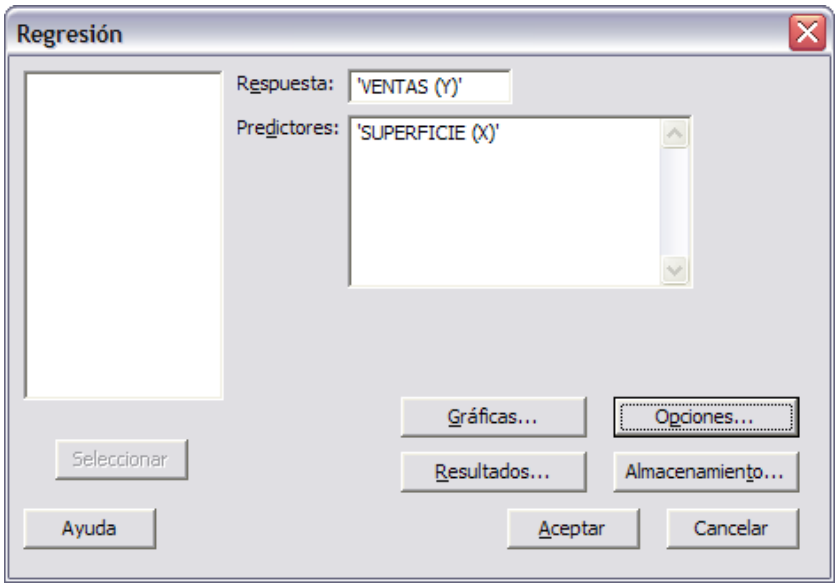
Cuadro de diálogo: Regresión-Gráficas.



Haga clic en ***Aceptar*** en el cuadro de dialogo para que lo regrese al primer cuadro de dialogo

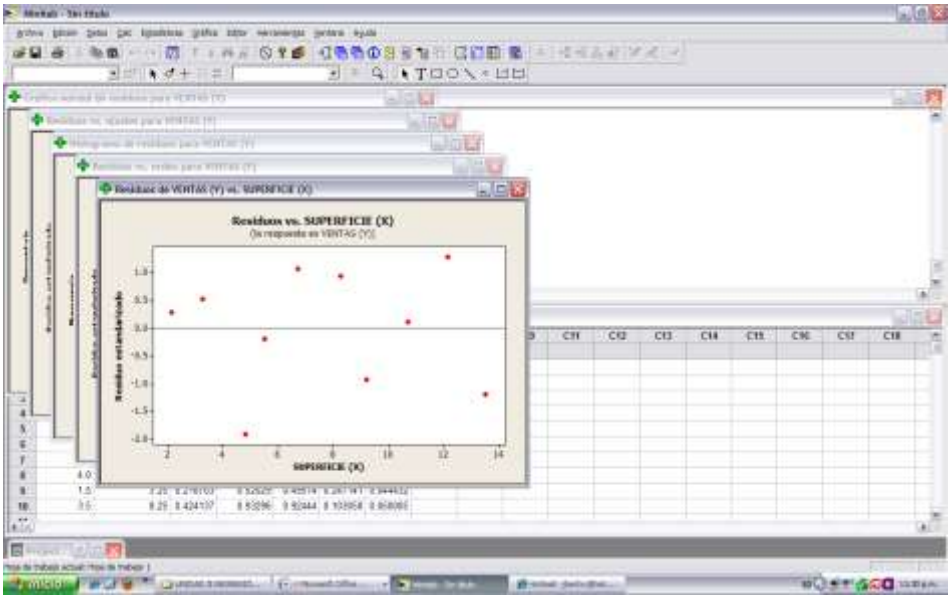
ISBN 978-607-8285-62-4

Cuadro de diálogo: Regresión.



Haga clic en ***Aceptar*** en este cuadro de dialogo.

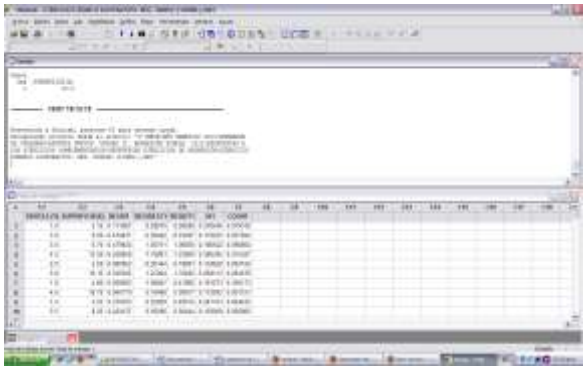
Salida de resultados y gráficas de Minitab 15



Para probar la hipótesis de normalidad debemos realizar la prueba de Anderson-Darling una vez calculamos los residuales estandarizados de la siguiente salida

ISBN 978-607-8285-62-4

Hoja de resultados de Minitab 15.

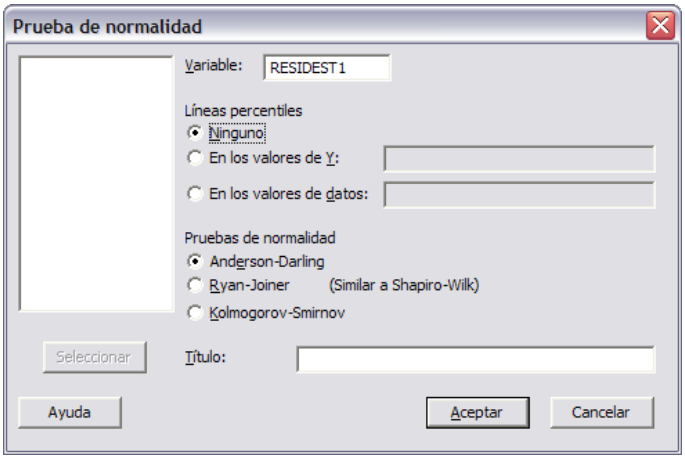


Seleccionamos la opción ***Estadísticas básicas y pruebas de normalidad*** del menú ***Estadísticas*** presentando el siguiente cuadro de dialogo

Cuadro de diálogo: Prueba de normalidad.

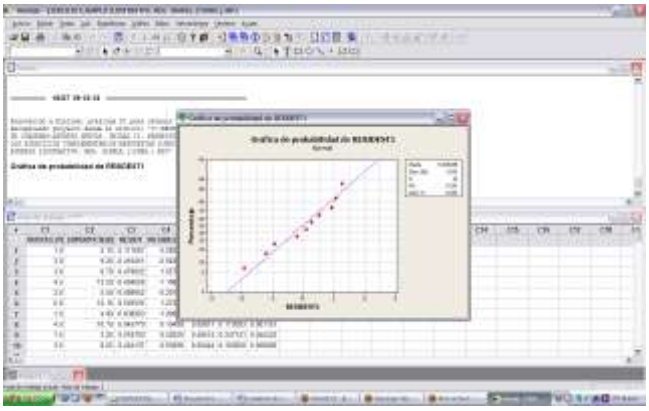
En el campo ***Variable*** selecciona **C4 RESISEST1.** En el campo ***Prueba de normalidad*** ya estará activada la opción **Anderson-Darling**

Cuadro de diálogo: Prueba de normalidad.



Haga clic en **Aceptar** en este cuadro de dialogo para que nos proporcione la siguiente gráfica

ISBN 978-607-8285-62-4



Salida del programa Minitab 15

Salida de la ventana Sesión de Minitab

Análisis de regresión: VENTAS (Y) vs. SUPERFICIE (X)

```
La ecuación de regresión es
VENTAS (Y) = 0.118 + 0.359 SUPERFICIE (X)


Predictor          Coef  Coef. de EE     T      P    VIF
Constante        0.1181       0.3551  0.33  0.748
SUPERFICIE (X)  0.35851      0.04214  8.51  0.000  1.000


S = 0.480023   R-cuad. = 90.0%   R-cuad.(ajustado) = 88.8%


Análisis de varianza

Fuente          GL      SC      MC      F      P
Regresión        1  16.682  16.682  72.40  0.000
Error residual   8   1.843   0.230
Total            9  18.525


Valores pronosticados para nuevas observaciones

Nueva
  Obs  Ajuste  Ajuste SE     IC de 95%       PI de 95%
    1   3.703      0.182  (3.284, 4.123)  (2.519, 4.887)


Valores de predictores para nuevas observaciones

Nueva  SUPERFICIE
  Obs         (X)
    1        10.0
```


ISBN 978-607-8285-62-4

El modelo ajustado se puede expresar como:

$$\boxed{\widehat{Y}_i = 0.11813 + 0.35851X_i}$$

Salida de la ventana Sesión de Minitab

Uso de la ecuación de regresión lineal múltiple para hacer la estimación

```
La ecuación de regresión es
VENTAS (Y) = 0.118 + 0.359 SUPERFICIE (X)
```

Solución al inciso d.

Interpretación de la ecuación

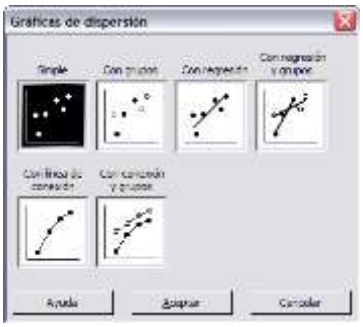
Con este modelo se podría llegar a la conclusión de que por **cada mil metros cuadrado que se incrementan a la superficie de piso** donde se exhibe la mercancía de la tienda ***(X)***, **el volumen de ventas (*Y*) se incrementa en 0.358513 millones de pesos** ó por cada metro cuadrado que se incrementa a la superficie de piso donde se exhibe la mercancía de la tienda, el volumen de ventas se incrementa (la pendiente es positiva) en **$ 358.51 pesos**. Esta pendiente también se puede contemplar como representante del volumen de ventas ***(Y)*** que varía de acuerdo a la superficie de piso de la tienda (***X)***. **La ordenada al origen** $\widehat{\beta}_o$ se calculó como **0.11813 millones de pesos ó $ 118,130 pesos** . **La ordenada al origen** representa el valor del volumen de ventas ***(Y)*** cuando la superficie de piso de la tienda ***(X)*** es igual a cero (en este problema **0.11813 millones de pesos ó $ 118,130 pesos**). Puesto que el resultado de la variable independiente ***(X)*** raramente puede ser cero, esta ordenada se puede considerar como expresión del volumen de ventas ***(Y)*** que varía con factores ajenos al resultado de la superficie de piso de la tienda ***(X).***

Solución al inciso e.

Comenzamos introduciendo nuevamente los datos en la hoja de Trabajo 1 de Minitab o bien borramos la información de la salida de la ventana Sesión.

Como tenemos un ***modelo de regresión y correlación lineal simple*** seleccionamos la opción **G*ráfica de dispersión*** del menú ***Gráficas***

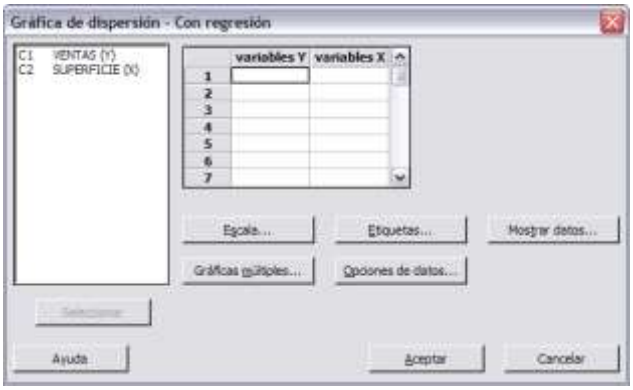
Dibujo, o "ajuste", de una recta a través de un diagrama de dispersión



ISBN 978-607-8285-62-4

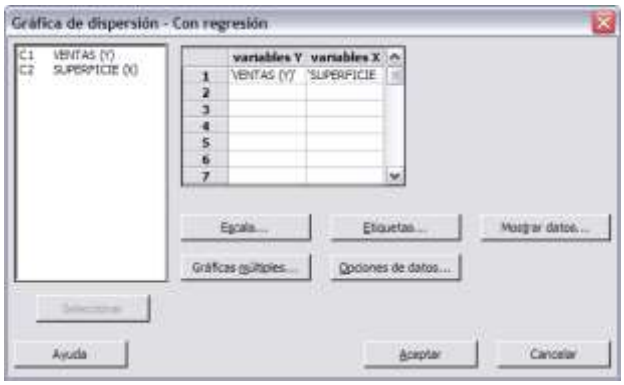
Seleccione el cuadro que dice ***Con regresión*** y haga clic en ***Aceptar.***

Cuadro de diálogo: Gráficas de dispersión.



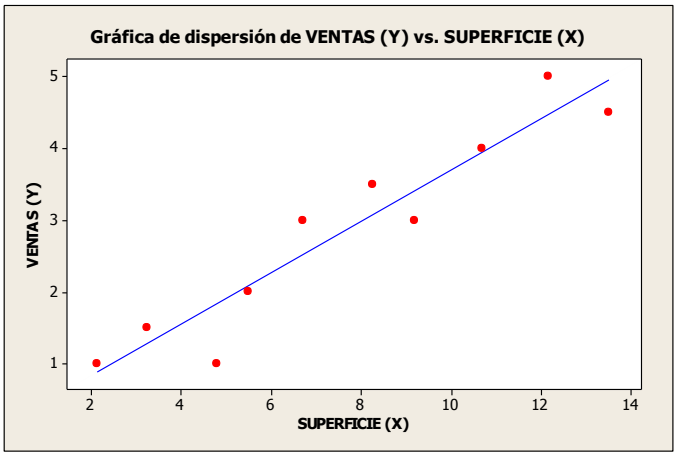
En la tabla donde dice ***variables Y*** seleccione del cuadro izquierdo C1 VENTAS (Y) y donde dice ***variables X*** seleccione del cuadro izquierdo C2 SUPERFICIE (X)

Cuadro de diálogo: Gráficas de dispersión-Con regresión.



Haga clic en ***Aceptar.***

Cuadro de diálogo: Gráficas de dispersión-Con regresión.



Gráficas de dispersió

ISBN 978-607-8285-62-4

Ecuación con que se calcula el error estándar de estimación

Solución al inciso f

El error estándar del estimador, proporcionado por el símbolo $S_{Y.X}$, se define como

$$S_{Y:X} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i}{n-2}} = \mathbf{0.480023}$$

Salida del programa Minitab 15

Salida de la ventana Sesión de Minitab

```
S = 0.480023   R-cuad. = 90.0%   R-cuad.(ajustado) = 88.8%
```

Nota: en la salida de la ventana Sesión el error estándar del estimador tiene la nomenclatura **(S)**

Prueba de una hipótesis con respecto a β_1

Solución al inciso g.

Se puede determinar **si hay relación significativa entre las variables *X* y *Y*** al probar si $\boldsymbol{\beta_1}(\boldsymbol{la\ pendiente\ real})$ **es igual a cero**. Si se rechaza esta hipótesis se concluiría que hay relación lineal. Si no se rechaza, un cambio en *X* no proporciona ningún cambio pronosticado en *Y*, y se sigue que *X* no tiene ningún valor para predecir *Y*.

La **distribución *t*** se puede **utilizar** para realizar **pruebas de significancia** y **construir intervalos de confianza** para la verdadera **pendiente de la regresión**

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0: β_1 =0 (no existe relación)
$H_{1:}$ $\beta_1 \neq 0$ (existe relación)

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$t_{calculada(n-2)} = \frac{\hat{\beta}_1}{S_{\beta_1}} = 8.51$$

Salida del programa Minitab 15

Salida de la ventana Sesión de Minitab

Predictor	Coef	Coef. de EE	T	P	VIF
Constante	0.1181	0.3551	0.33	0.748	
SUPERFICIE (X)	0.35851	0.04214	**8.51**	**0.000**	1.000

ISBN 978-607-8285-62-4

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

$$\text{p-level} \leq 0.05$$

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $t_{cal.}$ tiene un p-level ≤ 0.05

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como p-level de 0.000 es <0.05 y <0.01 se considera que la prueba es Altamente Significativa (AS) y se rechaza H_o.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación lineal significativa entre el volumen de ventas y la superficie de piso de las tiendas.

Obtención de los limites superior e inferior de la región de no rechazo de H_0

<u>Solución al inciso h.</u>

NOTA: Minitab no establece intervalos de confianza para la pendiente de la recta pero se puede establecer fácilmente en forma manual de la siguiente manera:

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de $\hat{\beta}_1$ y determinar si el valor hipotético $(\hat{\beta}_1 = 0)$ está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_1$ se obtendría de la siguiente manera:

$$\boldsymbol{\beta_1 = \hat{\beta}_1 \pm t_{n-2} S_{\hat{\beta}_1}}$$

$$\beta_1 = 0.35851 \pm 2.31(0.04214)$$

$$\boldsymbol{\beta_1} = 0.35851 \pm 0.09732 \begin{cases} \boldsymbol{LIC} = 0.35851 - 0.09734 = \mathbf{0.26117} \\ \boldsymbol{LSC} = 0.35851 + 0.09734 = \mathbf{0.45585} \end{cases}$$

Salida de la ventana Sesión de Minitab

Salida del programa Minitab 15

```
La ecuación de regresión es
Y = 0.118 + 0.359 X


Predictor     Coef  Coef. de EE     T      P
Constante   0.1181       0.3551  0.33  0.748
X          0.35851      0.04214  8.51  0.000
```


ISBN 978-607-8285-62-4

Otra forma de expresar el intervalo de confianza es:

$$\boldsymbol{\widehat{\beta}_1 - t_{n-2} S_{\widehat{\beta}_1} \leq \beta_1 \leq \widehat{\beta}_{1+} t_{n-2} S_{\widehat{\beta}_1}}$$

$$\boldsymbol{0.26117 \leq \beta_1 \leq 0.45585}$$

Interpretación: Se estima, con una confianza del 95%, que la pendiente real se encuentra entre 0.26119 y 0.45583. Puesto que estos valores son superiores a cero se puede concluir que hay relación lineal significativa entre el volumen de ventas y la superficie de piso de las tiendas.

Prueba *F* de la regresión como un todo.

Solución al inciso i.

Hay una prueba alternativa, una prueba ***F*** para la hipótesis nula de un valor predictivo nulo. Esta prueba proporciona el mismo resultado que una prueba ***t*** bilateral de $H_0: \beta_1 = 0$ en la regresión lineal simple. A continuación se presenta un resumen de la prueba ***F***:

Paso 1. Juego de hipótesis.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0: β_1 =0 (no existe relación)
$H_{1:}$ $\beta_1 \neq 0$ (existe relación)

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$F_{calculada} = \frac{CMR}{CME} = \frac{SCR/_{G.L.}}{SCE/_{G.L.}} = 72.40$$

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

```
Análisis de varianza

Fuente          GL      SC      MC      F      P
Regresión        1  16.682  16.682  72.40  0.000
Error residual   8   1.843   0.230
Total            9  18.525
```

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

p-level ≤ 0.05

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $f_{cal.}$ tiene un p-level ≤ 0.05

ISBN 978-607-8285-62-4

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como p-level de 0.000 es <0.05 y <0.01 se considera que la prueba es Altamente Significativa (AS) y se rechaza H_o.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación entre el volumen de ventas y la superficie se piso donde se exhiben los productos.

Nota importante:

Observe que $t^2 = 8.51^2 = 72.4 = F = 72.4$ por eso se dice que son equivalentes.

Intervalo de confianza del 95% para el verdadero valor de Y

<u>Solución al inciso j.</u>

Se puede desarrollar una estimación por intervalo de confianza para hacer inferencia sobre el valor predicho de ***Y,*** la fórmula es:

$$\boldsymbol{\mu_{Y:X} = \widehat{Y}_i \pm t_{\alpha/2.n-2} S_{Y:X}\sqrt{h_i}}$$

Donde

$$\boldsymbol{h_i = \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n} X^2 - n(\overline{X})^2}}$$

Por lo tanto

$$\mu_{Y:X} = \hat{Y}_i \pm t_{0.05.8} S_{Y:X}\sqrt{h_i} = \begin{cases} LIC = \mathbf{3.284} \\ LSC = \mathbf{4.123} \end{cases}$$

Otra forma de expresar el intervalo de confianza es:

$$\mathbf{3.284} \leq \mu_{Y.X} \leq \mathbf{4.123}$$

Salida de la ventana Sesión de Minitab

Salida del programa Minitab 15

```
Valores pronosticados para nuevas observaciones

Nueva
Obs  Ajuste  Ajuste SE     IC de 95%        PI de 95%
  1   3.703      0.182  (3.284, 4.123)  (2.519, 4.887)
```

Interpretación: En **95** de cada **100** muestras (95% de confianza) de tamaño **10**, **el verdadero volumen de ventas promedio mensuales de una tienda** que tiene una **superficie de piso de 10,000 metros cuadrados** oscilará entre **3'284,000 y 4'123,000 millones de pesos.**

ISBN 978-607-8285-62-4

Desarrollo del coeficiente muestral de determinación

Solución al inciso k.

El coeficiente de determinación mide la proporción que se explica por la variable independiente en el modelo de regresión y se puede expresar como el cociente de la suma explicada de cuadrados o suma del cuadrado de la regresión ***SCR (Variación explicada)*** entre la suma de cuadrados tota ***SCT(Variación Total).***

$$r_{Y:X}^2 = \frac{SCR}{SCT} = 0.90049x\ 100 = 90.05\%$$

Salida de la ventana Sesión de Minitab

Salida del programa Minitab 15

```
S = 0.480023   R-cuad. = 90.0%   R-cuad.(ajustado) =
88.8%
```

Interpretación: **El 90.05%** de la **variación del volumen de ventas** (en millones de pesos) se **puede explicar por la superficie de piso (en miles de metros cuadrados)**. Este es un ejemplo donde hay **una fuerte relación lineal entre dos variables**, dado que el uso de un modelo de regresión ha reducido la variabilidad en la predicción del volumen de ventas en 90%. Solo el 10% de la variabilidad en el volumen de ventas se puede explicar por factores distintos que los explicados por el modelo de regresión lineal simple.

NOTA: Minitab no establece el coeficiente de correlación pero puede ser fácilmente calculado manualmente de la siguiente manera:

Cálculo del coeficiente de correlación muestral

Por lo general la fuerza de una relación entre dos variables en una población se mide mediante el **coeficiente de correlación**, cuyos valores oscilan entre -1 para la correlación negativa perfecta hasta +1 para la correlación positiva perfecta. Se puede obtener con facilidad el coeficiente de correlación mediante la fórmula:

$$\boldsymbol{r_{Y.X} = \sqrt{r_{Y.X}^2} = \sqrt{0.90} = 0.94868\ x\ 100 = 94.87\%}$$

Interpretación de *r*

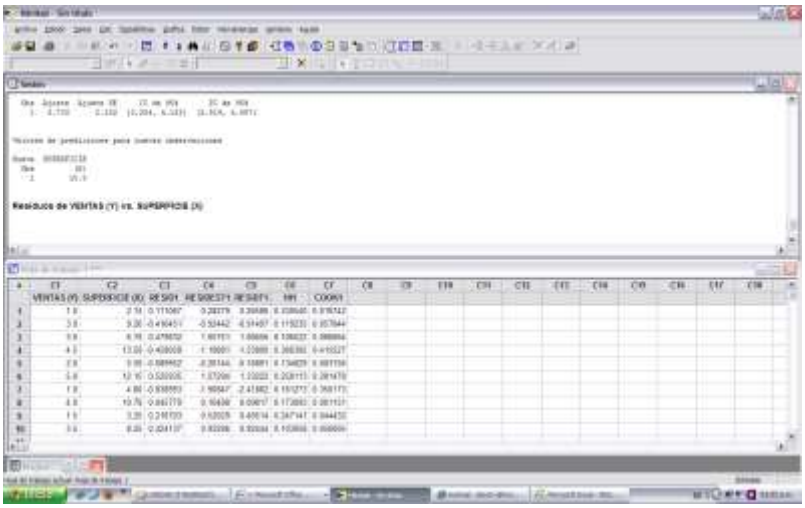
Interpretación: En este problema del volumen de ventas, puesto que $r^2 = 0.90$ y la pendiente $\hat{\beta}_1$ es positiva, el coeficiente de correlación se interpreta como ***+0.94868***. La cercanía del coeficiente de correlación con +1.0 implica una fuerte asociación entre el volumen de ventas y la superficie de piso en que se exhiben las mercancías.

ISBN 978-607-8285-62-4

Solución al inciso l.

Análisis de residuales

Cálculo de residuales:



Interpretación: Los residuales normales aparecen en la columna **RESID1** y los **residuales estandarizados** aparecen en la columna **RESIDEST1.**

Salida del programa Minitab 15

Salida de la ventana Sesión de Minitab

```
Estadístico de Durbin-Watson = 2.81773
```

El estadístico de Durbin-Watson determina si la correlación entre los términos de error adyacentes es cero

Interpretación: El estadístico de Durbin Watson es ***d*= 2.81773.** Este valor es mayor a 1.5 por lo que se puede pensar en que la autocorrelación no sea un problema.

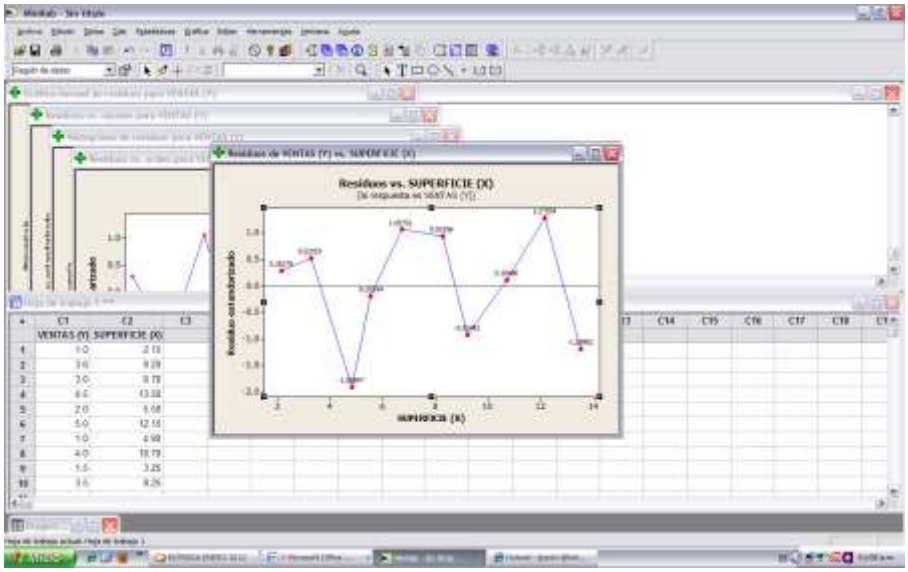
Solución al inciso m.

Diagnóstico de la regresión

Supuestos de linealidad y homoscedasticidad;

La figura siguiente presenta el **gráfico de la variable independiente contra los residuales**, que sirve para detectar problemas de **linealidad y homoscedasticidad** en el modelo de ajuste. **Lo ideal es que todas las gráficas presenten una estructura aleatoria en sus puntos**.

ISBN 978-607-8285-62-4



Interpretación:

Linealidad

Linealidad

Así, se puede observar que aunque haya una **amplia dispersión en la gráfica residual, no hay patrón ó relación aparente entre los residuales estandarizados y** X_i . Los **residuales** parecen estar **distribuidos en forma pareja por encima y por debajo de 0 para diferentes valores de *X***. Por lo tanto se puede concluir que el modelo ajustado **parece ser el apropiado**.

Homoscedasticidad

Homoscedasticidad

La suposición de **homoscedasticidad** se puede evaluar también de la gráfica de residuales estandarizados con X_i. Si parece haber un ***"efecto de abanico"*** en el cual aumenta ó disminuye la variabilidad de los residuales al aumentar ***X*** se demuestra la falta de homogeneidad en las varianzas de Y_i a cada nivel de ***X.*** **Para los datos del volumen de ventas de una tienda no parece haber diferencias importantes en la variabilidad de** SR_i para diferentes valores de X_i. Por lo tanto **se puede concluir** que para este modelo ajustado **no hay violación aparente a la suposición de igual varianza en cada nivel de *X*.**

Normalidad

Supuesto de Normalidad:

El **supuesto de normalidad** en la regresión es posible evaluarlo de un análisis residual colocando los **residuales estandarizados** en una **distribución de frecuencias** y mostrando los resultados en un **histograma**.

ISBN 978-607-8285-62-4

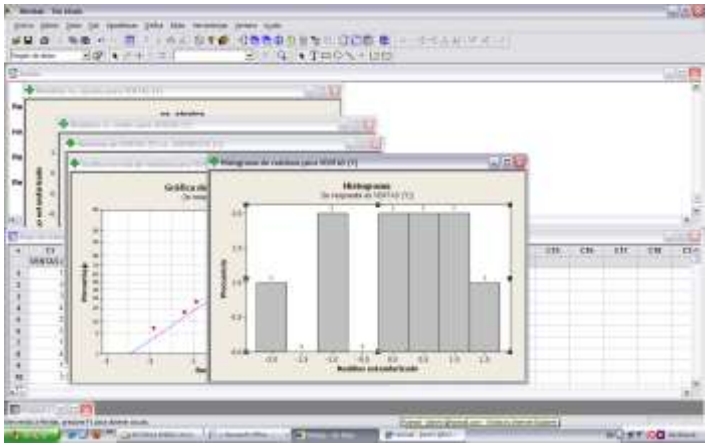
El histograma de residuos es una herramienta exploratoria que muestra las características generales de los datos, incluyendo:

- Valores típicos, dispersión o variación y forma
- Valores inusuales en los datos

La presencia de largas colas en la gráfica podrían indicar sesgo en los datos. Si una o dos barras están lejos de las demás, esos puntos pueden ser valores atípicos. Debido a que el aspecto del histograma cambia según el número de intervalos utilizados para agrupar los datos, utilice la gráfica de probabilidad normal y las pruebas de bondad de ajuste para evaluar la normalidad de los residuos.

Un valor atípico es un valor inusualmente grande o pequeño. Los valores atípicos pueden tener una influencia desproporcionada sobre los resultados estadísticos, como la media, lo que puede generar interpretaciones engañosas. Por ejemplo, un conjunto de datos incluye los valores: 1, 2, 3 y 34. El valor medio, 10, que es mayor que la mayoría de los datos (1, 2, 3), es influenciado considerablemente por el punto de dato extremo, 34. En este caso, el valor medio da la impresión de que los valores de los datos son superiores de lo que realmente son. Es necesario investigar los valores atípicos, porque pueden proporcionar información útil sobre sus datos o proceso.

La figura siguiente presenta el **Histograma de residuales estandarizados** para detectar hipótesis de normalidad en el modelo.

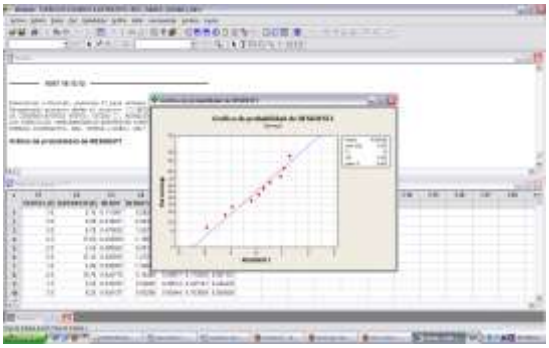


Interpretación:

Es difícil evaluar la **suposición de normalidad** para una muestra de tan sólo 10 observaciones y los procedimientos de pruebas disponibles quedan fuera del alcance del presente trabajo, sin embargo **se puede observar que los datos aunque no parecen tener una "forma de campana" exacta, la mayor parte de los residuales están ubicados cerca del centro de la distribución** por lo que parece razonable llegar a la conclusión de que **no hay en modo alguno violación a la suposición de normalidad.** El histograma indica que los datos podrían tener valores atípicos, lo cual se muestra mediante dos barras, en el extremo izquierdo de la gráfica.

Otra forma de detectar **hipótesis de normalidad** es mediante la **gráfica de probabilidad normal**. La gráfica ideal es **la diagonal del primer cuadrante.**

La figura siguiente presenta el gráfico para detectar hipótesis de normalidad en el modelo.



ISBN 978-607-8285-62-4

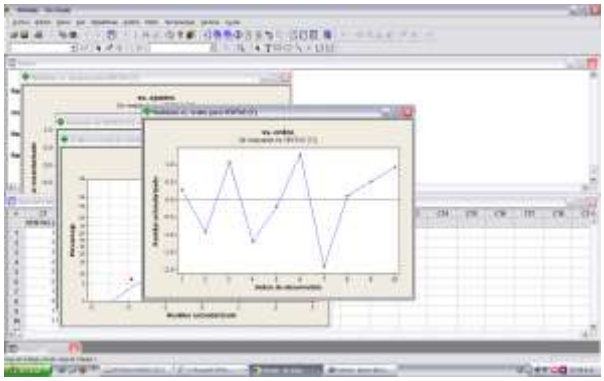
Interpretación:

La **gráfica de probabilidad normal** muestra un patrón **aproximadamente lineal** que concuerda con una distribución normal. **El último punto de la esquina inferior izquierda de la gráfica puede ser un valor atípico**. El destacado de la gráfica identifica este punto **como 7**, punto que deberá verificarse como observación **inusual ó identificación de valores atípicos**. A la derecha de la gráfica se presenta la **prueba de Anderson-Darling** que arroja un **estadístico de prueba de 0.261 con un nivel *p* de 0.625** que al ser mayor a **0.05** nos hace **no rechazar la hipótesis nula** concluyendo que la distribución de **los residuales estandarizados es normal.**

Independencia

Supuesto de Independencia

La suposición de **independencia** requiere que el **error** (diferencia "residual" entre un valor observado y uno predicho de Y) **sea independiente para cada valor de X**. Con frecuencia esta suposición se refiere a datos que se recopilan a lo largo de un periodo. Estos tipos de modelos caen bajo la denominación general de series de tiempo. La **suposición de independencia** se puede evaluar trazando **los residuales en el orden o la sucesión en que se obtuvieron los datos observados**. **Lo ideal es que la gráfica presente una estructura aleatoria en sus puntos.**



El estadístico de Durbin-Watson está condicionado según el orden de las observaciones (filas). Minitab parte del supuesto de que las observaciones están ordenadas significativamente como, por ejemplo, un orden en el tiempo. El estadístico de Durbin-Watson determina si la correlación entre los términos de error adyacentes es cero

Interpretación:

La **gráfica de residuos versus orden** no muestra **un efecto de "autocorrelación" entre observaciones sucesivas**, es decir **no hay correlación entre una observación en particular y aquellos valores que la precedieron y la siguieron no afectando la suposición de independencia**. Además el **estadístico de Durbin-Watson** es ***d*= 2.81773.** Este valor es **mayor a 1.5** por lo que **no se puede pensar en que la autocorrelación sea un problema.**

ISBN 978-607-8285-62-4

Elementos de la matriz sombrero h_i

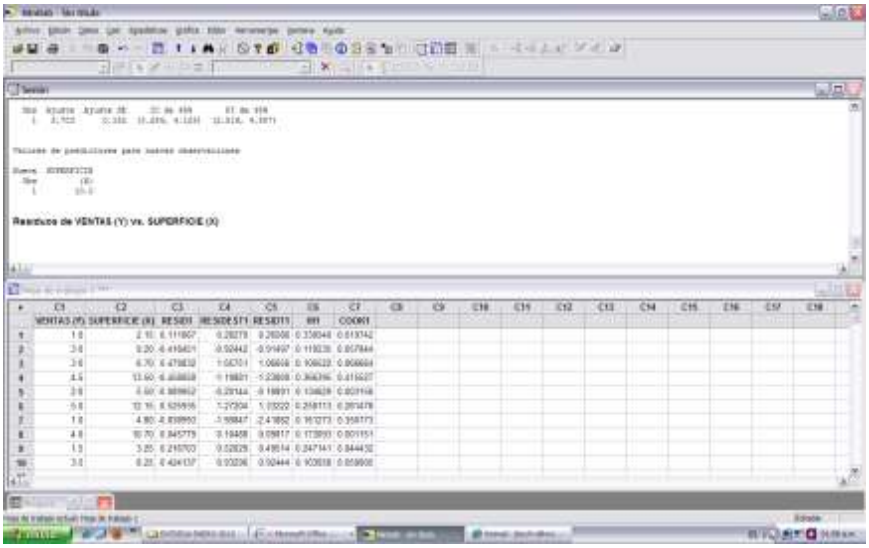
En modelos de regresión, h_i (apalancamiento) mide la distancia de un valor x de observación hasta el promedio de los valores x para todas las observaciones en un conjunto de datos. Las observaciones con valores de apalancamiento grandes pudieran ejercer una influencia desproporcionada sobre el modelo y producir resultados desviados. Por ejemplo, un coeficiente significativo pudiera parecer no significativo. Sin embargo, no todos los puntos con apalancamiento son observaciones influyentes.

Solución al inciso n.

Cada $\boldsymbol{h_i}$ refleja el **apalancamiento o la "influencia" de cada $\boldsymbol{X_i}$** sobre el modelo de regresión ajustado. **Si existen esos puntos de influencia** quizá sea **necesario evaluar de nuevo la necesidad de mantenerlos en el modelo**. En la regresión lineal simple Hoaglin y Welsch sugieren la siguiente regla de decisión: **Si $\boldsymbol{h_i > 4/n}$ entonces $\boldsymbol{X_i}$ es un punto de influencia y se puede considerar candidato a ser eliminado del modelo.**

Cuando se desarrollo una estimación **por intervalo de confianza $\boldsymbol{\mu_{Y.X}}$** , se definieron los **"elementos diagonales de la matriz sombrero" $\boldsymbol{h_i}$** como

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2}$$



Interpretación: (Ver la columna HI1). Para los datos del volumen de ventas, puesto que n=10, los criterios deben ser "destacar" cualquier valor h_i superior a 4/10=0.40. Consultando la tabla anterior se puede observar que ninguna observación es candidata potencial para ser removida del modelo del tiempo de entrega.

ISBN 978-607-8285-62-4

Residuales de Student eliminados, t_i^*

Los residuales eliminados studentizados son útiles en la detección de valores atípicos. Los residuos eliminados studentizados de una observación se calculan dividiendo un residuo eliminado de la observación entre un estimado de su desviación estándar. Un residuo eliminado t_i^* es la diferencia entre y_i y su valor ajustado en un modelo que omite la observación i^ésima de sus cálculos. La observación se omite para ver cómo se comporta el modelo sin este valor atípico potencial. Si una observación tiene un residuo eliminado studentizado grande (si su valor absoluto es mayor que 2), podría tratarse de un valor atípico en sus datos.

Solución al inciso o.

En el estudio del análisis de residuales se definieron **los residuales estandarizados** en la ecuación como:

$$SR_i = \frac{\varepsilon_i}{S_{Y:X}\sqrt{1-h_i}}$$

Para medir la **repercusión adversa sobre el modelo de cada caso individual**, Hoaglin y Welsch desarrollaron también el **residual de Student eliminado t_i^*** que se presenta en la siguiente ecuación:

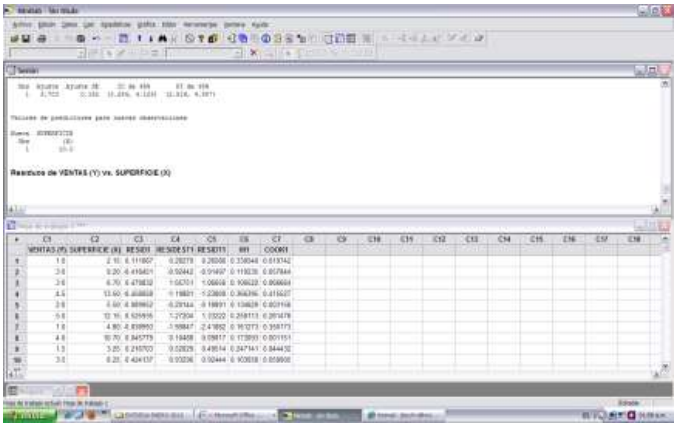
$$t_i^* = \frac{\varepsilon_i}{S_{(i)}\sqrt{1-h_i}}$$

Donde $S_{(i)}$ = **error estándar de la estimación** para un modelo que incluye **todas las observaciones excepto la observación i.**

En regresión lineal simple Hoaglin y Welsch sugieren que si

$$|t_i^*| > t_{0.10.n-3}$$

los valores Y observados y predichos son tan diferentes **que X_i es un punto de influencia que afecta de modo adverso el modelo y se puede considerar como un candidato para ser eliminado.**



Interpretación: (Ver la columna RESIDT1). Para los datos del volumen de ventas, puesto que n=10, **los criterios deben ser "destacar" cualquier valor superior a** $|t_i^*| > t_{0.10.7} = 1.89458$. Consultando la tabla anterior se puede visualizar que $t_7^* = -2.41882$. Por lo tanto la **séptima tienda puede tener un efecto adverso sobre el modelo** y se puede considerar candidato a ser retirado del modelo, **sin embargo como de acuerdo al criterio h_i la tienda 7 no presento un efecto adverso**, se debe tomar en cuenta otro criterio antes de tomar esa decisión como el criterio D_i de Cook, que se basa tanto en h_i como en el estadístico residual estandarizado t_i^*.

ISBN 978-607-8285-62-4

Estadístico de distancia de Cook, D_i

La distancia de Cook (D_i) es una medidas de la influencia de una observación sobre el conjunto de coeficientes de regresión en un modelo de regresión . Las observaciones influyentes tienen un impacto desproporcionado sobre el modelo y pueden generar resultados engañosos. Por ejemplo, un coeficiente significativo pudiera parecer no significativo. Las observaciones influyentes pueden ser puntos de apalancamiento, valores atípicos o ambos.

Las observaciones influyentes son observaciones que tienen un impacto desproporcionado en un modelo de regresión. Las observaciones influyentes, también conocidas como observaciones poco comunes, son importantes para identificar porque pueden producir resultados engañosos. Por ejemplo, un coeficiente significativo pudiera parecer no significativo.

Las observaciones influyentes pueden ser:

- Puntos de apalancamiento que se encuentran en el extremo de la dirección x
- Valores atípicos, que se encuentran en el extremo de la dirección y con respecto a la línea de regresión ajustada
- Puntos de apalancamiento y valores atípicos

<u>Solución al inciso p.</u>

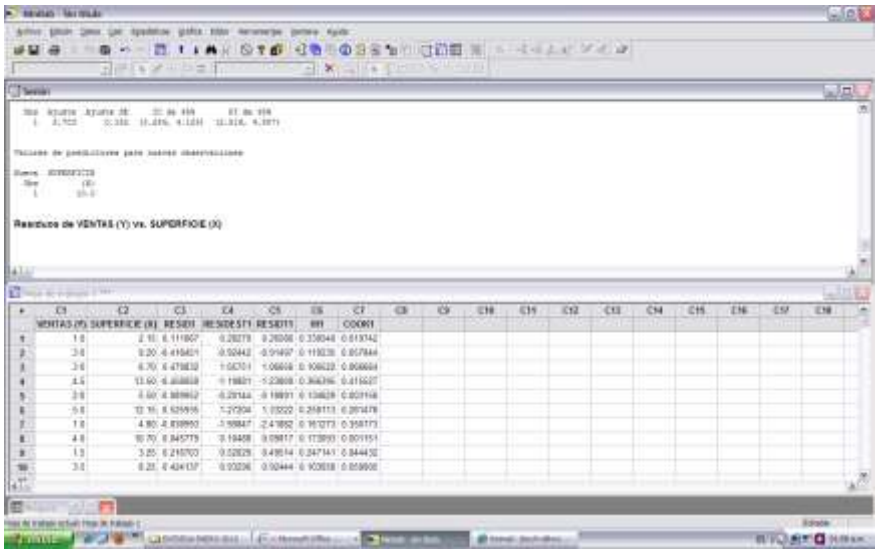
Para **decidir** si un punto **que ha sido destacado mediante h_i ó t_i está afectando indebidamente al modelo**, Cook y Weisberg sugieren **el uso del estadístico D_i**. En el modelo de regresión lineal simple se muestra $D_i$en la ecuación:

$$D_i = \frac{SR_i^2 h_i}{2(1-h_i)}$$

En la regresión lineal simple Cook y Weisberg sugieren que si

$$D_i > F_{0.50.2.n-2}$$

Lo que significaría que posiblemente la observación tenga una repercusión sobre los resultados del ajuste del modelo de regresión lineal simple.



Interpretación: (Ver la columna COOK1) Para los datos del volumen de ventas (en millones de pesos), puesto que n=10, **el criterio sería "destacar" cualquier $D_i > F_{0.50.2.8} = 0.756828$** . Consultando la tabla anterior se puede observar que **ninguna observación es candidata potencial para ser removida del modelo del volumen de ventas**. En caso de que **alguna observación una vez estudiados los tres criterios fuera necesario eliminar alguna(s) observación(es) se debería estudiar un modelo alterno en el que se hayan eliminado dichas observaciones que no fue el caso en este modelo.**

ISBN 978-607-8285-62-4

OBJETIVO 2.8 El alumno podrá definir las cuatro componentes de una serie de tiempo, calculará e interpretará los índices estacionales mediante promedios móviles, determinará una ecuación de tendencia lineal con los datos desestacionalizados y la utilizará para desarrollar pronósticos ajustados estacionalmente.

ANTECEDENTES

CONCEPTOS DE:

Diagrama de dispersión, Variable dependiente, variable independiente, principio de mínimos cuadrados, forma general de la ecuación de regresión lineal simple, pendiente de la línea de regresión, punto donde se intercepta la línea de regresión con el eje Y.

2.8.1 UTILIZACIÓN DE DATOS DESESTACIONALIZADOS PARA PRONÓSTICOS

CONCEPTOS BÁSICOS SERIES DE TIEMPO

Podemos definir a una **serie de tiempo** como un conjunto de valores numéricos obtenidos en periodos iguales en el tiempo; diarios, mensuales, trimestrales, anuales, etc.

La suposición fundamental del **análisis de series de tiempo** es que los factores que han influido en los **patrones de actividad en el pasado y el presente** tendrán más o menos **la misma influencia en el futuro**. Entonces la meta del análisis de series de tiempo es: **identificar y aislar los factores de influencia con el fin de realizar predicciones (pronósticos) con fines administrativos de planeación y control.**

Se han desarrollado varios **modelos matemáticos** que explican las **fluctuaciones entre los factores** que componen una serie de tiempo. Tal vez el más utilizado y esencial sea el **modelo multiplicativo clásico** para datos registrados cada **año, trimestre ó mes.**

El **modelo multiplicativo clásico de series de tiempo** para datos anuales se puede establecer de la siguiente manera:

ISBN 978-607-8285-62-4

$$Y_i = T_i x C_i x I_i$$

donde en el año *i*

T_i = valor de la componente de tendencia.

C_i = valor del componente cíclico.

I_i = valor del componente irregular

Cuando los datos se obtienen **por trimestre o mes**, una observación Y_i registrada en el periodo *i* puede estar dada por la ecuación:

$$Y_i = T_i x S_i x C_i x I_i$$

donde: T_i, C_i, e I_i son los valores respectivos de las componentes de Tendencia, Cíclico e Irregular en el periodo *i* y ,

S_i = valor del componente estacional en el periodo *i.*

Cuatro clases de variación en una serie de tiempo

Podemos entonces definir a las componentes de la siguiente manera:

Tendencia: Patrón de **movimiento global o persistente a largo plazo** hacia arriba ó hacia abajo por varios años.

Cíclico: **Oscilación o movimiento repetitivo arriba o abajo** en cuatro etapas: ***pico*** (auge ó prosperidad); ***contracción*** (recesión); ***fondo*** (depresión) y ***expansión*** (recuperación).

Estacional: **Fluctuación** más o menos regular que ocurre en **cada periodo de 12 meses de cada año.**

Irregular: **Fluctuación errática o residual** en una serie que está presente después de tomar en cuenta los efectos sistemáticos (de tendencia, estacional y cíclico), **es** de corta duración y sin repetición.

Si existe una **tendencia,** se podrán aplicar varios **métodos de pronóstico** de **series de tiempo** con base en si los datos son manejados en **forma anual ó en forma mensual o trimestral** como el método de razón a promedio móvil.

USOS DEL ÍNDICE ESTACIONAL Ó TEMPORAL PARA REALIZAR PRONÓSTICOS

Los **índices temporales ó estaciónales** se utilizan para **eliminar de una serie temporal los efectos de la estacionalidad**. A este

ISBN 978-607-8285-62-4

proceso se le denomina **desestacionalización ó destemporalización** de una serie de tiempo. Antes que podamos **identificar la componente de tendencia** para llevar a cabo **un pronóstico** de una serie temporal **debemos eliminar la variación estacional.**

Una vez que hemos **eliminado la variación temporal**, podemos **calcular una línea de tendencia desestacionalizada**, que luego podemos **proyectar hacia futuro** de la siguiente manera:

Uso del método de la razón al promedio móvil

Paso 1. Estimar la variable dependiente sin la variación estacional.
Paso 2. Incluir la estacionalidad en la estimación anterior.

MÉTODO DE RAZÓN A PROMEDIO MÓVIL:

Esta técnica proporciona un **índice** que describe el **grado de variación estacional eliminando el efecto de las componentes de tendencia, cíclica e irregula**r de los datos originales **(Y).** Este índice esta basado en una **media 100** con el grado de **estacionalidad** medido por las variaciones con respecto a la base.

El método se basa en seis pasos para calcular el **índice estacional ó temporal**

1) Calcular el total móvil.
2) Calcular el promedio móvil.
3) Centrar el promedio móvil.
4) Calcular el porcentaje del valor real.
5) Calcular la media modificada ó índices estacionales desajustados.
6) Ajuste de la media modificada y obtención del **índice estacional ó temporal ajustado.**

PROMEDIOS MÓVILES:

Los **promedios móviles** para un periodo determinado de **longitud L**, consiste en una **serie de promedios aritméticos** en el tiempo tales que cada uno se calcula a partir de una secuencia de L valores observados. Estos promedios móviles se representan por el símbolo **PM(L).**

El **método de promedios móviles** permite **suavizar** una serie de tiempo aunque el método es un poco subjetivo ya que depende de L, la longitud del periodo seleccionado para calcular los promedios, el cual deberá ser elegido considerando un ciclo en la serie para tratar de eliminar las fluctuaciones cíclicas.

A manera de ejemplo, si se desea calcular **promedios móviles de 4 trimestres** de una serie que contiene **20 trimestres (5 años)**, el primer promedio móvil de cuatro trimestres se calcula con la suma de los valores para los primeros 4 trimestres en la serie dividida entre 4,

ISBN 978-607-8285-62-4

$$\boldsymbol{PM1(4)} = \frac{Y_1 + Y_2 + Y_3 + Y_4}{4}$$

El primer promedio se centra en el valor medio, es decir, entre el **trimestre II y el trimestre III** de esta serie de tiempo.

El segundo promedio móvil de 4 trimestres se calcula con la suma de los trimestres II al V (es decir al trimestre I del siguiente año en la serie), dividida entre 4 y se centra entre el **trimestre III y el IV** de esta serie de tiempo:

$$\boldsymbol{PM2(4)} = \frac{Y_2 + Y_3 + Y_4 + Y_5}{4}$$

y así sucesivamente.

Como se trata de una serie de tiempo trimestral, L, la longitud del periodo elegido para construir los promedios móviles, debe ser un número de trimestres **par,** para un **promedio móvil de 4 trimestres**, no es posible hacer cálculos para los **primeros dos trimestres ó los últimos 2 trimestres de la serie.**

Cálculo de la línea de tendencia del mejor ajuste

AJUSTE DE TENDENCIA Y PRONÓSTICO MEDIANTE MÍNIMOS CUADRADOS:

Cuando **la tendencia** se describe mediante **una línea recta** se puede utilizar la **ecuación general** para estimar una línea recta:

$$\hat{Y}_i = a + bx_i$$

donde:

$\hat{Y}_i$ = Valor estimado de la variable dependiente.
x_i= Variable independiente **codificada** (tiempo en el análisis de tendencia)
a= Intersección con el eje Y.(el valor de Y cuando x=0)
b= Pendiente de la línea de tendencia.

Codificación de la variable tiempo para simplificar los cálculos

$$b = \frac{\sum xY - n\bar{x}\bar{Y}}{\sum x^2 - n\bar{x}^2} \qquad a = \bar{Y} - b\bar{x}$$

CODIFICACIÓN DEL TIEMPO:

En condiciones normales la variable independiente **TIEMPO** la medimos como **semanas, meses, trimestres y años**. Estas medidas las podemos codificar para simplificar los cálculos. Para utilizar este proceso, obtendremos el tiempo media y luego restaremos ese valor a cada tiempo muestra. Por ejemplo si las series se componen de solo

ISBN 978-607-8285-62-4

tres años 1993, 1994 y 1995 podemos transformar los tres valores **en -1, 0, 1** donde el 0 representa la media (1994), -1 representa el primer año (1993- 1994= -1) y 1 represente el último año (1995-1994=1). Si la serie se compone por ejemplo **de 8 trimestres (2 años)** podemos transformar los ocho valores en **1,2,...,8 y así sucesivamente.**

2.8.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 2.8.1.1 SERIES DE TIEMPO

La tabla siguiente muestra las ventas trimestrales de una cadena de autoservicio de 2008 a 2012. Las ventas se presentan en millones de pesos.

AÑO	TRIMESTRE I	TRIMESTRE II	TRIMESTRE III	TRIMESTRE IV
2008	9.38	6.44	14.00	17.78
2009	9.10	6.44	13.72	19.04
2010	9.66	7.00	14.56	19.74
2011	9.94	7.98	15.54	20.30
2012	11.20	8.68	15.96	20.86

a) Determine los índices estacionales y elimine la estacionalidad en esos datos (usando el método de razón a promedio móvil de 4 trimestres) e interprete sus resultados.
b) Calcule la línea de mínimos cuadrados que mejor describa estos datos e interprete sus resultados.
c) Estime las ventas trimestrales para los años 2013,2014,2015,2016 y 2017.
d) Grafique los datos originales, los datos sin la variación estacional, la tendencia y las ventas trimestrales futuras.

<u>Solución al inciso a.</u>

El método de razón a promedio móvil se basa en seis pasos para calcular el **índice temporal**

Eliminación de la estacionalidad en las series de tiempo

1) Calcular el total móvil.
2) Calcular el promedio móvil.
3) Centrar el promedio móvil.
4) Calcular el porcentaje del valor real.
5) Calcular la media modificada ó los índices estacionales desajustados
6) Ajuste de la media modificada y obtención del **índice estacional ó temporal ajustado.**

ISBN 978-607-8285-62-4

Paso1: cálculo del total móvil de 4 trimestres

Paso 1: Calcular el total móvil con base a cuatro trimestres. (Columna 4)

TM1=9.38+6.44+14+17.78=47.6

TM2=6.44+14+17.78+9.1=47.32

⋮

TM17=11.2+8.68+15.96+20.86=56.70

Paso 2: cálculo del promedio móvil de 4 trimestres

Paso 2: Calcular el promedio móvil con base a cuatro trimestres. (Columna 5)

PM1(4)=47.6/4=11.900

PM2(4)=47.32/4=11.830

⋮

PM17(4)=56.70/4=14.175

Paso 3: cómo centrar el promedio móvil de 4 trimestres

Paso 3: Centrar el total móvil con base a cuatro trimestres. (Columna 6)

PMC1=(11.900+11.830)/2=11.8650

PMC2=(11.830+11.830)/2=11.8300

⋮

PMC16=(14.035+14.175)/2=14.1050

Paso 4: calcular el porcentaje del valor actual respecto al valor promedio

Paso 4: Calcular el porcentaje del valor real con base a cuatro trimestres. (Columna 7)

IED1=(14/11.865)X100=117.99410

IED2=(17.78/11.8300)X100=150.29586

⋮

IED16=(8.68/14.1050)X100=61.53846

Con los cálculos anteriores llenamos la siguiente tabla:

AÑO (1)	TRIMESTRE (2)	VENTAS (3)	TOTAL MOVIL (4)	PROMEDIO MOVIL (5)	PROMEDIO MOVIL CENTRADO (6)	PORCENTAJE DEL VALOR REAL RESPECTO AL MOVIL (7)=(3)/(6) x100
2008	I	9.38				
	II	6.44				
	III	14	47.60	11.900	11.8650	117.99410
	IV	17.78	47.32	11.830	11.8300	150.29586
2009	I	9.1	47.32	11.830	11.7950	77.15134
	II	6.44	47.04	11.760	11.9175	54.03818
	III	13.72	48.30	12.075	12.1450	112.96830
	IV	19.04	48.86	12.215	12.2850	154.98575
2010	I	9.66	49.42	12.355	12.4600	77.52809
	II	7	50.26	12.565	12.6525	55.32503
	III	14.56	50.96	12.740	12.7750	113.97260
	IV	19.74	51.24	12.810	12.9325	152.63870
2011	I	9.94	52.22	13.055	13.1775	75.43161
	II	7.98	53.20	13.300	13.3700	59.68586
	III	15.54	53.76	13.440	13.5975	114.28571
	IV	20.3	55.02	13.755	13.8425	146.64981
2012	I	11.2	55.72	13.930	13.9825	80.10013
	II	8.68	56.14	14.035	14.1050	61.53846
	III	15.96	56.70	14.175		
	IV	20.86				

ISBN 978-607-8285-62-4

Paso 5: reunir las respuestas del paso 4 y calcular la media modificada

Paso 5: Calcular la media modificada ó índices estacionales desajustados.

Se trascriben los resultados del paso 4 a la siguiente tabla y de los cuatro índices estacionales desajustados de cada trimestre se elimina el valor más bajo y el valor más alto y se suman.

Reducción de las variaciones cíclicas extremas y las variaciones irregulares

AÑO	TRIMESTRE I	II	III	IV
2008			~~117.99410~~	150.29586
2009	~~77.15134~~	~~54.03818~~	~~112.96830~~	~~154.98575~~
2010	77.52809	55.32503	113.97260	152.63870
2011	75.43161	59.68586	114.28571	~~146.64981~~
2012	~~80.10013~~	~~61.53846~~		
SUMA MODIFICADA	154.67943	115.01090	228.25832	302.93456
MEDIA MODIFICADA*	154.67943/2= 77.33971	115.01090/2= 57.50545	228.25832/2= 114.12916	302.93456/2= 151.46728
SUMA DE LAS MEDIAS MODIFICADAS Ó INDICES TEMPORALES DESAJUSTADOS	77.33971+57.50545+114.12916+151.46728=400.44160			

* Si la serie de tiempo consta de cuatro años, la media modificada no se debe dividir entre 2, queda el mismo valor de la suma modificada ya que al quitar el valor mas bajo y el más alto queda un solo valor.

Paso 6: ajuste de la media modificada

Paso 6: Ajustar la media modificada (índice estacional desajustado) y obtención del índice estacional ó temporal ajustado.

Debido a que la suma de los cuatro índices estacionales debe sumar 400 ya que cada índice está sobre la base de 100%, se debe ajustar la media modificada de 404.06 para que resulte 400. Para ello se obtiene lo que se llama Factor de ajuste de la siguiente manera:

$$Factor\ de\ ajuste = \frac{400}{400.44160} = 0.99890$$

TRIMESTRE	ÍNDICES	X	FACTOR DE AJUSTE	=	INDICES ESTACIONALES
I	77.33971	X	0.99890	=	77.25442
II	57.50545	X	0.99890	=	57.44203
III	114.12916	X	0.99890	=	114.00330
IV	151.46728	X	0.99890	=	151.30024
				SUMA	400.00

ISBN 978-607-8285-62-4

Interpretación: A la vista de los resultados, podemos observar que en el primer trimestre de cada año existe una disminución en las ventas del 22.74558% (100-77.25442), en el segundo trimestre una disminución en las ventas del 42.55797% (100-57.44203), en el tercer trimestre nuevamente una sobre venta del 14.00330% (114.00330-100) y en el cuarto trimestre una sobreventa del 51.30024% (151.30024-100).

Cálculo de la línea de tendencia del mejor ajuste

Solución al inciso b.

Una vez que hemos eliminado la variación temporal, podemos calcular una línea de tendencia desestacionalizada, que luego podemos proyectar hacia futuro.

Codificación de la variable tiempo para simplificar los cálculos

AÑO (1)	TRIME-STRE (2)	VEN-TAS REALES (3)	INDICE ESTA-CIONAL ENTRE 100 (4)	VENTAS SIN VARIACIÓN ESTACIONAL (Y) (5)=(3)/(4)	CODIFI-CACIÓN DE X (6)	xY	X^2
2008	I	9.38	0.77254	12.14	1	12.14	1
	II	6.44	0.57442	11.21	2	22.42	4
	III	14	1.14003	12.28	3	36.84	9
	IV	17.78	1.51300	11.75	4	47.01	16
2009	I	9.1	0.77254	11.78	5	58.90	25
	II	6.44	0.57442	11.21	6	67.27	36
	III	13.72	1.14003	12.03	7	84.24	49
	IV	19.04	1.51300	12.58	8	100.67	64
2010	I	9.66	0.77254	12.50	9	112.54	81
	II	7	0.57442	12.19	10	121.86	100
	III	14.56	1.14003	12.77	11	140.49	121
	IV	19.74	1.51300	13.05	12	156.56	144
2011	I	9.94	0.77254	12.87	13	167.27	169
	II	7.98	0.57442	13.89	14	194.49	196
	III	15.54	1.14003	13.63	15	204.47	225
	IV	20.3	1.51300	13.42	16	214.67	256
2012	I	11.2	0.77254	14.50	17	246.46	289
	II	8.68	0.57442	15.11	18	272.00	324
	III	15.96	0.77254	12.14	19	230.66	361
	IV	20.86	0.57442	11.21	20	224.20	400
			SUMAS	257	210	2802	2870
			MEDIAS	257/20= 12.83527	210/20= 10.50		

Cuando la tendencia se describe mediante una línea recta se puede utilizar la ecuación general para estimar una línea recta:

$$\hat{Y}_i = a + bx_i$$

donde:

$\hat{Y}_i$ = Valor estimado de la variable dependiente.
x_i = Variable independiente codificada (tiempo en el análisis de tendencia)
a= Intersección con el eje Y.(el valor de Y cuando x=0)
b= Pendiente de la línea de tendencia.

ISBN 978-607-8285-62-4

$$b = \frac{\sum xY - n\bar{x}\bar{Y}}{\sum x^2 - n\bar{x}^2} \qquad a = \bar{Y} - b\bar{x}$$

$$b = \frac{2802 - 20(10.5)(12.83527)}{2870 - 20(10.50)^2} \cong 0.16030$$

$$a = 12.83527 - 0.16030(10.5) \cong 11.15176$$

$$\boxed{\boldsymbol{\hat{Y}_i = a + bx_i = 11.15176 + 0.16030x_i}}$$

Interpretación: Cuando un trimestre pudiera valer cero las ventas sería de 11.15176 millones de pesos una vez eliminada la estacionalidad. Asimismo por cada trimestre que se aumente al modelo las ventas se incrementarían en 0.16030 millones de pesos..

Empleo de la estacionalidad en los pronósticos

Solución al inciso c.

Una vez que hemos eliminado la variación temporal y hemos calculado una línea de tendencia desestacionalizada, podemos proyectar hacia futuro de la siguiente manera:

Paso 1. Estimar la ocupación sin la variación estacional.
Paso 2. Incluir la estacionalidad en la estimación anterior.

Paso 1: determinación del valor sin variación estacional

Paso 1.

Se calculará la ocupación sin la variación estacional para el primer trimestre del año 2013 cuyo código es 21.

$$\hat{Y}_{21(\text{sin } variación\ estacional)} = a + bx_i = 11.15176 + 0.16030(21) \cong 2116$$

Paso 2: estacionalización de la estimación inicial

Paso 2

Se incluirá la estacionalidad en la estimación anterior para el primer trimestre del año 2013 que en este caso es 0.903.

$$\hat{Y}_{21(\text{con } variación\ estacional)} = (a + bx_{i)}\ x\ índice\ estacional$$
$$= 2082.6 + 1.59(21) \cong 2116x0.903 \cong 1910.81$$

Se repite el procedimiento para los 19 trimestres restantes y se construye la siguiente tabla:

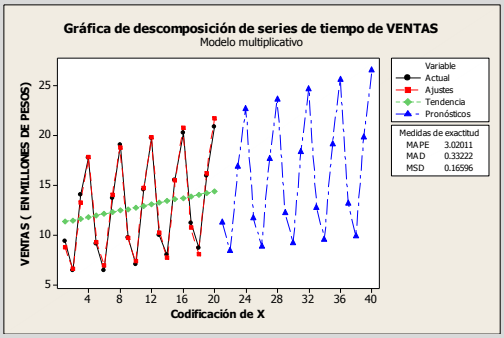
ISBN 978-607-8285-62-4

AÑO (1)	TRIMESTRE (2)	INDICE ESTACIONAL ENTRE 100 (3)	VENTAS SIN VARIACIÓN ESTACIONAL (Y) (4)	CODIFICACIÓN DE X (5)	$\hat{\hat{Y}}_t$ SIN VARIACIÓN ESTACIONAL (6)	PRONÓSTICO (7)=(6)X(3)
2013	I	0.77254	12.14	21	14.52	**11.22**
	II	0.57442	11.21	22	14.68	**8.43**
	III	1.14003	12.28	23	14.84	**16.92**
	IV	1.51300	11.75	24	15.00	**22.69**
2014	I	0.77254	11.78	25	15.16	**11.71**
	I	0.57442	11.21	26	15.32	**8.80**
	III	1.14003	12.03	27	15.48	**17.65**
	IV	1.51300	12.58	28	15.64	**23.67**
2015	I	0.77254	12.50	29	15.80	**12.21**
	II	0.57442	12.19	30	15.96	**9.17**
	III	1.14003	12.77	31	16.12	**18.38**
	IV	1.51300	13.05	32	16.28	**24.64**
2016	I	0.77254	12.87	33	16.44	**12.70**
	II	0.57442	13.89	34	16.60	**9.54**
	III	1.14003	13.63	35	16.76	**19.11**
	IV	1.51300	13.42	36	16.92	**25.61**
2017	I	0.77254	14.50	37	17.08	**13.20**
	II	0.57442	15.11	38	17.24	**9.91**
	III	0.77254	12.14	39	17.40	**19.84**
	IV	0.57442	11.21	40	14.52	**26.58**

INTERPRETACIÓN: El pronóstico de ventas incluyendo la variación estacional para el primer trimestre del año 2013 (Código 21) es de 11.22 millones de pesos, para el segundo trimestre del 2013 (Código 22) de 8.43 millones de pesos y así sucesivamente hasta el cuarto trimestre del 2017 (Código 40) que es de 26.58 millones de pesos.

Trazo de los datos, los datos sin variación estacional , la tendencia y pronósticos a futuro

Solución al inciso d.



INTERPRETACIÓN: La línea continua con círculo negro muestra los datos originales, la línea con un cuadrados los datos sin variación estacional, la línea con un rombo muestra la tendencia y la línea con triángulos muestra los pronósticos a futuro para los años de 2013, 2014, 2015, 2016 y 2017.

ISBN 978-607-8285-62-4

2.8.1.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 2.8.1.1
SERIES DE TIEMPO

La administración de un albergue para esquiadores tiene los siguientes datos acerca de la ocupación trimestral correspondiente a un periodo de cinco años. Desarrolle un modelo para realizar pronósticos a través del análisis de una serie de tiempo

AÑO	TRIMESTRE I	TRIMESTRE II	TRIMESTRE III	TRIMESTRE IV
2008	1861	2203	2415	1908
2009	1921	2343	2514	1986
2010	1834	2154	2098	1799
2011	1837	2025	2304	1965
2012	2073	2414	2339	1967

a) Determine los índices estacionales y elimine la estacionalidad en esos datos (empleando un promedio móvil de 4 trimestres) e interprete sus resultados.

b) Calcule la línea de mínimos cuadrados que mejor describa estos datos e interprete sus resultados.

c) Estime la ocupación trimestral para los años 2013,2014,2015,2016 y 2017.

d) Grafique los datos originales, los datos sin la variación estacional, la tendencia y la ocupación trimestral futura.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ISBN 978-607-8285-62-4

Uso del método de la razón al promedio móvil

Solución al inciso a.

El método se basa en seis pasos para calcular el **índice temporal**

1) Calcular el total móvil.
2) Calcular el promedio móvil.
3) Centrar el promedio móvil.
4) Calcular el porcentaje del valor real.
5) Calcular la media modificada ó índices estacionales desajustados.
6) Ajuste de la media modificada y obtención del **índice estacional ó temporal ajustado.**

Paso1: cálculo del total móvil de 4 trimestres

Paso 1: Calcular el total móvil con base a cuatro trimestres. (Columna 4)

TM1=
TM2=
⋮
TM17=

Paso 2: cálculo del promedio móvil de 4 trimestres

Paso 2: Calcular el promedio móvil con base a cuatro trimestres. (Columna 5)

PM1=
PM2=
⋮
PM17=

Paso 3: cómo centrar el promedio móvil de 4 trimestres

Paso 3: Centrar el total móvil con base a cuatro trimestres. (Columna 6)

PMC1=
PMC2=
⋮
PMC16=

Paso 4: calcular el porcentaje del valor actual respecto al valor promedio

Paso 4: Calcular el porcentaje del valor real con base a cuatro trimestres. (Columna 7)

IED1=
IED2=
⋮
IED16=

Con los cálculos anteriores llenamos la siguiente tabla:

ISBN 978-607-8285-62-4

AÑO (1)	TRIMESTRE (2)	OCUPACIÓN REAL (3)	TOTAL MOVIL (4)	PROMEDIO MOVIL (5)	PROMEDIO MOVIL CENTRADO (6)	PORCENTAJE DEL VALOR REAL RESPECTO AL MOVIL (7)=(3)/(6) x100
2008	I	1861				
	II	2203				
	III	2415				
	IV	1908				
2009	I	1921				
	II	2343				
	III	2514				
	IV	1986				
2010	I	1834				
	II	2154				
	III	2098				
	IV	1799				
2011	I	1837				
	II	2025				
	III	2304				
	IV	1965				
2012	I	2073				
	II	2414				
	III	2339				
	IV	1967				

Paso 5: reunir las respuestas del paso 4 y calcular la media modificada

Paso 5: Calcular la media modificada

Se trascriben los resultados del paso 4 a la siguiente tabla y de los cuatro índices estacionales desajustados de cada trimestre se elimina el valor más bajo y el valor más alto y se suman.

Reducción de las variaciones cíclicas extremas y las variaciones irregulares

AÑO	TRIMESTRE			
	I	II	III	IV
2008				
2009				
2010				
2011				
2012				
SUMA MODIFICADA				
MEDIA MODIFICADA				
SUMA DE LAS MEDIAS MODIFICADAS				

Paso 6: ajuste de la media modificada

Paso 6: Ajustar la media modificada (índices estacionales desajustados) y obtención del índice estacional ó temporal ajustado.

Debido a que la suma de los cuatro índices estacionales debe sumar 400 ya que cada índice está sobre la base de 100%, se debe ajustar la media modificada para que resulte 400. Para ello se obtiene lo que se llama Factor de ajuste de la siguiente manera:

ISBN 978-607-8285-62-4

$$Factor\ de\ ajuste =$$

TRIMESTRE	ÍNDICES	X	FACTOR DE AJUSTE	=	ÍNDICES ESTACIONALES
I		X		=	
II		X		=	
III		X		=	
IV		X		=	
				SUMA	400.00

Interpretación:

Eliminación de la estacionalidad en las series de tiempo

<u>Solución al inciso b.</u>

Una vez que hemos eliminado la variación temporal, podemos calcular una línea de tendencia desestacionalizada, que luego podemos proyectar hacia futuro.

AÑO (1)	TRIMESTRE (2)	OCUPACIÓN REAL (3)	INDICE ESTACIONAL ENTRE 100 (4)	OCUPACIÓN SIN VARIACIÓN ESTACIONAL (Y) (5)=(3)/(4)	CODIFICACIÓN DE X (6)	xY	X^2
2008	I	1861					
	II	2203					
	III	2415					
	IV	1908					
2009	I	1921					
	II	2343					
	III	2514					
	IV	1986					
2010	I	1834					
	II	2154					
	III	2098					
	IV	1799					
2011	I	1837					
	II	2025					
	III	2304					
	IV	1965					
2012	I	2073					
	II	2414					
	III	2339					
	IV	1967					
			SUMAS				
			MEDIAS				

ISBN 978-607-8285-62-4

Cuando la tendencia se describe mediante una línea recta se puede utilizar la ecuación general para estimar una línea recta:

$$\hat{Y}_i = a + bx_i$$

donde:

$\hat{Y}_i$ = Valor estimado de la variable dependiente.
x_i = Variable independiente codificada (tiempo en el análisis de tendencia)
a= Intersección con el eje Y.(el valor de Y cuando x=0)
b= Pendiente de la línea de tendencia.

$$b = \frac{\sum xY - n\bar{x}\bar{Y}}{\sum x^2 - n\bar{x}^2} \qquad a = \bar{Y} - b\bar{x}$$

$b =$ $\qquad$ $a =$

$$\boxed{\boldsymbol{\hat{Y}_i = a + bx_i =}}$$

Interpretación:

Empleo de la estacionalidad en los pronósticos

Solución al inciso c.

Una vez que hemos eliminado la variación temporal y hemos calculado una línea de tendencia desestacionalizada, podemos proyectar hacia futuro de la siguiente manera:

Paso 1. Estimar la ocupación sin la variación estacional.
Paso 2. Incluir la estacionalidad en la estimación anterior.

Paso 1: determinación del valor sin variación estacional

Paso 1.

Se calculará la ocupación sin la variación estacional para el primer trimestre del año 2013 cuyo código es 21.

$\hat{Y}_{21(\text{sin } variación\ estacional)} = a + bx_i =$

Paso 2: estacionalización de la estimación inicial

Paso 2

Se incluirá la estacionalidad en la estimación anterior para el primer trimestre del año 2013 que en este caso es _____

ISBN 978-607-8285-62-4

$$\hat{Y}_{21(\text{con } variación\ estacional)} = (a + bx_{i)}\ x\ iíndice\ estacional =$$

Se repite el procedimiento para los 19 trimestres restantes y se construye la siguiente tabla:

AÑO (1)	TRIMES-TRE (2)	INDICE ESTACIO-NAL ENTRE 100 (3)	OCUPA-CIÓN SIN VARIA-CIÓN ESTACIO-NAL (Y) (4)	CODIFICA-CIÓN DE X (5)	$\hat{\hat{Y}}_i$ SIN VARIA-CIÓN ESTACIO-NAL (6)	PRONÓS-TICO (7)=(6)X(3)
2013	I					
	II					
	III					
	IV					
2014	I					
	II					
	III					
	IV					
2015	I					
	II					
	III					
	IV					
2016	I					
	II					
	III					
	IV					
2017	I					
	II					
	III					
	IV					

INTERPRETACIÓN:

Trazo de los datos, los datos sin variación estacional y la tendencia

<u>Solución al inciso d.</u>

ISBN 978-607-8285-62-4

INTERPRETACIÓN:

2.8.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 2.8.1.1
SERIES DE TIEMPO

Una empresa vende una variedad de aparatos de línea blanca y del hogar. En los últimos cuatro años se reportaron las siguientes ventas trimestrales (en millones de pesos)

AÑO	TRIMESTRE I	TRIMESTRE II	TRIMESTRE III	TRIMESTRE IV
2009	6.72	5.32	7.84	9.52
2010	6.02	5.32	7.98	8.40
2011	7.84	6.44	8.96	8.26
2012	7.56	6.72	9.10	8.40

a) Determine los índices estacionales y elimine la estacionalidad en esos datos (usando el método de razón a promedio móvil de 4 trimestres) e interprete sus resultados.
b) Calcule la línea de mínimos cuadrados que mejor describa estos datos e interprete sus resultados.
c) Estime las ventas trimestrales para los cuatro trimestres de los años 2013,2014,2015 y 2016.
d) Grafique los datos originales, los datos sin la variación estacional, la tendencia y las ventas trimestrales futuras.

ISBN 978-607-8285-62-4

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Uso del método de la razón al promedio móvil

Solución al inciso a.

Paso1: cálculo del total móvil de 4 trimestres

Paso 1: Calcular el total móvil con base a cuatro trimestres. (Columna 4)

TM1=
TM2=
⋮
TM13=

Paso 2: cálculo del promedio móvil de 4 trimestres

Paso 2: Calcular el promedio móvil con base a cuatro trimestres. (Columna 5)

PM1=
PM2=
⋮
PM13=

Paso 3: cómo centrar el promedio móvil de 4 trimestres

Paso 3: Centrar el total móvil con base a cuatro trimestres. (Columna 6)

PMC1=
PMC2=
⋮
PMC12=

Paso 4: calcular el porcentaje del valor actual respecto al valor promedio

Paso 4: Calcular el porcentaje del valor real con base a cuatro trimestres. (Columna 7)

IED1=
IED2=
⋮
IED12=

ISBN 978-607-8285-62-4

Con los cálculos anteriores llenamos la siguiente tabla:

AÑO (1)	TRIMESTRE (2)	VENTAS (3)	TOTAL MOVIL (4)	PROMEDIO MOVIL (5)	PROMEDIO MOVIL CENTRADO (6)	PORCENTAJE DEL VALOR REAL RESPECTO AL MOVIL (7)=(3)/(6)x100
2009	I					
	II					
	III					
	IV					
2010	I					
	II					
	III					
	IV					
2011	I					
	II					
	III					
	IV					
2012	I					
	II					
	III					
	IV					

Paso 5: reunir las respuestas del paso 4 y calcular la media modificada

Paso 5: Calcular la media modificada

Se trascriben los resultados del paso 4 a la siguiente tabla y de los cuatro índices estacionales desajustados de cada trimestre se elimina el valor más bajo y el valor más alto y se suman.

Reducción de las variaciones cíclicas extremas y las variaciones irregulares

AÑO	TRIMESTRE			
	I	II	III	IV
2009				
2010				
2011				
2012				
SUMA MODIFICADA				
MEDIA MODIFICADA				
SUMA DE LAS MEDIAS MODIFICADAS				

ISBN 978-607-8285-62-4

Paso 6: ajuste de la media modificada

Paso 6: Ajustar la media modificada (índices estacionales desajustados) y obtención del índice estacional ó temporal ajustado.

$$Factor\ de\ ajuste =$$

TRIMESTRE	ÍNDICES	X	FACTOR DE AJUSTE	=	INDICES ESTACIONALES
I		X		=	
II		X		=	
III		X		=	
IV		X		=	
				SUMA	400.00

Interpretación:

<u>Solución al inciso b.</u>

Eliminación de la estacionalidad en las series de tiempo

Una vez que hemos eliminado la variación temporal, podemos calcular una línea de tendencia desestacionalizada, que luego podemos proyectar hacia futuro.

AÑO (1)	TRIMESTRE (2)	VENTAS (3)	INDICE ESTACIONAL ENTRE 100 (4)	OCUPACIÓN SIN VARIACIÓN ESTACIONAL (Y) (5)=(3)/(4)	CODIFICACIÓN DE X (6)	xY	X^2
2009	I						
	II						
	III						
	IV						
2010	I						
	II						
	III						
	IV						
2011	I						
	II						
	III						
	IV						
2012	I						
	II						
	III						
	IV						
			SUMAS				
			MEDIAS				

ISBN 978-607-8285-62-4

$b =$ $a =$

Cálculo de los valores sin variación estacional

$$\boxed{\hat{Y}_i = a + bx_i =}$$

Interpretación:

Empleo de la estacionalidad en los pronósticos

Solución al inciso c.

Paso 1: determinación del valor sin variación estacional

Paso 1.

Se calculará la ocupación sin la variación estacional para el primer trimestre del año 2013 cuyo código es ___.

$\hat{Y}_{\_\_(\text{sin } variación\ estacional)} =$

Paso 2: estacionalización de la estimación inicial

Paso 2

Se incluirá la estacionalidad en la estimación anterior para el primer trimestre del año 2013 que en este caso es _____

$\hat{Y}_{\_\_(\text{con } variación\ estacional)} = (a + bx_{i)}\ x\ índice\ estacional =$

Se repite el procedimiento para los 16 trimestres restantes y se construye la siguiente tabla:

AÑO (1)	TRIMES-TRE (2)	INDICE ESTACIO-NAL ENTRE 100 (3)	OCUPA-CIÓN SIN VARIA-CIÓN ESTACION AL (Y) (4)	CODIFICACIÓN DE X (5)	$\hat{Y}_i$ SIN VARIACIÓN ESTACIONAL (6)	PRONÓSTICO (7)=(6)X3)
2013	I					
	II					
	III					
	IV					
2014	I					
	II					
	III					
	IV					
2015	I					
	II					
	III					
	IV					
2016	I					
	II					
	III					
	IV					

ISBN 978-607-8285-62-4

INTERPRETACIÓN:

Trazo de los datos, los datos sin variación estacional y la tendencia

Solución al inciso d.

INTERPRETACIÓN:

ISBN 978-607-8285-62-4

2.8.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 2.8.1
SERIES DE TIEMPO

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

2.8.1.1 Un gran centro comercial tiene tiendas departamentales, restaurantes y locales. En los últimos cuatro años se informaron las siguientes ventas trimestrales (en millones de pesos).

AÑO	TRIMESTRE I	TRIMESTRE II	TRIMESTRE III	TRIMESTRE IV
2009	16.38	11.30	18.15	10.66
2010	16.61	11.55	16.95	10.78
2011	15.96	11.82	16.78	10.52
2012	16.89	11.23	18.29	9.75

a) Determine los índices estacionales y elimine la estacionalidad en esos datos (usando el método de razón a promedio móvil de 4 trimestres) e interprete sus resultados.
b) Calcule la línea de mínimos cuadrados que mejor describa estos datos e interprete sus resultados.
c) Estime las ventas trimestrales del centro comercial para los años 2013,2014,2015 y 2016.
d) Grafique los datos originales, los datos sin la variación estacional, la tendencia y la ventas trimestrales futuras.

2.8.1.2 El propietario de una empresa analiza el ausentismo entre sus trabajadores. En los últimos cinco años registro los siguientes números de inasistencias de sus trabajadores, en días, para cada trimestre del año:

AÑO	TRIMESTRE I	TRIMESTRE II	TRIMESTRE III	TRIMESTRE IV
2008	6	12	9	5
2009	7	14	11	6
2010	8	18	14	6
2011	6	17	13	5
2012	8	19	15	7

a) Determine los índices estacionales y elimine la estacionalidad en esos datos (usando el método de razón a promedio móvil de 4 trimestres) e interprete sus resultados.
b) Calcule la línea de mínimos cuadrados que mejor describa estos datos e interprete sus resultados.
c) Estime el número de inasistencias de los trabajadores para los años 2013,2014,2015, 2016 y 2017.
d) Grafique los datos originales, los datos sin la variación estacional, la tendencia y número de inasistencias trimestrales futuras.

ISBN 978-607-8285-62-4

S.T.

EJEMPLO **ILUSTRATIVO EN MINITAB 15**

EJEMPLO ILUSTRATIVO INTEGRAL DE SERIES DE TIEMPO EN MINITAB 15

La tabla siguiente muestra las ventas trimestrales de una cadena de autoservicio de 2008 a 2012. Las ventas se presentan en millones de pesos.

AÑO	TRIMESTRE I	TRIMESTRE II	TRIMESTRE III	TRIMESTRE IV
2008	9.38	6.44	14.00	17.78
2009	9.10	6.44	13.72	19.04
2010	9.66	7.00	14.56	19.74
2011	9.94	7.98	15.54	20.30
2012	11.20	8.68	15.96	20.86

a) Determine los índices estacionales y elimine la estacionalidad en esos datos (usando el método de razón a promedio móvil de 4 trimestres) e interprete sus resultados.
b) Calcule la línea de mínimos cuadrados que mejor describa estos datos e interprete sus resultados.
c) Estime las ventas trimestrales para los años 2013,2014,2015,2016 y 2017.
d) Grafique los datos originales, los datos sin la variación estacional, la tendencia y las ventas trimestrales futuras.

Uso del método de la razón al promedio móvil.

Solución al inciso a.

Cuando el número de observaciones en cada variable es extenso, los cálculos manuales son tediosos. Existen muchos paquetes de software que pueden mostrar los resultados entre ellos Minitab (Versión 15)

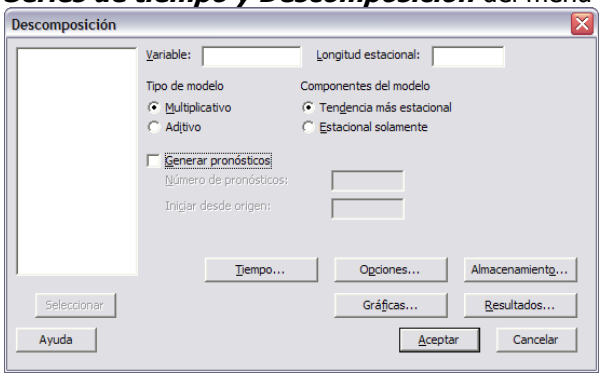
Comenzamos introduciendo los datos en la hoja de Trabajo 1 de Minitab, tal y como se muestra a continuación:

Hoja de trabajo Minitab 15.

ISBN 978-607-8285-62-4

Como tenemos un ***modelo de series de tiempo*** seleccionamos la opción ***Series de tiempo y Descomposición*** del menú ***Estadísticas***,

Cuadro de diálogo: Descomposición.



En ***Variable***, ingrese **C1 VENTAS,** en ***Longitud estacional***, teclee **4.**

Cuadro de diálogo: Descomposición.

De un clic en el botón ***Gráficas*** y seleccione la opción ***no mostrar gráficas***. Haga clic en ***Aceptar*** para regresar a la primera ventana de dialogo ***Descomposición***.

Haga clic en ***Aceptar*** en el cuadro de dialogo ***Descomposición.***

ISBN 978-607-8285-62-4

Salida del programa Minitab 15

Salida de la ventana Sesión

Descomposición de series de tiempo para VENTAS

```
Modelo multiplicativo


Datos     VENTAS
Longitud  20
NMissing  0


Ecuación de tendencia ajustada

Yt = 11.152 + 0.160*t
```

Índices estacionales

Período	Índice
1	0.77254
2	0.57442
3	1.14003
4	1.51300

Interpretación: A la vista de los resultados, podemos observar que en el **primer trimestre de cada año** existe una **disminución** en las ventas del **22.74558%** (100-77.25442), **en el segundo trimestre** una **disminución** en las ventas del **42.55797**% (100-57.44203), **en el tercer trimestre** una **sobre venta** del **14.00330%** (114.00330-100) y en el **cuarto trimestre** una **sobreventa del 51.30024**% (151.30024-100).

Obtención de la pendiente y la intersección en Y

Solución al inciso b.

Con la misma salida de la ventana Sesión del inciso anterior podemos establecer la línea de mínimos cuadrados:

Salida del programa Minitab 15

Salida de la ventana Sesión

```
Ecuación de tendencia ajustada

Yt = 11.152 + 0.160*t
```

Interpretación: Cuando un trimestre **pudiera valer cero** las **ventas sería de 11.15176 millones de pesos** una vez eliminada la estacionalidad. Asimismo **por cada trimestre que se aumente** al modelo **las ventas se incrementarían en 0.16030 millones de pesos.**

ISBN 978-607-8285-62-4

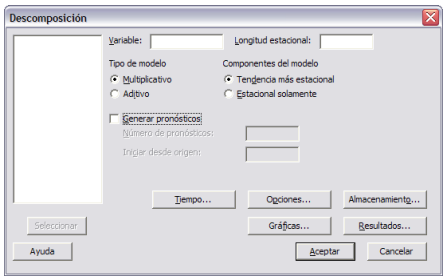
Empleo de la estacionalidad en el cálculo de los pronósticos

Solución al inciso c.

Comenzamos introduciendo nuevamente los datos en la hoja de Trabajo 1 de Minitab o bien borramos la información de la salida de la ventana Sesión.

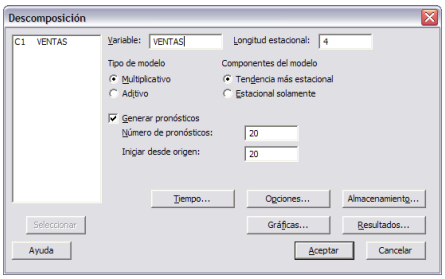
Como tenemos un *modelo de series de tiempo* seleccionamos la opción ***Series de tiempo y Descomposición*** del menú ***Estadísticas,***

Cuadro de diálogo: Descomposición.



En ***Variable***, ingrese **C1 VENTAS,** en ***Longitud estacional***, teclee **4.** Active la casilla ***Generar pronósticos*** y en ***número de pronósticos*** teclee **20** y en ***iniciar desde origen*** vuelva a teclear **20.**

Cuadro de diálogo: Descomposición.



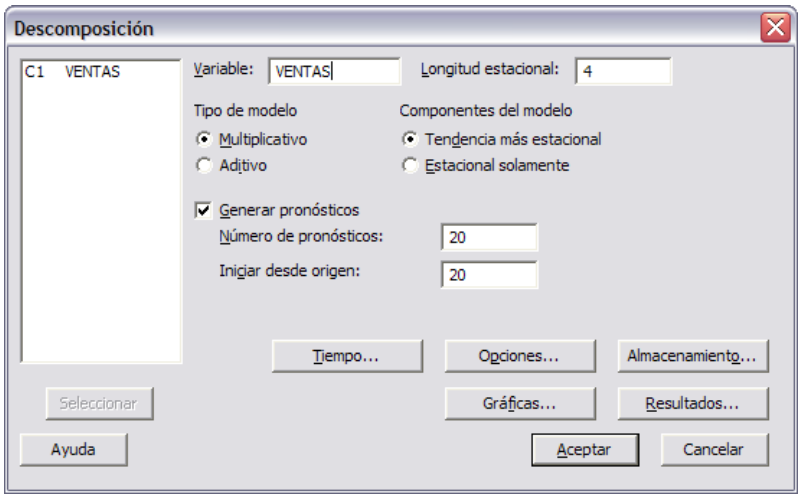
Haga clic en el botón ***Gráficas***. Active la casilla que dice ***no mostrar gráfica***.

Cuadro de diálogo: Descomposición-Gráficas.

Haga clic en ***Aceptar*** en el cuadro de dialogo para que lo regrese al primer cuadro de dialogo ***Descomposición.***

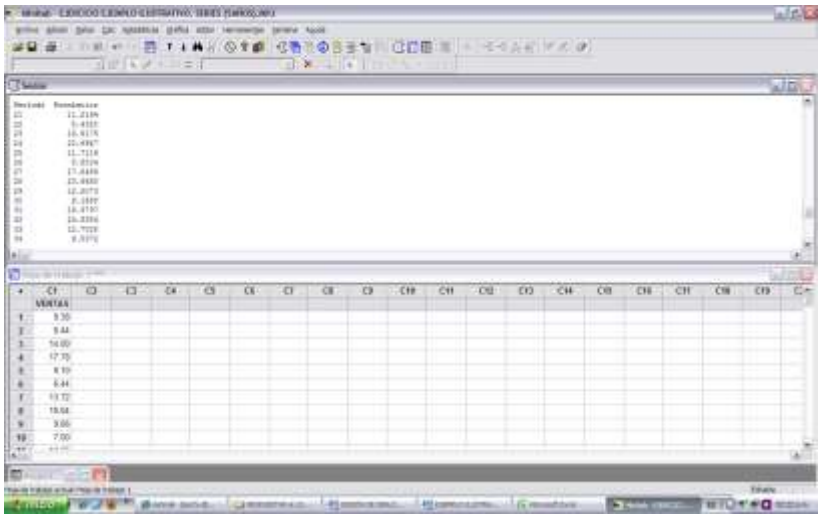
ISBN 978-607-8285-62-4

Cuadro de diálogo: Descomposición.



Haga clic en ***Aceptar*** en este cuadro de dialogo.

Indices estacionales proporcionados por Minitab 15.



Salida del programa Minitab 15

Salida de la ventana Sesión de Minitab

Descomposición de series de tiempo para VENTAS

```
Modelo multiplicativo


Datos      VENTAS
Longitud   20
NMissing   0


Ecuación de tendencia ajustada

Yt = 11.152 + 0.160*t
```


ISBN 978-607-8285-62-4

Índices estacionales

Período	Índice
1	0.77254
2	0.57442
3	1.14003
4	1.51300

Medidas de exactitud

MAPE	3.02011
MAD	0.33222
MSD	0.16596

Pronósticos

Período	Pronóstico
21	11.2164
22	8.4320
23	16.9175
24	22.6947
25	11.7119
26	8.8004
27	17.6486
28	23.6650
29	12.2073
30	9.1688
31	18.3797
32	24.6354
33	12.7028
34	9.5372
35	19.1109
36	25.6057
37	13.1982
38	9.9056
39	19.8420
40	26.5761

INTERPRETACIÓN: El **pronóstico de ventas incluyendo la variación estacional** para el **primer trimestre del año 2013** (Código 21) es de **11.22 millones de pesos,** para el **segundo trimestre del 2013** (Código 22) **de 8.43 millones de pesos** y así sucesivamente hasta el **cuarto trimestre del 2017** (Código 40) que es de **26.58 millones de pesos.**

ISBN 978-607-8285-62-4

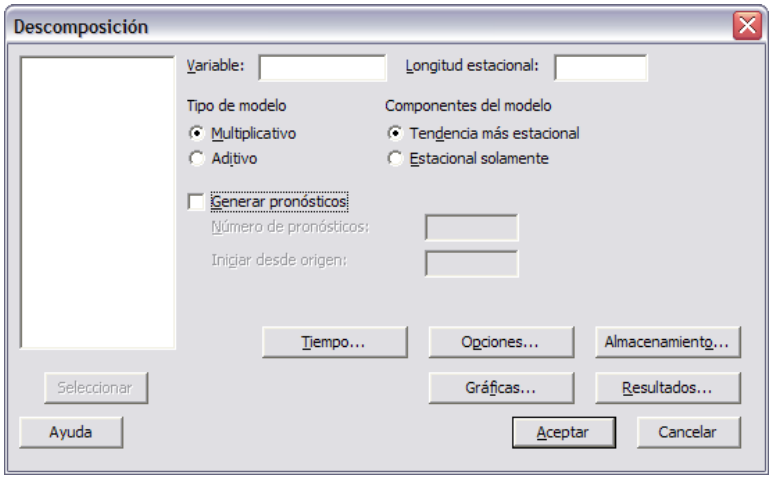
Trazo de los datos, los datos sin variación estacional y la tendencia

Solución al inciso d.

Comenzamos introduciendo nuevamente los datos en la hoja de Trabajo 1 de Minitab o bien borramos la información de la salida de la ventana Sesión.

Como tenemos un *modelo de series de tiempo* seleccionamos la opción ***Series de tiempo y Descomposición*** del menú ***Estadísticas***,

Cuadro de diálogo: Descomposición.



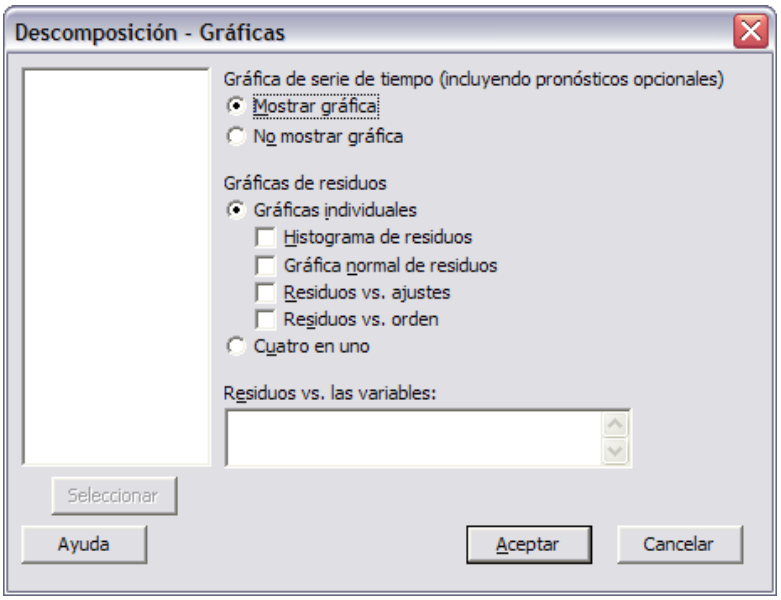
En ***Variable***, ingrese **C1 VENTAS,** en ***Longitud estacional***, teclee **4.** Active la casilla ***Generar pronósticos*** y en ***número de pronósticos*** teclee **20** y en ***iniciar desde origen*** vuelva a teclear **20.**

Cuadro de diálogo: Descomposición.

Haga clic en el botón ***Gráficas***. Active la casilla que dice ***Mostrar gráfica***.

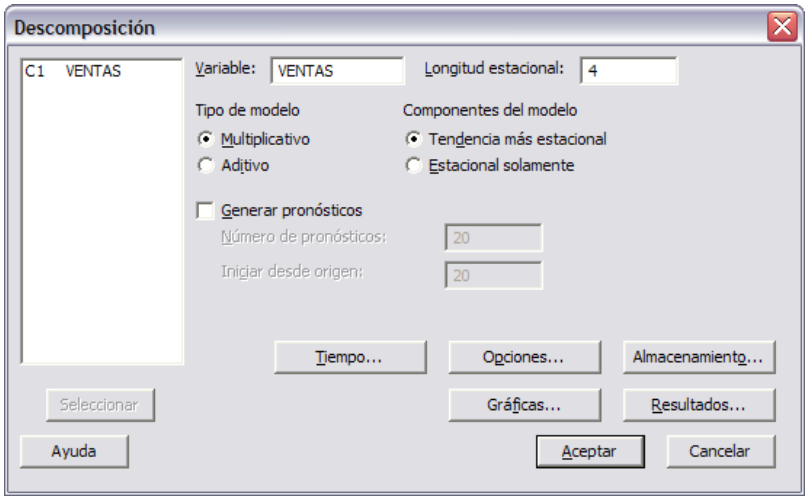
ISBN 978-607-8285-62-4

Cuadro de diálogo: Descomposición-Gráficas.



Haga clic en ***Aceptar*** en el cuadro de dialogo para que lo regrese al primer cuadro de dialogo ***Descomposición.***

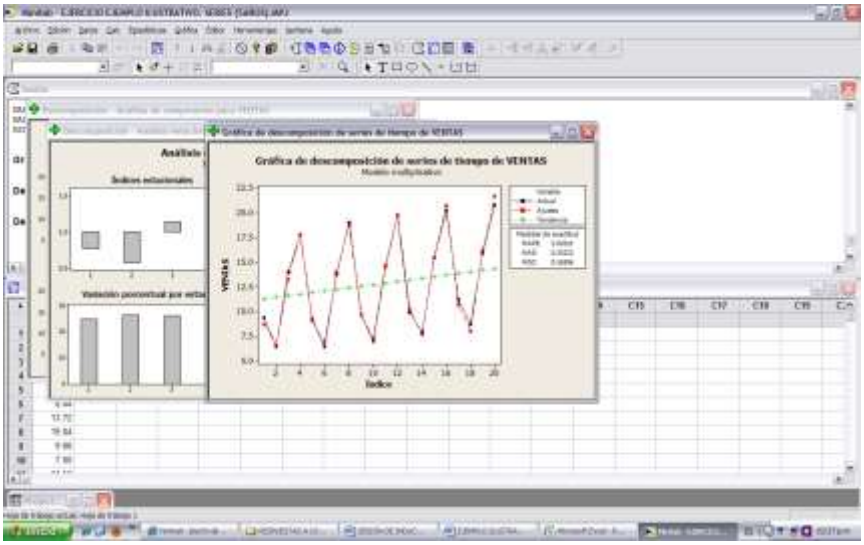
Cuadro de diálogo: Descomposición.



Haga clic en ***Aceptar*** en este cuadro de dialogo.

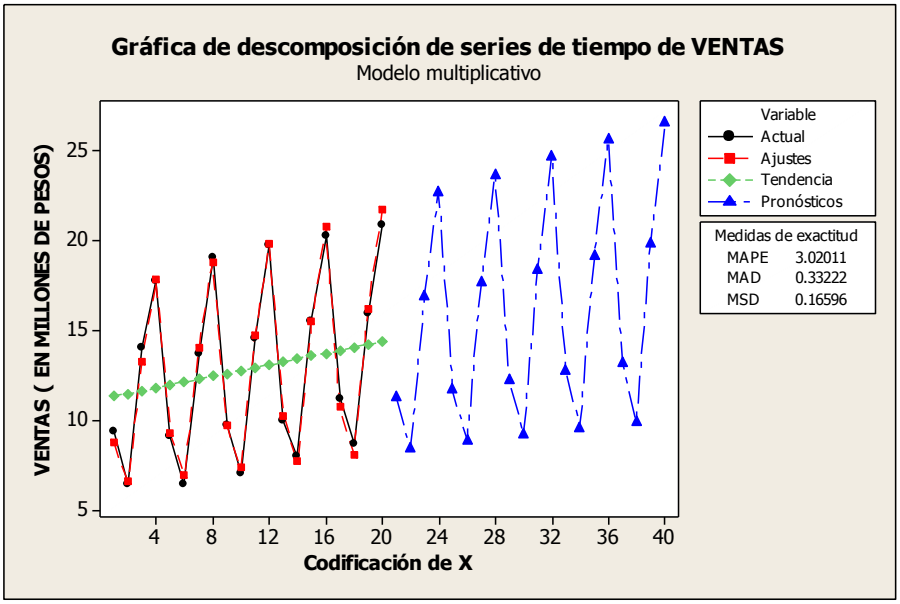
ISBN 978-607-8285-62-4

Resultados generados por Minitab 15.



Aparece la ***Gráfica de descomposición*** de series de tiempo de **VENTAS**:

Gráfica generada por Minitab 15.



INTERPRETACIÓN: La línea continua con **círculo negro** muestra los **datos originales,** la línea con un **cuadrado** los **datos sin variación estacional**, la línea con **un rombo** muestra **la tendencia** y la línea **con triángulos** muestra **los pronósticos a futuro** para los años de **2013, 2014, 2015, 2016 y 2017.**

ISBN 978-607-8285-62-4

EJERCICIOS COMPLEMENTARIOS

1 EJERCICIO COMPLEMENTARIO

EJERCICIO COMPLEMENTARIO 1

El gerente de mercadotecnia de una cadena de supermercados querría investigar el efecto del espacio del aparador sobre la venta de alimento para mascotas. Se seleccionó una muestra aleatoria de 12 tiendas de igual tamaño con los resultados siguientes:

Tienda	1	2	3	4	5	6	7	8	9	10	11	12
Espacio del aparador en m^2 (X)	0.5	0.5	0.5	1.0	1.0	1.0	1.5	1.5	1.5	2.0	2.0	2.5
Ventas semanales en miles de pesos (Y)	16	22	14	19	24	26	23	27	28	29	31	20

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación
c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 1.2$
g) Determine el error estándar de estimación.
h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***
k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 1.2 unidades o sea $X_0 = 1.2$.
l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.

ISBN 978-607-8285-62-4

n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.
p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

2 EJERCICIO COMPLEMENTARIO

EJERCICIO COMPLEMENTARIO 2

El consultor en comportamiento organizacional de una empresa ha diseñado una prueba para mostrar a los supervisores los peligros de una vigilancia excesiva a los subordinados. Un trabajador de la línea de montaje recibe una serie de tareas muy complicadas para que las realice. Durante la ejecución, un supervisor constantemente lo interrumpe para ayudarle a terminarlas. El trabajador, una vez concluidas las tareas, se somete a un test psicológico que mide la hostilidad ante la autoridad (una puntuación alta denota poca hostilidad). A diez trabajadores se les asignaron tareas y luego se les interrumpió varias veces (*X*). En (*Y*) se indican las puntuaciones correspondientes a la prueba de hostilidad

Obs.	1	2	3	4	5	6	7	8	9	10
Interrupciones (X)	5	5	10	10	15	15	20	20	30	35
Puntuación (Y)	58	54	41	45	27	26	13	9	8	5

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación
c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos *X* y *Y* y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 8$
g) Determine el error estándar de estimación.
h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***
k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 8 unidades o sea $X_0 = 8$.
l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.
n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin- Watson.
o) Determine lo adecuado del ajuste del modelo.
p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

3 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 3**

El gerente de personal de una empresa intuye que quizá haya relación entre el ausentismo y la edad de un trabajador. Desea tomar la edad de un trabajador para desarrollar un modelo de predicción de días de ausencia durante un año laboral. Seleccionó una muestra aleatoria de 10 trabajadores con los resultados siguientes:

Trabajador	1	2	3	4	5	6	7	8	9	10
Edad en años (X)	31	65	41	18	50	62	33	40	65	44
Días de aus. (Y)	15	6	10	21	9	7	14	11	5	8

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación
c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 50$
g) Determine el error estándar de estimación.
h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***
k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 50 unidades o sea $X_0 = 50$.
l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.
n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.
p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

4 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 4**

El jefe de redacción de un gran diario metropolitano ha estado tratando de persuadir al dueño para que mejore las condiciones de trabajo en el taller de prensa. Está convencido, de que el nivel de ruido cuando las prensas están funcionando, produce niveles nocivos de tensión y ansiedad. Hace poco hizo que se administrara un test psicológico durante el cual los trabajadores del taller fueron puestos en cuartos con diversos niveles de ruido y luego se sometieron a un test que mide el estado de ánimo y los niveles de ansiedad. La siguiente tabla muestra el índice de su nivel de ansiedad y el nivel de ruido a que fueron expuestos. (1.0 es un nivel bajo y 10.0 es un nivel alto.)

Trabajador	1	2	3	4	5	6	7	8	9	10
Nivel de ruido (X)	4	3	1	2	6	8	2	3	1	6
Nivel de ansiedad (Y)	39	35	16	18	41	38	25	36	12	40

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación
c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 5$
g) Determine el error estándar de estimación.
h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***
k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 5 unidades o sea $X_0 = 5$.
l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.
n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.
p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

5 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 5**

Una compañía fabricante de partes quiere desarrollar un modelo para estimar el número de horas-trabajador requeridas para corridas de producción de tamaños de lotes diferentes. Se seleccionó una muestra aleatoria de 10 corridas de producción, con los siguientes resultados:

Corrida	1	2	3	4	5	6	7	8	9	10
Tamaño del lote (X)	20	20	30	30	40	40	50	50	60	70
Horas-trabajador (Y)	50	55	73	67	87	95	108	112	120	132

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación
c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 35$
g) Determine el error estándar de estimación.
h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***
k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 35 unidades o sea $X_0 = 35$.
l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.
n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.
p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

6 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 6**

El gerente de una escuela de computación quiere desarrollar un modelo para predecir el número de visitas anuales de mantenimiento para terminales interactivas, sobre la base de la antigüedad de la terminal. Se seleccionó una muestra aleatoria de 10 terminales con los siguientes datos:

Terminal	1	2	3	4	5	6	7	8	9	10
Antigüedad en años (X)	1	1	2	2	2	3	3	3	5	6
Visitas de servicio (Y)	3	3	4	5	6	7	8	10	11	10

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación
c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 3.5$
g) Determine el error estándar de estimación.
h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***
k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 3.5 unidades o sea $X_0 = 3.5$
l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.
n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.
p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

7 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 7**

En la contabilidad de costos, con frecuencia se trata de estimar los gastos indirectos basándose en el número de unidades producidas. La gerencia de la empresa ha reunido información sobre esos gastos y las unidades producidas en diferentes plantas y le gustaría estimar una ecuación de regresión para predecir los gastos indirectos en el futuro.

Planta	1	2	3	4	5	6	7	8	9	10
Unidades producidas(X)	40	42	53	35	56	39	48	30	37	40
Gastos Indirectos(Y)	191	168	272	155	256	173	234	116	153	178

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación
c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 50$
g) Determine el error estándar de estimación.
h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***
k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 50 unidades o sea $X_0 = 50$
l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.
n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.
p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

8 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 8**

Un consultor quiere averiguar la exactitud con que un nuevo índice de rendimiento en el trabajo mide lo que es importante para una corporación. Una manera de verificarlo es examinar la relación existente entre dicho índice y el sueldo de un empleado. Se escogió una muestra de 12 empleados y se reunió información sobre el sueldo (en miles) y el índice (1- 10; 10 es optimo).

Empleado	1	2	3	4	5	6	7	8	9	10	11	12
Indice (X)	9	7	8	4	7	5	5	6	8	4	9	6
Sueldo en miles de pesos (Y)	36	25	33	15	28	19	20	22	35	13	34	24

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación
c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 7$
g) Determine el error estándar de estimación.
h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***
k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 7 unidades o sea $X_0 = 7$
l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.
n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.
p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

9 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 9**

Supongamos que usted tiene a su cargo el dinero de cierta región. Se le dan los siguientes datos de antecedentes sobre el suministro de dinero y el Producto Nacional Bruto (ambos en millones de dólares).

Obs.	1	2	3	4	5	6	7	8	9	10
Suministro de dinero en millones de pesos (X)	20	25	32	36	33	40	42	46	48	50
PNB en millones de pesos (Y)	53	55	60	70	72	77	84	90	97	100

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación
c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 30$
g) Determine el error estándar de estimación.
h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***
k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 30 unidades o sea $X_0 = 30$
l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.
n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson.
o) Determine lo adecuado del ajuste del modelo.
p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

10 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 10**

Una compañía administra a sus vendedores una prueba en adiestramiento de ventas antes de permitirles salir a trabajar. La administración de la compañía está interesada en determinar la relación entre las calificaciones de la prueba y las ventas hechas por esos vendedores al final de un año de trabajo. Los siguientes datos se recolectaron de 10 agentes de ventas que han estado en el campo durante un año:

Vendedor	1	2	3	4	5	6	7	8	9	10
Calif. prueba (X)	2.6	3.7	2.4	4.5	2.6	5.0	2.8	3.0	4.0	3.4
Unidades vendidas (Y)	95	140	85	180	100	182	115	136	175	150

a) Calcule la covarianza muestral.
b) Convierta el valor de la covarianza en el coeficiente de correlación
c) Encuentre la estimación mínimo cuadrática para la recta de regresión.
d) Interprete los coeficientes de regresión $\hat{\beta}_0$ y $\hat{\beta}_1$:
e) Represente gráficamente los datos X y Y y la ecuación de predicción.
f) Calcule el valor de $\hat{Y}$ cuando $X_0 = 4.2$
g) Determine el error estándar de estimación.
h) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***t.***
i) Determine un intervalo de confianza de 95% para la pendiente de ***Y***
j) Pruebe la significación de la relación entre la variable dependiente y la independiente utilizando el estadístico de prueba ***F.***
k) Estime e interprete un intervalo de confianza del 95% para el verdadero valor del de la variable dependiente Y cuando se tenga un valor de la variable independiente X de 4.2 unidades o sea $X_0 = 4.2$
l) Determine e interprete el coeficiente de determinación.
m) Determine e interprete el coeficiente de correlación.
n) Realice un análisis de residuales sobre sus resultados incluyendo el estadístico de Durbin-Watson
o) Determine lo adecuado del ajuste del modelo.
p) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
q) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
r) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

11 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 11**

El director general de cierta casa bursátil ha recopilado las siguientes cifras trimestrales con respecto al nivel de cuentas recibidas durante los últimos cinco años (x $ 1,000)

AÑO	TRIMESTRE I	TRIMESTRE II	TRIMESTRE III	TRIMESTRE IV
2008	102	120	90	78
2009	110	126	95	83
2010	111	128	97	83
2011	115	135	103	91
2012	122	144	110	98

a) Determine los índices estacionales y elimine la estacionalidad en esos datos (usando el método de razón a promedio móvil de 4 trimestres) e interprete sus resultados.
b) Calcule la línea de mínimos cuadrados que mejor describa estos datos e interprete sus resultados.
c) Estime el nivel de cuentas recibidas para los cuatro trimestres de los años 2013,2014,2015, 2016 y 2017.
d) Grafique los datos originales, los datos sin la variación estacional, la tendencia y el nivel de cuentas recibidas trimestrales futuras.

12 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 12**

Los siguientes datos representan la cantidad trimestral de dinero en efectivo en circulación durante un periodo de cuatro años

AÑO	TRIMESTRE I	TRIMESTRE II	TRIMESTRE III	TRIMESTRE IV
2009	87	106	86	125
2010	85	110	83	127
2011	84	105	87	128
2012	88	104	88	124

a) Determine los índices estacionales y elimine la estacionalidad en esos datos (usando el método de razón a promedio móvil de 4 trimestres) e interprete sus resultados.
b) Calcule la línea de mínimos cuadrados que mejor describa estos datos e interprete sus resultados.
c) Estime la cantidad de dinero en efectivo en circulación para los cuatro trimestres de los años 2013,2014,2015 y 2016
d) Grafique los datos originales, los datos sin la variación estacional, la tendencia y la cantidad de dinero en efectivo en circulación trimestral futura.

ISBN 978-607-8285-62-4

13 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 13**

Las tasas de rotación de personal de una empresa de la industria automotriz, por trimestre, son:

AÑO	TRIMESTRE I	TRIMESTRE II	TRIMESTRE III	TRIMESTRE IV
2008	4.6	6.4	11.9	7.4
2009	4.3	6.8	11.4	8.9
2010	4.2	6.9	12.2	9.8
2011	5.3	7.3	12.9	9.3
2012	4.4	5.4	10.6	7.5

a) Determine los índices estacionales y elimine la estacionalidad en esos datos (usando el método de razón a promedio móvil de 4 trimestres) e interprete sus resultados.
b) Calcule la línea de mínimos cuadrados que mejor describa estos datos e interprete sus resultados.
c) Estime las tasas de rotación de personal para los cuatro trimestres de los años 2013,2014,2015, 2016 y 2017.
d) Grafique los datos originales, los datos sin la variación estacional, la tendencia y las tasas de rotación de personal trimestrales futuras.

14 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 14**

Las ventas de helados, por trimestre en una cadena de heladerías desde 2009 se indican a continuación (en millones de pesos):

AÑO	TRIMESTRE I	TRIMESTRE II	TRIMESTRE III	TRIMESTRE IV
2009	24	22	9	28
2010	26	25	11	30
2011	28	27	12	30
2012	30	29	11	31

a) Determine los índices estacionales y elimine la estacionalidad en esos datos (usando el método de razón a promedio móvil de 4 trimestres) e interprete sus resultados.
b) Calcule la línea de mínimos cuadrados que mejor describa estos datos e interprete sus resultados.
c) Estime las ventas para los cuatro trimestres de los años 2013,2014,2015 y 2016
d) Grafique los datos originales, los datos sin la variación estacional, la tendencia y las ventas trimestrales futuras.

ISBN 978-607-8285-62-4

AUTOEVALUACIÓN CON REACTIVOS DE **FALSO Ó VERDADERO**

EN CADA UNO DE LOS REACTIVOS, CONTESTE CON UNA F SI CONSIDERA QUE LA AFIRMACIÓN ES FALSA Y CON UNA V SI CONSIDERA QUE LA AFIRMACIÓN ES VERDADERA.

1. La variación estacional es repetitiva y predecible alrededor de una línea de tendencia durante un año. **( )**

2. Si en la gráfica de residuales parece haber un "efecto de abanico" en el cual disminuye la variabilidad de los residuales al aumentar X demuestra falta de homogeneidad en las varianzas de Y_i a cada nivel de X, es decir heteroscedasticidad. **( )**

3. El método de razón a promedio móvil para obtener los índices estacionales elimina solamente la componente cíclica e irregular los datos originales (Y) en una serie de tiempo. **( )**

4. Una prueba formal para verificar el supuesto de normalidad en el análisis de regresión es el estadístico de Durbin-Watson. **( )**

5. La variación irregular en una serie de tiempo suele ser un movimiento impredecible y aleatorio y suele ocurrir en intervalos cortos. **( )**

6. El análisis de las series de tiempo se utiliza para estudiar los patrones de cambio en la información durante intervalos regulares de tiempo. **( )**

7. Los índices estacionales se emplean para suprimir los efectos de la ciclicidad en una serie de tiempo. **( )**

8. La variación explicada en un modelo de regresión se refiere a la variación que puede explicarse por medio de la variable independiente. **( )**

9. Cuando se codifican los valores del tiempo, se resta a cada valor el valor mínimo del tiempo en la serie; por tanto, el código del valor más pequeño es negativo. **( )**

10. Si $r = 0.80$, entonces la ecuación de regresión explica 80% de la variación total de la variable dependiente. **( )**

11. La tendencia a largo plazo sin alteraciones de una serie de tiempo es una componente de las series de tiempo. **( )**

ISBN 978-607-8285-62-4

12. Un valor r^2 cercano a 1 indica una estrecha correlación entre *X* y *Y*. **( )**

13. La asimetría al colocar los residuales estandarizados en una distribución de frecuencias y mostrando los resultados en un histograma puede significar la violación a la suposición de normalidad en la regresión. **( )**

14. Si la suma de los cuadrados de las desviaciones es pequeña, esto significa que la línea de regresión es representativa de los datos. **( )**

15. En el análisis de regresión la variable que estamos intentando predecir es la variable independiente. **( )**

16. La relación entre las variables dependiente e independiente puede ser directa cuando la variable dependiente disminuye al aumentar la variable independiente. **( )**

17. La relación entre las variables dependiente e independiente puede ser directa cuando la pendiente de la recta de regresión es negativa. **( )**

18. La relación entre las variables dependiente e independiente puede ser inversa cuando la variable dependiente disminuye al aumentar la variable independiente. **( )**

19. El método del promedio móvil es el método básico para medir la fluctuación estacional en una serie de tiempo. **( )**

20. Supóngase que la pendiente de una ecuación de estimación sea positiva. Entonces el valor de *r* debe ser la raíz cuadrada positiva de r^2. **( )**

21. La suposición de linealidad en la regresión se puede evaluar colocando los residuales estandarizados en una distribución de frecuencia y mostrando los resultados en un histograma. **( )**

22. En el análisis de regresión podemos emplear más de una variable dependiente. **( )**

23. Cuando se codifican los valores del tiempo en trimestres, el código del valor más pequeño es negativo. **( )**

24. El supuesto de normalidad en el análisis de regresión, requiere que la variación en torno a la línea de regresión sea constante para todos los valores de X. **( )**

25. A la técnica empleada para desarrollar la ecuación y dar las estimaciones se conoce como análisis de regresión. **( )**

26. El análisis de regresión se utiliza para determinar el grado de relación que hay entre las variables. **( )**

27. Las tendencias seculares representan la dirección a largo plazo de una serie de tiempo. **( )**

28. Para mejorar la exactitud de la predicción en el análisis de regresión podemos agregar más variables dependientes en el modelo. **( )**

29. El procedimiento para identificar la tendencia y los ajustes estacionales se combina para producir pronósticos. **( )**

ISBN 978-607-8285-62-4

30. Si el error estándar del estimador es grande, la recta de regresión puede no representar a los datos.

( )

31. El suposición de normalidad se puede probar con una prueba ji cuadrada de bondad de ajuste de lo adecuado del ajuste del modelo.

( )

32. Si en la gráfica de residuales parece haber un "efecto de abanico" en el cual aumenta la variabilidad de los residuales al aumentar X demuestra falta de normalidad en el modelo ajustado para los datos.

( )

33. En el análisis de influencia cualquier valor D_i superior a 4/20 debe ser destacado como un punto de influencia y se puede considerar candidato a ser retirado del modelo.

( )

34. El análisis de influencia se utiliza para evaluar lo adecuado del modelo de regresión ajustado a los datos.

( )

35. Una serie de tiempo es un grupo de datos cuantitativos que se obtienen en periodos regulares.

( )

36. Cada "elemento diagonal de la matriz sombrero" h_i refleja la "influencia" de cada X_i sobre el modelo de regresión ajustado.

( )

37. Cuando se obtiene $S_{Y.X} = 0$ en una ecuación de estimación, la variable dependiente en los puntos observados se debe estimar perfectamente.

( )

38. El análisis de residuales se utiliza para evaluar lo adecuado del modelo de regresión ajustado a los datos.

( )

39. Cuando se codifican los valores del tiempo, se resta a cada valor el valor promedio del tiempo en la serie; por tanto, el código del valor más pequeño es cero.

( )

40. El movimiento repetitivo alrededor de una línea de tendencia durante un periodo de dos o más años se describe como cíclico.

( )

41. A la técnica empleada para obtener la ecuación de regresión, minimizando la suma de cuadrados de las distancias verticales entre los verdaderos valores de Y y los valores pronosticados de Y se le denomina principio de mínimos cuadrados.

( )

42. El primer paso en el cálculo del índice estacional consiste en obtener el total móvil de cuatro trimestres.

( )

43. Si en la gráfica de residuales no existe un patrón aparente de los residuales contra X quiere decir que el modelo ajustado es apropiado para los datos.

( )

44. Una serie de tiempo es un grupo de datos cualitativos que se obtienen en periodos regulares.

( )

45. El error estándar de estimación se basa en los cuadrados de las desviaciones respecto a la línea de regresión.

( )

46. Una técnica que se puede utilizar como ayuda para controlar las operaciones presentes y planear las necesidades futuras es el análisis de regresión.

( )

ISBN 978-607-8285-62-4

47. Un residual estandarizado muy grande sugiere que el dato puede ser un valor atípico o aberrante.

( )

48. Si la ecuación de una línea es $Y = 26 - 24X$, podemos afirmar que la relación entre Y y X es lineal e inversa.

( )

49. La relación entre las variables dependiente e independiente puede ser inversa cuando la variable dependiente aumenta al aumentar la variable independiente.

( )

50. Al utilizar el método de mínimos cuadrados en una línea para ajustar un conjunto de puntos, los errores individuales, tanto positivos como negativos, de la línea deben dar un total de cero.

( )

51. Un intervalo de confianza informa acerca de la gama de valores de Y para un valor particular de X.

( )

52. A la medida de la dispersión de los valores observados, con respecto a la línea de regresión se le denomina desviación estándar de estimación.

( )

53. La línea de regresión esta derivada de una población entera y no de una muestra.

( )

54. Si la pendiente de la recta de regresión es positiva entonces la relación entre las variables dependiente e independiente es directa.

( )

55. Podemos interpretar el coeficiente de determinación como la variación de Y que es explicada por la variación de X.

( )

56. En el análisis de regresión la variable conocida recibe el nombre de variable independiente.

( )

57. El análisis de la serie de tiempo nos ayuda a analizar las tendencias históricas, pero no puede ayudarnos a resolver las incertidumbres del futuro.

( )

58. El coeficiente de correlación explica el porcentaje de la variación total de la variable dependiente.

( )

59. En la ecuación $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1$ para la variable dependiente Y y la variable independiente X, la intersección en el eje Y es $\hat{\beta}_i$.

( )

60. Un valor r cercano a 0 indica una estrecha correlación entre X y Y.

( )

61. En el análisis de influencia cualquier valor t_i^* superior a 4/20 debe ser destacado como un punto de influencia y se puede considerar candidato a ser retirado del modelo.

( )

62. Un valor r^2 cercano a 1 indica una estrecha correlación entre X y Y.

( )

63. A la porción de la variación total en la variable dependiente Y, que se explica por la variación en la variable independiente X se le denomina coeficiente de correlación.

( )

64. El supuesto de independencia del error, requiere que el error sea independiente para la mayoría de los valores de X.

( )

ISBN 978-607-8285-62-4

65. En el análisis de influencia cualquier valor h_i superior a 4/20 debe ser destacado como un punto de influencia y se puede considerar candidato a ser retirado del modelo.

( )

66. A la medida de la intensidad de la relación entre dos variables de le denomina coeficiente de determinación.

( )

67. El análisis de regresión nos ayuda a analizar las tendencias históricas, pero no puede ayudarnos a resolver las incertidumbres del futuro.

( )

68. El error estándar de la estimación mide la variabilidad de los valores alrededor de la ecuación de regresión.

( )

69. La fuerza de la correlación depende de si la relación es positiva ó negativa.

( )

70. Si un diagrama de residuales contra el tiempo muestra un patrón cíclico no aleatorio, es probable que exista autocorrelación.

( )

71. El movimiento repetitivo alrededor de una línea de tendencia durante un periodo menor a un año se describe como cíclico.

( )

72. En el análisis de regresión la ecuación de la estimación es válida solo sobre el mismo intervalo que el que está dado por los datos de la muestra original a partir de los cuales se desarrollo.

( )

73. Los análisis de regresión y correlación muestran como determinar la fuerza y la naturaleza de una relación entre dos variables.

( )

74. Si los verdaderos errores en el análisis de regresión son en realidad independientes, el valor esperado del estadístico de Durbin-Watson oscilará alrededor de 2.0.

( )

75 El análisis de regresión se usa para detectar los patrones de cambio en la información estadística durante intervalos regulares de tiempo.

( )

76. Cualquier valor del estadístico de Durbin-Watson mayor que a 1.6 nos lleva a sospechar que hay autocorrelación violando el supuesto de independencia del error.

( )

77. Las líneas trazadas en ambos lados de la línea de regresión ± 1, ± 2, ± 3 multiplicadas por el valor del error estándar de la estimación reciben el nombre de límites de confianza.

( )

78. Al ascenso y descenso de una serie de tiempo en periodos mayores de un año se le denomina como variación estacional.

( )

79. Un intervalo de confianza, presenta el valor medio de Y para una valor dado de X.

( )

80. Un valor *r* cercano a -1 indica una estrecha correlación positiva entre *X* y *Y*.

( )

81. Si en la gráfica de residuales parece haber un "efecto de abanico" en el cual aumenta la variabilidad de los residuales al aumentar X demuestra falta de homogeneidad en las varianzas de Y_i a cada nivel de X, es decir heteroscedasticidad.

( )

ISBN 978-607-8285-62-4

82. La relación entre las variables dependiente e independiente puede ser directa cuando la variable dependiente aumenta al aumentar la variable independiente. **( )**

83. El análisis de correlación se utiliza para medir la fuerza de asociación entre las variables cualitativas. **( )**

84. La variación no explicada en un modelo de regresión se refiere a la variación que no puede explicarse por medio de la variable dependiente. **( )**

85. El análisis de residuales estudia el efecto potencial o la "influencia" de cada punto sobre el modelo ajustado. **( )**

86. Una técnica que se puede utilizar como ayuda para controlar las operaciones presentes y planear las necesidades futuras es el análisis de series de tiempo. **( )**

87. El error estándar de la estimación se mide perpendicularmente desde la línea de regresión y no sobre el eje *Y*. **( )**

88. La razón para desestacionalizar las series de tiempo es eliminar las fluctuaciones estacionales a fin de estudiar la tendencia y el ciclo. **( )**

89. A los patrones de cambio en una serie de tiempo en un año y que tienden a repetirse cada año se le llama variación estacional. **( )**

90. La relación entre las variables dependiente e independiente puede ser inversa cuando la pendiente de la recta de regresión es negativa. **( )**

ISBN 978-607-8285-62-4

AUTOEVALUACIÓN CON REACTIVOS DE **OPCIÓN MÚLTIPLE**

EN CADA UNO DE LOS REACTIVOS SIGUIENTES, SELECCIONE LA OPCIÓN QUE CONSIDERE CORRECTA.

1. Supóngase que nos dicen que existe una relación inversa entre la demanda de un producto y el precio de éste. Puede afirmarse que:
a) Los ventas tienden a ser altas cuando el precio es alto es grande
b) Un gran incremento en los precios hace que las ventas disminuyan
c) Si el precio se congela las ventas disminuyen
d) Las ventas tienden a ser altas cuando el precio se mantiene estable

2. Se cuenta con datos trimestrales de las ventas de una empresa durante un periodo de cinco años. Se quiere utilizar el método de razón a promedio móvil para determinar los índices estacionales, el sexto paso será:
a) Ajustar la media modificada y obtención del índice estacional
b) Obtener la media modificada
c) Obtener el porcentaje de promedio real respecto al móvil
d) Calcular el promedio móvil centrado de 4 trimestres

3. En una ecuación de regresión lineal simple se calcula, $\hat{\beta}_0 = 4$ y $\hat{\beta}_1 = 2$ para determinada estimación de las ventas, con una variable independiente. Si la estimación de las ventas es de 10, ¿qué valor cabe esperar que tenga la variable independiente?
a) 1.75
b) 3
c) 8
d) 4

4. La suma total de los cuatro índices desajustados es 399.15. Si el índice desajustado del tercer trimestre es 85.36, Cuál es el índice estacional ajustado del primer semestre aproximado a dos cifras?:
a) 85.18
b) 21.39
c) 85.54
d) No puede determinarse con la información disponible

ISBN 978-607-8285-62-4

5. Supóngase que nos dicen que existe una relación directa entre el precio de las zanahorias y la lluvia durante la estación del cultivo. Puede afirmarse que:
a) Los precios tienden a ser altos cuando la precipitación pluvial es grande
b) Los precios tienden a ser bajos cuando la precipitación pluvial es grande
c) Una gran cantidad de lluvia hace que los precios disminuyan
d) La falta de lluvia hace que los precios aumenten

6. Se cuenta con datos trimestrales de las ventas de una empresa durante un periodo de cinco años. Se quiere utilizar el método de razón a promedio móvil para determinar los índices estacionales, el tercer paso será:
a) Calcular el promedio móvil de 4 trimestres.
b) Obtener la media modificada.
c) Calcular el total móvil de 4 trimestres.
d) Calcular el promedio móvil centrado de 4 trimestres.

7. En una ecuación de regresión lineal simple se calcula, $\hat{\beta}_0 = 20$ y $\hat{\beta}_1 = -2$ para determinada estimación de las ventas, con una variable independiente. Si esta variable toma el valor de 10, ¿qué valor cabe esperar que tenga la variable dependiente?
a) 24
b) 40
c) 0
d) 198

8. Se calculó la línea de tendencia sin considerar la variación estacional para estimar las ventas trimestrales en un periodo de 2009-2012 y la ecuación que resulto fué: $\hat{Y}_i = 350 + 3x_i$. ¿Cuál será el pronóstico de las ventas para el segundo trimestre de 2013 si los índices estacionales para los cuatro trimestres son 90.30; 106.62; 112.11 y 90.97 respectivamente?
a) 356
b) 430.74
c) 379.57
d) 404

9. Supóngase que la ecuación de estimación $\hat{Y}_i = 7 + 5X_i$ se ha calculado para conocer el comportamiento de las ventas. ¿Cuál de los siguientes enunciados es verdadero en esta situación?
a) La pendiente de la recta es positiva
b) La relación entre las variables es directa
c) La ordenada al origen es 7
d) Todas los anteriores

10. Se cuenta con datos trimestrales de las ventas de una empresa durante un periodo de cinco años. Se quiere utilizar el método de razón a promedio móvil para determinar los índices estacionales, el quinto paso será:
a) Ajustar la media modificada y obtención del índice estacional
b) Obtener la media modificada
c) Obtener el porcentaje de promedio real respecto al móvil
d) Calcular el promedio móvil centrado de 4 trimestres

ISBN 978-607-8285-62-4

11. El error estándar del estimador es el mismo en todas las observaciones sobre una línea de regresión porque suponemos que:
a) Los errores entre un valor observado y uno predicho son independientes
b) Los valores observados de *Y* están normalmente distribuidos alrededor de cada valor estimado de Ŷ
c) La variancia de la distribución alrededor de cada valor posible de Ŷ es la misma
d) Todos los datos disponibles se obtuvieron en forma aleatoria

12. Se calculó la línea de tendencia sin considerar la variación estacional para estimar las ventas trimestrales en un periodo de 2007-2012 y la ecuación que resulto fué: $\hat{Y}_i = 100 + 5x_i$. ¿Cuál será el pronóstico de las ventas para el primer trimestre de 2013 si los índices estacionales para los cuatro trimestres son 90.30; 106.62; 112.11 y 90.97 respectivamente?
a) 105
b) 225
c) 185.12
d) 203.18

13. Para que la ecuación de estimación sea un estimador perfecto de la variable dependiente, ¿Cuál de los siguientes enunciados deberá ser verdadero?
a) El coeficiente de determinación es cero
b) Todos los puntos de datos se hallan sobre la línea de regresión
c) El error estándar de la estimación es cero
d) Todos los anteriores

14. Una serie de tiempo con datos anuales de los años 2010-2012 está bien descrita por la ecuación: $\widehat{Y}_i = 10 + 6x_i$. Basándose en esta tendencia secular, ¿Cuál será el valor del pronóstico para 2014?
a) 16
b) 64
c) 34
d) 28

15. En la ecuación $Y_i = \beta_0 + \beta_{1i} + \varepsilon_i$, la ε_i representa:
a) Los efectos de los otros factores conocidos o desconocidos
b) Las desviaciones de los verdaderos valores de Y de los valores pronosticados o estimados
c) Los efectos combinados de los factores impredecibles e ignorados
d) Todos los anteriores

16. Una serie de tiempo de datos anuales puede incluir uno de los siguientes componentes. Señale cuál.
a) Tendencia secular
b) Fluctuación cíclica
c) Variación estacional
d) Todas las anteriores

ISBN 978-607-8285-62-4

17. Supóngase que conocemos la altura de un estudiante pero no su peso. Utilizamos una ecuación de regresión para obtener una estimación de su peso basándonos en su altura. Por tanto, podemos suponer que:

a) El peso es la variable independiente
b) La altura es la variable independiente
c) La altura es la variable dependiente
d) a y c pero no b.

18. Suponga que se está estudiando una serie de tiempo de datos durante los trimestres de 2011 y 2012. El tercer trimestre de 2012 será codificado como:

a) 3
b) 7
c) 2
d) 6

19. Se cuenta con datos trimestrales de las ventas de una empresa durante un periodo de cinco años. Se quiere utilizar el método de razón a promedio móvil para determinar los índices estacionales, el cuarto paso será:

a) Ajustar la media modificada y obtención del índice estacional
b) Obtener la media modificada
c) Obtener el porcentaje de promedio real respecto al móvil
d) Calcular el promedio móvil centrado de 4 trimestres

20. Suponga que una serie de tiempo debe ajustarse con una recta. La forma de la ecuación es $\hat{Y}i_{=}a + bx_i$. ¿Qué representa la x minúscula en esta fórmula?

a) Estimaciones de la variable dependiente
b) Valores codificados de la variable de tiempo
c) Una constante numérica
d) Valores de la variable de tiempo

21. Los porcentajes del valor real respecto al móvil en el segundo trimestre de cada año en datos trimestrales de 2009 a 2012 son: 2010:108.9; 2011:104.1; 2012:102.5. ¿Cuál es el índice desajustado del segundo trimestre?:

a) 52.05
b) 102.5
c) 104.1
d) 54.45

22. Supóngase que queremos comparar el supuesto valor de β_1 con un valor muestral de $\hat{\beta}_1$ que se ha calculado. ¿Cuál de los siguientes valores debe ser calculado antes que otros?

a) $S_{\widehat{\beta}_1}$
b) $S_{Y.X}$
c) $S_{\widehat{\beta}_0}$
d) Los cálculos pueden hacerse en cualquier orden

ISBN 978-607-8285-62-4

23. Se calculó la línea de tendencia sin considerar la variación estacional para estimar las ventas trimestrales en un periodo de 2008-2012 y la ecuación que resulto fué: $\hat{Y}_i = 250 + 2x_i$. ¿Cuál será el pronóstico de las ventas para el tercer trimestre de 2013 si los índices estacionales para los cuatro trimestres son 90.30; 106.62; 112.11 y 90.97 respectivamente?

a) 331.85
b) 296
c) 287
d) 322.88

24. Se cuenta con datos trimestrales de las ventas de una empresa durante un periodo de cinco años. Se quiere utilizar el método de razón a promedio móvil para determinar los índices estacionales, el primer paso será:

a) Calcular el promedio móvil de 4 trimestres
b) Obtener la media modificada
c) Calcular el total móvil de 4 trimestres
d) Calcular el promedio móvil centrado de 4 trimestres

25. El valor de r en una situación particular es 70%. ¿Cuál será el coeficiente de determinación en este caso?

a) 4900
b) 49
c) 0.49
d) No puede determinarse con la información disponible

26. Se cuenta con datos trimestrales de las ventas de una empresa durante un periodo de cinco años. Se quiere utilizar el método de razón a promedio móvil para determinar los índices estacionales, el segundo paso será:

a) Calcular el promedio móvil de 4 trimestres
b) Obtener la media modificada
c) Calcular el total móvil de 4 trimestres
d) Calcular el promedio móvil centrado de 4 trimestres

27. La suma total de los cuatro índices desajustados es 405.22. Si el índice desajustado del segundo trimestre es 125.64, Cuál es el índice estacional ajustado del primer semestre aproximado a dos cifras?:

a) 124.02
b) 127.28
c) 31.01
d) No puede determinarse con la información disponible

28. Los porcentajes del valor real respecto al móvil en el cuarto trimestre de cada año en datos trimestrales de 2008 a 2012 son 2008:89.62; 2009:92.52; 2010:91.98; 2011:91.82. ¿Cuál es el índice desajustado del cuarto trimestre?:

a) 182.14
b) 91.07
c) 183.80
d) 91.90

ISBN 978-607-8285-62-4

29. Se calculó la línea de tendencia sin considerar la variación estacional para estimar las ventas trimestrales en un periodo de 2009-2012 y la ecuación que resultó fué: $\hat{Y}_i = 300 + 5x_i$. ¿Cuál será el pronóstico de las ventas para el segundo trimestre de 2013 si el índice estacional del segundo trimestre es 90?

a) 310
b) 351
c) 390
d) 279

30. Se calculó la línea de tendencia sin considerar la variación estacional para estimar las ventas trimestrales en un periodo de 2008-2012 y la ecuación que resulto fué: $\hat{Y}_i = 500 + 6x_i$. ¿Cuál será el pronóstico de las ventas para el primer trimestre de 2014 si el índice estacional del primer trimestre es 62?

a) 650.
b) 506.
c) 313.72
d) 403.

ISBN 978-607-8285-62-4

GLOSARIO DE REGRESIÓN LINEAL SIMPLE. PARTE 1

ANÁLISIS DE CORRELACIÓN. Técnica de con que se determina el grado de relación lineal que hay entre variables.

ANÁLISIS DE RESIDUALES. Respecto a regresión, análisis de las diferencias entre Y y $\hat{Y}$ para valorar las premisas y proporciona guías sobre qué tan bien se ajusta la ecuación a los datos.

ANÁLISIS DE LA VARIANCIA PARA LA REGRESIÓN. Procedimiento con que se calcula la razón F; se emplea para probar la significancia de la regresión como un todo. Se relaciona con el análisis de variancia

COEFICIENTE (r) DE CORRELACIÓN . Una medida de la relación lineal entre dos mediciones numéricas hechas en el mismo conjunto de sujetos. Oscila de -1 a +1, con el cero indicando la ausencia de relación**.** Raíz cuadrada del coeficiente de determinación. Su signo indica la dirección de la relación entre dos variables, directa o inversa.

COEFICIENTE (R^2) DE DETERMINACIÓN . Medida de la proporción de variación de *Y*, la variable dependiente, que se explica con la línea de regresión; esto es, por la relación de las *Y* con la variable independiente. Se interpreta como la cantidad de variación en una variable que puede definirse por el conocimiento de una segunda variable.

COEFICIENTES DE REGRESIÓN. La constante β_1 en la ecuación de regresión lineal simple, $Y = \beta_0 + \beta_1 X$. ,se interpreta como la pendiente de la línea de regresión y β_0 como la ordenada al origen.

DIAGRAMA DE DISPERSIÓN. Grafica de puntos sobre una rejilla rectangular; las coordenadas *X* y *Y* de cada punto de corresponden a las dos mediciones hachas en algún elemento particular de la muestra, y el patrón de puntos indica la relación existente entre las dos variables.

ECUACIÓN DE ESTIMACIÓN. Fórmula matemática que relaciona la variable desconocida con las variables conocidas en el análisis de regresión.

ECUACIÓN DE REGRESIÓN. Es una ecuación que define la relación lineal entre dos variables.

ERROR ESTÁNDAR DE ESTIMACIÓN. Medida de al confiabilidad de la ecuación de estimación, que indica la variabilidad de los puntos observados alrededor de la línea de regresión; es decir, hasta qué punto los valores observados difieren de los predichos en la línea de regresión.

F CALCULADA. Estadístico que se usa como prueba de la significancia de una variable explicatoria individual.

ISBN 978-607-8285-62-4

GLOSARIO DE REGRESIÓN LINEAL SIMPLE. PARTE 2

INTERSECCIÓN EN *Y*. Constante de cualquier recta, cuyo valor representa el valor de la variable *Y* cuando la variable *X* tiene un valor de 0.

LÍNEA DE REGRESIÓN. Línea ajustada a un conjunto de puntos de datos para estimar la relación entre dos variables.

MÉTODO DE MÍNIMOS CUADRADOS. Técnica con que se ajusta una recta mediante un conjunto de puntos, de manera que se minimice la suma de los cuadrados de las distancias verticales entre *n* puntos y la línea.

MULTICOLINEALIDAD. Problema estadístico que en ocasiones se presenta en el análisis de regresión múltiple; en él se reduce la confiabilidad de los coeficientes de regresión, a causa de un alto nivel de correlación entre las variables independientes.

PENDIENTE. Constante de cualquier recta, cuyo valor representa en qué medida el cambio de cada unidad de la variable independiente modifica la variable dependiente.

PRINCIPIO DE MÍNIMOS CUADRADOS. Técnica empleada para obtener la ecuación de regresión, minimizando la suma de los cuadrados de las distancias verticales entre los valores verdaderos de ***Y*** y los valores pronosticados de ***Y.***

RAZÓN F CALCULADA. Estadístico que se usa para probar la significancia de la regresión como un todo.

REGRESIÓN. (de **Y** en **X**) proceso por el cual se determina una ecuación para predecir **Y** a partir de **X**. Proceso general de predecir una variable a partir de otra con medios estadísticos, usando datos anteriores.

REGRESIÓN MÚLTIPLE. Proceso estadístico por medio del cual se utilizan, algunas variables para predecir otra variable.

RELACIÓN CURVILÍNEA. Nexo de dos variables que es descrito por una línea curva.

RELACIÓN DIRECTA. Relación entre dos variables en la cual, al aumentar un valor de la variable independiente, también aumenta el de la variable dependiente.

RELACIÓN INVERSA: Relación entre dos variables en la cual, al aumentar la variable independiente, disminuye la variable dependiente.

RELACIÓN LINEAL. Tipo particular de asociación entre dos variables, que puede ser descrita matemáticamente con una recta.

RESIDUAL. Diferencia entre el valor probable (predicción) y el valor real de la variable dependiente (resultado o respuesta) en regresión.

TÉCNICAS DE MODELADO. Métodos con que se decide que variables incluir en un modelo de regresión y las diferentes maneras de incluirlas.

VARIABLE DEPENDIENTE. Aquella que estamos tratando de predecir en el análisis de regresión.

VARIABLE INDEPENDIENTE. La variable, o variables, conocidas en el análisis de regresión.

ISBN 978-607-8285-62-4

αA

SIMBOLOGÍA

$=$	Igual	Σ	Letra griega mayúscula signa; símbolo que indica una suma
$\neq$	Desigual	$g.l.$	Grados de libertad
$<$	Menor que	ε	Letra griega épsilon; usada para simbolizar el error experimental
$\leq$	Menor que o igual a	F	Símbolo para la prueba y la distribución F
$>$	Mayor que	n	Tamaño de la muestra
$\geq$	Mayor que o igual que	r	Correlación de la muestra
H_o	Hipótesis nula	r^2	Correlación al cuadrado, llamado coeficiente de determinación
H_1	Hipótesis alterna	S	Desviación estándar de la muestra
α	Letra griega alfa; probabilidad de un error tipo I	SE	Error estándar de la muestra
β	Letra griega beta; probabilidad de un error tipo II	$S_{Y.X}$	Error estándar de la estimación en regresión
β_0	Valor poblacional de la ordenada al origen de la línea de regresión	t	Símbolo para la razón t (la razón crítica que sigue a una distribución t)
β_1	Valor poblacional de la pendiente de la línea de regresión.	X	Variable independiente (explicatoria, predictora) en regresión
$\hat{\beta}_0$	Valor estimado de la ordenada al origen de la línea de regresión	$\bar{X}$	Media de la muestra; X con barra
$\hat{\beta}_1$	Valor estimado de la pendiente de la línea de regresión	Y	Variable dependiente (resultado, respuesta, criterio) en regresión
μ	Letra griega mu; media de la población	$\hat{y}$	Valor probable (predicción) de Y en regresión
ρ	Letra griega rho, correlación de la población	SCT	Suma de Cuadrados Total
σ	Legra griega minúscula sigma; desviación estándar de población	SCR	Suma de Cuadrados de la Regresión
τ	Legra griega tau; usada para simbolizar términos en el modelo ANOVA	SCE	Suma de Cuadrados del Error
CMR	Cuadrado Medio de la Regresión	CME	Cuadrado Medio del Error

ISBN 978-607-8285-62-4

FÓRMULAS CLAVE. PARTE 1

• **Covarianza** $cov(X,Y) = S_{XY} \frac{\sum[(X-\bar{X})(Y-\bar{Y})]}{n-1}$	**(1)**	• **Coeficiente de correlación muestral** $r = \frac{cov(X,Y)}{S_X S_Y} = \frac{S_{XY}}{S_X S_Y}$	**(2)**
• **Desviación estándar muestral de X** $S_X = \sqrt{\frac{\sum_{i=1}^{n} X^2 - n\bar{X}^2}{n-1}}$	**(3)**	• **Desviación estándar muestral de Y** $S_Y = \sqrt{\frac{\sum_{i=1}^{n} Y^2 - n\bar{Y}^2}{n-1}}$	**(4)**
• **Pendiente de la recta de regresión** $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} XY - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X^2 - n\bar{X}^2}$	**(5)**	• **Ordenada de la recta de regresión** $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$	**(6)**
• **Varianza de la población** $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$	**(7)**	• **Desviación estándar de la población** $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}}$	**(8)**
• **Varianza muestral** $S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$	**(9)**	• **Desviación estándar muestral** $S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}$	**(10)**
• **Error estándar del estimador** $S_{Y:X} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i}{n-2}}$	**(11)**	• **Estadístico *t* para los coeficientes** $\boldsymbol{t_{\alpha/2,(n-2)} = \frac{\hat{\beta}_i}{S_{\beta_i}}}$	**(12)**
• **Error estándar del coeficiente** $S_{\beta_i} = \frac{S_{y.x}}{\sqrt{\sum_{i=1}^{n} X^2 - n\bar{X}^2}}$	**(13)**	• **Razón *F* calculada** $\boldsymbol{F_{calculada} = \frac{CMR}{CME} = \frac{SCR/_{g.l.}}{SCE/_{g.l.}}}$	**(14)**

ISBN 978-607-8285-62-4

FÓRMULAS CLAVE. PARTE 2

• **Suma de cuadrados de la Regresión** $$SCR = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} XY - n\bar{Y}^2$$	**(15)**	• **Suma de cuadrados del Error** $$SCE = \sum_{i=1}^{n} Y^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y - \hat{\beta}_1 \sum_{i=1}^{n} XY$$	**(16)**
• **Suma de Cuadrados Total** $$SCT = \sum_{i=1}^{n} Y^2 - n\bar{Y}^2 =$$	**(17)**	• **Intervalo de confianza** $$\boldsymbol{\mu_{Y:X} = \hat{Y}_i \pm t_{\alpha/2:n-2} S_{Y:X}\sqrt{h_i}}$$ $$\boldsymbol{h_i = \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2}}$$	**(18)**
• **Coeficiente de Determinación** $$\boldsymbol{r^2_{Y:X} = \frac{SCR}{SCT} = \frac{\hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} XY - n\bar{Y}^2}{\sum_{i=1}^{n} Y^2 - n\bar{Y}^2}}$$	**(19)**	• **Coeficiente de correlación** $$\boldsymbol{r_{y.x} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}}$$	**(20)**
• **Error ó residual** $$\boldsymbol{\varepsilon_i = Y_i - \hat{Y}_i}$$	**(21)**	• **Residual estandarizado** $$\boldsymbol{SR_i = \frac{\varepsilon_i}{S_{Y:X}\sqrt{1 - h_i}}}$$ $$\boldsymbol{h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2}}$$	**(22)**
• **Estadístico Durbin-Watson** $$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} \varepsilon_i^2}$$	**(23)**	• **Elementos matriz sombrero** $$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2}$$	**(24)**
• **Residual de Student eliminado** $$t_i^* = \frac{\varepsilon_i}{S_{(i)}\sqrt{1 - h_i}}$$	**(25)**	• **Estadístico D_i** $$D_i = \frac{SR_i^2 h_i}{2(1 - h_i)}$$	**(26)**

ISBN 978-607-8285-62-4

USO DE LA CALCULADORA CASIO *fx*-82MS. PARTE 1

USO DE LA CALCULADORA CASIO fx.82MS
FIX, SCI, RND

- Para cambiar los ajustes para el número de lugares decimales, el número de dígitos significantes, o el formato de presentación exponencial, presione varias veces (en general tres veces) la tecla $\boxed{\textbf{MODE}}$ hasta alcanzar la pantalla de ajustes mostrada a continuación.

Fix	Sci	Norm
1	2	3

- Presione la tecla numérica ($\boxed{1}$, $\boxed{2}$ o $\boxed{3}$) que corresponde al elemento de ajuste que sea cambiar.
  $\boxed{1}$ (**Fix**): Número de lugares decimales.
  $\boxed{2}$ (**Sci**): Número de dígitos significativos.
  $\boxed{3}$ (**Norm**): Formato de presentación exponencial.

- Ejemplo: 200 ÷ 7.5 x 14 =

  $200\ \boxed{\div}\ 7.5\ \boxed{x}\ 14\ \boxed{=}\ \boxed{373.3333333}$

  (Especifica cinco lugares decimales) $\boxed{\text{MODE}}$ ... $\boxed{1}$(Fix)$\boxed{5}$ $\boxed{\overbrace{373.33333}^{\text{FIX}}}$

- Presione $\boxed{\textbf{MODE}}$ **...**$\boxed{3}$ (**Norm**) $\boxed{1}$ para borrar la especificación Fix.

Cálculos de regresión

Utilice la tecla $\boxed{\textbf{MODE}}$ para ingresar el modo **REG** cuando desea realizar cálculos estadísticos usando la regresión.

- Presione una vez la tecla $\boxed{\textbf{MODE}}$ hasta alcanzar la pantalla de ajustes mostrada a continuación.

COMP	SD	REG
1	2	3

- En el modo SD y modo REG, la tecla $\boxed{\textbf{M}+}$ opera como la tecla $\boxed{\textbf{DT}}$**.**
- Ingresando el modo REG visualiza pantallas similares a las mostradas a continuación.

Lin	Log	Exp →
1	2	3

- Presione la tecla numérica ($\boxed{1}$, $\boxed{2}$ o $\boxed{3}$) que corresponde al tipo de regresión que desea usar.
  $\boxed{1}$ (**Lin**): Regresión lineal
  $\boxed{2}$ (**Log**): Regresión logarítmica
  $\boxed{3}$ (**Exp**): Regresión exponencial

ISBN 978-607-8285-62-4

USO DE LA CALCULADORA CASIO *fx*-82MS. PARTE 2

- Inicie siempre el ingreso de datos con $\boxed{\textbf{MODE}}\boxed{\textbf{CLR}}\boxed{\textbf{1}}(\textbf{Scl})\boxed{=}$ para borra la memoria estadística.
- Ingrese los datos en forma de pares ordenados usando la secuencia de tecla siguiente. $< datos\ \boldsymbol{X} >$ $\boxed{,}$ $< datos\ \boldsymbol{Y} >$ $\boxed{\textbf{M}+}$. Ingreso los demás pares ordenados en la misma forma.
- Los valores producidos por un cálculo de regresión lineal dependen de los valores ingresados, y los resultados pueden ser vueltos a llamar usando las operaciones de tecla mostrados en la tabla siguiente.

Para llamar este tipo de valor:	**Realice esta operación de tecla:**
$\sum X^2$	$\boxed{\textbf{SHIFT}}\boxed{\textbf{S}-\textbf{SUM}}\boxed{\textbf{1}}$
$\sum X$	$\boxed{\textbf{SHIFT}}\boxed{\textbf{S}-\textbf{SUM}}\boxed{\textbf{2}}$
n	$\boxed{\textbf{SHIFT}}\boxed{\textbf{S}-\textbf{SUM}}\boxed{\textbf{3}}$
$\sum Y^2$	$\boxed{\textbf{SHIFT}}\boxed{\textbf{S}-\textbf{SUM}}\boxed{\textbf{REPLAY}\rightarrow}\boxed{\textbf{1}}$
$\sum Y$	$\boxed{\textbf{SHIFT}}\boxed{\textbf{S}-\textbf{SUM}}\boxed{\textbf{REPLAY}\rightarrow}\boxed{\textbf{2}}$
$\sum XY$	$\boxed{\textbf{SHIFT}}\boxed{\textbf{S}-\textbf{SUM}}\boxed{\textbf{REPLAY}\rightarrow}\boxed{\textbf{3}}$
$\bar{X}$	$\boxed{\textbf{SHIFT}}\boxed{\textbf{S}-\textbf{VAR}}\boxed{\textbf{1}}$
$X\sigma_n$	$\boxed{\textbf{SHIFT}}\boxed{\textbf{S}-\textbf{VAR}}\boxed{\textbf{2}}$
$X\sigma_{n-1}$	$\boxed{\textbf{SHIFT}}\boxed{\textbf{S}-\textbf{VAR}}\boxed{\textbf{3}}$
$\bar{Y}$	$\boxed{\textbf{SHIFT}}\boxed{\textbf{S}-\textbf{VAR}}\boxed{\textbf{REPLAY}\rightarrow}\boxed{\textbf{1}}$
$Y\sigma_n$	$\boxed{\textbf{SHIFT}}\boxed{\textbf{S}-\textbf{VAR}}\boxed{\textbf{REPLAY}\rightarrow}\boxed{\textbf{2}}$
$Y\sigma_{n-1}$	$\boxed{\textbf{SHIFT}}\boxed{\textbf{S}-\textbf{VAR}}\boxed{\textbf{REPLAY}\rightarrow}\boxed{\textbf{3}}$

ISBN 978-607-8285-62-4

USO DE LA CALCULADORA CASIO *fx*-82MS. PARTE 3

Para llamar este tipo de valor:	Realice esta operación de tecla:
Coeficiente de regresión **A** (Ordenada al origen) $\widehat{\beta}_0$	SHIFT S – VAR REPLAY → REPLAY → 1
Coeficiente de regresión **B** (Pendiente de la recta) $\widehat{\beta}_1$	SHIFT S – VAR REPLAY → REPLAY → 2
Coeficiente de correlación **r**	SHIFT S – VAR REPLAY → REPLAY → 3
$\widehat{X}$	SHIFT S – VAR REPLAY → REPLAY → REPLAY → 1
$\widehat{Y}$	SHIFT S – VAR REPLAY → REPLAY → REPLAY → 2

- **Regresión lineal**
- La fórmula de regresión lineal simple es : $y = A + Bx$. ó $\boxed{\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i}$
- **Ejemplo:** Ventas vs. Superficie

Observación No.	Superficie(X)	Ventas (Y)
1	2.15	1.0
2	9.20	3.0
3	6.70	3.0
4	13.50	4.5
5	5.50	2.0
6	12.15	5.0
7	4.80	1.0
8	10.70	4.0
9	3.25	1.5
10	8.25	3.5

ISBN 978-607-8285-62-4

USO DE LA CALCULADORA CASIO *fx*-82MS. PARTE 4

Para los datos anteriores, realice la regresión lineal simple para los coeficientes de la ecuación de regresión lineal simple, **A** (equivalente a $\hat{\beta}_0$), **B** (equivalente a $\hat{\beta}_1$), el coeficiente de correlación ***r*** y el coeficiente de determinación r^2 . Luego, utilice la regresión para estimar el volumen de ventas $\hat{Y}$ para una superficie ***X*** de **10**. **Nota:** Ajuste los resultados a **cinco decimales.**

En el modo **REG**

$\boxed{\mathbf{1\ (Lin)}}$

$\boxed{\mathbf{SHIFT}}\boxed{\mathbf{CLR}}\boxed{\mathbf{1}}(\mathbf{Scl})\boxed{=}$ (para borra la memoria estadística)

$\mathbf{2.15}\boxed{,}\mathbf{1.0}\boxed{\mathbf{M+}}$ $\boxed{\begin{array}{ll} & \mathbf{REG} \\ \mathbf{n} = & \mathbf{1.} \end{array}}$

Cada vez que presiona $\boxed{M+}$ para registrar un ingreso (par ordenado) , el número de dato ingresado (par ordenado) hasta este punto se indica sobre la presentación (valor *n*) .

$\mathbf{9.20}\boxed{,}\mathbf{3.0}\boxed{\mathbf{M+}}\ \mathbf{6.70}\boxed{,}\mathbf{3.0}\boxed{\mathbf{M+}}\ \ldots\ldots\ldots\mathbf{8.25}\boxed{,}\ \mathbf{3.5}\boxed{\mathbf{M+}}$ $\boxed{\begin{array}{ll} & \mathbf{REG} \\ \mathbf{n} = & \mathbf{10.} \end{array}}$

Coeficiente de regresión A= 0.118129074 ≅ 0.11813

$\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-VAR}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{1}}\boxed{=}$

(Especifica cinco lugares decimales) $\boxed{\mathbf{MODE}}\boxed{\mathbf{MODE}}\boxed{\mathbf{MODE}}\boxed{\mathbf{1}}(\mathbf{Fix})\boxed{\mathbf{5}}$ $\boxed{\begin{array}{c}\mathbf{FIX}\\ \mathbf{0.11813}\end{array}}$

Coeficiente de regresión B= 0.35851 $\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-VAR}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{2}}\boxed{=}$

Coeficiente de correlación *r*=0.94894 $\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-VAR}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{REPLAY\rightarrow}}\ \boxed{\mathbf{3}}\boxed{=}$

Coeficiente de determinación $r^2 = 0.90049$

$\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-VAR}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{3}}\boxed{\mathbf{X^2}}\boxed{=}$

Volumen de ventas de $\hat{Y}$ **cuando** $X_0\ es\ 10 = 3.70326$

$\mathbf{10}\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-VAR}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{2}}\boxed{=}$

Error estándar del estimador $S_{Y.X} = 0.48002$

$\boxed{\sqrt{}}\boxed{(}\boxed{(}\boxed{SHIFT}\boxed{S-SUM}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{1}}\boxed{-}\boxed{SHIFT}\boxed{S-VAR}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{1}}\boxed{X}\boxed{SHIFT}$
$\boxed{S-SUM}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{2}}\boxed{-}\boxed{SHIFT}\boxed{S-VAR}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{REPLAY\rightarrow}}\boxed{\mathbf{2}}\boxed{X}\boxed{SHIFT}\boxed{S-SUM}\boxed{\mathbf{REPLAY\rightarrow}}$
$\boxed{\mathbf{3}}\boxed{)}\boxed{\div}\ \mathbf{8}\boxed{)}\boxed{=}$

Error estándar del coeficiente de regresión $S_{\beta_1} = 0.04214$

$\mathbf{0.48002}\boxed{\div}\boxed{\sqrt{}}\boxed{(}\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-SUM}}\boxed{\mathbf{1}}\boxed{-}\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-SUM}}\boxed{\mathbf{3}}\boxed{\mathbf{x}}\boxed{\mathbf{SHIFT}}\boxed{\mathbf{S-VAR}}\boxed{\mathbf{1}}\boxed{\mathbf{X^2}}\boxed{)}\boxed{=}$

y S_{β_i} *se refiere al error estándar del coeficiente de regresión cuya fórmula matemática es:*

$$S_{\beta_i} = \frac{S_{y.x}}{\sqrt{\sum_{i=1}^{n} X^2 - n\bar{X}^2}}$$

ISBN 978-607-8285-62-4

USO DE LA CALCULADORA CASIO *fx*-82MS. PARTE 5

Estadístico $t_{calculada} = 8.51$

$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{\text{REPLAY}\rightarrow}\boxed{\text{REPLAY}\rightarrow}\boxed{2}\boxed{\div}\ 0.04214\ \boxed{=}$

$$t_{calculada(n-2)} = \frac{\hat{\beta}_1}{S_{\beta_1}} = \frac{0.35851}{0.042143} \cong 8.50759 \cong 8.51$$

Suma de cuadrados de la Regresión SCR=16.68162

$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{\text{REPLAY}\rightarrow}\boxed{\text{REPLAY}\rightarrow}\boxed{1}\boxed{\text{X}}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{\text{REPLAY}\rightarrow}\boxed{2}\boxed{+}$
$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{\text{REPLAY}\rightarrow}\boxed{\text{REPLAY}\rightarrow}\boxed{2}\boxed{\text{X}}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{\text{REPLAY}\rightarrow}\boxed{3}\boxed{-}$
$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{3}\boxed{\text{X}}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{\text{REPLAY}\rightarrow}\boxed{1}\boxed{\text{X}^2}\boxed{=}$

$$SCR = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} XY - n\bar{Y}^2 = 0.11813(28.5) + 0.35851(263.7) - 10(2.85)^2 \cong 16.68162$$

Suma de cuadrados de la Regresión SCE=1.84338

$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{\text{REPLAY}\rightarrow}\boxed{1}\boxed{-}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{\text{REPLAY}\rightarrow}\boxed{\text{REPLAY}\rightarrow}\boxed{1}\boxed{\text{X}}$
$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{\text{REPLAY}\rightarrow}\boxed{2}\boxed{-}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{\text{REPLAY}\rightarrow}\boxed{\text{REPLAY}\rightarrow}\boxed{2}\boxed{\text{X}}$
$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{\text{REPLAY}\rightarrow}\boxed{3}\boxed{=}$

$$SCE = \sum_{i=1}^{n} Y^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y - \hat{\beta}_1 \sum_{i=1}^{n} XY = 99.75 - 0.11813(28.5) - 0.35851(263.7) \cong 1.84338$$

Suma de cuadrados total SCT=18.52500

$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{\text{REPLAY}\rightarrow}\boxed{1}\boxed{-}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{3}\boxed{\text{X}}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{\text{REPLAY}\rightarrow}\boxed{1}\boxed{\text{X}^2}\boxed{=}$

$$SCT = \sum_{i=1}^{n} Y^2 - n\bar{Y}^2 = 99.75 - 10(2.85)^2 = 99.75 - 81.225 = 18.525$$

Los elementos de la matriz sombrero , h_i =0.14364

$\boxed{1}\boxed{\div}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{3}\boxed{=}\boxed{+}\boxed{(}\boxed{(}\boxed{\text{X}_0}\boxed{-}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{1}\boxed{)}\boxed{\text{X}^2}\boxed{\div}$
$\boxed{(}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{1}\boxed{-}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{SUM}}\boxed{3}\boxed{\text{X}}\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{1}\boxed{\text{X}^2}\boxed{)}\boxed{=}$

$$h_i = \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} X^2 - n(\bar{X})^2}$$

Residual $\boldsymbol{Y - \hat{Y} = 0.11107}$

$\mathbf{1 - 2.15}\ \boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{\text{REPLAY}\rightarrow}\boxed{\text{REPLAY}\rightarrow}\boxed{\text{REPLAY}\rightarrow}\boxed{2}\boxed{=}$

$\varepsilon_1 = Y_1 - \hat{Y}_1$ donde $\hat{Y}_1 = 0.11813 + 0.35851(2.15) = 0.88893$ entonces,

$$\varepsilon_1 = 1.0 - 0.88893 = 0.11107$$

ISBN 978-607-8285-62-4

ESTADÍSTICA II

CUADERNO DE TRABAJO

CAPÍTULO 3
REGRESIÓN LÍNEAL MÚLTIPLE

Javier Bech Vertti

ISBN 978-607-8285-62-4

ESTADÍSTICA II CAPÍTULO 3

Av. Universidad No. 940
Ciudad Universitaria
C.P. 20131, Aguascalientes, Ags.
http://www.uaa.mx/direcciones/dgdv/editorial/

ISBN 978-607-8285-62-4

Hecho en México / *Made in Mexico*

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

CONTENIDO

CAPÍTULO 3
ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

ISBN 978-607-8285-62-4

ISBN 978-607-8285-62-4

ISBN 978-607-8285-62-4

CAPÍTULO 3. ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

OBJETIVO 3.1 El alumno podrá calcular e interpretar la recta de regresión por el método de mínimos cuadrados.

ANTECEDENTES

CONCEPTOS DE:

Variable aleatoria. Tipos de variable. Variable dependiente. Variable independiente. Ecuación de tendencia lineal. Ordenada al origen. Pendiente de la recta. Relación directa de dos variables. Relación inversa de dos variables. Curva normal. Normal estándar. Estimador de punto. Distribución de probabilidad.

3.1.1 EL MÉTODO DE MÍNIMOS CUADRADOS. COEFICIENTES DE REGRESIÓN

CONCEPTOS BÁSICOS
COEFICIENTES DE REGRESIÓN

El análisis de regresión múltiple es una extensión del análisis de regresión simple que nos **permite utilizar una mayor parte de la información** de que disponemos para **estimar el valor de la variable dependiente** a aplicaciones que implican dos o más variables independientes.

En ocasiones la correlación entre dos variables puede ser insuficiente para determinar una adecuada ecuación de estimación, sin embargo, si

ISBN 978-607-8285-62-4

agregamos los datos de mas variables independientes, podremos obtener **una ecuación de estimación** que describa la relación **con mayor precisión.**

Los pasos de la regresión y correlación lineal múltiple

El **análisis de regresión y correlación múltiple** son un **proceso** que consta de los siguientes **pasos**:

1.- Definir **la ecuación** de regresión múltiple.
2.- Examinar el **error estándar de estimación** para la regresión múltiple.
3.- Probar la **significación de la relación entre la variable dependiente y las variables explicativas.**
4.- Construir **intervalos de confianza** para $\mu_{Y.X} = Y$
5.- Determinar la **contribución de cada variable explicatoria** mediante la comparación de diferentes modelos de regresión mediante el método conocido como **criterio de la prueba parcial *F.***
6.- Calcular el **coeficiente de determinación** para medir la proporción de la variación en la variable dependiente que se explica por las variables independientes en el modelo de regresión múltiple y aplicar el análisis de **correlación lineal múltiple** para medir la fuerza de la asociación en el modelo de regresión lineal múltiple
7.- Determinarlos **coeficientes de determinación parcial** para medir la proporción de la variación en la variable dependiente que se explica por cada variable explicatoria.
8.- Verificar la existencia de **multicolinearidad** para analizar la correlación entre las variables independientes.
9.- Realizar un **diagnóstico de la regresión** mediante el **análisis de los residuales estandarizados** para estudiar **posibles violaciones a las suposiciones** del modelo de regresión.
10.- Realizar un **diagnóstico de la regresión** mediante el **análisis de influencias** para evaluar lo apropiado de un modelo en particular en relación con el **efecto potencial o la "influencia"** de cada punto sobre ese modelo ajustado.

El análisis de regresión genera una ecuación para describir la relación estadística entre uno o más predictores y la variable de respuesta y para predecir nuevas observaciones

El **modelo de regresión lineal múltiple** está dado por la **función:**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

En el caso de dos variables independientes, que se denotan con X_1 y X_2, el modelo algebraico lineal es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Donde:

Y_i = Variable dependiente
X_i = Variables independientes
β_0 = Primer parámetro de la regresión u ordenada al origen

ISBN 978-607-8285-62-4

β_1 =Segundo parámetro de la regresión ó pendiente de ***Y*** con la variable ***X₁***, manteniendo constante la variable ***X₂***
β_2 =Tercer parámetro de la regresión ó pendiente de **Y** con la variable ***X₂***, manteniendo constante la variable ***X₁***
ε_i = Error aleatorio de muestreo en ***Y*** para la observación ***i***

Para **estimar los parámetros de la regresión** se utiliza el **método de mínimos cuadrados**.

El **método de mínimos cuadrados** determina la **ecuación del plano de regresión** minimizando la suma de los cuadrados de las distancias verticales entre los valores reales de *Y* y los valores pronosticados para *Y.*

Así, con base en los **datos muestrales, la ecuación de regresión lineal múltiple** para el caso de dos variables independientes quedaría de la siguiente forma:

$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_{1i} + \hat{\beta}_{2i} X_{2i}$$

Las **ecuaciones normales** para estimar los **parámetros de la regresión múltiple** con **dos variables independientes** son:

$$\sum_{i=1}^{n} Y = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_2$$

$$\sum_{i=1}^{n} X_1 Y = \hat{\beta}_0 \sum_{i=1}^{n} X_1 + \hat{\beta}_1 \sum_{i=1}^{n} X_1^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_1 X_2$$

$$\sum_{i=1}^{n} X_2 Y = \hat{\beta}_0 \sum_{i=1}^{n} X_2 + \hat{\beta}_1 \sum_{i=1}^{n} X_1 X_2 + \hat{\beta}_2 \sum_{i=1}^{n} X_2^2$$

Los **valores** de los **tres coeficientes de regresión**, $\boldsymbol{\hat{\beta}_0, \hat{\beta}_1 y \hat{\beta}_2}$ se pueden obtener solucionando este grupo de ecuaciones simultáneas. En este caso se utilizará **notación matricial** para bosquejar parte de las matemáticas en la que se basa la regresión múltiple dado que en cálculos subsecuentes se requieren datos de la **matriz inversa**.

La regresión generalmente utiliza el método de mínimos cuadrados ordinarios, del cual se obtiene la ecuación al minimizar la suma de los residuos cuadrados.

El punto de partida para el uso de la notación matricial es **el modelo mismo de regresión múltiple**. El **modelo lineal general** relaciona una respuesta ***Y*** con un conjunto de **variables independientes** de la forma

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

ISBN 978-607-8285-62-4

Las **estimaciones** de **mínimos cuadrados** $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_k$ del término constante u ordenada al origen y las pendientes parciales en el modelo lineal general se pueden obtener utilizando **matrices.**

Sea el vector columna ***Y*** de tamaño (*n* x 1)

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

El vector de observaciones de ***Y***, y sea la matriz ***X*** de tamaño *n* x (k+1)

$$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}$$

La **matriz de valores de las variables independientes aumentada con una columna de unos**. La **primera fila de X** contiene un **1** y los valores para las ***k*** variables independientes de la primera observación $\boldsymbol{Y_1}$. La **fila 2** contiene un **1** y los valores para las ***k*** variables independientes de la segunda observación $\boldsymbol{Y_2}$. Análogamente, las otras filas contienen valores para las observaciones restantes.

Para encontrar **las estimaciones mínimo cuadráticas** $\boldsymbol{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k}$ del **término constante u ordenada al origen y las pendientes parciales en el modelo de regresión múltiple** recuerde que el **principio de mínimos cuadrados** incluye elegir las estimaciones que **minimicen las sumas de los cuadrados de los residuos**. Las **ecuaciones normales** que resultan de ello son, en **notación matricial**,

$$(X'X)\hat{\beta} = X'Y$$

Empleo de las ecuaciones al resolver para las constantes

donde

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

es el **vector buscado de de coeficientes estimados**. Suponiendo que la matriz $X'X$ tiene una inversa, **la solución es**

$$\hat{\beta} = (X'X)^{-1}X'Y$$

ISBN 978-607-8285-62-4

donde:

$$(X'X)^{-1} = \frac{1}{|X'X|}\alpha(X'X)'$$

Interpretación de la ecuación

Para la interpretación de los **coeficientes de la regresión lineal múltiple**, la **ordenada en el origen o intersección con el eje Y** es la intersección ***Y***. Es el valor estimado de la variable dependiente $\hat{Y}_i$ cuando las X_i **=0.** En otras palabras, $\hat{\beta}_0$ es el valor estimado de $\hat{Y}_i$ cuando la línea de regresión cruza el eje ***Y*** cuando las ***X*** son ceros y las **pendientes parciales estimadas en el modelo de regresión múltiple** representan el cambio promedio en la variable dependiente $\hat{Y}_i$ para cada cambio de una unidad (ya sea aumento o reducción) en cada una de las variables independientes X_i cuando mantenemos constantes las demás variables independientes.

3.1.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 3.1.1.1 COEFICIENTES DE REGRESIÓN

Samuel Prado, propietario y director general de un establecimiento quiere conocer el comportamiento de las ventas (en miles de pesos) de un equipo de sonido que se expende en el establecimiento. Se percata de que existen muchos factores que podrían ayudarle a explicar las ventas pero piensa que la inversión en publicidad (en miles de pesos) y el precio (en cientos de pesos) son los principales factores determinantes. Samuel ha reunido los datos que se anexan a continuación.

Observación	Ventas Y	Publicidad X_1	Precio X_2
1	33	3	125
2	61	6	115
3	70	10	140
4	82	13	130
5	17	9	145
6	24	6	140
7	75	11	138
8	80	12	127
9	35	4	128
10	20	8	145

ISBN 978-607-8285-62-4

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1$ y $\hat{\beta}_2$:
c) Calcule el volumen de ventas $\hat{Y}$ cuando la inversión en publicidad X_{1i} es de $ 11,000 pesos y el precio del equipo de sonido X_{2i} es de $ 13,500.00 pesos.

Empleo de las ecuaciones al resolver para las constantes

Respuesta al inciso a.

El modelo algebraico lineal es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_{2i} X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

En el caso de dos variables independientes, que se denotan con X_1 y X_2, el modelo algebraico lineal es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_{2i} X_{2i} + \varepsilon_i$$

Con base en los datos muestrales, la ecuación de regresión lineal para el caso de dos variables independientes es:

$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_{1i} + \hat{\beta}_{2i} X_{2i}$$

Las ecuaciones normales para estimar los parámetros de la regresión múltiple con dos variables independientes son:

Ecuaciones normales

$$\sum_{i=1}^{n} Y = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_2$$

$$\sum_{i=1}^{n} X_1 Y = \hat{\beta}_0 \sum_{i=1}^{n} X_1 + \hat{\beta}_1 \sum_{i=1}^{n} X_1^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_1 X_2$$

$$\sum_{i=1}^{n} X_2 Y = \hat{\beta}_0 \sum_{i=1}^{n} X_2 + \hat{\beta}_1 \sum_{i=1}^{n} X_1 X_2 + \hat{\beta}_2 \sum_{i=1}^{n} X_2^2$$

Donde:

$$X = A^{-1}K \quad \text{ó} \quad \hat{\beta} = (X'X)^{-1}X'Y$$

Y

$$(X'X)^{-1} = \frac{1}{|X'X|}\alpha(X'X)'$$

ISBN 978-607-8285-62-4

El punto de partida para el uso de la notación matricial es el modelo mismo de regresión múltiple. Un modelo que relaciona una respuesta Y con un conjunto de variables independientes se la forma

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_{2i} X_{2i} + \cdots + \beta_{ki} X_{ki} + \varepsilon_i$$

Se llama modelo lineal general. Las estimaciones de mínimos cuadrados $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_k$ del término constante y las pendientes parciales en el modelo lineal general se pueden obtener utilizando matrices.

Sea el vector columna ***Y*** de tamaño (*n* x 1)

Vector columna Y.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

El vector de observaciones de ***Y***, y sea la matriz ***X*** de tamaño *n* x (k+1)

$$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}$$

La matriz de valores de las variables independientes aumentada con una columna de unos. La primera fila de X contiene un 1 y los valores para las k variables independientes de la primera observación. La fila 2 contiene un 1 y los valores de las variables independientes para Y_2 o la segunda observación. Análogamente, las otras filas contienen valores para las observaciones restantes.

Para encontrar las estimaciones mínimo cuadráticas $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_k$ del término constante y las pendientes parciales en el modelo de regresión múltiple recuerde que el principio de mínimos cuadrados incluye elegir las estimaciones que minimicen las sumas de los cuadrados de los residuos. Las ecuaciones normales que resultan de ello son, en notación matricial,

$$(X'X)\hat{\beta} = X'Y$$

ISBN 978-607-8285-62-4

Donde

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

es el vector buscado de de coeficientes estimados. Suponiendo que la matriz $X'X$ tiene una inversa, la solución es

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Para los datos anteriores,

Vector Y de tamaño (nx1)

Vector de observaciones de *y*, y sea la matriz X de tamaño *n* x (k+1)

$$\boldsymbol{Y} = \begin{bmatrix} 33 \\ 61 \\ 70 \\ 82 \\ 17 \\ 24 \\ 75 \\ 80 \\ 35 \\ 20 \end{bmatrix} \quad \text{y} \quad \mathbf{X} = \begin{bmatrix} 1 & 3 & 125 \\ 1 & 6 & 115 \\ 1 & 10 & 140 \\ 1 & 13 & 130 \\ 1 & 9 & 145 \\ 1 & 6 & 140 \\ 1 & 11 & 138 \\ 1 & 12 & 127 \\ 1 & 4 & 128 \\ 1 & 8 & 145 \end{bmatrix}$$

Entonces,

Traspuesta de X *X

$$\boldsymbol{X'X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 6 & 10 & 13 & 9 & 6 & 11 & 12 & 4 & 8 \\ 125 & 115 & 140 & 130 & 145 & 140 & 138 & 127 & 128 & 145 \end{bmatrix} \begin{bmatrix} 1 & 3 & 125 \\ 1 & 6 & 115 \\ 1 & 10 & 140 \\ 1 & 13 & 130 \\ 1 & 9 & 145 \\ 1 & 6 & 140 \\ 1 & 11 & 138 \\ 1 & 12 & 127 \\ 1 & 4 & 128 \\ 1 & 8 & 145 \end{bmatrix}$$

$$= \begin{bmatrix} 10 & 82 & 1{,}333 \\ 82 & 776 & 11{,}014 \\ 1{,}333 & 11{,}014 & 178{,}557 \end{bmatrix}$$

ISBN 978-607-8285-62-4

Traspuesta de X*Y

$$\boldsymbol{X'Y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 6 & 10 & 13 & 9 & 6 & 11 & 12 & 4 & 8 \\ 125 & 115 & 140 & 130 & 145 & 140 & 138 & 127 & 128 & 145 \end{bmatrix} \begin{bmatrix} 33 \\ 61 \\ 70 \\ 82 \\ 17 \\ 24 \\ 75 \\ 80 \\ 35 \\ 20 \end{bmatrix} = \begin{bmatrix} 497 \\ 4{,}613 \\ 65{,}315 \end{bmatrix}$$

$$(\boldsymbol{X'X})^{-\mathbf{1}} = \frac{1}{|X'X|}\alpha(X'X)'$$

Para obtener la inversa de $\boldsymbol{X'X}$ calculamos primero el determinante de $\boldsymbol{X'X}$ utilizando la regla se Sarrus, bajando los dos primeros renglones,

Determinante de X´X.

$$|\boldsymbol{X'X}| = \frac{\begin{vmatrix} 10 & 82 & 1{,}333 \\ 82 & 776 & 11{,}014 \\ 1{,}333 & 11{,}014 & 178{,}557 \end{vmatrix}}{\begin{vmatrix} 10 & 82 & 1{,}333 \\ 82 & 776 & 11{,}014 \end{vmatrix}}$$

$$= 10(776)(178{,}557) + 82(11{,}014)(1{,}333) + 1{,}333(82)(11{,}014) - 1{,}333(776)(1{,}333) - 10(11{,}014)(11{,}014) - 82(82)(178{,}557) = \mathbf{829,796}$$

Posteriormente obtenemos la matriz de cofactores de la transpuesta de $\boldsymbol{X'X}$ cuidando los signos de la matriz que en este caso por ser una matriz de 3 x 3

seria: $\begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}$

Matriz de cofactores de la transpuesta de X*X

$$\boldsymbol{\alpha(X'X)'} = \begin{bmatrix} \begin{vmatrix} 776 & 11{,}014 \\ 11.014 & 178{,}557 \end{vmatrix} = 17'252{,}036 & \begin{vmatrix} 82 & 11{,}014 \\ 1{,}333 & 178{,}557 \end{vmatrix} = 39{,}988 & \begin{vmatrix} 82 & 776 \\ 1{,}333 & 11{,}014 \end{vmatrix} = -131{,}260 \\ \begin{vmatrix} 82 & 1{,}333 \\ 11{,}014 & 178{,}557 \end{vmatrix} = 39{,}988 & \begin{vmatrix} 10 & 1{,}333 \\ 1{,}333 & 178{,}557 \end{vmatrix} = 8{,}681 & \begin{vmatrix} 10 & 82 \\ 1{,}333 & 11{,}014 \end{vmatrix} = -834 \\ \begin{vmatrix} 82 & 1{,}333 \\ 776 & 11{,}014 \end{vmatrix} = -131{,}260 & \begin{vmatrix} 10 & 1{,}333 \\ 82 & 11{,}014 \end{vmatrix} = -834 & \begin{vmatrix} 10 & 82 \\ 82 & 776 \end{vmatrix} = 1{,}036 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{17'252,036} & \mathbf{39,988} & \mathbf{-131,260} \\ \mathbf{39,988} & \mathbf{8,681} & \mathbf{-834} \\ \mathbf{-131,260} & \mathbf{-834} & \mathbf{1,036} \end{bmatrix}$$

ISBN 978-607-8285-62-4

Finalmente obtenemos la inversa de $\boldsymbol{X'X}$

Matriz inversa de la transpuesta de X*X

$$(\boldsymbol{X'X})^{-1} = \frac{1}{|X'X|}\alpha(X'X)' = \frac{1}{829{,}796}\begin{bmatrix} 17'252{,}036 & 39{,}988 & -131{,}260 \\ 39{,}988 & 8{,}681 & -834 \\ -131{,}260 & -834 & 1{,}036 \end{bmatrix}$$

$$= \begin{bmatrix} 17'252{,}036/829{,}796 & 39{,}988/829{,}796 & -131{,}260/829{,}796 \\ 39{,}988/829{,}796 & 8{,}681/829{,}796 & -834/829{,}796 \\ -131{,}260/829{,}796 & -834/829{,}796 & 1{,}036/829{,}796 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{20.79070} & \mathbf{0.04819} & \mathbf{-0.15818} \\ \mathbf{0.04819} & \mathbf{0.01046} & \mathbf{-0.00101} \\ \mathbf{-0.15818} & \mathbf{-0.00101} & \mathbf{0.00125} \end{bmatrix}$$

Con la matriz inversa y la matriz de cofactores calculamos los coeficientes de la regresión:

Vector de coeficientes estimados

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$$

$$= \begin{bmatrix} 17'252{,}036/829{,}796 & 39{,}988/829{,}796 & -131{,}260/829{,}796 \\ 39{,}988/829{,}796 & 8{,}681/829{,}796 & -834/829{,}796 \\ -131{,}260/829{,}796 & -834/829{,}796 & 1{,}036/829{,}796 \end{bmatrix}\begin{bmatrix} 497 \\ 4{,}613 \\ 65{,}315 \end{bmatrix}$$

$$= \begin{bmatrix} 223.52438 \\ 6.56400 \\ -1.70780 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

La ecuación de regresión lineal múltiple se puede expresar como

Ecuación predictiva

$$\boxed{\widehat{\boldsymbol{Y}}_{\boldsymbol{i}} = \mathbf{223.52438} + \mathbf{6.56400}\boldsymbol{X_{1i}} - \mathbf{1.70780}\boldsymbol{X_{2i}}}$$

Donde

Y_i = volumen de ventas (en miles de pesos) para la observación *i*.

X_{1i}=Inversión en publicidad (en miles de pesos) para la observación *i*.

X_{2i}= precio del equipo (en cientos de pesos) para la observación *i*.

ISBN 978-607-8285-62-4

Interpretación de la ecuación

Respuesta al inciso b.

La ordenada al origen $\hat{\beta}_0$, calculada como 223.52438, representa el volumen de ventas (en miles de pesos) que se generaría cuando la inversión en publicidad fuera de $ 0.00 pesos y el precio del equipo de sonido fuera de $ 0.00 pesos.

La pendiente de la inversión en publicidad $\hat{\beta}_1$, calculada como 6.56400, significa que para un equipo de sonido con *determinado* precio fijo (constante), el volumen de ventas se incrementará en $ 6.56400 por cada peso de aumento en la inversión en publicidad.

Asimismo la pendiente del precio del equipo de sonido $\hat{\beta}_2$, calculada como -1.70780, significa que para un equipo de sonido con *determinada* inversión fija en publicidad (constante), el volumen de ventas se disminuirá en $ 1.70780 por cada peso de aumento en el precio del equipo de sonido.

Respuesta al inciso c.

Uso de la ecuación de regresión lineal múltiple para hacer la estimación

$$\hat{Y}_{11.135} = 223.5238 + 6.56400(11) - 1.70780(135) = 65.17538$$

Volumen de ventas de $\hat{\boldsymbol{Y}}$ cuando $\boldsymbol{X_1\ es\ 11\ y\ X_2\ es\ 135 = 65.17538}$, es decir, **$ 65,175.80 pesos.**

3.1.1.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 3.1.1.1 COEFICIENTES DE REGRESIÓN

Para el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	11	9	4
2	30	28	14
3	26	17	11
4	20	19	9
5	15	20	8
6	25	24	11
7	35	32	12
8	17	13	21
9	39	36	9
10	45	40	17

ISBN 978-607-8285-62-4

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1$ y $\hat{\beta}_2$:
c) Calcule el valor de $\hat{Y}$ cuando X_{1i} es 22 y X_{2i} es de 10.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es muy importante que primero resuelva el ejercicio en forma manual y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Empleo de las ecuaciones al resolver para las constantes

Respuesta al inciso a.

El modelo algebraico lineal es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_{2i} X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

En el caso de dos variables independientes, que se denotan con X_1 y X_2, el modelo algebraico lineal es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_{2i} X_{2i} + \varepsilon_i$$

Con base en los datos muestrales, la ecuación de regresión lineal para el caso de dos variables independientes es:

$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_{1i} + \hat{\beta}_{2i} X_{2i}$$

Las ecuaciones normales para estimar los parámetros de la regresión múltiple con dos variables independientes son:

Ecuaciones normales

$$\sum_{i=1}^{n} Y = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_2$$

$$\sum_{i=1}^{n} X_1 Y = \hat{\beta}_0 \sum_{i=1}^{n} X_1 + \hat{\beta}_1 \sum_{i=1}^{n} X_1^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_1 X_2$$

$$\sum_{i=1}^{n} X_2 Y = \hat{\beta}_0 \sum_{i=1}^{n} X_2 + \hat{\beta}_1 \sum_{i=1}^{n} X_1 X_2 + \hat{\beta}_2 \sum_{i=1}^{n} X_2^2$$

ISBN 978-607-8285-62-4

Donde:

$$X = A^{-1}K \quad \text{ó} \quad \hat{\beta} = (X'X)^{-1}X'Y$$

Y

$$(X'X)^{-1} = \frac{1}{|X'X|}\alpha(X'X)'$$

El punto de partida para el uso de la notación matricial es el modelo mismo de regresión múltiple. Un modelo que relaciona una respuesta Y con un conjunto de variables independientes se la forma

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_{2i} X_{2i} + \cdots + \beta_{ki} X_{ki} + \varepsilon_i$$

Se llama modelo lineal general. Las estimaciones de mínimos cuadrados $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_k$ del término constante y las pendientes parciales en el modelo lineal general se pueden obtener utilizando matrices.

Sea el vector columna *Y* de tamaño (*n* x 1)

Vector Y de tamaño (nx1)

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

El vector de observaciones de *Y*, y sea la matriz *X* de tamaño *n* x (k+1)

Vector de observaciones de *y*, y sea la matriz X de tamaño *n* x (k+1)

$$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}$$

La matriz de valores de las variables independientes aumentada con una columna de unos. La primera fila de X contiene un 1 y los valores para las k variables independientes de la primera observación. La fila 2 contiene un 1 y los valores de las variables independientes para Y_2 o la segunda observación. Análogamente, las otras filas contienen valores para las observaciones restantes.

Para encontrar las estimaciones mínimo cuadráticas $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_k$ del término constante y las pendientes parciales en el modelo de regresión múltiple recuerde que el principio de mínimos cuadrados incluye elegir las estimaciones que minimicen las sumas de los cuadrados de los

ISBN 978-607-8285-62-4

residuos. Las ecuaciones normales que resultan de ello son, en notación matricial,

$$(X'X)\hat{\beta} = X'Y$$

donde

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

es el vector buscado de de coeficientes estimados. Suponiendo que la matriz $X'X$ tiene una inversa, la solución es

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Para los datos anteriores,

Vector Y de tamaño (nx1)

Vector de observaciones de *y*, y sea la matriz X de tamaño *n* x (k+1)

$Y =$ y $X =$

Entonces,

Transpuesta de X *X

$X'X =$

Transpuesta de X*Y

$X'Y =$

ISBN 978-607-8285-62-4

$$(X'X)^{-1} = \frac{1}{|X'X|}\alpha(X'X)'$$

Para obtener la inversa de $X'X$ calculamos primero el determinante de $X'X$ utilizando la regla se Sarrus, bajando los dos primeros renglones,

Determinante de la transpuesta de X´X.

$|X'X| =$

Posteriormente obtenemos la matriz de cofactores de la transpuesta de $X'X$ cuidando los signos de la matriz que en este caso por ser una matriz de 3 x 3

seria: $\begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}$

Matriz de cofactores de la transpuesta de X*X

$\alpha(X'X)' =$

ISBN 978-607-8285-62-4

Matriz inversa de la transpuesta de X´X.

Finalmente obtenemos la inversa de $X'X$

$$(X'X)^{-1} = \frac{1}{|X'X|}\alpha(X'X)' =$$

Vector de coeficientes estimados

Con la matriz inversa y la matriz de cofactores calculamos los coeficientes de la regresión:

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

ISBN 978-607-8285-62-4

Ecuación predictiva

La ecuación de regresión lineal múltiple se puede expresar como

$\boxed{\hat{Y}_i =}$

Donde

$Y_i =$
$X_{1i} =$
$X_{2i} =$

Interpretación de la ecuación

Respuesta al inciso b.

La ordenada al origen $\hat{\beta}_0$, calculada como _____, representa el valor de la variable dependiente Y que se generaría cuando la variable independiente X_1 fuera cero y el valor de la variable independiente X_2 también fuera cero.

La pendiente de la variable independiente X_1, $\hat{\beta}_1$ calculada como _____, significa que para un conjunto de datos con *determinado* valor de la variable independiente X_2 fijo (constante), el valor de la variable dependiente Y se _____en _____ por cada unidad de medida que se incremente al valor de la variable independiente X_1.

Asimismo la pendiente de la variable independiente X_2, $\hat{\beta}_2$, calculada como _____, significa que para un conjunto de datos con *determinado* valor de la variable independiente X_1 fijo (constante), el valor de la variable independiente X_2 se _____ en _____por cada unidad de medida que se incremente al valor de la variable independiente X_2.

Uso de la ecuación de regresión lineal múltiple para hacer la estimación

Los valores ajustados son estimaciones de puntos de la respuesta media para los valores dados de los predictores, niveles de factores o componentes

Respuesta al inciso c.

$\hat{Y}_{22.10} =$

El valor estimado de la variable dependiente $\hat{Y}$ cuando $X_1\ es\ 22\ y\ X_2\ es\ 10 =$, es decir, _____unidades.

ISBN 978-607-8285-62-4

3.1.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 3.1.1.1
COEFICIENTES DE REGRESIÓN

Si tenemos el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	15	7.2	3.0
2	33	4.2	5.0
3	32	1.3	5.0
4	42	1.4	9.5
5	20	6.3	3.7
6	27	3.6	5.5
7	30	2.1	5.2
8	25	5.0	3.1
9	37	1.5	7.8
10	10	8.5	2.5

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.

b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$:

c) Calcule el valor de $\hat{Y}$ cuando X_{1i} es 4.5 y X_{2i} es de 4.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ISBN 978-607-8285-62-4

Empleo de las ecuaciones al resolver para las constantes

Respuesta al inciso a.

Para los datos anteriores,

Vector Y de tamaño (nx1)

$\boldsymbol{Y} =$ y $\mathbf{X} =$

Vector de observaciones de *y*, y sea la matriz X de tamaño *n* x (k+1)

Entonces,

Transpuesta de X *X

$\boldsymbol{X}'\boldsymbol{X} =$

Transpuesta de X*Y

$\boldsymbol{X}'\boldsymbol{Y} =$

ISBN 978-607-8285-62-4

$$(\boldsymbol{X'X})^{-\mathbf{1}} = \frac{1}{|X'X|}\alpha(X'X)'$$

Para obtener la inversa de $\boldsymbol{X'X}$ calculamos primero el determinante de $\boldsymbol{X'X}$ utilizando la regla se Sarrus, bajando los dos primeros renglones,

Determinante de la transpuesta de X*X

$|\boldsymbol{X'X}| =$

Posteriormente obtenemos la matriz de cofactores de la transpuesta de $\boldsymbol{X'X}$ cuidando los signos de la matriz que en este caso por ser una matriz de 3 x 3

$$\text{seria: } \begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}$$

Matriz de cofactores de la transpuesta de X*X

$\boldsymbol{\alpha}(\boldsymbol{X'X})' =$

ISBN 978-607-8285-62-4

Finalmente obtenemos la inversa de $\boldsymbol{X'X}$

Matriz inversa de la transpuesta de X*X

$$(\boldsymbol{X'X})^{-\mathbf{1}} = \frac{1}{|X'X|}\alpha(X'X)' =$$

Con la matriz inversa y la matriz de cofactores calculamos los coeficientes de la regresión:

Vector de coeficientes estimados

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-\mathbf{1}}\boldsymbol{X'Y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

ISBN 978-607-8285-62-4

Ecuación predictiva

La ecuación de regresión lineal múltiple se puede expresar como

$$\boxed{\hat{\boldsymbol{Y}}_{\boldsymbol{i}} = \qquad\qquad\qquad\qquad}$$

Donde

$Y_i =$
$X_{1i} =$
$X_{2i} =$

Interpretación de la ecuación

<u>Respuesta al inciso b.</u>

Uso de la ecuación de regresión lineal múltiple para hacer la estimación

<u>Respuesta al inciso c.</u>

$$\hat{Y}_{4.5.4} =$$

El valor estimado de la variable dependiente $\hat{\boldsymbol{Y}}$ cuando $\boldsymbol{X_1\ es\ 4.5\ \ y\ X_2\ es\ 4} =$ **,** es decir, **_____unidades.**

ISBN 978-607-8285-62-4

3.1.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 3.1.1 COEFICIENTES DE REGRESIÓN

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

3.1.1.1 **S**uponga que una compañía de productos de consumo quisiera medir la efectividad de la publicidad en radio y televisión y la publicidad en periódicos en la promoción de sus productos. Se seleccionó una muestra aleatoria de 10 ciudades con poblaciones aproximadamente iguales para el estudio durante un periodo de prueba de un mes. Se registraron las ventas (en miles de pesos) durante el mes de prueba, junto con los niveles de gastos en los medios de publicidad, con los resultados siguientes:

Ciudad	Ventas ($000,000) Y	Publicidad en radio y televisión ($000) X_1	Publicidad en periódicos ($000) X_2
1	9.73	0	400
2	6.25	250	250
3	9.71	300	300
4	9.31	350	350
5	11.77	350	350
6	9.82	400	250
7	16.28	450	450
8	13.29	550	250
9	14.36	600	300
10	17.41	650	350

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.

b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$.

c) Calcule las ventas $\hat{Y}$ cuando la publicidad en radio y televisión X_{1i} es de $ 380,000 pesos y la publicidad en periódicos X_{2i} es de $ 280,000 pesos.

ISBN 978-607-8285-62-4

3.1.1.2 El gerente distrital de ventas de de un fabricante de automóviles estudia las ventas de éstos. En forma específica, quiere determinar qué factores influyen en el número de automóviles vendidos en una distribuidora.

Para investigarlo, seleccionó al azar 10 distribuidoras. De éstas, obtiene el número de automóviles vendidos el mes pasado, los minutos de publicidad en radio comprados el mes pasado y el número de vendedores de tiempo completo contratados. La información es la siguiente:

Distribuidora	Automóviles vendidos el mes pasado Y	Publicidad (en minutos) X_1	Fuerza de ventas X_2
1	127	18	10
2	159	22	14
3	144	23	12
4	139	17	12
5	128	16	12
6	180	26	17
7	102	15	7
8	163	24	16
9	106	18	10
10	149	30	11

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.

b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$.

c) Calcule el número de automóviles vendidos el mes pasado $\hat{Y}$ cuando se compra publicidad X_{1i} por 20 minutos y la contratación de vendedores X_{2i} es de 15.

ISBN 978-607-8285-62-4

3.1.1.3 Los siguientes datos representan las calificaciones de estadística para una muestra aleatoria de 10 estudiantes de primer año de determinada institución de enseñanza superior, junto con sus calificaciones en un examen de inteligencia aplicado cuando aún cursaban el último año de secundaria y el número de periodos de clase perdidos por los 10 estudiantes que tomaron el curso de estadística.

Estudiante	Calificación de estadística Y	Calificación del examen X_1	Clases perdidas X_2
1	98	70	5
2	74	50	7
3	87	70	3
4	81	55	4
5	85	65	1
6	90	65	2
7	74	55	4
8	76	55	5
9	91	70	3
10	94	65	2

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.

b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$.

c) Calcule la calificación de estadística $\hat{Y}$ cuando la calificación del examen de inteligencia X_{1i} es de 60 y el número de clases perdidas X_{2i} de 2.

ISBN 978-607-8285-62-4

OBJETIVO 3.2. El alumno podrá calcular e interpretar el error estándar del estimador, probar la significancia entre la variable dependiente e independientes y elaborar e interpretar intervalos de confianza para el verdadero valor de la variable dependiente *Y*.

ANTECEDENTES

CONCEPTOS DE:

Varianza poblacional. Desviación estándar poblacional. Varianza muestral. Desviación estándar de la muestra, Error estándar de la muestra. Tamaño de la muestra.

3.2.1 ERROR ESTÁNDAR DE ESTIMACIÓN

CONCEPTOS BÁSICOS
ERROR ESTÁNDAR DEL ESTIMADOR

El error estándar de la estimación calcula la variación en la respuesta media estimada para un conjunto determinado de valores predictores, niveles

El error estándar describe la variación con respecto a la recta de regresión en el caso de la regresión lineal simple. Este mismo concepto se aplica en la regresión múltiple. Si se tienen **dos variables independientes**, puede pensarse en **la variación respecto a un plano de regresión**. Si hay más de dos variables independientes, no se tiene una interpretación geométrica de la ecuación, **pero el error estándar del estimador sigue siendo una medida del "error" o variabilidad de la predicción**.

El **error estándar del estimador**, proporcionado por el símbolo $S_{Y.12\ldots k}$, se define en forma general para *k* número de variables independientes como:

ISBN 978-607-8285-62-4

de factores o componentes, y se utiliza para generar el intervalo de confianza para la predicción. Cuanto menor sea el error estándar, más precisa será la respuesta media estimada.

$$S_{Y.12\ldots k} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-(k+1)}} = \sqrt{\frac{\sum_{i=1}^{n} Y^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y - \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y - \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y}{n-k-1}}$$

$$= \sqrt{\frac{SCE}{g.l.}} = \sqrt{\frac{Y'Y - \hat{\beta}'(X'Y)}{n-k-1}} = \sqrt{CME}$$

Observe que la ecuación en su estructura es muy parecida a la que utilizamos para la desviación estándar de una muestra.

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Medición de la dispersión alrededor del plano de regresión múltiple: el error estándar de estimación

Para el caso de sólo **dos variables independientes** se puede resumir de la siguiente manera:

$$S_{Y.12} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-3}} = \sqrt{\frac{\sum_{i=1}^{n} Y^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y - \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y - \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y}{n-3}}$$

$$= \sqrt{\frac{SCE}{g.l.}} = \sqrt{\frac{Y'Y - \hat{\beta}'(X'Y)}{n-3}} = \sqrt{CME}$$

3.2.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 3.2.1.1
ERROR ESTÁNDAR DEL ESTIMADOR

Samuel Prado, propietario y director general de un establecimiento quiere conocer el comportamiento de las ventas (en miles de pesos) de un equipo de sonido que se expende en el establecimiento. Se percata de que existen muchos factores que podrían ayudarle a explicar las ventas pero piensa que la inversión en publicidad (en miles de pesos) y el precio (en cientos de pesos) son los principales factores determinantes. Samuel ha reunido los datos que se anexan a continuación.

Observación	Ventas Y	Publicidad X_1	Precio X_2
1	33	3	125
2	61	6	115
3	70	10	140
4	82	13	130
5	17	9	145
6	24	6	140
7	75	11	138
8	80	12	127
9	35	4	128
10	20	8	145

ISBN 978-607-8285-62-4

d) Determine el error estándar del estimador para toda la regresión lineal múltiple

Medición de la dispersión alrededor del plano de regresión múltiple: el error estándar de estimación

Solución al inciso d.

$$S_{Y.12} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-3}} = \sqrt{\frac{\sum_{i=1}^{n} Y^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y - \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y - \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y}{n-k-1}}$$

$$= \sqrt{\frac{SCE}{g.l.}} = \sqrt{\frac{Y'Y - \hat{\beta}'(X'Y)}{n-k-1}} = \sqrt{CME}$$

Para calcular el error estándar del estimador primero es calcular la **SC(residual)=SCE**

$$SCE = Y'Y - \hat{\beta}'(X'Y)$$

$\hat{\beta}$ y $X'Y$ se calcularon, respectivamente, como

Vector de coeficientes

Transpuesta de X*Y

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 223.52438 \\ 6.56400 \\ -1.70780 \end{bmatrix} \quad \text{y} \quad \boldsymbol{X'Y} = \begin{bmatrix} 497 \\ 4{,}613 \\ 65{,}315 \end{bmatrix}$$

Entonces

Transpuesta de Y*Y

$$Y'Y = [33 \quad 61 \quad 70 \quad 82 \quad 17 \quad 24 \quad 75 \quad 80 \quad 35 \quad 20] \begin{bmatrix} 33 \\ 61 \\ 70 \\ 82 \\ 17 \\ 24 \\ 75 \\ 80 \\ 35 \\ 20 \end{bmatrix} = 30{,}949$$

Finalmente

$$\hat{\beta}'(X'Y) = [223.52438 \quad 6.56400 \quad -1.70780] \begin{bmatrix} 497 \\ 4{,}613 \\ 65{,}315 \end{bmatrix} \cong 29{,}826.39186$$

Suma de cuadrados del error

$$SCE = Y'Y - \hat{\beta}'(X'Y) = 30{,}949 - 29{,}826.39186 \cong \mathbf{1{,}122.60814}$$

El error estándar de estimación

$$\boldsymbol{S_{Y.12}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-3}} = \sqrt{\frac{SCE}{n-k-1}} = \sqrt{\frac{1122.60814}{7}} \cong \mathbf{12.66383}$$

ISBN 978-607-8285-62-4

3.2.1.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 3.2.1.1
ERROR ESTÁNDAR DEL ESTIMADOR

Para el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	11	9	4
2	30	28	14
3	26	17	11
4	20	19	9
5	15	20	8
6	25	24	11
7	35	32	12
8	17	13	21
9	39	36	9
10	45	40	17

d) Determine el error estándar del estimador para toda la regresión lineal múltiple.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Medición de la dispersión alrededor del plano de regresión múltiple: el error estándar de estimación

Solución al inciso d.

$$S_{Y.12} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-3}} = \sqrt{\frac{\sum_{i=1}^{n} Y^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y - \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y - \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y}{n-k-1}}$$

$$= \sqrt{\frac{SCE}{g.l.}} = \sqrt{\frac{Y'Y - \hat{\beta}'(X'Y)}{n-k-1}} = \sqrt{CME}$$

Para calcular el error estándar del estimador primero es calcular la **SC(residual)=SCE**

ISBN 978-607-8285-62-4

$$SCE = Y'Y - \hat{\beta}'(X'Y)$$

Vector de coeficientes estimados

$\hat{\beta}$ y $X'Y$ se calcularon, respectivamente, como

Transpuesta de X*Y

$\boldsymbol{\hat{\beta}} =$ y $\boldsymbol{X'Y} =$

Entonces

Transpuesta de Y*Y

$Y'Y =$

Finalmente

$\hat{\beta}'(X'Y) =$

Suma de cuadrados del error

$SCE =$

El error estándar de estimación

$\boldsymbol{S_{Y.12}} =$

ISBN 978-607-8285-62-4

3.2.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 3.2.1.1
ERROR ESTÁNDAR DEL ESTIMADOR

Si tenemos el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	15	7.2	3.0
2	33	4.2	5.0
3	32	1.3	5.0
4	42	1.4	9.5
5	20	6.3	3.7
6	27	3.6	5.5
7	30	2.1	5.2
8	25	5.0	3.1
9	37	1.5	7.8
10	10	8.5	2.5

d) Determine el error estándar de estimación.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ISBN 978-607-8285-62-4

Medición de la dispersión alrededor del plano de regresión múltiple: el error estándar de estimación

<u>Solución al inciso d.</u>

Para calcular el error estándar del estimador primero es calcular la **SC(residual)=SCE**

$$SCE = Y'Y - \hat{\beta}'(X'Y)$$

Vector de coeficientes estimados

$\widehat{\beta}\ y\ X'Y$ se calcularon, respectivamente, como

Transpuesta de X*Y

$\boldsymbol{\widehat{\beta}} =$ y $\boldsymbol{X'Y} =$

Entonces

Transpuesta de Y*Y

$Y'Y =$

Finalmente

$\hat{\beta}'(X'Y) =$

Suma de cuadrados del error

$SCE =$

El error estándar de estimación

$\boldsymbol{S_{Y.12}} =$

ISBN 978-607-8285-62-4

3.2.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 3.2.1 ERROR ESTÁNDAR DEL ESTIMADOR

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

3.2.1.1 **S**uponga que una compañía de productos de consumo quisiera medir la efectividad de la publicidad en radio y televisión y la publicidad en periódicos en la promoción de sus productos. Se seleccionó una muestra aleatoria de 10 ciudades con poblaciones aproximadamente iguales para el estudio durante un periodo de prueba de un mes. Se registraron las ventas (en miles de pesos) durante el mes de prueba, junto con los niveles de gastos en los medios de publicidad, con los resultados siguientes:

Ciudad	Ventas ($000,000) Y	Publicidad en radio y televisión ($000) X_1	Publicidad en periódicos ($000) X_2
1	9.73	0	400
2	6.25	250	250
3	9.71	300	300
4	9.31	350	350
5	11.77	350	350
6	9.82	400	250
7	16.28	450	450
8	13.29	550	250
9	14.36	600	300
10	17.41	650	350

d) Determine el error estándar del estimador para toda la regresión lineal múltiple.

3.2.1.2 **E**l gerente distrital de ventas de de un fabricante de automóviles estudia las ventas de éstos. En forma específica, quiere determinar qué factores influyen en el número de automóviles vendidos en una distribuidora. Para investigarlo, seleccionó al azar 10 distribuidoras. De éstas, obtiene el número de automóviles vendidos el mes pasado, los minutos de publicidad en radio comprados el mes pasado y el número de vendedores de tiempo completo contratados. La información es la siguiente:

ISBN 978-607-8285-62-4

Distribuidora	Automóviles vendidos el mes pasado Y	Publicidad (en minutos) X_1	Fuerza de ventas X_2
1	127	18	10
2	159	22	14
3	144	23	12
4	139	17	12
5	128	16	12
6	180	26	17
7	102	15	7
8	163	24	16
9	106	18	10
10	149	30	11

d) Determine el error estándar del estimador para toda la regresión lineal múltiple

3.2.1.3 Los siguientes datos representan las calificaciones de estadística para una muestra aleatoria de 10 estudiantes de primer año de determinada institución de enseñanza superior, junto con sus calificaciones en un examen de inteligencia aplicado cuando aún cursaban el último año de secundaria y el número de periodos de clase perdidos por los 10 estudiantes que tomaron el curso de estadística.

Estudiante	Calificación de estadística Y	Calificación del examen X_1	Clases perdidas X_2
1	98	70	5
2	74	50	7
3	87	70	3
4	81	55	4
5	85	65	1
6	90	65	2
7	74	55	4
8	76	55	5
9	91	70	3
10	94	65	2

d) Determine el error estándar del estimador para toda la regresión lineal múltiple.

ISBN 978-607-8285-62-4

CONCEPTOS DE:

Variables aleatorias. Variable dependiente. Variable independiente. Población, marco y muestra. Parámetro. La función de probabilidad, Las distribuciones de probabilidad, Características de la forma de una distribución de probabilidad. Prueba de hipótesis. Estructura de las hipótesis nula y alternativa, Error tipo I y tipo II. Distribución t de Student. Prueba t. Nivel de significancia. Distribución *F*. Prueba *F* para la razón de varianzas. Estadístico de prueba. Análisis de Varianza. La significancia observada (valor p). Estimador puntual. Varianza poblacional. Desviación estándar poblacional. Varianza muestral. Desviación estándar de la muestra. Error estándar de la muestra. Estructura de un intervalo de confianza.

3.2.2 PRUEBAS DE SIGNIFICANCIA. RELACIÓN EXISTENTE ENTRE VARIABLES

CONCEPTOS BÁSICOS PRUEBAS DE SIGNIFICANCIA

Significancia de la regresión como un todo.

Una vez ajustado un modelo de regresión a un grupo de datos se debe determinar **si hay relación significativa entre la variable dependiente y el grupo de variables explicatorias**. Las hipótesis se pueden establecer de la siguiente manera:

Juego de hipótesis:

$$H_o: \beta_1 = \beta_2 = 0 \ (no\ existe\ relación)$$
$$H_1: (Por\ lo\ menos\ un\ coeficiente\ de\ regresión\ no\ es\ igual\ a\ cero)$$

Se puede probar la hipótesis nula utilizando una **prueba *F***. Cuando se prueba la significación de los coeficientes de regresión, a la medida del error aleatorio de le conoce como la **varianza del error**, por lo que la prueba ***F*** es la razón de la varianza debida a la regresión dividida entre la varianza del error.

ISBN 978-607-8285-62-4

Los cuadrados medios representan una estimación de la varianza de la población. Se calcula dividiendo la suma correspondiente de los cuadrados entre los grados de libertad.

En regresión, los cuadrados medios se utilizan para determinar si los términos de un modelo son significativos.

TABLA DE ANOVA

Fuente de variación	g.l.	Suma de cuadrados	Cuadrado Medio	Cociente F
Regresión	k	SCR	$CMR = {SCR}/{k}$	$F_{CALC.} = \frac{CMR}{CME}$
Error	$n-k-1$	SCE	$CME = {SCE}/{n-k-1}$	
Total	$n-1$	SCT		

Donde:

n= número de observaciones.
k= número de variables independientes

$$SCR = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y + \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y - n\bar{Y}^2 \ (Variación\ explicada)$$

$$SCE = \sum_{i=1}^{n} Y^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y - \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y - \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y \ (Variación\ NO\ explicada)$$

El cuadrado medio de regresión se obtiene dividiendo la suma de los cuadrados de la regresión entre los grados de libertad.

$$SCT = SCR + SCE = \sum_{i=1}^{n} Y^2 - n\bar{Y}^2 \ (Variación\ Total)$$

En forma matricial:

$$\boldsymbol{SCR = \hat{\beta}'(X'Y) - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} (Variación\ explicada)}$$

$$\boldsymbol{SCE = Y'Y - \hat{\beta}'(X'Y)(Variación\ NO\ explicada)}$$

$$\boldsymbol{SCT = Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} (Variación\ total)}$$

El cuadrado medio del error se obtiene dividiendo la suma de los cuadrados del error entre los grados de libertad. El cuadrado medio del error es la varianza (s^2) alrededor de la regresión ajustada.

La regla de decisión es de rechazar Ho si F calculada es mayor o igual a un valor crítico determinado para alfa de 0.05 y para V_1= k g.l. y V_2 = n-k-1 g.l.

Dividiendo CMR entre CME se obtiene *F*, que sigue la distribución *F* con grados de libertad para la regresión y grados de libertad para el error.

Si la hipótesis nula $\beta_i = 0$ es verdadera, la razón es:

$$F_{calculada} = \frac{CMR}{CME} = \frac{SCR/_{g.l.}}{SCE/_{g.l.}}$$

ISBN 978-607-8285-62-4

En caso de que la prueba anterior haya resultado **significativa o altamente significativa** sólo se ha mostrado que **alguno, pero no necesariamente todos los coeficientes de regresión, no son iguales a cero** y, por tanto, son útiles para las predicciones.

El siguiente paso consiste en utilizar **la prueba *t*** para probar individualmente las variables para determinar **cuáles coeficientes de regresión pueden ser 0 y cuáles no**. Si una β_i puede ser cero, ello implica que esta variable independiente en particular no tiene ningún valor para explicar cualquier variación en el valor dependiente. **Si hay coeficientes para los cuales no se puede rechazar H_0, se pueden eliminar de la ecuación de regresión.**

Prueba de una hipótesis con respecto a β_1.

Prueba de una hipótesis con respecto a β_2.

Las hipótesis se pueden establecer de la siguiente manera para el caso de dos variables independientes:

Juego de hipótesis:

Para la variable independiente X_1:

H_0: $\hat{B}_1 = 0$ (no existe relación)

H_1: $\hat{B}_1 \neq 0$ (existe relación)

Para la variable independiente X_2:

H_0: $\hat{B}_2 = 0$ (no existe relación)

H_1: $\hat{B}_2 \neq 0$ (existe relación)

Podemos probar coeficientes de regresión individuales utilizando la distribución ***t.*** La fórmula es :

$$t_{\alpha/2.n-k-1} = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}}$$

$\boldsymbol{\hat{\beta}_j}$ se refiere a cualquiera de los coeficientes de regresión y $\boldsymbol{S_{\hat{\beta}_j}}$ se refiere al error estándar del coeficiente de regresión cuya fórmula en forma matricial es:

El error estándar del coeficiente de regresión de β_j es decir S_{β_j}.

$$S_{\hat{\beta}_j} = S_{Y.12}\sqrt{v_{jj}}$$

Donde $S_{Y.12}$ es el error estándar del estimador a partir de la ecuación de regresión y v_{jj} es el elemento en la fila $j+1$, columna $j+1$ de $(X'X)^{-1}$

$$(X'X)^{-1} = \begin{bmatrix} v_{00} & & \\ & v_{11} & \\ & & v_{22} \end{bmatrix}$$

Un **segundo y equivalente método** para probar la existencia de una relación lineal entre las variables, es establecer un **estimado de intervalo de confianza de $\boldsymbol{\hat{\beta}_j}$** y determinar si el valor hipotético $(\hat{\beta}_j = 0)$ está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_j$ se obtendría de la siguiente manera:

Intervalo de confianza para β_j.

$$\boldsymbol{\beta_j = \hat{\beta}_j \mp t_{\alpha/2.n-k-1} S_{\hat{\beta}j_1}}$$

ISBN 978-607-8285-62-4

3.2.2.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 3.2.2.1 PRUEBAS DE SIGNIFICANCIA

Samuel Prado, propietario y director general de un establecimiento quiere conocer el comportamiento de las ventas (en miles de pesos) de un equipo de sonido que se expende en el establecimiento. Se percata de que existen muchos factores que podrían ayudarle a explicar las ventas pero piensa que la inversión en publicidad (en miles de pesos) y el precio (en cientos de pesos) son los principales factores determinantes. Samuel ha reunido los datos que se anexan a continuación.

Observación	Ventas Y	Publicidad X_1	Precio X_2
1	33	3	125
2	61	6	115
3	70	10	140
4	82	13	130
5	17	9	145
6	24	6	140
7	75	11	138
8	80	12	127
9	35	4	128
10	20	8	145

e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)

f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.

Significancia de la regresión como un todo.

Solución al inciso e.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

$$\boldsymbol{H_o: \beta_1 = \beta_2 = 0\ (no\ existe\ relación)}$$

$$\boldsymbol{H_1: (Por\ lo\ menos\ un\ coeficiente\ de\ regresión\ no\ es\ igual\ a\ cero)}$$

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

Prueba *F*. Análisis de Varianza.

$$F_{calculada} = \frac{CMR}{CME} = \frac{SCR/_{G.L.}}{SCE/_{G.L.}} = 15.98$$

SCT= SCR+SCE

En forma matricial:

$$\boldsymbol{SCR = \widehat{\beta}'(X'Y) - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} (Variación\ explicada)}$$

$$\boldsymbol{SCE = Y'Y - \widehat{\beta}'(X'Y)(Variación\ NO\ explicada)}$$

$$\boldsymbol{SCT = Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} (Variación\ total)}$$

Suma de cuadrados de la regresión.

$$\boldsymbol{SCR = \widehat{\beta}'(X'Y) - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}}$$

$$\boldsymbol{= [223.52438 \quad 6.56400 \quad -1.70780] \begin{bmatrix} 497 \\ 4,613 \\ 65,315 \end{bmatrix} - \frac{497^2}{10}}$$

$$\boldsymbol{= 29,826.39186 - 24,700.9 \cong 5,125.49186}$$

Suma de cuadrados del error.

$$\boldsymbol{SCE = Y'Y - \widehat{\beta}'(X'Y)}$$

$$\boldsymbol{= [33 \quad 61 \quad 70 \quad 82 \quad 17 \quad 24 \quad 75 \quad 80 \quad 35 \quad 20] \begin{bmatrix} 33 \\ 61 \\ 70 \\ 82 \\ 17 \\ 24 \\ 75 \\ 80 \\ 35 \\ 20 \end{bmatrix}}$$

$$\boldsymbol{- [223.52438 \quad 6.56400 \quad -1.70780] \begin{bmatrix} 497 \\ 4,613 \\ 65,315 \end{bmatrix}}$$

$$\boldsymbol{= 30,949 - 29,826.39186 \cong 1,122.60814}$$

ISBN 978-607-8285-62-4

Suma de cuadrados total.

$$SCT = Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

$$= [33 \quad 61 \quad 70 \quad 82 \quad 17 \quad 24 \quad 75 \quad 80 \quad 35 \quad 20]\begin{bmatrix}33\\61\\70\\82\\17\\24\\75\\80\\35\\20\end{bmatrix} - \frac{497^2}{10}$$

$$= 30,949 - 24,700.9 \cong 6,248.1$$

Los cuadrados medios representan una estimación de la varianza de la población. Se calcula dividiendo la suma correspondiente de los cuadrados entre los grados de libertad.
El cuadrado medio de regresión se obtiene dividiendo la suma de los cuadrados de la regresión entre los grados de libertad.

El cuadrado medio del error se obtiene dividiendo la suma de los cuadrados del error entre los grados de libertad. El cuadrado medio del error es la varianza (s^2) alrededor de la línea de regresión ajustada.

Tabla de Anova

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***F calculada***
Regresión (X_1 y X_2)	*K*=2	*SCR*=5,125.49186	*SCR/G.L.*=2,562.74593	$\frac{CMR}{CME} = \frac{2,562.74593}{160.37259} \cong 15.98$
Error	*n-k-1*= 7	*SCE*=1,122.60814	*SCE/G.L.*=160.37259	
Total	*n-1*=9	*SCT*=6,248.1		

Paso3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

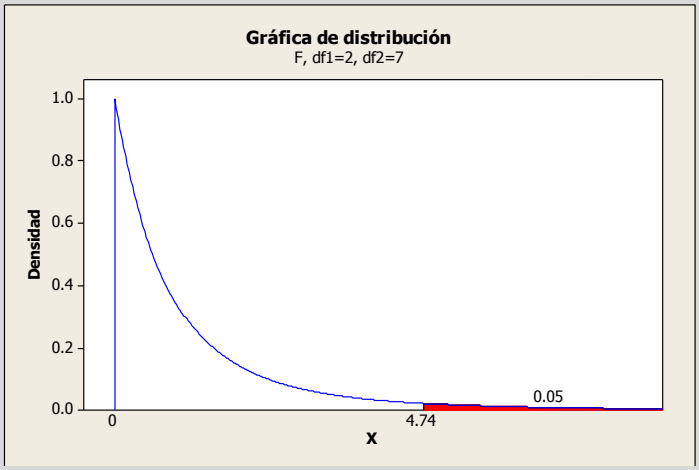
Para determinar la región de rechazo, se necesita el valor crítico. El valor crítico en el estadístico ***F*** se encuentra en las tablas de valores críticos de ***F.*** Para utilizar esta tabla se necesita conocer los grados de libertad en el numerador y en el denominador. Los grados de libertad en el numerador son iguales al número de variables independientes, designados como "k". Los grados de libertad en el denominador son el número de observaciones menos el número de variables independientes menos 1. Para este problema existen dos variables independientes, por lo tanto los grados de libertad en el numerador son*: k= 2 g.l.* y los grados de libertad del denominador para 10 observaciones y dos variables independientes son: *n-k-1=10-2-1=7 g.l.*

ISBN 978-607-8285-62-4

Como existen tablas para niveles de Alfa diferentes, busque la que corresponda al nivel de significancia solicitada, en este caso 0.05, y desplácese horizontalmente sobre la parte superior de la página hasta llegar a los 2 grados de libertad del numerador. Luego descienda en esa columna hasta llegar a la fila que presenta 7 grados de libertad. El valor en esta intersección es **4.74** que en este caso es el valor crítico.

Grados de libertad en el denominador	**Grados de libertad en el numerador**			
	1	**2**	3	4
7	5.59	**4.74**	4.35	4.12
8	5.32	4.46	4.07	3.84
9	5.12	4.26	3.86	3.63
10	4.96	4.10	3.71	3.48
11	4.84	3.98	3.59	3.36
12	4.75	3.89	3.49	3.26
13	4.67	3.81	3.41	3.18
14	4.60	3.74	3.34	3.11

Esta información se presenta en el siguiente diagrama



ISBN 978-607-8285-62-4

Si el *valor p* es menor que o igual al nivel Alfa (α), rechace la hipótesis nula en favor de la hipótesis alternativa. Si $0.01 \leq p \leq 0.05$ se dice que la prueba es significativa (S). Si $p < 0.01$ se dice que la prueba es altamente significativa (AS).

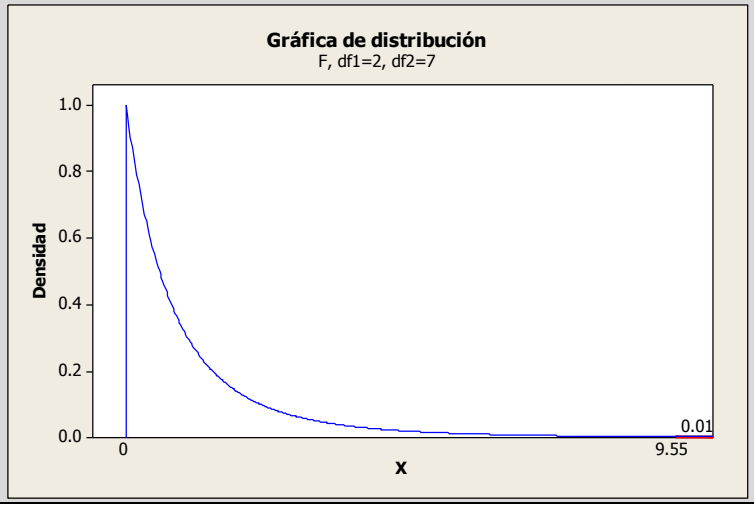
Si el *valor p* es mayor que el nivel Alfa (α), no rechace la hipótesis nula. Si $p > 0.05$ se dice que la prueba es no significativa (NS).

Los niveles de significancia más utilizado son 0.05 y 0.01.

Si se desea aplicar el criterio *p-level* en la conclusión busque en las tablas de valores críticos de ***F*** la que corresponda al nivel de significancia de 0.01, y desplácese horizontalmente sobre la parte superior de la página hasta llegar a los 2 grados de libertad del numerador. Luego descienda en esa columna hasta llegar a la fila que presenta 7 grados de libertad. El valor en esta intersección es **9.55** que en este caso es el valor crítico.

Grados de libertad en el denominador	**Grados de libertad en el numerador**			
	1	**2**	3	4
7	12.2	**9.55**	8.45	7.85
8	11.26	8.65	7.59	7.01
9	10.6	8.02	6.99	6.42
10	10.0	7.56	6.55	5.99
11	9.65	7.21	6.22	5.67
12	9.33	6.93	5.95	5.41
13	9.07	6.70	5.74	5.21
14	8.86	6.51	5.56	5.04

Esta información se presenta en el siguiente diagrama



ISBN 978-607-8285-62-4

Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $f_{cal.} \geq 4.74$

Paso5. Interpretación.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:
$Como\ 15.98 > 4.74\ y\ > 9.55\ \therefore la\ prueba\ es\ (AS)\ y\ se\ rechaza H_0$.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación entre el volumen de ventas y al menos una de las variables independientes, ya sea la inversión en publicidad ó el precio del equipo de sonido.

¿Es significativa una variable explicatoria?.

<u>Solución al inciso f.</u>

Hasta este punto se ha mostrado que alguno, pero no necesariamente todos los coeficientes de regresión, no son iguales a cero y, por tanto, son útiles para las predicciones. El siguiente paso consiste en probar individualmente las variables para determinar cuáles coeficientes de regresión pueden ser 0 y cuáles no. Si una β puede ser cero, ello implica que esta variable independiente en particular no tiene ningún valor para explicar cualquier variación en el valor dependiente. Si hay coeficientes para los cuales no se puede rechazar H_0, se pueden eliminar de la ecuación de regresión.

Ahora se realizarán dos pruebas de hipótesis: para la inversión en publicidad y para el precio del equipo de sonido.

Prueba de una hipótesis con respecto a β_1.

1. Prueba para la inversión en publicidad.

La distribución t se puede utilizar para realizar pruebas de significancia y construir intervalos de confianza para la verdadera pendiente

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

$$H_o: \beta_1 = 0\ (\ no\ existe\ relación)$$
$$H_1: \beta_1 \neq 0\ (\ existe\ relación)$$

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$t_{calculada} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

$$S_{\hat{\beta}_1} \textit{ es la desviación estándar de la distribución muestral del coeficiente de regresión neta de la variable independiente inversión en publicidad}$$

En forma matricial

$$S_{\hat{\beta}_j} = S_{Y.12}\sqrt{v_{jj}}$$

La matriz inversa de la transpuesta de X*X, cuando se multiplica por el error estándar del estimador, es la matriz de varianza-covarianza de los coeficientes.

Donde $S_{Y.12}$ es el error estándar del estimador a partir de la ecuación de regresión y v_{jj} es el elemento en la fila $j+1$, columna $j+1$ de $(X'X)^{-1}$

La matriz de varianza – covarianza es una matriz cuadrada que contiene las varianzas y covarianzas asociadas a diversas variables. Los elementos de la diagonal de la matriz contienen las varianzas de las variables, y los elementos que se encuentran fuera de la diagonal contienen las covarianzas entre todos los pares posibles de variables.

$$(X'X)^{-1} = \begin{bmatrix} v_{00} & & \\ & v_{11} & \\ & & v_{22} \end{bmatrix}$$

Dado que para obtener los coeficientes $\hat{\beta}_j$ se calculó anteriormente la matriz $(X'X)^{-1}$, entonces

$$(X'X)^{-1} = \begin{bmatrix} 20.79070 & & \\ & 0.01046 & \\ & & 0.00125 \end{bmatrix}$$

El error estándar del coeficiente de regresión de β_1 es decir S_{β_1}

$$\boldsymbol{S_{\hat{\beta}_1}} = S_{Y.12}\sqrt{v_{11}} = 12.66383\sqrt{0.01046} = \mathbf{1.29502}$$

$$\boldsymbol{t_{calculada}} = \frac{\hat{\beta}_1}{S_{\beta_1}} = \frac{6.564}{1.29502} \cong 5.06865 \cong \mathbf{5.07}$$

Paso3. Región de rechazo.

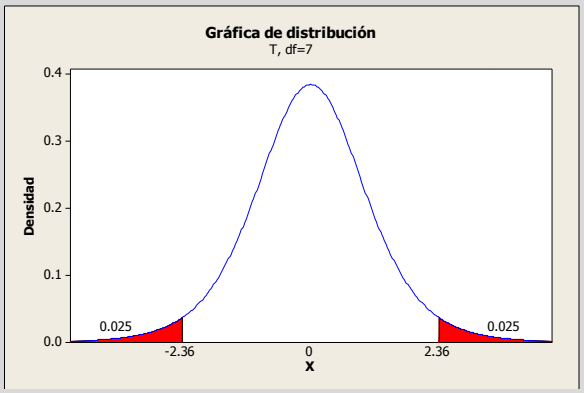
Paso 3.- Establecer la región de rechazo de (H_0).

La hipótesis alternativa no indica una dirección por lo que esta es una prueba de dos colas. Hay 7 grados de libertad, obtenidos de n-k-1 = 10-2-1=7. El valor de ***t*** es **2.3646 ó 2.36** que se obtiene buscando en la tabla de valores críticos de ***t*** bajo prueba de dos colas, usando .05 como nivel de significancia y por tanto 0.025 como área de la cola superior, con 7 grados de libertad de la siguiente manera:

ISBN 978-607-8285-62-4

Grados de libertad	Áreas de la cola superior					
	.25	.10	.05	.025	.01	.005
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	**2.3646**	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1993

Esta información se presenta en el siguiente diagrama



Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $t_{calculada} \leq -2.36$ *ó* ≥ 2.36

Paso5. Interpretación.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: como 5.07 > 2.36, se rechaza H_0.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación lineal significativa entre el volumen de ventas y la inversión en publicidad, es decir se concluye que el coeficiente de regresión no es cero. La variable independiente "inversión en publicidad" debe incluirse en el análisis.

ISBN 978-607-8285-62-4

Intervalo de confianza para β_1.

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de $\hat{\beta}_1$ y determinar si el valor hipotético $(\hat{\beta}_1 = 0)$ está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_1$ se obtendría de la siguiente manera:

$$\boldsymbol{\beta_1} = \boldsymbol{\hat{\beta}_1} \pm \boldsymbol{t_{n-k-1} S_{\hat{\beta}_1}}$$

$$\beta_1 = 6.564 \mp 2.36(1.29502)$$

Obtención de los limites superior e inferior de la región de no rechazo de H_0.

$$\boldsymbol{\beta_1} = 6.564 \mp 3.05625 \begin{cases} LIC = 6.564 - 3.05625 = \mathbf{3.50} \\ LSC = 6.564 + 3.05625 = \mathbf{9.62} \end{cases}$$

Otra forma de expresar el intervalo de confianza es:

$$\boldsymbol{\hat{\beta}_1 - t_{n-k-1} S_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_{1+} t_{n-k-1} S_{\hat{\beta}_1}}$$

$$6.564 - 2.36(1.29502) \leq \beta_1 \leq 6.564 + 2.36(1.29502)$$

$$6.564 - 3.05625 \leq \beta_1 \leq 6.564 + 3.05625$$

$$3.50 \leq \boldsymbol{\beta_1} \leq 9.62$$

Interpretación

Interpretación: Se estima, con una confianza del 95%, que la pendiente real se encuentra entre 3.50 y 9.62. Puesto que estos valores son superiores a cero se puede concluir que hay relación lineal significativa entre el volumen de ventas y la inversión en publicidad. La variable independiente "inversión en publicidad" debe incluirse en el análisis.

Prueba de una hipótesis con respecto a β_2.

2. Prueba para el precio del equipo de sonido

La distribución t se puede utilizar para realizar pruebas de significancia y construir intervalos de confianza para la verdadera pendiente

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

$$H_o: \beta_2 = 0 \ (\ no\ existe\ relación)$$
$$H_1: \beta_2 \neq 0 \ (\ existe\ relación)$$

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$t_{calculada} = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}}$$

$$S_{\hat{\beta}_2} \text{ es la desviación estándar de la distribución muestral del coeficiente de regresión neta de la variable independiente precio del equipo de sonido}$$

La matriz inversa de la transpuesta de X*X, cuando se multiplica por el error estándar del estimador, es la matriz de varianza-covarianza de los coeficientes.

En forma matricial

$$S_{\hat{\beta}_j} = S_{Y.12}\sqrt{v_{jj}}$$

Donde $S_{Y.12}$ es el error estándar del estimador a partir de la ecuación de regresión y v_{jj} es el elemento en la fila $j+1$, columna $j+1$ de $(X'X)^{-1}$

La matriz de varianza – covarianza es una matriz cuadrada que contiene las varianzas y covarianzas asociadas a diversas variables. Los elementos de la diagonal de la matriz contienen las varianzas de las variables, y los elementos que se encuentran fuera de la diagonal contienen las covarianzas entre todos los pares posibles de variables.

$$(X'X)^{-1} = \begin{bmatrix} v_{00} & & \\ & v_{11} & \\ & & v_{22} \end{bmatrix}$$

Dado que para obtener los coeficientes $\hat{\beta}_j$ se calculó anteriormente la matriz $(X'X)^{-1}$, entonces

$$(X'X)^{-1} = \begin{bmatrix} 20.79070 & & \\ & 0.01046 & \\ & & 0.00125 \end{bmatrix}$$

$$\boldsymbol{S_{\hat{\beta}_2}} = S_{Y.12}\sqrt{v_{22}} = 12.66383\sqrt{0.00125} \cong \mathbf{0.44773}$$

$$\boldsymbol{t_{calculada}} = \frac{\hat{\beta}_2}{S_{\beta_2}} = \frac{-1.70780}{0.44773} \cong -3.81432 \cong \mathbf{-3.81}$$

Paso3. Región de rechazo.

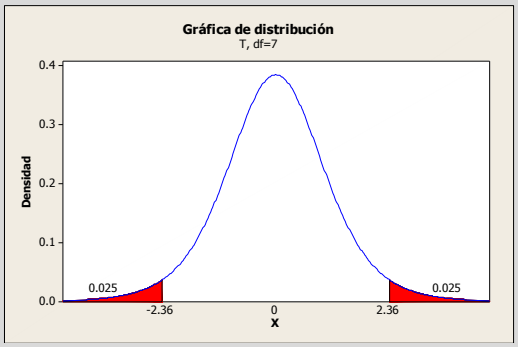
Paso 3.- Establecer la región de rechazo de (H_0).

La hipótesis alternativa no indica una dirección por lo que esta es una prueba de dos colas. Hay 7 grados de libertad, obtenidos de n-k-1 = 10-2-1=7. El valor de ***t*** es **2.3646 ó 2.36** que se obtiene buscando en la tabla de valores críticos de ***t*** bajo prueba de dos colas, usando .05 como nivel de significancia y por tanto 0.025 como área de la cola superior, con 7 grados de libertad de la siguiente manera:

ISBN 978-607-8285-62-4

Grados de libertad	Áreas de la cola superior					
	.25	.10	.05	.025	.01	.005
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	**2.3646**	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1993

Esta información se presenta en el siguiente diagrama



Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $t_{calculada} \leq -2.36$ *ó* ≥ 2.36

Paso5. Interpretación.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: como -3.81< -2.36, se rechaza H_0.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación lineal significativa entre el volumen de ventas y el precio del equipo de sonido, es decir se concluye que el coeficiente de regresión no es cero. La variable independiente "precio del equipo de sonido" debe incluirse en el análisis.

ISBN 978-607-8285-62-4

Intervalo de confianza para β_2.

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de $\hat{\beta}_2$ y determinar si el valor hipotético ($\hat{\beta}_2 = 0$) está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_1$ se obtendría de la siguiente manera:

$$\boldsymbol{\beta_2 = \widehat{\beta}_2 \mp t_{n-k-1} S_{\widehat{\beta}_2}}$$

$$\beta_2 = -1.7078 \mp 2.36(0.44773)$$

Obtención de los limites superior e inferior de la región de no rechazo de H_0.

$$\boldsymbol{\beta_2} = -1.7078 \mp 1.05664 \begin{cases} LIC = -1.7078 - 1.05664 = \mathbf{-2.76} \\ LSC = -1.7078 + 1.05664 = \mathbf{-0.65} \end{cases}$$

Otra forma de expresar el intervalo de confianza es:

$$\boldsymbol{\widehat{\beta}_2 - t_{n-k-1} S_{\widehat{\beta}_2} \leq \beta_2 \leq \widehat{\beta}_{2+} t_{n-k-1} S_{\widehat{\beta}_2}}$$

$$-1.7078 - 2.36(0.44773) \leq \beta_2 \leq -1.7078 + 2.36(0.44773)$$

$$-1.7078 - 1.05664 \leq \beta_2 \leq -1.7078 + 1.05664$$

$$\boldsymbol{-2.76 \leq \beta_2 \leq -0.65}$$

Interpretación

Interpretación: Se estima, con una confianza del 95%, que la pendiente real se encuentra entre -2.76 y -0.65. Puesto que estos valores son inferiores a cero se puede concluir que hay relación lineal significativa entre el volumen de ventas y el precio del equipo de sonido. La variable independiente "precio del equipo de sonido" debe incluirse en el análisis.

ISBN 978-607-8285-62-4

3.2.2.1 ACTIVIDAD DE APRENDIZAJE

ACTIVIDAD DE APRENDIZAJE 3.2.2.1 PRUEBAS DE SIGNIFICANCIA

Para el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	11	9	4
2	30	28	14
3	26	17	11
4	20	19	9
5	15	20	8
6	25	24	11
7	35	32	12
8	17	13	21
9	39	36	9
10	45	40	17

e) Prueba la significancia de la relación entre la variable dependiente (Y) y las variables explicatorias (independientes)

f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Significancia de la regresión como un todo

Solución al inciso e.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

$\boldsymbol{H_o}$:

$\boldsymbol{H_1}$:

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

Prueba *F*. Análisis de Varianza.

$$F_{calculada} = \frac{CMR}{CME} = \frac{SCR/_{G.L.}}{SCE/_{G.L.}} =$$

SCT= SCR+SCE

En forma matricial:

$$\boldsymbol{SCR = \widehat{\beta}'(X'Y) - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} (Variación\ explicada)}$$
$$\boldsymbol{SCE = Y'Y - \widehat{\beta}'(X'Y)(Variación\ NO\ explicada)}$$
$$\boldsymbol{SCT = Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} (Variación\ total)}$$

Suma de cuadrados de la regresión

$$\boldsymbol{SCR = \widehat{\beta}'(X'Y) - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}} =$$

Suma de cuadrados del error

$$\boldsymbol{SCE = Y'Y - \widehat{\beta}'(X'Y)} =$$

Suma de cuadrados total

$$\boldsymbol{SCT = Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}} =$$

ISBN 978-607-8285-62-4

Tabla de ANOVA

Tabla de Anova

El cuadrado medio de regresión se obtiene dividiendo la suma de los cuadrados de la regresión entre los grados de libertad.

El cuadrado medio del error se obtiene dividiendo la suma de los cuadrados del error entre los grados de libertad. El cuadrado medio del error es la varianza (s^2) alrededor de la línea de regresión ajustada.

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada}$***
Regresión (X_1 y X_2)	*k=*	*SCR=*	*SCR/G.L.=*	$\frac{CMR}{CME}=$
Error	*n-k-1=*	*SCE=*	*SCE/G.L.=*	
Total	*n-1=*	*SCT=*		

Paso3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso5. Interpretación.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

ISBN 978-607-8285-62-4

¿Es significativa una variable explicatoria?.

Solución al inciso f.

Ahora se realizarán dos pruebas de hipótesis: para la variable independiente $\mathbf{X_1}$ y para la variable independiente $\mathbf{X_2}$

Prueba de una hipótesis con respecto a β_1.

PRUEBA PARA LA VARIABLE INDEPENDIENTE X_1

La distribución t se puede utilizar para realizar pruebas de significancia y construir intervalos de confianza para la verdadera pendiente.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0:

$H_{1:}$

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$t_{calculada} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

$$S_{\hat{\beta}_1} \; es\ la\ desviación\ estándar\ de\ la\ distribución\ muestral\ del\ coeficiente\ de\ regresión\ neta\ de\ la\ variable\ independiente\ inversión\ en\ publicidad$$

La matriz inversa de la transpuesta de X*X, cuando se multiplica por el error estándar del estimador, es la matriz de varianza-covarianza de los coeficientes.

En forma matricial

$$S_{\hat{\beta}_j} = S_{Y.12}\sqrt{v_{jj}}$$

Muchas aplicaciones estadísticas calculan la matriz de varianza-covarianza para los estimadores de parámetros en un modelo estadístico. Se utiliza con frecuencia para calcular los errores estándar de los estimadores o funciones de los estimadores. Por ejemplo, la regresión logística crea esta matriz para los coeficientes estimados, lo que permite ver las varianzas de los coeficientes y las covarianzas entre todos los pares posibles de coeficientes.

Donde $S_{Y.12}$ es el error estándar del estimador a partir de la ecuación de regresión y v_{jj} es el elemento en la fila $j+1$, columna $j+1$ de $(X'X)^{-1}$

$$(X'X)^{-1} = \begin{bmatrix} v_{00} & & \\ & v_{11} & \\ & & v_{22} \end{bmatrix}$$

Dado que para obtener los coeficientes $\hat{\beta}_j$ se calculó anteriormente la matriz $(X'X)^{-1}$, entonces

$$(X'X)^{-1} =$$

ISBN 978-607-8285-62-4

El error estándar del coeficiente de regresión de β_1 es decir S_{β_1}

$$\boldsymbol{S_{\hat{\beta}_1}} = S_{Y.12}\sqrt{v_{11}} =$$

Estadístico de prueba t

$$\boldsymbol{t_{calculada}} = \frac{\hat{\beta}_1}{S_{\beta_1}} =$$

Paso3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso5. Interpretación.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

Intervalo de confianza para β_1.

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de $\hat{\beta}_1$ y determinar si el valor hipotético $(\hat{\beta}_1 = 0)$ está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_1$ se obtendría de la siguiente manera:

$$\boldsymbol{\beta_1} = \boldsymbol{\hat{\beta}_1} \pm \boldsymbol{t_{n-k-1}S_{\hat{\beta}_1}}$$

$$\beta_1 =$$

Obtención de los limites superior e inferior de la región de no rechazo de H_0.

$$\boldsymbol{\beta_1} = \qquad \begin{cases} LIC = \\ LSC = \end{cases}$$

ISBN 978-607-8285-62-4

Otra forma de expresar el intervalo de confianza es:

$$\boldsymbol{\hat{\beta}_1 - t_{n-k-1} S_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_{1+} t_{n-k-1} S_{\hat{\beta}_1}}$$

$$\leq \boldsymbol{\beta_1} \leq$$

Interpretación.

Interpretación:

PRUEBA PARA LA VARIABLE INDEPENDIENTE X_2

Prueba de una hipótesis con respecto a $\beta_{2.}$

La distribución t se puede utilizar para realizar pruebas de significancia y construir intervalos de confianza para la verdadera pendiente.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0:

$H_{1:}$

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

La matriz inversa de la transpuesta de X*X, cuando se multiplica por el error estándar del estimador, es la matriz de varianza-covarianza de los coeficientes.

$$t_{calculada} = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}}$$

$$S_{\hat{\beta}_2} \; es\ la\ desviación\ estándar\ de\ la\ distribución\ muestral\ del\ coeficiente\ de\ regresión\ neta\ de\ la\ variable\ independiente\ precio\ del\ equipo\ de\ sonido$$

En forma matricial

$$S_{\hat{\beta}_j} = S_{Y.12}\sqrt{v_{jj}}$$

La matriz de varianza – covarianza es una matriz cuadrada que contiene las varianzas y covarianzas asociadas a diversas variables.

ISBN 978-607-8285-62-4

Los elementos de la diagonal de la matriz contienen las varianzas de las variables, y los elementos que se encuentran fuera de la diagonal contienen las covarianzas entre todos los pares posibles de variables.

Donde $S_{Y.12}$ es el error estándar del estimador a partir de la ecuación de regresión y v_{jj} es el elemento en la fila $j+1$, columna $j+1$ de $(X'X)^{-1}$

$$(X'X)^{-1} = \begin{bmatrix} v_{00} & & \\ & v_{11} & \\ & & v_{22} \end{bmatrix}$$

Dado que para obtener los coeficientes $\hat{\beta}_j$ se calculó anteriormente la matriz $(X'X)^{-1}$, entonces

$$(X'X)^{-1} =$$

El error estándar del coeficiente de regresión de β_2 es decir S_{β_2}.

$$\boldsymbol{S_{\hat{\beta}_2}} = S_{Y.12}\sqrt{v_{22}} =$$

$$\boldsymbol{t_{calculada}} = \frac{\hat{\beta}_2}{S_{\beta_2}} =$$

Paso3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso5. Interpretación.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

ISBN 978-607-8285-62-4

Administrativa:

Intervalo de confianza para β_2.

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de $\hat{\beta}_2$ y determinar si el valor hipotético $(\hat{\beta}_2 = 0)$ está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_1$ se obtendría de la siguiente manera:

Obtención de los limites superior e inferior de la región de no rechazo de H_0.

$$\boldsymbol{\beta_2} = \boldsymbol{\widehat{\beta}_2 \mp t_{n-k-1} S_{\widehat{\beta}_2}}$$

$$\beta_2 =$$

$$\boldsymbol{\beta_2} = \qquad \begin{cases} LIC = \\ LSC = \end{cases}$$

Otra forma de expresar el intervalo de confianza es:

$$\boldsymbol{\widehat{\beta}_2 - t_{n-k-1} S_{\widehat{\beta}_2} \leq \beta_2 \leq \widehat{\beta}_{2+} t_{n-k-1} S_{\widehat{\beta}_2}}$$

$$\leq \boldsymbol{\beta_2} \leq$$

Interpretación

Interpretación:

ISBN 978-607-8285-62-4

3.2.2.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 3.2.2.1
PRUEBAS DE SIGNIFICANCIA

Si tenemos el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	15	7.2	3.0
2	33	4.2	5.0
3	32	1.3	5.0
4	42	1.4	9.5
5	20	6.3	3.7
6	27	3.6	5.5
7	30	2.1	5.2
8	25	5.0	3.1
9	37	1.5	7.8
10	10	8.5	2.5

e) Prueba la significancia de la relación entre la variable dependiente (Y) y las variables explicatorias (independientes)

f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ISBN 978-607-8285-62-4

Significancia de la regresión como un todo.

Solución al inciso e.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

$\boldsymbol{H_o}$:
$\boldsymbol{H_1}$:

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

Prueba F. Análisis de Varianza.

$$F_{calculada} =$$

SCT= SCR+SCE

Suma de cuadrados de la regresión.

$\boldsymbol{SCR} =$

Suma de cuadrados del error.

$\boldsymbol{SCE} =$

Suma de cuadrados total.

$\boldsymbol{SCT} =$

ISBN 978-607-8285-62-4

Tabla de ANOVA

Tabla de Anova

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada}$***
Regresión (X_1 y X_2)				
Error				
Total				

Paso3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso5. Interpretación.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

¿Es significativa una variable explicatoria?

<u>Solución al inciso f.</u>

Ahora se realizarán dos pruebas de hipótesis: para la variable independiente $\mathbf{X_1}$ y para la variable independiente $\mathbf{X_2}$

ISBN 978-607-8285-62-4

Prueba de una hipótesis con respecto a β_1

PRUEBA PARA LA VARIABLE INDEPENDIENTE X_1

La distribución t se puede utilizar para realizar pruebas de significancia y construir intervalos de confianza para la verdadera pendiente

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0:

$H_{1:}$

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

Dado que para obtener los coeficientes $\hat{\beta}_j$ se calculó anteriormente la matriz $(X'X)^{-1}$, entonces

$$(X'X)^{-1} =$$

El error estándar del coeficiente de regresión de β_1 es decir S_{β_1}.

$$\boldsymbol{S_{\hat{\beta}_1}} = S_{Y.12}\sqrt{v_{11}} =$$

Estadístico de prueba t

$$\boldsymbol{t_{calculada}} = \frac{\hat{\beta}_1}{S_{\beta_1}} =$$

Paso3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso5. Interpretación.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

ISBN 978-607-8285-62-4

Estadística:

Administrativa:

Intervalo de confianza para β_1.

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de $\hat{\beta}_1$ y determinar si el valor hipotético $(\hat{\beta}_1 = 0)$ está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_1$ se obtendría de la siguiente manera:

$$\boldsymbol{\beta_1} = \boldsymbol{\hat{\beta}_1 \pm t_{n-k-1} S_{\hat{\beta}_1}}$$

$$\beta_1 =$$

Obtención de los limites superior e inferior de la región de no rechazo de H_0.

$$\boldsymbol{\beta_1} = \qquad \begin{cases} LIC = \\ LSC = \end{cases}$$

Otra forma de expresar el intervalo de confianza es:

$$\boldsymbol{\hat{\beta}_1 - t_{n-k-1} S_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_{1+} t_{n-k-1} S_{\hat{\beta}_1}}$$

$$\leq \boldsymbol{\beta_1} \leq$$

Interpretación.

Interpretación:

Prueba de una hipótesis con respecto a β_2.

PRUEBA PARA LA VARIABLE INDEPENDIENTE X_2

La distribución t se puede utilizar para realizar pruebas de significancia y construir intervalos de confianza para la verdadera pendiente.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0:
$H_{1:}$

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

Dado que para obtener los coeficientes $\hat{\beta}_j$ se calculó anteriormente la matriz $(X'X)^{-1}$, entonces

$$(X'X)^{-1} =$$

$$\boldsymbol{S_{\hat{\beta}_2}} = S_{Y.12}\sqrt{v_{22}} =$$

$$\boldsymbol{t_{calculada}} = \frac{\hat{\beta}_2}{S_{\beta_2}} =$$

Paso3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso5. Interpretación.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

ISBN 978-607-8285-62-4

Intervalo de confianza para β_2.

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de $\hat{\beta}_2$ y determinar si el valor hipotético ($\hat{\beta}_2 = 0$) está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_1$ se obtendría de la siguiente manera:

$$\boldsymbol{\beta_2 = \widehat{\beta}_2 \mp t_{n-k-1} S_{\widehat{\beta}_2}}$$

$$\beta_2 =$$

Obtención de los limites superior e inferior de la región de no rechazo de H_0.

$$\boldsymbol{\beta_2 =} \qquad \begin{cases} LIC = \\ LSC = \end{cases}$$

Otra forma de expresar el intervalo de confianza es:

$$\boldsymbol{\widehat{\beta}_2 - t_{n-k-1} S_{\widehat{\beta}_2} \leq \beta_2 \leq \widehat{\beta}_{2+} t_{n-k-1} S_{\widehat{\beta}_2}}$$

$$\leq \boldsymbol{\beta_2} \leq$$

Interpretación

Interpretación:

ISBN 978-607-8285-62-4

3.2.2 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 3.2.2 PRUEBAS DE SIGNIFICANCIA

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

3.2.2.1 **S**uponga que una compañía de productos de consumo quisiera medir la efectividad de la publicidad en radio y televisión y la publicidad en periódicos en la promoción de sus productos. Se seleccionó una muestra aleatoria de 10 ciudades con poblaciones aproximadamente iguales para el estudio durante un periodo de prueba de un mes. Se registraron las ventas (en miles de pesos) durante el mes de prueba, junto con los niveles de gastos en los medios de publicidad, con los resultados siguientes:

Ciudad	Ventas ($000,000) Y	Publicidad en radio y televisión ($000) X_1	Publicidad en periódicos ($000) X_2
1	9.73	0	400
2	6.25	250	250
3	9.71	300	300
4	9.31	350	350
5	11.77	350	350
6	9.82	400	250
7	16.28	450	450
8	13.29	550	250
9	14.36	600	300
10	17.41	650	350

e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)

f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.

3.2.2.2 **E**l gerente distrital de ventas de de un fabricante de automóviles estudia las ventas de éstos. En forma específica, quiere determinar qué factores influyen en el número de automóviles vendidos en una distribuidora. Para investigarlo, seleccionó al azar 10 distribuidoras. De éstas, obtiene el número de automóviles vendidos el mes pasado, los minutos de publicidad en radio comprados el mes pasado y el número de vendedores de tiempo completo contratados. La información es la siguiente:

ISBN 978-607-8285-62-4

Distribuidora	Automóviles vendidos el mes pasado Y	Publicidad (en minutos) X_1	Fuerza de ventas X_2
1	127	18	10
2	159	22	14
3	144	23	12
4	139	17	12
5	128	16	12
6	180	26	17
7	102	15	7
8	163	24	16
9	106	18	10
10	149	30	11

e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)

f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.

3.2.2.3 **L**os siguientes datos representan las calificaciones de estadística para una muestra aleatoria de 10 estudiantes de primer año de determinada institución de enseñanza superior, junto con sus calificaciones en un examen de inteligencia aplicado cuando aún cursaban el último año de secundaria y el número de periodos de clase perdidos por los 10 estudiantes que tomaron el curso de estadística.

Estudiante	Calificación de estadística Y	Calificación del examen X_1	Clases perdidas X_2
1	98	70	5
2	74	50	7
3	87	70	3
4	81	55	4
5	85	65	1
6	90	65	2
7	74	55	4
8	76	55	5
9	91	70	3
10	94	65	2

ISBN 978-607-8285-62-4

e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)

f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.

ANTECEDENTES

CONCEPTOS DE:

La función de probabilidad, Las distribuciones de probabilidad, Características de la forma de una distribución de probabilidad. Distribución t de Student. Nivel de significancia. La significancia observada (valor p). Estimador puntual. Varianza poblacional. Desviación estándar poblacional. Varianza muestral. Desviación estándar de la muestra. Error estándar de la estimación. Estructura de un intervalo de confianza.

3.2.3 INTERVALOS DE CONFIANZA PARA LA MEDIA *Y*, CON NUEVOS VALORES DE LAS VARIABLES INDEPENDIENTES.

CONCEPTOS BÁSICOS INTERVALO DE CONFIANZA DE LA MEDIA "Y"

Un intervalo de confianza

El error estándar del estimador se utiliza también para establecer intervalos de confianza para reportar el valor ***medio*** de ***Y*** con **nuevos valores de las variables independientes**, si el tamaño de la muestra es suficientemente grande y la dispersión alrededor del plano de regresión se aproxima a la distribución normal.

Se puede desarrollar **una estimación por intervalo de confianza** para hacer inferencia sobre el valor predicho de ***Y,*** la fórmula es:

$$\mu_{Y.X} = Y = \hat{Y} \mp t_{\alpha/2,n-k-1} S_{Y.12..k}\sqrt{h_i}$$

donde:

$$h_i = X_i'(X'X)^{-1}X_i$$

$$\hat{Y}_i = Valor\ predicho\ de\ Y = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

ISBN 978-607-8285-62-4

representa un rango en el que probablemente una nueva observación individual se incluya en la configuración especificada dada de los predictores.

$$S_{Y.12\ldots k} = Error\ estándar\ del\ estimador$$

$$X_i^{\prime} = Matriz\ del\ i-ésimo\ renglón\ de\ la\ matriz\ \boldsymbol{X}$$

$$(\boldsymbol{X}^{\prime}\boldsymbol{X})^{-1} = \boldsymbol{Matriz\ inversa\ de\ X^{\prime}X}$$

$$X_i = Matriz\ columna\ del\ i-ésimo\ renglón\ de\ la\ matriz\ \boldsymbol{X}$$

3.2.3.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 3.2.3.1 INTERVALO DE CONFIANZA DE LA MEDIA "Y"

Samuel Prado, propietario y director general de un establecimiento quiere conocer el comportamiento de las ventas (en miles de pesos) de un equipo de sonido que se expende en el establecimiento. Se percata de que existen muchos factores que podrían ayudarle a explicar las ventas pero piensa que la inversión en publicidad (en miles de pesos) y el precio (en cientos de pesos) son los principales factores determinantes. Samuel ha reunido los datos que se anexan a continuación.

Observación	Ventas Y	Publicidad X_1	Precio X_2
1	33	3	125
2	61	6	115
3	70	10	140
4	82	13	130
5	17	9	145
6	24	6	140
7	75	11	138
8	80	12	127
9	35	4	128
10	20	8	145

g) Construya un intervalo de confianza para las verdaderas ventas cuando se destina una inversión en publicidad de $ 11, 000 y se fija un precio al producto de $ 13,500.00

ISBN 978-607-8285-62-4

Intervalo de confianza para $\hat{Y}$.

Solución al inciso g.

Se da otro uso de la notación matricial cuando se utiliza un modelo de regresión múltiple para estimar el valor esperado de $\boldsymbol{Y}$, con nuevos valores de las variables independientes. Las fórmulas en notación matricial incluyen nuevamente a la matriz $(\boldsymbol{X'X})^{-1}$.

$$\mu_{Y.12} = \hat{Y}_i \pm t_{\alpha/2.n-k-1} S_{Y:12}\sqrt{h_i}$$

Donde

$$h_i = X_i'(X'X)^{-1}X_i$$

Como $X_1 = 11\ y\ X_2 = 135$, entonces

Uso de la ecuación de regresión lineal múltiple para hacer la estimación.

$$\hat{Y}_{11.135} = 223.52438 + 6.56400X_{1i} - 1.70780X_{2i}$$
$$= 223.52438 + 6.564(11) - 1.7078(135) \cong 65.17538$$

Y

En modelos de regresión h_i mide la distancia de un valor x de observación hasta el promedio de los valores x para todas las observaciones en un conjunto de datos.

$$h_i = X_i'(X'X)^{-1}X_i$$
$$= [1 \quad 11 \quad 135]\begin{bmatrix} 17'252{,}036/829{,}796 & 39{,}988/829{,}796 & -131{,}260/829{,}796 \\ 39{,}988/829{,}796 & 8{,}681/829{,}796 & -834/829{,}796 \\ -131{,}260/829{,}796 & -834/829{,}796 & 1{,}036/829{,}796 \end{bmatrix}\begin{bmatrix} 1 \\ 11 \\ 135 \end{bmatrix}$$
$$= [-0.03398 \quad 0.02758 \quad -0.00069]\begin{bmatrix} 1 \\ 11 \\ 135 \end{bmatrix} \cong 0.17625$$

Sustituyendo en la ecuación:

Obtención de los limites superior e inferior del intervalo de confianza para $\mu_{Y.X1.X2}$

$$\mu_{Y.11.135} = \hat{Y}_i \pm t_{0.05,7} S_{Y:12}\sqrt{h_i} = 65.17538 \mp 2.36(12.66383)\sqrt{0.17625}$$
$$= 65.175 \mp 12.54708 \begin{cases} LIC = 65.17538 - 12.54705 \cong 52.62833\ (Miles\ de\ pesos) \\ LSC = 65.17538 + 12.54705 \cong 77.72243 (Miles\ de\ pesos) \end{cases}$$

Interpretación del intervalo de confianza.

Interpretación: En **95** de cada **100** muestras (95% de confianza) de tamaño **10**, el verdadero volumen de ventas promedio cuando se invierte en publicidad **$ 11,000.00 pesos** y se fija un precio al producto de **$13,500.00** oscilará aproximadamente entre **$ 52,628.33** y **$ 77,722.43** pesos.

ISBN 978-607-8285-62-4

3.2.3.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 3.2.3.1 INTERVALO DE CONFIANZA DE LA MEDIA "Y"

Para el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	11	9	4
2	30	28	14
3	26	17	11
4	20	19	9
5	15	20	8
6	25	24	11
7	35	32	12
8	17	13	21
9	39	36	9
10	45	40	17

g) Construya un intervalo de confianza para el verdadero valor de Y cuando X_{1i} es 22 y X_{2i} es de 10.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Intervalo de confianza para $\hat{Y}$.

<u>Solución al inciso g.</u>

$$\mu_{Y.12} = \hat{Y}_i \pm t_{\alpha/2.n-k-1} S_{Y:12}\sqrt{h_i}$$

Donde

$$h_i = X_i^{'}(X^{'}X)^{-1}X_i$$

ISBN 978-607-8285-62-4

Como $X_1 =$ $y\ X_2 =$, entonces

Uso de la ecuación de regresión lineal múltiple para hacer la estimación.

$$\hat{Y}_{22.10} =$$

Y

En modelos de regresión h_i mide la distancia de un valor x de observación hasta el promedio de los valores x para todas las observaciones en un conjunto de datos.

$$h_i = X_i^{'}(X^{'}X)^{-1}X_i =$$

Obtención de los limites superior e inferior del intervalo de confianza para $\mu_{Y.X1.X2}$

Sustituyendo en la ecuación:

$$\mu_{Y.22.10} = \hat{Y}_i \pm t_{0.05,7}S_{Y:12}\sqrt{h_i} = \begin{cases} LIC = \\ LSC = \end{cases}$$

Interpretación del intervalo de confianza.

Interpretación:

ISBN 978-607-8285-62-4

3.2.3.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 3.2.3.1
INTERVALO DE CONFIANZA DE LA MEDIA "Y"

Si tenemos el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	15	7.2	3.0
2	33	4.2	5.0
3	32	1.3	5.0
4	42	1.4	9.5
5	20	6.3	3.7
6	27	3.6	5.5
7	30	2.1	5.2
8	25	5.0	3.1
9	37	1.5	7.8
10	10	8.5	2.5

g) Construya un intervalo de confianza para el verdadero valor de Y cuando X_{1i} es 4.5 y X_{2i} es de 4.0

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ISBN 978-607-8285-62-4

Intervalo de confianza para $\hat{Y}$.

Uso de la ecuación de regresión lineal múltiple para hacer la estimación.

En modelos de regresión h_i mide la distancia de un valor x de observación hasta el promedio de los valores x para todas las observaciones en un conjunto de datos.

Obtención de los limites superior e inferior del intervalo de confianza para $\mu_{Y.X1.X2}$

Interpretación del intervalo de confianza.

<u>Solución al inciso g.</u>

Como $X_1 = 4.5 \quad y\ X_2 = 4.0$, entonces

$$\hat{Y}_{4.5.4} =$$

Y

$$h_i =$$

Sustituyendo en la ecuación:

$$\mu_{Y.4.5.4} = \qquad \begin{cases} LIC = \\ LSC = \end{cases}$$

Interpretación:

ISBN 978-607-8285-62-4

3.2.3 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 3.2.3 INTERVALO DE CONFIANZA DE LA MEDIA "Y"

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

3.2.3.1 **S**uponga que una compañía de productos de consumo quisiera medir la efectividad de la publicidad en radio y televisión y la publicidad en periódicos en la promoción de sus productos. Se seleccionó una muestra aleatoria de 10 ciudades con poblaciones aproximadamente iguales para el estudio durante un periodo de prueba de un mes. Se registraron las ventas (en miles de pesos) durante el mes de prueba, junto con los niveles de gastos en los medios de publicidad, con los resultados siguientes:

Ciudad	Ventas ($000,000) Y	Publicidad en radio y televisión ($000) X_1	Publicidad en periódicos ($000) X_2
1	9.73	0	400
2	6.25	250	250
3	9.71	300	300
4	9.31	350	350
5	11.77	350	350
6	9.82	400	250
7	16.28	450	450
8	13.29	550	250
9	14.36	600	300
10	17.41	650	350

g) Construya un intervalo de confianza para las verdaderas ventas cuando se destina una inversión en publicidad en radio y televisión de $ 380, 000 pesos y se fija una publicidad en periódicos de $ $ 280,000 pesos.

3.2.3.2 **E**l gerente distrital de ventas de de un fabricante de automóviles estudia las ventas de éstos. En forma específica, quiere determinar qué factores influyen en el número de automóviles vendidos en una distribuidora. Para investigarlo, seleccionó al azar 10 distribuidoras. De éstas, obtiene el número de automóviles vendidos el mes pasado, los minutos de publicidad en radio comprados el mes pasado y el número de vendedores de tiempo completo contratados. La información es la siguiente:

ISBN 978-607-8285-62-4

Distribuidora	Automóviles vendidos el mes pasado Y	Publicidad (en minutos) X_1	Fuerza de ventas X_2
1	127	18	10
2	159	22	14
3	144	23	12
4	139	17	12
5	128	16	12
6	180	26	17
7	102	15	7
8	163	24	16
9	106	18	10
10	149	30	11

g) Construya un intervalo de confianza para el número de automóviles vendidos el mes pasado cuando se compran 20 minutos de publicidad y se emplean a 15 vendedores.

3.2.3.3 Los siguientes datos representan las calificaciones de estadística para una muestra aleatoria de 10 estudiantes de primer año de determinada institución de enseñanza superior, junto con sus calificaciones en un examen de inteligencia aplicado cuando aún cursaban el último año de secundaria y el número de periodos de clase perdidos por los 10 estudiantes que tomaron el curso de estadística.

Estudiante	Calificación de estadística Y	Calificación del examen X_1	Clases perdidas X_2
1	98	70	5
2	74	50	7
3	87	70	3
4	81	55	4
5	85	65	1
6	90	65	2
7	74	55	4
8	76	55	5
9	91	70	3
10	94	65	2

g) Construya un intervalo de confianza para la verdadera calificación de estadística para una calificación del examen de inteligencia de 60 y 2 clases perdidas. Interprete los resultados.

ISBN 978-607-8285-62-4

OBJETIVO 3.3 El alumno podrá calcular y utilizar el criterio de las "*f*" parciales para determinar la contribución de las variables explicatorias. Asimismo podrá calcular e interpretar el coeficiente de determinación múltiple, el coeficiente de correlación múltiple, así como los coeficientes de determinación parcial.

ANTECEDENTES

CONCEPTOS DE:

Variables aleatorias. Variable dependiente. Variable independiente. Población, marco y muestra. Parámetro. La función de probabilidad, Las distribuciones de probabilidad, Características de la forma de una distribución de probabilidad. Prueba de hipótesis. Estructura de las hipótesis nula y alternativa, Error tipo I y tipo II. Distribución t de Student. Prueba t. Nivel de significancia. Distribución *F*. Prueba *F* para la razón de varianzas. Estadístico de prueba. Análisis de Varianza. La significancia observada (valor p). Estimador puntual. Varianza poblacional. Desviación estándar poblacional. Varianza muestral. Desviación estándar de la muestra. Error estándar de la muestra.

3.3.1 PRUEBA DE PORCIONES DE UN MODELO DE REGRESIÓN MÚLTIPLE. CRITERIO PARA PRUEBA *F* PARCIAL

CONCEPTOS BÁSICOS PRUEBA *F* PARCIAL

Cuando se desarrolla un **modelo de regresión lineal múltiple** uno de los objetivos es **utilizar sólo aquellas variables explicatorias que sean útiles** para predecir el valor de la variable dependiente.

Un **método** para determinar **la contribución de una variable explicatoria** es conocido como **criterio para prueba *F* parcial.** Consiste en determinar la **contribución a la regresión de la suma de cuadrados por cada variable explicatoria** después de haber incluido

ISBN 978-607-8285-62-4

Contribuciones individuales.

todas las otras variables explicatorias del modelo. **La nueva variable explicatoria sólo será incluida si el modelo mejora en forma significativa.**

La **contribución de cada variable explicatoria** se evaluará al tomar en cuenta la suma de regresión de los cuadrados de un modelo que incluye todas las variables explicatorias excepto la de interés, SCR (*Todas las variables excepto k*). De esta manera, en general, para determinar la contribución de la variable *k* , sabiendo que ya todas las otras variables están incluidas, se tendría:

SCR (X_k|todas las variables excepto *k*)=SCT(todas las variables incluyendo *k*) –SCR(todas las variables excepto *k*)

Por ejemplo **si sólo hay dos variables explicatorias**, la contribución de cada una se puede determinar de la siguiente manera:

Contribución de la variable X_1 sabiendo que X_2 está incluida:

$$SCR(X_1|X_2) = SCR(X_1 \text{ Y } X_2) - SCR(X_2)$$

Contribución de la variable X_2 sabiendo que X_1 está incluida:

$$SCR(X_2|X_1) = SCR(X_1 \text{ Y } X_2) - SCR(X_1)$$

La **hipótesis nula y alternativa** para probar la **contribución de X_1** al modelo serían:

Prueba *F* parcial para probar la contribución de X_1.

H_0: la variable $\boldsymbol{X_1}$ no mejora en forma significativa el modelo, una vez incluida la variable $\boldsymbol{X_2}$.

H_1: la variable $\boldsymbol{X_1}$ mejora en forma significativa el modelo, una vez incluida la variable $\boldsymbol{X_2}$.

El criterio para la **prueba *F* parcial** se expresa de la siguiente manera:

$$F_{1.\ n-k-1} = SCR(X_k \text{ todas las variables excepto k}) / CME$$
$$\alpha = 0.05 \text{ (Extremo derecho)}$$

Si hay **dos variables explicatorias** el criterio para la **prueba *F* parcial** quedaría de la siguiente manera:

$$F_{1.\ n-k-1} = SCR(X_1|X_2)/CME = SCR(X_1 \text{ Y } X_2) - SCR(X_2) / CME$$

Dado que los grados de libertad de la $SCR(X_1|X_2)$ es 1 se puede escribir la expresión anterior como :

$$F_{1.\ n-k-1} = CMR(X_1|X_2)/CME$$

ISBN 978-607-8285-62-4

Prueba *F* parcial para probar la contribución de X_2.

La **hipótesis nula y alternativa** para probar la **contribución de X_2** al modelo serían:

H_0: la variable X_2 no mejora en forma significativa el modelo, una vez incluida la variable X_1

H_1: la variable X_2 mejora en forma significativa el modelo, una vez incluida la variable X_1.

El criterio para la **prueba *F* parcial** se expresa de la siguiente manera:

$$F_{1.\ n-k-1} = SCR(X_k \text{ todas las variables excepto k}) / CME$$

$$\alpha = 0.05 \text{ (Extremo derecho)}$$

Si hay **dos variables explicatorias** el criterio para la **prueba *F* parcial** quedaría de la siguiente manera:

$$F_{1.\ n-k-1} = SCR(X_2 | X_1)/CME = SCR(X_1 \text{ Y } X_2) - SCR(X_1) / CME$$

Dado que los grados de libertad de la $SCR(X_2 | X_1)$ es 1 se puede escribir la expresión anterior como :

$$F_{1.\ n-k-1} = CMR(X_2 | X_1)/CME$$

ISBN 978-607-8285-62-4

3.3.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 3.3.1.1
PRUEBA *F* PARCIAL

Samuel Prado, propietario y director general de un establecimiento quiere conocer el comportamiento de las ventas (en miles de pesos) de un equipo de sonido que se expende en el establecimiento. Se percata de que existen muchos factores que podrían ayudarle a explicar las ventas pero piensa que la inversión en publicidad (en miles de pesos) y el precio (en cientos de pesos) son los principales factores determinantes. Samuel ha reunido los datos que se anexan a continuación.

Observación	Ventas Y	Publicidad X_1	Precio X_2
1	33	3	125
2	61	6	115
3	70	10	140
4	82	13	130
5	17	9	145
6	24	6	140
7	75	11	138
8	80	12	127
9	35	4	128
10	20	8	145

h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.

Contribuciones individuales.

Solución al inciso h.

CONTRIBUCIÓN DE X_1 (INVERSIÓN EN PUBLICIDAD) UNA VEZ INCLUIDA X_2 (PRECIO DEL PRODUCTO) EN EL MODELO :

Prueba *F* parcial para probar la contribución de X_1.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0: la variable X_1 no mejora en forma significativa el modelo, una vez incluida la variable X_2.

H_1: la variable X_1 mejora en forma significativa el modelo, una vez incluida la variable X_2.

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$F_{1.n-k-1} = F_{1.10-2-1} = F_{1.7} = \frac{SCR(X_1 \mid X_2)}{CME} = \frac{SCR(X_1 y X_2) - SCR(X_2)}{CME}$$
$$= \frac{CMR(X_1 \mid X_2)}{CME} = 25.68$$

Donde, utilizando solamente Y y X_2 como si fuera un modelo de regresión lineal simple, se tiene:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 X_2 = 193.28810 - 1.07718 X_2$$

Suma de cuadrados de la regresión para X_2.

$$SCR(X_2) = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y - n(\bar{Y})^2$$
$$= 193.28810(497) - 1.07718(65{,}315) - 10(49.7)^2 = 1{,}007.27106$$

$$SCR(X_1|X_2) = \mathrm{SCR}(X_1 \text{ y } X_2) - \mathrm{SCR}(X_2) = 5{,}125.49186 - 1{,}007.27106$$
$$\cong \mathbf{4,118.22080}$$

Suma de cuadrados de la regresión para $(X_1|X_2)$.

OPCIONAL :

Otra forma alternativa para evaluar la contribución realizada por una variable explicatoria se basa en el error estándar de su coeficiente de regresión. Puesto que los errores estándar de los coeficientes de regresión $\boldsymbol{S_{bk}}$ **y el** ***CME*** ya se calcularon anteriormente, la contribución de una variable independiente en particular a la suma de regresión de los cuadrados se puede determinar en la forma siguiente:

$$\boxed{SCR(X_1|X_2) = \frac{\hat{\beta}_1^2 \mathrm{CME}}{S_{\hat{\beta}_1}^2} = \frac{6.564^2(160.3)}{1.295^2} \cong 4{,}118.4247}$$

ISBN 978-607-8285-62-4

Tabla de ANOVA parcial .

Tabla de ANOVA parcial :

Para este caso se usará la primera forma de cálculo :

$$SCR(X_1|X_2) \cong \mathbf{4,118.22080}$$

Fuente de Variación	***Grados de libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada\ parcial}$***
Regresión (X_1 y X_2)	*k*=2	*SCR(X_1 y X_2)* =5,125.49186	*CMR(X_1 y X_2) =SCR/G.L.*=2,562.74593	$\frac{CMR(X_1\|X_2)}{CME} = \frac{4,118.22080}{160.37259} \cong 25.68$
Regresión (X_2)	*K=1*	*SCR(X_2)* =1,007.27106	*CMR(X_2)= SCR(X_2)/G.L.*=1,007.27106/1=1,007.27106	
Regresión (X_1 | X_2)	*K=1*	*SCR(X_1 | X_2)* =4,118.22080	*CMR(X_1 | X_2)=SCR(X_1 | X_2) /G.L.*=4,118.22080/1 =4,118.22080	
Error	*n-k-1*= 7	*SCE*=1,122.60814	*CME=SCE/G.L.*=160.37259	
Total	*n-1*=9	*SCT*=6,248.1		

Paso3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Para determinar la región de rechazo, se necesita el valor crítico. El valor crítico en el estadístico ***F*** se encuentra en las tablas de valores críticos de ***F***. Para utilizar esta tabla se necesita conocer los grados de libertad en el numerador y en el denominador. Los grados de libertad en el numerador son iguales al número de variables independientes, designados como "k".

ISBN 978-607-8285-62-4

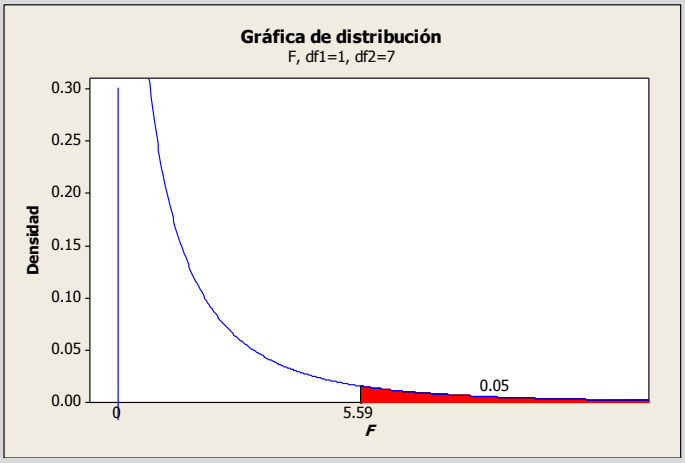
Los grados de libertad en el denominador son el número de observaciones menos el número de variables independientes menos 1. Para este problema al particionar la tabla de ANOVA sólo existe una solo variable independiente X_1, por lo tanto los grados de libertad en el numerador son*: k= 1 g.l.* y los grados de libertad del denominador, en este caso el error, para 10 observaciones y dos variable independientes con: *n-k-1=10-2-1=7 g.l.*

Como existen tablas para niveles de Alfa diferentes, busque la que corresponda al nivel de significancia solicitada, en este caso 0.05, y desplácese horizontalmente sobre la parte superior de la página hasta llegar a los 1 grados de libertad del numerador. Luego descienda en esa columna hasta llegar a la fila que presenta 7 grados de libertad. El valor en esta intersección es **5.59** que en este caso es el valor crítico.

Grados de libertad en el denominador	**Grados de libertad en el numerador**			
	1	2	3	4
7	**5.59**	4.74	4.35	4.12
8	5.32	4.46	4.07	3.84
9	5.12	4.26	3.86	3.63
10	4.96	4.10	3.71	3.48
11	4.84	3.98	3.59	3.36
12	4.75	3.89	3.49	3.26
13	4.67	3.81	3.41	3.18
14	4.60	3.74	3.34	3.11

ISBN 978-607-8285-62-4

Esta información se presenta en el siguiente diagrama:



Si se desea aplicar el criterio *p-level* en la conclusión busque en las tablas de valores críticos de ***F*** la que corresponda al nivel de significancia de 0.01, y desplácese horizontalmente sobre la parte superior de la página hasta llegar a los 1 grados de libertad del numerador. Luego descienda en esa columna hasta llegar a la fila que presenta 7 grados de libertad. El valor en esta intersección es **12.25** que en este caso es el valor crítico.

Si el *valor p* es menor que o igual al nivel Alfa (α), rechace la hipótesis nula en favor de la hipótesis alternativa. Si $0.01 \leq p \leq 0.05$ se dice que la prueba es significativa (S). Si $p < 0.01$ se dice que la prueba es altamente significativa (AS).

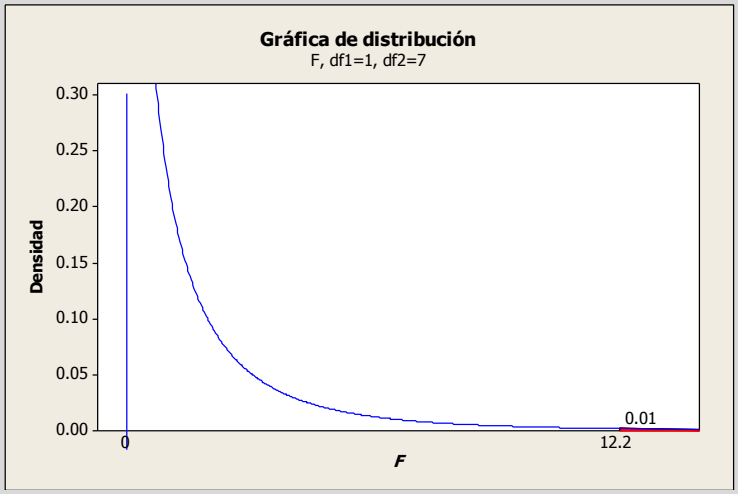
Si el *valor p* es mayor que el nivel Alfa (α), no rechace la hipótesis nula. Si $p > 0.05$ se dice que la prueba es no significativa (NS).

Los niveles de significancia más utilizado son 0.05 y 0.01

Grados de libertad en el denominador	**Grados de libertad en el numerador**			
	1	2	3	4
7	**12.25**	9.55	8.45	7.85
8	11.26	8.65	7.59	7.01
9	10.6	8.02	6.99	6.42
10	10.0	7.56	6.55	5.99
11	9.65	7.21	6.22	5.67
12	9.33	6.93	5.95	5.41
13	9.07	6.70	5.74	5.21
14	8.86	6.51	5.56	5.04

ISBN 978-607-8285-62-4

Esta información se presenta en el siguiente diagrama



Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso5. Interpretación.

Se rechaza H_0 si fcalc ≥ 5.59

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:
$Como\ 25.68 > 5.59 > 12.25\ \therefore la\ prueba\ es\ (AS)\ y\ se\ rechaza H_0$.

Administrativa: Existe evidencia suficiente para decir que estadísticamente la variable X_1 (inversión en publicidad) si contribuye significativamente en el modelo, una vez incluida la variable X_2 (precio del producto).

Prueba *F* parcial para probar la contribución de X_2.

CONTRIBUCIÓN DE X_2 (PRECIO DEL PRODUCTO) UNA VEZ INCLUIDA EN EL MODELO X_1 (INVERSIÓN EN PUBLICIDAD) :

Paso 1. Juego de hipótesis.

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).
H_0: la variable X_2 no mejora en forma significativa el modelo, una vez incluida la variable X_1.

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

H_1: la variable X_2 mejora en forma significativa el modelo, una vez incluida la variable X_1.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$F_{1.n-k-1} = F_{1.10-2-1} = F_{1.7} = \frac{SCR(X_2 \mid X_1)}{CME} = \frac{SCR(X_1 y X_2) - SCR(X_1)}{CME}$$
$$= \frac{CMR(X_2 \mid X_1)}{CME} = 14.56$$

Suma de cuadrados de la regresión para X_1.

Donde, utilizando solamente Y y X_1 como si fuera un modelo de regresión lineal simple, se tiene:

Suma de cuadrados de la regresión para $(X_2|X_1)$.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 = 7.14865 + 5.18919 X_1$$
$$SCR(X_1) = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y - n(\bar{Y})^2$$
$$= 7.14865(497) + 5.18919(4{,}613) - 10(49.7)^2 = 2{,}789.70811$$

$$SCR(X_2|X_1) = \mathrm{SCR}(X_1 \text{ y } X_2) - \mathrm{SCR}(X_1) = 5{,}125.49186 - 2{,}78970811$$
$$\cong \mathbf{2{,}335.78375}$$

OPCIONAL :

Otra forma alternativa para evaluar la contribución realizada por una variable explicatoria se basa en el error estándar de su coeficiente de regresión. Puesto que los errores estándar de los coeficientes de regresión $\boldsymbol{S_{bk}}$ **y el** ***CME*** ya se calcularon anterioremente, la contribución de una variable independiente en particular a la suma de regresión de los cuadrados se puede determinar en la forma siguiente:

$$\boxed{SCR(X_2|X_1) = \frac{\hat{\beta}_2^2 \mathrm{CME}}{S_{\hat{\beta}_2}^2} = \frac{(-1.7078)^2(160.3)}{(0.4474)^2} \cong 2{,}335.69204}$$

ISBN 978-607-8285-62-4

Tabla de ANOVA parcial.

Tabla de ANOVA parcial :
Para este caso se usará la primera forma de cálculo :

$$SCR(X_2|X_1) \cong \mathbf{2,335.78375}$$

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrad***	***Cuadrado Medio***	***$F_{calculada\ parcial}$***
Regresión (X_1 y X_2)	*k*=2	*SCR(X_1 y X_2)* =5,125.49186	*SCR/G.L.*=2,562.74593	$\frac{CMR(X_2\|X_1)}{CME} = \frac{2,335.78375}{160.37259} \cong 14.56$
Regresión (X_1)	*K=1*	*SCR(X_1)* =2,789.70811	*SCR(X_1)/G.L.*=2,789.70811/1=2,789.70811	
Regresión ($X_2 \mid X_1$)	*K=1*	*SCR($X_2 \mid X_1$)* =2,335.78375	*SCR($X_2 \mid X_1$) /G.L.*=2,335.78375	
Error	*n-k-1*=7	*SCE*=1,122.60814	*SCE/G.L.*=160.37259	
Total	*n-1*=9	*SCT*=6,248.1		

Paso3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

Para determinar la región de rechazo, se necesita el valor crítico. El valor crítico en el estadístico ***F*** se encuentra en las tablas de valores críticos de ***F***. Para utilizar esta tabla se necesita conocer los grados de libertad en el numerador y en el denominador. Los grados de libertad en el numerador son iguales al número de variables independientes, designados como "k". Los grados de libertad en el denominador son el número de observaciones menos el número de variables independientes menos 1. Para este problema al particionar la tabla de ANOVA sólo existe una solo variable independiente X_1, por lo tanto los grados de libertad en el numerador son*: k= 1 g.l.* y los grados de libertad del denominador, en este caso el error, para 10 observaciones y dos variable independientes con: *n-k-1=10-2-*

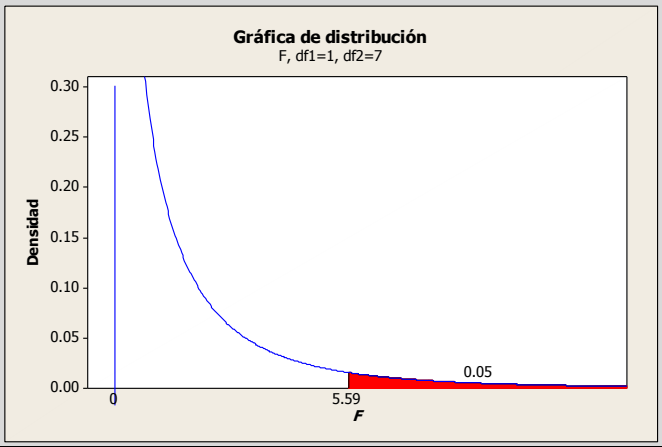
ISBN 978-607-8285-62-4

1=7 g.l.

Como existen tablas para niveles de Alfa diferentes, busque la que corresponda al nivel de significancia solicitada, en este caso 0.05, y desplácese horizontalmente sobre la parte superior de la página hasta llegar a los 1 grados de libertad del numerador. Luego descienda en esa columna hasta llegar a la fila que presenta 7 grados de libertad. El valor en esta intersección es **5.59** que en este caso es el valor crítico.

Grados de libertad en el denominador	**Grados de libertad en el numerador**			
	1	2	3	4
7	**5.59**	4.74	4.35	4.12
8	5.32	4.46	4.07	3.84
9	5.12	4.26	3.86	3.63
10	4.96	4.10	3.71	3.48
11	4.84	3.98	3.59	3.36
12	4.75	3.89	3.49	3.26
13	4.67	3.81	3.41	3.18
14	4.60	3.74	3.34	3.11

Esta información se presenta en el siguiente diagrama



ISBN 978-607-8285-62-4

Si el *valor p* es menor que o igual al nivel Alfa (a), rechace la hipótesis nula en favor de la hipótesis alternativa. Si $0.01 \leq p \leq 0.05$ se dice que la prueba es significativa (S). Si $p < 0.01$ se dice que la prueba es altamente significativa (AS).

Si el *valor p* es mayor que el nivel Alfa (a), no rechace la hipótesis nula. Si $p > 0.05$ se dice que la prueba es no significativa (NS).

Los niveles de significancia más utilizado son 0.05 y 0.01

Si se desea aplicar el criterio *p-level* en la conclusión busque en las tablas de valores críticos de ***F*** la que corresponda al nivel de significancia de 0.01, y desplácese horizontalmente sobre la parte superior de la página hasta llegar a los 1 grados de libertad del numerador. Luego descienda en esa columna hasta llegar a la fila que presenta 7 grados de libertad. El valor en esta intersección es **12.25** que en este caso es el valor crítico.

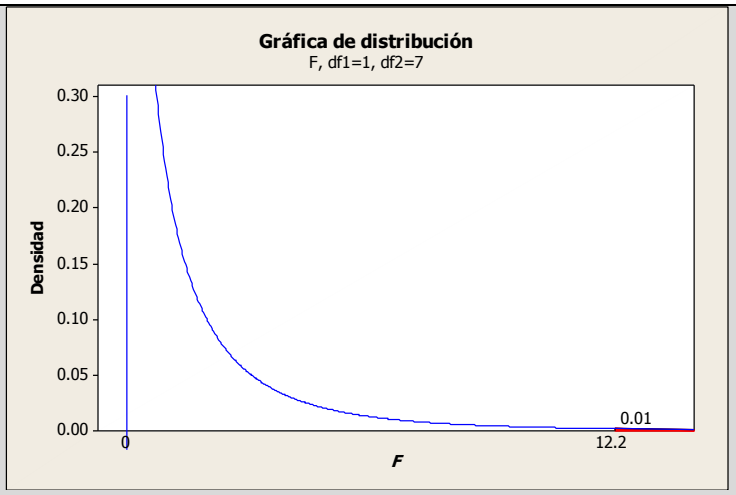
Grados de libertad en el denominador	**Grados de libertad en el numerador**			
	1	2	3	4
7	**12.25**	9.55	8.45	7.85
8	11.26	8.65	7.59	7.01
9	10.6	8.02	6.99	6.42
10	10.0	7.56	6.55	5.99
11	9.65	7.21	6.22	5.67
12	9.33	6.93	5.95	5.41
13	9.07	6.70	5.74	5.21
14	8.86	6.51	5.56	5.04

Esta información se presenta en el siguiente diagrama

ISBN 978-607-8285-62-4

Paso4. Regla de decisión.

Paso5. Interpretación.



Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si fcalc ≥ 5.59

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

$$Como\ 14.56 > 5.59 > 12.25\ \therefore la\ prueba\ es\ (AS)\ y\ se\ rechaza H_0\ .$$

Administrativa: Existe evidencia suficiente para decir que estadísticamente la variable X_2 (precio del producto) si contribuye significativamente en el modelo, una vez incluida la variable X_1 (inversión en publicidad).

Por lo tanto, mediante la prueba de la contribución de cada variable explicatoria, luego que ha incluida la otra en el modelo, se determinó que cada una de las dos variables explicatorias contribuyen a mejorar en forma significativa el modelo. Por consiguiente, el modelo de regresión múltiple debe incluir tanto la inversión en publicidad X_1 como el precio del equipo de sonido X_2.

ISBN 978-607-8285-62-4

3.3.1.1 ACTIVIDAD DE APRENDIZAJE

ACTIVIDAD DE APRENDIZAJE 3.3.1.1
PRUEBA *F* PARCIAL

Para el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	11	9	4
2	30	28	14
3	26	17	11
4	20	19	9
5	15	20	8
6	25	24	11
7	35	32	12
8	17	13	21
9	39	36	9
10	45	40	17

h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Contribuciones individuales.

Prueba F parcial para probar la contribución de X_1.

<u>Solución al inciso h.</u>

CONTRIBUCIÓN DE X_1 UNA VEZ INCLUIDA X_2 EN EL MODELO :

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$F_{1.n-k-1} =$$

Donde, utilizando solamente Y y X_2 como si fuera un modelo de regresión lineal simple, se tiene:

$\hat{Y}_i =$

Suma de cuadrados de la regresión para X_2.

$$SCR(X_2) = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y - n(\bar{Y})^2 =$$

Suma de cuadrados de la regresión para $(X_1|X_2)$.

$$SCR(X_1|X_2) = \mathrm{SCR}(X_1 \text{ y } X_2) - \mathrm{SCR}(X_2) =$$

OPCIONAL :

Otra forma alternativa para evaluar la contribución realizada por una variable explicatoria se basa en el error estándar de su coeficiente de regresión. Puesto que los errores estándar de los coeficientes de regresión $\boldsymbol{S_{bk}}$ **y el** ***CME*** ya se calcularon anteriormente, la contribución de una variable independiente en particular a la suma de regresión de los cuadrados se puede determinar en la forma siguiente:

$$\boxed{SCR(X_1|X_2) = \frac{\hat{\beta}_1^2 \mathrm{CME}}{S_{\hat{\beta}_1}^2} =}$$

ISBN 978-607-8285-62-4

Tabla de ANOVA parcial

Tabla de ANOVA parcial :

Para este caso se usará la primera forma de cálculo :

$SCR(X_1|X_2) \cong$

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada\ parcial}$***
Regresión (X_1 y X_2)	*k=*	*SCR(X_1 y X_2) =*	*CMR(X_1 y X_2) =SCR/G.L.=*	$\frac{CMR(X_1 \mid X_2)}{CME} =$
Regresión (X_2)	*K=*	*SCR(X_2) =*	*CMR(X_2)= SCR(X_2)/G.L.=*	
Regresión ($X_1 \mid X_2$)	*K=*	*SCR($X_1 \mid X_2$) =*	*CMR($X_1 \mid X_2$)=SCR($X_1 \mid X_2$) /G.L.=*	
Error	*n-k-1=*	*SCE=*	*CME=SCE/G.L.=*	
Total	*n-1=*	*SCT=*		

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

$$F_{crítica} =$$

ISBN 978-607-8285-62-4

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

Prueba F parcial para probar la contribución de X_2.

CONTRIBUCION DE X_2 UNA VEZ INCLUIDA EN EL MODELO X_1 :
Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$F_{1.n-k-1} =$

Donde, utilizando solamente Y y X_1 como si fuera un modelo de regresión lineal simple, se tiene:

$\hat{Y}_i =$

Suma de cuadrados de la regresión para X_1.

$$SCR(X_1) = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y - n(\bar{Y})^2 =$$

Suma de cuadrados de la regresión para $(X_2|X_1)$.

$$SCR(X_2|X_1) = \mathrm{SCR}(X_1 \text{ y } X_2) - \mathrm{SCR}(X_1) =$$

OPCIONAL :

Otra forma alternativa para evaluar la contribución realizada por una variable explicatoria se basa en el error estándar de su coeficiente de regresión. Puesto que los errores estándar de los coeficientes de regresión $\boldsymbol{S_{bk}}$ **y el** ***CME*** ya se calcularon anteriormente, la contribución de una variable independiente en particular a la suma de regresión de los cuadrados se puede determinar en la forma siguiente:

Suma de cuadrados de la regresión para $(X_2|X_1)$.

$$\boxed{SCR(X_2|X_1) = \frac{\hat{\beta}_2^2 \mathrm{CME}}{S_{\hat{\beta}_2}^2} =}$$

ISBN 978-607-8285-62-4

Tabla de ANOVA parcial.

Tabla de ANOVA parcial :
Para este caso se usará la primera forma de cálculo :

$SCR(X_2|X_1) \cong$

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada\ parcial}$***
Regresión (X_1 y X_2)	*k=*	*SCR(X_1 y X_2) =*	*SCR/G.L.=*	$\frac{CMR(X_2 \mid X_1)}{CME} =$
Regresión (X_1)	*K=*	*SCR(X_1) =*	*SCR(X_1)/G.L.=*	
Regresión ($X_2 \mid X_1$)	*K=1*	*SCR($X_2 \mid X_1$) =*	*SCR($X_2 \mid X_1$) /G.L.=*	
Error	*n-k-1= 7*	*SCE=*	*SCE/G.L.=*	
Total	*n-1=9*	*SCT=*		

Paso3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

$$F_{crítica} =$$

Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso5. Interpretación.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

ISBN 978-607-8285-62-4

Administrativa:

3.3.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 3.3.1.1
PRUEBA *F* PARCIAL

Si tenemos el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	15	7.2	3.0
2	33	4.2	5.0
3	32	1.3	5.0
4	42	1.4	9.5
5	20	6.3	3.7
6	27	3.6	5.5
7	30	2.1	5.2
8	25	5.0	3.1
9	37	1.5	7.8
10	10	8.5	2.5

h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.

ISBN 978-607-8285-62-4

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Contribuciones individuales.

Solución al inciso h.

Prueba F parcial para probar la contribución de X_1.

CONTRIBUCIÓN DE X_1 UNA VEZ INCLUIDA X_2 EN EL MODELO :

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$F_{1.n-k-1} =$$

Donde, utilizando solamente Y y X_2 como si fuera un modelo de regresión lineal simple, se tiene:

$\hat{Y}_i =$

Suma de cuadrados de la regresión para X_2.

$SCR(X_2) =$

Suma de cuadrados de la regresión para $(X_1|X_2)$.

$SCR(X_1|X_2) =$

ISBN 978-607-8285-62-4

OPCIONAL :

Otra forma alternativa para evaluar la contribución realizada por una variable explicatoria se basa en el error estándar de su coeficiente de regresión. Puesto que los errores estándar de los coeficientes de regresión S_{bk} **y el *CME*** ya se calcularon anterioremente, la contribución de una variable independiente en particular a la suma de regresión de los cuadrados se puede determinar en la forma siguiente:

Suma de cuadrados de la regresión para $(X_1|X_2)$.

$$SCR(X_1|X_2) =$$

Tabla de ANOVA parcial :

Para este caso se usará la primera forma de cálculo :

$SCR(X_1|X_2) \cong$

Tabla de ANOVA parcial.

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	***$F_{calculada\ parcial}$***
Regresión (X_1 y X_2)				
Regresión (X_2)				
Regresión ($X_1 \mid X_2$)				
Error				
Total				

Paso3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

$$F_{crítica} =$$

Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

ISBN 978-607-8285-62-4

Paso5. Interpretación.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

Administrativa:

Prueba F parcial para probar la contribución de X_2.

CONTRIBUCION DE X_2 UNA VEZ INCLUIDA EN EL MODELO X_1 :

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$F_{1.n-k-1} =$

Donde, utilizando solamente Y y X_1 como si fuera un modelo de regresión lineal simple, se tiene:

$\hat{Y}_i =$

Suma de cuadrados de la regresión para X_1.

$SCR(X_1) =$

Suma de cuadrados de la regresión para $(X_2|X_1)$.

$SCR(X_2|X_1) =$

ISBN 978-607-8285-62-4

OPCIONAL :

Otra forma alternativa para evaluar la contribución realizada por una variable explicatoria se basa en el error estándar de su coeficiente de regresión. Puesto que los errores estándar de los coeficientes de regresión S_{bk} **y el** ***CME*** ya se calcularon anteriormente, la contribución de una variable independiente en particular a la suma de regresión de los cuadrados se puede determinar en la forma siguiente:

$SCR(X_2|X_1) =$

Tabla de ANOVA parcial.

Tabla de ANOVA parcial :

Para este caso se usará la primera forma de cálculo :

$$SCR(X_2|X_1) \cong$$

Fuente de Variación	***Grados de Libertad***	***Suma de Cuadrados***	***Cuadrado Medio***	$F_{calculada\ parcial}$

Paso3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

$$F_{crítica} =$$

Paso4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Paso5. Interpretación.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

ISBN 978-607-8285-62-4

Estadística:

Administrativa:

3.3.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 3.3.1

PRUEBA *F* PARCIAL

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice**

3.3.1.1 **S**uponga que una compañía de productos de consumo quisiera medir la efectividad de la publicidad en radio y televisión y la publicidad en periódicos en la promoción de sus productos. Se seleccionó una muestra aleatoria de 10 ciudades con poblaciones aproximadamente iguales para el estudio durante un periodo de prueba de un mes. Se registraron las ventas (en miles de pesos) durante el mes de prueba, junto con los niveles de gastos en los medios de publicidad, con los resultados siguientes:

Ciudad	Ventas ($000,000) Y	Publicidad en radio y televisión ($000) X_1	Publicidad en periódicos ($000) X_2
1	9.73	0	400
2	6.25	250	250
3	9.71	300	300
4	9.31	350	350
5	11.77	350	350
6	9.82	400	250
7	16.28	450	450
8	13.29	550	250
9	14.36	600	300
10	17.41	650	350

h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.

ISBN 978-607-8285-62-4

un software para comparar sus resultados. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

3.3.1.2 **E**l gerente distrital de ventas de de un fabricante de automóviles estudia las ventas de éstos. En forma específica, quiere determinar qué factores influyen en el número de automóviles vendidos en una distribuidora. Para investigarlo, seleccionó al azar 10 distribuidoras. De éstas, obtiene el número de automóviles vendidos el mes pasado, los minutos de publicidad en radio comprados el mes pasado y el número de vendedores de tiempo completo contratados. La información es la siguiente:

Distribuidora	Automóviles vendidos el mes pasado Y	Publicidad (en minutos) X_1	Fuerza de ventas X_2
1	127	18	10
2	159	22	14
3	144	23	12
4	139	17	12
5	128	16	12
6	180	26	17
7	102	15	7
8	163	24	16
9	106	18	10
10	149	30	11

h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.

ISBN 978-607-8285-62-4

3.3.1.3 Los siguientes datos representan las calificaciones de estadística para una muestra aleatoria de 10 estudiantes de primer año de determinada institución de enseñanza superior, junto con sus calificaciones en un examen de inteligencia aplicado cuando aún cursaban el último año de secundaria y el número de periodos de clase perdidos por los 10 estudiantes que tomaron el curso de estadística.

Estudiante	Calificación de estadística Y	Calificación del examen X_1	Clases perdidas X_2
1	98	70	5
2	74	50	7
3	87	70	3
4	81	55	4
5	85	65	1
6	90	65	2
7	74	55	4
8	76	55	5
9	91	70	3
10	94	65	2

h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.

ISBN 978-607-8285-62-4

ANTECEDENTES

CONCEPTOS DE:

Población. Muestra. Variable. Tipos de variable. Escalas de medición de las variables. La media de la población. Tamaño de la muestra. Ejes cartesianos. Diagrama de dispersión. Covarianza de la muestra. Desviación estándar de la muestra para X. Desviación estándar de la muestra para Y. Sumas de cuadrados de la Regresión. Suama de Cuadrados Total.

3.3.2 COEFICIENTE DE DETERMINACIÓN Y DE CORRELACIÓN MÚLTIPLE

CONCEPTOS BÁSICOS COEFICIENTES DE DETERMINACIÓN Y CORRELACIÓN

El coeficiente de determinación se utiliza en el análisis de regresión para indicar en qué medida el modelo es capaz de predecir las respuestas de nuevas observaciones, en tanto que R^2 indica en qué medida el modelo se ajusta a sus datos.

R^2 pronosticada se calcula eliminando sistemáticamente cada una de las observaciones del conjunto de datos, estimando la ecuación de regresión y

En la **regresión múltiple**, ya que existen por lo menos dos variables explicativas, **el coeficiente de determinación múltiple** representa la proporción de la variación en Y que se explica por el grupo de variables explicativas seleccionadas.

En el caso de dos variables explicativas, el coeficiente de determinación múltiple $(r^2_{Y.12})$ se obtiene de la siguiente manera:

$$r^2_{Y.12} = \frac{SCR}{SCT}$$

Donde :

$$\boldsymbol{SCR = \widehat{\beta}_0 \sum_{i=1}^{n} Y + \widehat{\beta}_1 \sum_{i=1}^{n} X_1 Y + \widehat{\beta}_2 \sum_{i=1}^{n} X_2 Y - n\overline{Y}^2 = \widehat{\beta}'(X'Y) - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} (Variación\ explicada)}$$

y

$$SCT = SCR + SCE = \sum_{i=1}^{n} Y^2 - n\overline{Y}^2 = Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} (Variación\ total)$$

No obstante al tratar con **modelos de regresión múltiple**, algunos investigadores o analistas sugieren que se calcule **un R^2 "ajustado"** que refleje tanto el **número de variables explicativas** en el modelo como el tamaño de la muestra. En la **regresión múltiple** se puede representar un R^2 ajustado como:

$$r^2_{adj} = 1 - \left[(1 - r^2_{Y.12\ldots k})\frac{n-1}{n-k-1}\right]$$

ISBN 978-607-8285-62-4

determinando en qué medida el modelo es capaz de predecir la observación que se eliminó. R^2 pronosticada varía entre 0 y 100%.

Por lo general la fuerza de una relación entre **una variable dependiente Y** y **dos ó más variables independientes X** en una población se mide mediante el **coeficiente de correlación**, cuyos valores oscilan **entre -1** para la **correlación negativa** perfecta hasta **+1** para la **correlación positiva** perfecta.

El coeficiente de correlación.

Se puede obtener con facilidad el **coeficiente de correlación** mediante la fórmula:

$$r = \sqrt{r^2}$$

3.3.2.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 3.3.2.1 COEFICIENTES DE DETERMINACIÓN Y CORRELACIÓN

Samuel Prado, propietario y director general de un establecimiento quiere conocer el comportamiento de las ventas (en miles de pesos) de un equipo de sonido que se expende en el establecimiento. Se percata de que existen muchos factores que podrían ayudarle a explicar las ventas pero piensa que la inversión en publicidad (en miles de pesos) y el precio (en cientos de pesos) son los principales factores determinantes. Samuel ha reunido los datos que se anexan a continuación.

Observación	Ventas Y	Publicidad X_1	Precio X_2
1	33	3	125
2	61	6	115
3	70	10	140
4	82	13	130
5	17	9	145
6	24	6	140
7	75	11	138
8	80	12	127
9	35	4	128
10	20	8	145

i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados

ISBN 978-607-8285-62-4

El coeficiente de determinación múltiple.

Solución al inciso i.

$(r^2_{Y.12})$ se obtiene de la siguiente manera:

$$r^2_{Y.12} = \frac{SCR}{SCT}$$

Donde :

$$\boldsymbol{SCR = \widehat{\beta}_0 \sum_{i=1}^{n} Y + \widehat{\beta}_1 \sum_{i=1}^{n} X_1 Y + \widehat{\beta}_2 \sum_{i=1}^{n} X_2 Y - n\bar{Y}^2 = \widehat{\beta}'(X'Y)} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} (\boldsymbol{Variación\ explicada})$$

y

$$SCT = SCR + SCE = \sum_{i=1}^{n} Y^2 - n\bar{Y}^2 = Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} (Variación\ total)$$

De cálculos anteriores tenemos que

Suma de cuadrados de la regresión.

$$\boldsymbol{SCR} = \hat{\beta}'(X'Y) - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = [223.52438 \quad 6.56400 \quad -1.70780] \begin{bmatrix} 497 \\ 4{,}613 \\ 65{,}315 \end{bmatrix} - \frac{497^2}{10}$$

$$= 29{,}826.39186 - 24{,}700.9 \cong \mathbf{5,125.49186}$$

y

Suma de cuadrados total.

$$\boldsymbol{SCT = Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}}$$

$$= [\mathbf{33} \quad \mathbf{61} \quad \mathbf{70} \quad \mathbf{82} \quad \mathbf{17} \quad \mathbf{24} \quad \mathbf{75} \quad \mathbf{80} \quad \mathbf{35} \quad \mathbf{20}] \begin{bmatrix} \mathbf{33} \\ \mathbf{61} \\ \mathbf{70} \\ \mathbf{82} \\ \mathbf{17} \\ \mathbf{24} \\ \mathbf{75} \\ \mathbf{80} \\ \mathbf{35} \\ \mathbf{20} \end{bmatrix} - \frac{\mathbf{497^2}}{\mathbf{10}}$$

$$= \mathbf{30,949 - 24,700.9 \cong 6,248.1}$$

Por lo tanto

Coeficiente de determinación múltiple.

$$\boldsymbol{r^2_{Y.12}} = \frac{SCR}{SCT} = \frac{5{,}125.49186}{6{,}248.1} = 0.8203x\ 100 = \mathbf{82.03}\%$$

Interpretación: Este coeficiente de determinación múltiple, significa que el 82.03% de la variación del volumen de ventas se puede explicar mediante la variación de la inversión en publicidad y la variación en el precio del equipo de sonido.

ISBN 978-607-8285-62-4

¿Porqué R^2 ajustada?. Por ejemplo, usted trabaja en una firma de asesores financieros y está desarrollando un modelo para predecir condiciones futuras en los mercados. El modelo por el que se decidió luce prometedor porque tiene una R^2 de 87%. Sin embargo, al calcular R^2 ajustada, usted observa que ésta desciende a 52%. Esto podría indicar que el modelo está sobreajustado, y sugiere que el nivel de precisión con que su modelo predecirá nuevas observaciones no es, ni cercanamente, el mismo nivel de precisión con que se ajusta a los datos actuales.

No obstante al tratar con modelos de regresión múltiple, algunos investigadores o analistas sugieren que se calcule un R^2 "ajustado" que refleje tanto el número de variables explicativas en el modelo como el tamaño de la muestra. En la regresión múltiple se puede representar un R^2 ajustado como:

$$r^2_{adj} = 1 - \left[(1 - r^2_{Y.12\ldots k})\frac{n-1}{n-k-1}\right]$$

Por lo tanto

$$\boldsymbol{r^2_{adj}} = 1 - \left[(1 - r^2_{Y.12})\frac{n-1}{n-k-1}\right] = 1 - \left[(1 - 0.8203)\frac{9}{7}\right] =$$

$$0.769x100 = \mathbf{76.90}\%$$

Interpretación: Lo anterior nos dice que el 76.90% de la variación en el volumen de ventas se puede explicar mediante el modelo de regresión lineal múltiple-ajustado por el número de variables de predicción y el tamaño de la muestra.

Por lo general la fuerza de una relación entre dos variables en una población se mide mediante el **coeficiente de correlación**, cuyos valores oscilan entre -1 para la correlación negativa perfecta hasta +1 para la correlación positiva perfecta. Se puede obtener con facilidad el coeficiente de correlación mediante la fórmula:

Coeficiente de correlación múltiple

$$\boldsymbol{r_{Y.12..k}} = \sqrt{\boldsymbol{r^2_{Y.12\ldots k}}}$$

Entonces

$$\boldsymbol{r_{Y.12}} = \sqrt{r^2_{y.12}} = \sqrt{0.8203} = 0.90570\ x\ 100 = \mathbf{90.57}\%$$

Interpretación: En este problema del volumen de ventas, puesto que $r^2 = 0.8203$, el coeficiente de correlación se interpreta como ***+0.9057***. La cercanía del coeficiente de correlación con +1.0 implica una fuerte asociación del volumen de ventas (Y) con respecto a la inversión en publicidad(X_1) y el precio del equipo de sonido (X_2).

ISBN 978-607-8285-62-4

3.3.21 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 3.3.2.1 COEFICIENTES DE DETERMINACIÓN Y CORRELACIÓN

Para el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	11	9	4
2	30	28	14
3	26	17	11
4	20	19	9
5	15	20	8
6	25	24	11
7	35	32	12
8	17	13	21
9	39	36	9
10	45	40	17

i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

El coeficiente de determinación múltiple.

Solución al inciso i.

De cálculos anteriores tenemos que

Suma de cuadrados de la regresión.

$$\boldsymbol{SCR} = \hat{\beta}'(X'Y) - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} =$$

ISBN 978-607-8285-62-4

y

Suma de cuadrados total.

$$\boldsymbol{SCT = Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}} =$$

Por lo tanto

Coeficiente de determinación múltiple.

$$\boldsymbol{r^2_{Y.12}} = \frac{SCR}{SCT} =$$

Interpretación del coeficiente de determinación no ajustado

Interpretación:

No obstante al tratar con modelos de regresión múltiple, algunos investigadores o analistas sugieren que se calcule un R^2 "ajustado" que refleje tanto el número de variables explicativas en el modelo como el tamaño de la muestra. En la regresión múltiple se puede representar un R^2 ajustado como:

$$r^2_{adj} = 1 - \left[(1 - r^2_{Y.12\ldots k})\frac{n-1}{n-k-1}\right]$$

Por lo tanto

Coeficiente de determinación ajustado.

$$\boldsymbol{r^2_{adj}} =$$

Interpretación del coeficiente de determinación ajustado.

Interpretación:

Por lo general la fuerza de una relación entre dos variables en una población se mide mediante el **coeficiente de correlación**, cuyos valores oscilan entre -1 para la correlación negativa perfecta hasta +1 para la correlación positiva perfecta. Se puede obtener con facilidad el coeficiente de correlación mediante la fórmula:

ISBN 978-607-8285-62-4

Coeficiente de correlación múltiple.

$$r_{Y.12..k} = \sqrt{r^2_{Y.12...k}}$$

Entonces

$$r_{Y.12} = \sqrt{r^2_{y.12}} =$$

Interpretación del coeficiente de correlación múltiple

Interpretación:

3.3.2.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 3.3.2.1 COEFICIENTES DE DETERMINACIÓN Y CORRELACIÓN

Si tenemos el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	15	7.2	3.0
2	33	4.2	5.0
3	32	1.3	5.0
4	42	1.4	9.5
5	20	6.3	3.7
6	27	3.6	5.5
7	30	2.1	5.2
8	25	5.0	3.1
9	37	1.5	7.8
10	10	8.5	2.5

ISBN 978-607-8285-62-4

i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

El coeficiente de determinación múltiple.

<u>Solución al inciso i.</u>

De cálculos anteriores tenemos que

Suma de cuadrados de la regresión.

$$\boldsymbol{SCR} =$$

y

Suma de cuadrados total.

$$\boldsymbol{SCT} =$$

Por lo tanto

Coeficiente de determinación múltiple.

$$\boldsymbol{r^2_{Y.12}} =$$

Interpretación del coeficiente de determinación

Interpretación:

ISBN 978-607-8285-62-4

Coeficiente de determinación ajustado. Interpretación del coeficiente de determinación ajustado	No obstante al tratar con modelos de regresión múltiple, algunos investigadores o analistas sugieren que se calcule un R^2 "ajustado" que refleje tanto el número de variables explicativas en el modelo como el tamaño de la muestra. En la regresión múltiple se puede representar un R^2 ajustado como: $r^2_{adj} =$ **Interpretación:**
Coeficiente de correlación múltiple. Interpretación del coeficiente de correlación múltiple .	Por lo general la fuerza de una relación entre dos variables en una población se mide mediante el **coeficiente de correlación**, cuyos valores oscilan entre -1 para la correlación negativa perfecta hasta +1 para la correlación positiva perfecta. $\boldsymbol{r_{Y.12}} =$ **Interpretación:**

ISBN 978-607-8285-62-4

3.3.2 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 3.3.2 COEFICIENTES DE DETERMINACIÓN Y CORRELACIÓN

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

3.3.2.1 **S**uponga que una compañía de productos de consumo quisiera medir la efectividad de la publicidad en radio y televisión y la publicidad en periódicos en la promoción de sus productos. Se seleccionó una muestra aleatoria de 10 ciudades con poblaciones aproximadamente iguales para el estudio durante un periodo de prueba de un mes. Se registraron las ventas (en miles de pesos) durante el mes de prueba, junto con los niveles de gastos en los medios de publicidad, con los resultados siguientes:

Ciudad	Ventas ($000,000) Y	Publicidad en radio y televisión ($000) X_1	Publicidad en periódicos ($000) X_2
1	9.73	0	400
2	6.25	250	250
3	9.71	300	300
4	9.31	350	350
5	11.77	350	350
6	9.82	400	250
7	16.28	450	450
8	13.29	550	250
9	14.36	600	300
10	17.41	650	350

i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados.

3.3.2.2 **E**l gerente distrital de ventas de de un fabricante de automóviles estudia las ventas de éstos. En forma específica, quiere determinar qué factores influyen en el número de automóviles vendidos en una distribuidora. Para investigarlo, seleccionó al azar 10 distribuidoras. De éstas, obtiene el número de automóviles vendidos el mes pasado, los minutos de publicidad en radio comprados el mes pasado y el número de vendedores de tiempo completo contratados. La información es la siguiente:

ISBN 978-607-8285-62-4

Distribuidora	Automóviles vendidos el mes pasado Y	Publicidad (en minutos) X_1	Fuerza de ventas X_2
1	127	18	10
2	159	22	14
3	144	23	12
4	139	17	12
5	128	16	12
6	180	26	17
7	102	15	7
8	163	24	16
9	106	18	10
10	149	30	11

i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados.

3.3.2.3 **L**os siguientes datos representan las calificaciones de estadística para una muestra aleatoria de 10 estudiantes de primer año de determinada institución de enseñanza superior, junto con sus calificaciones en un examen de inteligencia aplicado cuando aún cursaban el último año de secundaria y el número de periodos de clase perdidos por los 10 estudiantes que tomaron el curso de estadística.

Estudiante	Calificación de estadística Y	Calificación del examen X_1	Clases perdidas X_2
1	98	70	5
2	74	50	7
3	87	70	3
4	81	55	4
5	85	65	1
6	90	65	2
7	74	55	4
8	76	55	5
9	91	70	3
10	94	65	2

i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados.

ISBN 978-607-8285-62-4

ANTECEDENTES

CONCEPTOS DE:

Población. Muestra. Variable. Tipos de variable. Escalas de medición de las variables. La media de la población. Coeficiente de determinación múltiple. Prueba *F* parcial. Desviación estándar de la muestra para X. Desviación estándar de la muestra para Y. Sumas de cuadrados de la Regresión. Suma de cuadrados parcial de la regresión. Suma de Cuadrados Total.

3.3.3 COEFICIENTES DE DETERMINACIÓN PARCIAL. PROPORCIÓN DE LA VARIACIÓN EN LA VARIABLE DEPENDIENTE

CONCEPTOS BÁSICOS
COEFICIENTES DE DETERMINACIÓN PARCIAL

Los **coeficientes de determinación parcial ($r^2_{y1.2}$ y $r^2_{y2.1}$)** miden la **proporción de la variación en la variable dependiente** que se explica **por cada variable explicativa**, al mismo tiempo que se **controlan o se mantienen constantes** las otras **variables explicativas**.

Para un modelo de regresión múltiple con diversas variables explicativas (k) resulta que:

$$r^2_{Yk}todas\ las\ variables\ exepto\ k = \frac{SCR(X_k todas\ las\ variables\ excepto\ k)}{SCT - SCR(todas\ las\ variables\ incluso\ k + SCR(X_k todas\ las\ variables\ excepto\ k)}$$

En un **modelo** con **dos variables explicativas** resultaría de la siguiente manera:

$$r^2_{Y1.2} = \frac{SCR(X_1 \mid X_2)}{SCT - SCR(X_1\ y\ X_2) + SCR(X_1 \mid X_2)}$$

Coeficiente de determinación parcial de $r^2_{Y1.2}$.

$$r^2_{Y2.1} = \frac{SCR(X_2 \mid X_1)}{SCT - SCR(X_1\ y\ X_2) + SCR(X_2 \mid X_1)}$$

ISBN 978-607-8285-62-4

Coeficiente de determinación parcial de $r^2_{Y2.1}$.

Donde:

SCR ($X_1 | X_2$) = suma de los cuadrados de la contribución de la variable X_1 al modelo de regresión conociendo que la variable X_2 ha sido incluida en el modelo.
SCT = Suma total de los cuadrados para Y.
SCR (X_1 y X_2) = suma de regresión de los cuadrados cuando tanto la variable X_1 como la X_2 están incluidas en el modelo de regresión múltiple.
SCR ($X_2 | X_1$) = suma de los cuadrados de la contribución de la variable X_2 al modelo de regresión, sabiendo que la variable X_1 ha sido incluida en el modelo.

3.3.3.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 3.3.3.1 COEFICIENTES DE DETERMINACIÓN PARCIAL

Samuel Prado, propietario y director general de un establecimiento quiere conocer el comportamiento de las ventas (en miles de pesos) de un equipo de sonido que se expende en el establecimiento. Se percata de que existen muchos factores que podrían ayudarle a explicar las ventas pero piensa que la inversión en publicidad (en miles de pesos) y el precio (en cientos de pesos) son los principales factores determinantes. Samuel ha reunido los datos que se anexan a continuación.

Observación	Ventas Y	Publicidad X_1	Precio X_2
1	33	3	125
2	61	6	115
3	70	10	140
4	82	13	130
5	17	9	145
6	24	6	140
7	75	11	138
8	80	12	127
9	35	4	128
10	20	8	145

j) Determine los coeficientes de determinación parcial e intérprete sus resultados.

ISBN 978-607-8285-62-4

Solución al inciso j.

Coeficientes de determinación parcial.

Se puede calcular el coeficiente de determinación parcial de $\mathbf{X_1}$ como,

Coeficiente de determinación parcial de $r^2_{Y1.2}$.

$$\boldsymbol{r^2_{Y1.2}} = \frac{\boldsymbol{SCR(X_1 \mid X_2)}}{\boldsymbol{SCT - SCR(X_1\ y\ X_2) + SCR(X_1 \mid X_2)}} = \frac{\mathbf{4,118.22080}}{\mathbf{6,248.1 - 5,125.49186 + 4,118.22080}} = \mathbf{0.7858}\boldsymbol{x}\mathbf{100} = \mathbf{78.58\%}$$

Interpretación

INTERPRETACIÓN: El coeficiente de determinación parcial de la variable dependiente Y, volumen de ventas, cuando se mantiene constante ***X₂***, el precio del equipo de sonido, significa que para un precio fijo (constante) en el equipo de sonido, el **78.58%** de la variación en el volumen de ventas se puede explicar por **la inversión en publicidad *X₁*.**

Y el coeficiente de determinación parcial de $\mathbf{X_2}$ como,

Coeficiente de determinación parcial de $r^2_{Y2.1}$.

$$\boldsymbol{r^2_{Y2.1}} = \frac{SCR(X_2 \mid X_1)}{SCT - SCR(X_1\ y\ X_2) + SCR(X_2 \mid X_1)} = \frac{2{,}335.78375}{6{,}248.1 - 5{,}125.49186 + 2{,}335.78375} = 0.6754x100 = \mathbf{67.54\%}$$

Interpretación

INTERPRETACIÓN: El coeficiente de determinación parcial de la variable dependiente Y, volumen de ventas, cuando se mantiene constante ***X₁***, la inversión en publicidad, significa que para una inversión en publicidad fija (constante), el **67.54%** de la variación en el volumen de ventas se puede explicar por **el precio del equipo de sonido.**

ISBN 978-607-8285-62-4

3.3.3.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE APRENDIZAJE 3.3.3.1 COEFICIENTES DE DETERMINACIÓN PARCIAL

Para el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	11	9	4
2	30	28	14
3	26	17	11
4	20	19	9
5	15	20	8
6	25	24	11
7	35	32	12
8	17	13	21
9	39	36	9
10	45	40	17

j) Determine los coeficientes de determinación parcial e intérprete sus resultados.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Coeficientes de determinación parcial.

Solución al inciso j.

Se puede calcular el coeficiente de determinación parcial de $\mathbf{X_1}$ como,

Coeficiente de determinación parcial de $r^2_{Y1.2}$.

$$r^2_{Y1.2} = \frac{SCR(X_1 \mid X_2)}{SCT - SCR(X_1\ y\ X_2) + SCR(X_1 \mid X_2)} =$$

ISBN 978-607-8285-62-4

Interpretación

INTERPRETACIÓN:

Coeficiente de determinación parcial de $r^2_{Y2.1}$.

Y el coeficiente de determinación parcial de $\mathbf{X_2}$ como,

$$\boldsymbol{r^2_{Y2.1}} = \frac{SCR(X_2|X_1)}{SCT - SCR(X_1\ y\ X_2) + SCR(X_2|X_1)} =$$

Interpretación

INTERPRETACIÓN:

ISBN 978-607-8285-62-4

3.3.3.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 3.3.3.1
COEFICIENTES DE DETERMINACIÓN PARCIAL

Si tenemos el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	15	7.2	3.0
2	33	4.2	5.0
3	32	1.3	5.0
4	42	1.4	9.5
5	20	6.3	3.7
6	27	3.6	5.5
7	30	2.1	5.2
8	25	5.0	3.1
9	37	1.5	7.8
10	10	8.5	2.5

j) Determine los coeficientes de determinación parcial e intérprete sus resultados.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Coeficientes de determinación parcial.

Solución al inciso j.

Se puede calcular el coeficiente de determinación parcial de $\mathbf{X_1}$ como,

Coeficiente de determinación parcial de $r^2_{Y1.2}$.

$$r^2_{Y1.2} =$$

ISBN 978-607-8285-62-4

Interpretación

INTERPRETACIÓN:

Coeficiente de determinación parcial de $r^2_{Y2.1}$.

Y el coeficiente de determinación parcial de $\mathbf{X_2}$ como,

$$\boldsymbol{r^2_{Y2.1}} =$$

Interpretación

INTERPRETACIÓN:

ISBN 978-607-8285-62-4

3.3.3 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 3.3.3 COEFICIENTES DE DETERMINACIÓN PARCIAL

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

3.3.3.1 **S**uponga que una compañía de productos de consumo quisiera medir la efectividad de la publicidad en radio y televisión y la publicidad en periódicos en la promoción de sus productos. Se seleccionó una muestra aleatoria de 10 ciudades con poblaciones aproximadamente iguales para el estudio durante un periodo de prueba de un mes. Se registraron las ventas (en miles de pesos) durante el mes de prueba, junto con los niveles de gastos en los medios de publicidad, con los resultados siguientes:

Ciudad	Ventas ($000,000) Y	Publicidad en radio y televisión ($000) X_1	Publicidad en periódicos ($000) X_2
1	9.73	0	400
2	6.25	250	250
3	9.71	300	300
4	9.31	350	350
5	11.77	350	350
6	9.82	400	250
7	16.28	450	450
8	13.29	550	250
9	14.36	600	300
10	17.41	650	350

j) Determine los coeficientes de determinación parcial e intérprete sus resultados.

3.3.3.2 **E**l gerente distrital de ventas de de un fabricante de automóviles estudia las ventas de éstos. En forma específica, quiere determinar qué factores influyen en el número de automóviles vendidos en una distribuidora. Para investigarlo, seleccionó al azar 10 distribuidoras. De éstas, obtiene el número de automóviles vendidos el mes pasado, los minutos de publicidad en radio comprados el mes pasado y el número de vendedores de tiempo completo contratados. La información es la siguiente:

ISBN 978-607-8285-62-4

Distribuidora	Automóviles vendidos el mes pasado Y	Publicidad (en minutos) X_1	Fuerza de ventas X_2
1	127	18	10
2	159	22	14
3	144	23	12
4	139	17	12
5	128	16	12
6	180	26	17
7	102	15	7
8	163	24	16
9	106	18	10
10	149	30	11

j) Determine los coeficientes de determinación parcial e intérprete sus resultados.

3.3.3.3 Los siguientes datos representan las calificaciones de estadística para una muestra aleatoria de 10 estudiantes de primer año de determinada institución de enseñanza superior, junto con sus calificaciones en un examen de inteligencia aplicado cuando aún cursaban el último año de secundaria y el número de periodos de clase perdidos por los 10 estudiantes que tomaron el curso de estadística.

Estudiante	Calificación de estadística Y	Calificación del examen X_1	Clases perdidas X_2
1	98	70	5
2	74	50	7
3	87	70	3
4	81	55	4
5	85	65	1
6	90	65	2
7	74	55	4
8	76	55	5
9	91	70	3
10	94	65	2

j) Determine los coeficientes de determinación parcial e intérprete sus resultados.

ISBN 978-607-8285-62-4

ANTECEDENTES

CONCEPTOS DE:

Variable aleatoria. Tipos de variable. Variable dependiente. Variable independiente. Ecuación de tendencia lineal. Ordenada al origen. Pendiente de la recta. Relación directa de dos variables. Coeficiente de determinación múltiple. Covarianza de X_1X_2. Desviación estándar de X_1. Desviación estándar de X_2.

3.3.4 MULTICOLINEARIDAD. FACTOR DE VARIANZA INFLACIONARIA (VIF)

CONCEPTOS BÁSICOS MULTICOLINEARIDAD

En regresión, la multicolinealidad se refiere a los predictores que están correlacionados con otros predictores. La multicolinealidad moderada pudiera no representar un problema. Sin embargo, la multicolinealidad severa es problemática porque puede incrementar la varianza de los coeficientes de regresión, haciéndolos inestables y difíciles de interpretar.

En general no hay relación importante entre el coeficiente de determinación múltiple R^2 de la ecuación de regresión y los coeficientes individuales de determinación. Si todos las **variables independientes no están correlacionadas** entre sí, se pueden ir **añadiendo coeficientes individuales** de determinación; sin embargo, **si las X están correlacionadas**, es difícil **separar el valor predictivo global de X_1, X_2,...,X^k**, tal como se mide con $R^2_{Y.X1...Xk}$, en partes separadas que se puedan atribuir solamente a X_1,...,solamente a X_k. Por lo tanto, un problema importante en la aplicación del análisis de regresión múltiple incluye **la posible correlación de las variables independientes ó explicativas** (llamada en ocasiones **multicolinealidad ó multicolinearidad)**. Esta condición se refiere a situaciones en que algunas variables explicativas estén **altamente correlacionadas entre sí**. En esas situaciones las **variables correlacionadas** no proporcionan información nueva y resulta difícil separar el efecto de esas variables sobre la variable dependiente o de respuesta. En esos casos los valores de los coeficientes de regresión para las variables correlacionadas pueden fluctuar en forma importante, dependiendo de qué variables estén incluidas en el modelo.

Un método de medir la **colinealidad ó colinearidad** usa el factor de varianza inflacionaria **(VIF)** para cada variable explicativa. Este **VIF** se define en la siguiente ecuación:

ISBN 978-607-8285-62-4

¿Porqué medir la colinealidad?. Por ejemplo, un fabricante de juguetes desea pronosticar la satisfacción del cliente por medio de los resultados de un sondeo que inicialmente incluye "resistencia" y "falta de roturas" como variables predictoras en el modelo de regresión. El investigador encuentra que estas dos variables están fuerte y negativamente relacionadas y tienen un VIF mayor que 5. En este punto, el investigador podría tratar de retirar la variable .

$$VIF_j = \frac{1}{1 - r_j^2}$$

Donde r_j^2 representa el coeficiente de determinación múltiple de la variable explicativa X_j con todas las otras variables X.

Cuando sólo hay dos variables explicativas r_j^2 es el coeficiente de determinación entre X_1 y X_2. Si hubiera tres variables explicativas, entonces r_j^2 sería el coeficiente de determinación múltiple de X_1 con X_2 y X_3.

Cuando un grupo de variables explicativas no están correlacionadas, entonces **VIF_j** será **del orden de 1**. Si el grupo presenta una alta correlación entre sí, entonces **VIF_j** podría **exceder a 10** aunque algunos analistas o investigadores sugieren un criterio más conservador donde se emplearían alternativas a la regresión de mínimos cuadrados si el **VIF_j máximo excediera a 5.**

Se sugiere **aplicar las siguientes directrices** para **interpretar el VIF:**

VIF ≅ 1	No correlacionados
1<VIF≤5	Ligeramente correlacionados
5<VIF≤10	Moderadamente correlacionados
VIF>10	Altamente correlacionados

Los **valores de VIF mayores que 10** podrían indicar que la multicolinealidad estaría incidiendo excesivamente en los resultados de su regresión. En este caso, convendría **reducir la multicolinealidad eliminando los predictores irrelevantes de su modelo**.

Puesto que solo hay dos variables explicatorias en el modelo, se puede calcular el **VIF_j** de la siguiente manera

$$VIF_1 = VIF_2 = \frac{1}{1 - r_{X_1X_2}^2}$$

Primero se debe calcular $r_{X_1X_2}^2$, es decir el coeficiente de determinación utilizando únicamente las dos variables independientes **X_1** y **X_2** mediante las siguientes fórmulas:

ISBN 978-607-8285-62-4

Coeficiente de determinación de X_1 y X_2.

$$r^2_{X_1X_2} = \left[\frac{cov(X_1X_2)}{S_{X_1}S_{X_2}}\right]^2$$

Covarianza de X_1 y X_2.

$$cov(X_1X_2) = \frac{\sum[(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)]}{n-1}$$

Desviación estándar de X_1

$$S_{X_1} = \sqrt{\frac{\sum_{i=1}^{n} X_1^2 - n\bar{X}_1^2}{n-1}}$$

Desviación estándar de X_2

$$S_{X_2} = \sqrt{\frac{\sum_{i=1}^{n} X_2^2 - n\bar{X}_2^2}{n-1}}$$

NOTA: Si el cálculo se hace con calculadora usando el módulo de regresión lineal simple, se debe calcular el modelo suponiendo un modelo de regresión lineal simple utilizando como **Y** a $\mathbf{X_1}$ y como $\mathbf{X_1}$ a $\mathbf{X_2}$ y posteriormente encontrando el valor de **r** y elevándolo al cuadrado para obtener $r^2_{X_1X_2}$.

ISBN 978-607-8285-62-4

3.3.4.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 3.3.4.1 MULTICOLINEARIDAD

Multicolinearidad.

Samuel Prado, propietario y director general de un establecimiento quiere conocer el comportamiento de las ventas (en miles de pesos) de un equipo de sonido que se expende en el establecimiento. Se percata de que existen muchos factores que podrían ayudarle a explicar las ventas pero piensa que la inversión en publicidad (en miles de pesos) y el precio (en cientos de pesos) son los principales factores determinantes. Samuel ha reunido los datos que se anexan a continuación.

Observación	Ventas Y	Publicidad X_1	Precio X_2
1	33	3	125
2	61	6	115
3	70	10	140
4	82	13	130
5	17	9	145
6	24	6	140
7	75	11	138
8	80	12	127
9	35	4	128
10	20	8	145

k) Verifique la existencia de multicolinearidad

Solución al inciso k.

Construimos la siguiente tabla:

Obs.	**Publicidad**	**Precio**					
	X_1	X_2	$X_1 - \bar{X}_1$	$X_2 - \bar{X}_2$	$(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)$	X_1^2	X_2^2
1	3	125	-5.2	-8.3	43.16	9	15,625
2	6	115	-2.2	-18.3	40.26	36	13,225
3	10	140	1.8	6.7	12.06	100	19,600
4	13	13	4.8	-3.3	-15.84	169	16,900

ISBN 978-607-8285-62-4

5	9	145	0.8	11.7	9.36	81	21,025
6	6	140	-2.2	6.7	-14.74	36	19,600
7	11	138	2.8	4.7	13.16	121	19,044
8	12	127	3.8	-6.3	-23.94	144	16,129
9	4	128	-4.2	-5.3	22.26	16	16,384
10	8	145	-0.2	11.7	-2.34	64	21,025
SUMA	**82**	**1,333**			**83.4**	**776**	**178,557**
	PROMEDIO						
	8.2	**133.3**					

Covarianza de X_1 y X_2.

$$cov(X_1X_2) = \frac{\sum[(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)]}{n-1} = \frac{83.4}{9} = 9.26667$$

Desviación estándar de X_1.

$$S_{X_1} = \sqrt{\frac{\sum_{i=1}^{n} X_1^2 - n\bar{X}_1^2}{n-1}} = \sqrt{\frac{776 - 10(8.2)^2}{9}} = 3.39280$$

Desviación estándar de X_2.

$$S_{X_2} = \sqrt{\frac{\sum_{i=1}^{n} X_2^2 - nX_2^2}{n-1}} = \sqrt{\frac{178{,}557 - 10(133.3)^2}{9}} = 9.82118$$

Coeficiente de determinación de X_1 y X_2.

$$r_{X_1X_2}^2 = \left[\frac{cov(X_1X_2)}{S_{X_1}S_{X_2}}\right]^2 = \left[\frac{9.26667}{(3.39280)(9.82118)}\right]^2 = [0.27810]^2 = 0.07734$$

Factor de varianza inflacionaria VIF.

Finalmente,

$$VIF_1 = VIF_2 = \frac{1}{1 - r_{X_1X_2}^2} = \frac{1}{1 - 0.07734} = 1.08382$$

Se sugiere aplicar las siguientes directrices para interpretar el VIF:

VIF ≅ 1	No correlacionados.
1<VIF≤5	Ligeramente correlacionados.
5<VIF≤10	Moderadamente correlacionados.
VIF>10	Altamente correlacionados.

INTERPRETACIÓN: como el valor de **VIF** de **1.08382 < 5** podemos llegar a la conclusión de que no hay razones para sospechar multicolinearidad alguna entre la variable $\mathbf{X_1}$ (Inversión en publicidad) y $\mathbf{X_2}$ (Precio del producto), por lo que no se debe eliminar ninguna de las dos variables.

NOTA: Si el cálculo se hace con calculadora usando el módulo de regresión lineal simple, se debe calcular el modelo suponiendo un modelo de regresión lineal simple utilizando como **Y** a $\mathbf{X_1}$ y como $\mathbf{X_1}$ a $\mathbf{X_2}$ y posteriormente encontrando el valor de **r** y elevándolo al cuadrado para obtener $\boldsymbol{r_{X_1X_2}^2}$.

ISBN 978-607-8285-62-4

Usando la calculadora:

$\boxed{\text{SHIFT}}\ \boxed{\text{S} - \text{VAR}}\ \boxed{\text{REPLAY} \rightarrow}\ \boxed{\text{REPLAY} \rightarrow}\ \boxed{3}\ \boxed{X^2}\ \boxed{=}\ 0.07734$

3.3.4.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE **APRENDIZAJE 3.3.4.1 M**ULTI**COLINEARIDAD**

Para el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	11	9	4
2	30	28	14
3	26	17	11
4	20	19	9
5	15	20	8
6	25	24	11
7	35	32	12
8	17	13	21
9	39	36	9
10	45	40	17

k) Verifique la existencia de multicolinearidad e interprete su resultado.

ISBN 978-607-8285-62-4

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Multicolinearidad.

Solución al inciso k.

Construimos la siguiente tabla:

Obs.							
	X_1	X_2	$X_1-\bar{X}_1$	$X_2-\bar{X}_2$	$(X_1-\bar{X}_1)(X_2-\bar{X}_2)$	X_1^2	X_2^2
1							
2							
3							
5							
6							
7							
8							
9							
10							
SUMA							
	PROMEDIO						

Covarianza de X_1 y X_2.

$$cov(X_1X_2)=\frac{\sum[(X_1-\bar{X}_1)(X_2-\bar{X}_2)]}{n-1}=$$

Desviación estándar de X_1 .

$$S_{X_1}=\sqrt{\frac{\sum_{i=1}^{n}X_1^2-n\bar{X}_1^2}{n-1}}=$$

Desviación estándar de X_2 .

$$S_{X_2}=\sqrt{\frac{\sum_{i=1}^{n}X_2^2-nX_2^2}{n-1}}=$$

ISBN 978-607-8285-62-4

Coeficiente de determinación de X_1 y X_2.

$$r^2_{X_1X_2} = \left[\frac{cov(X_1X_2)}{S_{X_1}S_{X_2}}\right]^2 =$$

Finalmente,

Factor de varianza inflacionaria VIF.

$$VIF_1 = VIF_2 = \frac{1}{1 - r^2_{X_1X_2}} =$$

NOTA: Si el cálculo se hace con calculadora usando el módulo de regresión lineal simple, se debe calcular el modelo suponiendo un modelo de regresión lineal simple utilizando como **Y** a $\mathbf{X_1}$ y como $\mathbf{X_1}$ a $\mathbf{X_2}$ y posteriormente encontrando el valor de **r** y elevándolo al cuadrado para obtener $\boldsymbol{r^2_{X_1X_2}}$.

Se sugiere aplicar las siguientes directrices para interpretar el VIF:

VIF ≅ 1	No correlacionados.
1<VIF≤5	Ligeramente correlacionados.
5<VIF≤10	Moderadamente correlacionados.
VIF>10	Altamente correlacionados.

Usando la calculadora:

$\boxed{\textbf{SHIFT}}\ \boxed{\textbf{S}-\textbf{VAR}}\ \boxed{\textbf{REPLAY}\rightarrow}\ \boxed{\textbf{REPLAY}\rightarrow}\ \boxed{\textbf{3}}\ \boxed{\textbf{X}^2}\ \boxed{=}$

INTERPRETACIÓN:

ISBN 978-607-8285-62-4

3.3.4.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 3.3.4.1 MULTICOLINEARIDAD

Si tenemos el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	15	7.2	3.0
2	33	4.2	5.0
3	32	1.3	5.0
4	42	1.4	9.5
5	20	6.3	3.7
6	27	3.6	5.5
7	30	2.1	5.2
8	25	5.0	3.1
9	37	1.5	7.8
10	10	8.5	2.5

k) Verifique la existencia de multicolinearidad e interprete su resultado.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ISBN 978-607-8285-62-4

Multicolinearidad.

Solución al inciso k.

Construimos la siguiente tabla:

Obs.							
	X_1	**X_2**	$X_1 - \bar{X}_1$	$X_2 - \bar{X}_2$	$(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)$	X_1^2	X_2^2
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
SUMA							
	PROMEDIO						

Covarianza de X_1 y X_2..

$$cov(X_1X_2) =$$

Desviación estándar de X_1 .

$$S_{X_1} =$$

Desviación estándar de X_2 .

$$S_{X_2} =$$

Coeficiente de determinación de X_1 y X_2..

$$r^2_{X_1X_2} =$$

Factor de varianza inflacionaria VIF.

Finalmente,

$$VIF_1 = VIF_2 =$$

ISBN 978-607-8285-62-4

NOTA: Si el cálculo se hace con calculadora usando el módulo de regresión lineal simple, se debe calcular el modelo suponiendo un modelo de regresión lineal simple utilizando como **Y** a $\mathbf{X_1}$ y como $\mathbf{X_1}$ a $\mathbf{X_2}$ y posteriormente encontrando el valor de **r** y elevándolo al cuadrado para obtener $r^2_{X_1X_2}$.

Usando la calculadora:

$\boxed{\text{SHIFT}}\boxed{\text{S}-\text{VAR}}\boxed{\text{REPLAY}\rightarrow}\boxed{\text{REPLAY}\rightarrow}\boxed{3}\boxed{X^2}\boxed{=}$

INTERPRETACIÓN:

ISBN 978-607-8285-62-4

3.3.4 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 3.3.4 MULTICOLINEARIDAD

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

3.3.4.1 **S**uponga que una compañía de productos de consumo quisiera medir la efectividad de la publicidad en radio y televisión y la publicidad en periódicos en la promoción de sus productos. Se seleccionó una muestra aleatoria de 10 ciudades con poblaciones aproximadamente iguales para el estudio durante un periodo de prueba de un mes. Se registraron las ventas (en miles de pesos) durante el mes de prueba, junto con los niveles de gastos en los medios de publicidad, con los resultados siguientes:

Ciudad	Ventas ($000,000) Y	Publicidad en radio y televisión ($000) X_1	Publicidad en periódicos ($000) X_2
1	9.73	0	400
2	6.25	250	250
3	9.71	300	300
4	9.31	350	350
5	11.77	350	350
6	9.82	400	250
7	16.28	450	450
8	13.29	550	250
9	14.36	600	300
10	17.41	650	350

k) Verifique la existencia de multicolinearidad e interprete su resultado.

3.3.4.2 **E**l gerente distrital de ventas de de un fabricante de automóviles estudia las ventas de éstos. En forma específica, quiere determinar qué factores influyen en el número de automóviles vendidos en una distribuidora. Para investigarlo, seleccionó al azar 10 distribuidoras. De éstas, obtiene el número de automóviles vendidos el mes pasado, los minutos de publicidad en radio comprados el mes pasado y el número de vendedores de tiempo completo contratados. La información es la siguiente:

ISBN 978-607-8285-62-4

Distribuidora	Automóviles vendidos el mes pasado Y	Publicidad (en minutos) X_1	Fuerza de ventas X_2
1	127	18	10
2	159	22	14
3	144	23	12
4	139	17	12
5	128	16	12
6	180	26	17
7	102	15	7
8	163	24	16
9	106	18	10
10	149	30	11

k) Verifique la existencia de multicolinearidad e interprete su resultado.

3.3.4.3 Los siguientes datos representan las calificaciones de estadística para una muestra aleatoria de 10 estudiantes de primer año de determinada institución de enseñanza superior, junto con sus calificaciones en un examen de inteligencia aplicado cuando aún cursaban el último año de secundaria y el número de periodos de clase perdidos por los 10 estudiantes que tomaron el curso de estadística.

Estudiante	Calificación de estadística Y	Calificación del examen X_1	Clases perdidas X_2
1	98	70	5
2	74	50	7
3	87	70	3
4	81	55	4
5	85	65	1
6	90	65	2
7	74	55	4
8	76	55	5
9	91	70	3
10	94	65	2

k) Verifique la existencia de multicolinearidad e interprete su resultado.

ISBN 978-607-8285-62-4

OBJETIVO 3.4. El alumno podrá calcular los residuales estandarizados y determinará lo apropiado del ajuste del modelo.

ANTECEDENTES

CONCEPTOS DE:

Variable dependiente. Variable independiente. Coeficientes de regresión. Valor estimado de Y. Error aleatorio. Tipos de relación. Relación lineal. Correlación. Tablas de frecuencia. Histograma. Diagrama de tallo y hojas. Diagrama de caja y brazos. Valores atípicos. Distribuciones de probabilidad. Características de la distribución normal. Varianza muestral. Desviación estándar muestral. Elementos de la matriz sombrero.

3.4.1 ANÁLISIS DE RESIDUALES. DIAGNÓSTICO DE LA REGRESIÓN

CONCEPTOS BÁSICOS ANÁLISIS DE RESIDUALES

Un residuo es la diferencia entre un valor observado (y)

En análisis de regresión un **residual** es: $\varepsilon_i = Y_i - \hat{Y}$

La **gráfica de residuales** se puede definir como una gráfica de los residuales e_i con respecto a la variable independiente X_i. En el **análisis de regresión lineal simple**, puede utilizarse un diagrama de dispersión o una gráfica de residuales para observar si "parecen" satisfacerse las suposiciones de l**inealidad, normalidad y homocedasticidad de la regresión**. Sin embargo en el **análisis de regresión múltiple**, el único tipo de gráfica que permite abordar este análisis para el modelo global es la gráfica de residuales con respecto al valor ajustado $\hat{Y}$, porque esta es la única gráfica bidimensional que puede incluir el uso de ***varias variables independientes*** (solo se puede construir con la computadora). Si se observa en una de esas gráficas que existe algún problema en los supuestos de la regresión, entonces pueden elaborarse **gráficas individuales de residuales para cada variable independiente del modelo**, con el objeto de buscar la fuente del problema, en cuyo caso es conveniente calcular los "residuales estandarizados"; éstos representan cada residual dividido entre su error

ISBN 978-607-8285-62-4

y su valor ajustado correspondiente (ŷ). Los valores residuales son útiles especialmente en procedimientos de regresión y ANOVA porque ellos indican el grado hasta el cual un modelo representa la variación en los datos observados.

estándar. El residual estandarizado en la regresión lineal múltiple se presenta como la ecuación:

$$SR_i = \frac{\varepsilon_i}{S_{Y.12\ldots k}\sqrt{1-h_i}}$$

Donde:

$$h_i = X_i^{'}(X^{'}X)^{-1}X_i$$

Los residuales estandarizados son útiles en la detección de valores atípicos. El residuo estandarizado es igual al valor de un residuo, ε_i, divido entre el error estándar. Los residuos estandarizados mayores que 2 y menores que -2 usualmente son considerados grandes .

En particular si el **modelo tiene dos variables explicatorias** el **residual estandarizado** se presentaría como la ecuación:

$$SR_i = \frac{\varepsilon_i}{S_{Y.12}\sqrt{1-h_i}}$$

Donde:

$$h_i = X_i^{'}(X^{'}X)^{-1}X_i$$

Estos valores estandarizados permiten considerar la magnitud de los residuales en unidades que reflejan la variación estandarizada en torno al plano de regresión. **Los residuales estandarizados** se trazan con respecto al valor ajustado $\hat{Y}$. Si parece que los **residuales estandarizados varían para diferentes niveles de $\hat{Y}$** , hay un **posible efecto curvilíneo en por lo menos una variable explicatoria** y/o la necesidad **de transformar la variable dependiente.**

Los **patrones** en el **diagrama de los residuales estandarizados**, **en contraste con una variable explicatoria**, pueden señalar la existencia de un **efecto curvilíneo** y por consiguiente, llevar a la posible **transformación de esa variable explicatoria.**

El estadístico de Durbin-Watson es una prueba para detectar la presencia de autocorrelación en los residuos. La autocorrelación significa que las observaciones adyacentes están correlacionadas. Si están correlacionadas, la regresión de los cuadrados mínimos subestima el error estándar de los coeficientes; sus predictores podrían parecer significativos, cuando en realidad es posible que no lo sean.

Por otro lado una de las **hipótesis más importantes del análisis de regresión** es que los **términos de error** ($\boldsymbol{\varepsilon_i}$) , que se podrían llamar los **"residuos verdaderos"** , **son independientes**. Gran parte de la teoría estadística de la regresión depende de esta hipótesis. Los datos **de series temporales**, medidos en periodos sucesivos, a menudo muestran un **comportamiento más o menos cíclico**. Este problema restringido principalmente a los datos de series temporales, se llama **autocorrelación.** Una prueba formal para la autocorrelación se apoya en el **estadístico de Durbin-Watson**. El estadístico de Durbin-Watson es:

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2}$$

donde: e_i = residual del periodo *i*.

ISBN 978-607-8285-62-4

Si los **verdaderos errores son en realidad independientes**, el valor esperado de ***d*** es alrededor de 2.0. Cualquier valor de ***d*** menor que **1.5 o 1.6** nos lleva a sospechar que **hay autocorrelación.**

Evaluación de las suposiciones

Linealidad.

Linealidad

Se puede evaluar lo apropiado del modelo de regresión, trazando **los "residuales estandarizados"** con respecto al valor ajustado $\hat{Y}$. Si parece que los residuales estandarizados varían para diferentes niveles de $\hat{Y}$, hay un posible efecto curvilíneo en por lo menos una variable explicatoria y/o la necesidad de transformar la variable dependiente. Si se observa en la gráfica que existe algún problema, entonces pueden elaborarse **gráficas individuales de residuales para cada variable independiente del modelo**, con el objeto de buscar la fuente del problema.

Homoscedasticidad.

Homoscedasticidad

La suposición de **homoscedasticidad** se puede evaluar también de la gráfica de residuales estandarizados con respecto al valor ajustado $\hat{Y}$.. Si parece haber un **"efecto de abanico"** en el cual **aumenta ó disminuye la variabilidad de los residuales al aumentar** $\hat{Y}$ se demuestra la **falta de homogeneidad en las varianzas de** Y_i **a cada nivel de** $\hat{Y}$**.** Si se observa en la gráfica que existe algún problema, entonces pueden elaborarse **gráficas individuales de residuales para cada variable independiente del modelo**, con el objeto de buscar la fuente del problema.

Normalidad.

Normalidad

El supuesto de **normalidad** en la regresión es posible evaluarlo de un análisis residual colocando los **residuales estandarizados** en una **distribución de frecuencias** y mostrando los resultados en un **histograma**. Si el **histograma de frecuencias** de los residuales **no se ajusta al de una normal pueden existir valores atípicos**. Un **valor atípico** es un valor inusualmente grande o pequeño. Los **valores atípicos** pueden tener una **influencia desproporcionada** sobre los resultados estadísticos, como la media, lo que puede generar interpretaciones engañosas.

ISBN 978-607-8285-62-4

Un histograma de residuales es una herramienta exploratoria que muestra las características generales de los datos, incluyendo:

- Valores típicos, dispersión o variación y forma
- Valores inusuales en los datos

La presencia de largas colas en la gráfica podría indicar sesgo en los datos. Si una o dos barras están lejos de las demás, esos puntos pueden ser valores atípicos. Debido a que el aspecto del histograma cambia según el número de intervalos utilizados para agrupar los datos, utilice la gráfica de probabilidad normal y las pruebas de bondad de ajuste para evaluar la normalidad de los residuos.

A menudo es más fácil identificar gráficamente los valores atípicos mediante un diagrama de caja y brazos. Por ejemplo, una compañía rastrea los pagos atrasados sobre la base de la fecha de vencimiento en número de días. La gráfica de caja y brazos muestra dos valores atípicos, indicando dos cuentas que tienen un atraso exagerado. Un analista investiga las cuentas y descubre que los clientes se mudaron y nunca recibieron sus estados de cuenta.

Es necesario **investigar los valores atípicos**, porque pueden proporcionar información útil sobre sus datos o proceso. Existen varias explicaciones de los valores atípicos:

· **Error de entrada de datos**: Corrija el error y vuelva a analizar los datos

· **Problema del proceso**: Investigue el proceso para determinar la causa del valor atípico

· **Factor faltante**: Determine si no consideró un factor que tiene influencia sobre el proceso

· **Probabilidad aleatoria**: Investigue el proceso y el valor atípico para determinar si éste ocurrió por casualidad; realice el análisis con y sin el valor atípico para ver su impacto sobre los resultados .

A menudo es más fácil **identificar gráficamente los valores atípicos** mediante **una gráfica de caja**, al etiquetar las observaciones que son **por lo menos 1.5 veces el rango intercuartil (Q3 – Q1)** desde el **borde de la caja**. Por ejemplo, una compañía rastrea los pagos atrasados sobre la base de la fecha de vencimiento en número de días. La gráfica de caja siguiente muestra dos valores atípicos, indicando dos cuentas que tienen un atraso exagerado. Un analista investiga las cuentas y descubre que los clientes se mudaron y nunca recibieron sus estados de cuenta. Eliminando estos valores atípicos, se puede conseguir **normalidad** en los residuos.

Si contamos con papel normal o acceso a la computadora, podemos construir una gráfica de probabilidad normal de residuos: **Los puntos de esta gráfica deben generalmente formar una línea recta si los residuos se están normalmente distribuidos.** Si los puntos en la gráfica salen de una línea recta, el supuesto de normalidad puede ser inválido. Si sus datos tienen menos de 50 observaciones, la gráfica podría mostrar una curvatura en las colas, aun si los residuos están normalmente distribuidos. A medida que el número de observaciones disminuye, la gráfica de probabilidad podría mostrar una variación sustancial no linealidad, aun si los residuos están normalmente distribuidos. Utilice la gráfica de probabilidad y las pruebas de bondad de ajuste, tales como el **estadístico de Anderson-Darling**, para evaluar si los residuos están normalmente distribuidos.

ISBN 978-607-8285-62-4

Estadístico de Anderson-Darling

El estadístico de **Anderson-Darling** mide en este caso si los datos siguen una distribución normal. Mientras **mejor se ajuste la distribución** a los datos, **menor será este estadístico**. Utilice el **estadístico de Anderson-Darling** para comparar el ajuste de varias distribuciones para ver cual es la mejor o probar si una muestra de datos proviene de una población con una distribución normal. Las hipótesis para la prueba de Anderson-Darling son:

H_0: Los datos siguen una distribución normal

H_1: Los datos no siguen una distribución normal

Si el **valor *p*** (al estar disponible) para la **prueba de Anderson-Darling** es **inferior al nivel de significación seleccionado** (generalmente 0.05 ó 0.10), **concluya** que **los datos no siguen la distribución normal**.

Independencia

Independencia

La suposición de **independencia** requiere que el error (diferencia "residual" entre un valor observado y uno predicho de Y) sea independiente para cada valor de X. Con frecuencia esta suposición se refiere a datos que se recopilan a lo largo de un periodo. Estos tipos de modelos caen bajo la denominación general de series de tiempo. **La suposición de independencia se puede evaluar trazando los residuales en el orden o la sucesión en que se obtuvieron los datos observados.**

El estadístico de Durbin-Watson es una prueba para detectar la presencia de autocorrelación en los residuos. La autocorrelación significa que las observaciones adyacentes están correlacionadas. Si están correlacionadas, la regresión de los cuadrados mínimos subestima el error estándar de los coeficientes; sus predictores podrían parecer significativos, cuando en realidad es posible que no lo sean.

Una prueba formal para la autocorrelación se apoya en el **estadístico de Durbin-Watson**. El estadístico de Durbin-Watson es:

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2}$$

donde: e_i = residual del periodo i.

Si los verdaderos errores son en realidad independientes, el valor esperado de ***d* es alrededor de 2.0**. Cualquier valor de ***d*** menor que 1.5 o 1.6 nos lleva a sospechar que **hay autocorrelación.**

ISBN 978-607-8285-62-4

3.4.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 3.4.1.1 ANÁLISIS DE RESIDUALES

Samuel Prado, propietario y director general de un establecimiento quiere conocer el comportamiento de las ventas (en miles de pesos) de un equipo de sonido que se expende en el establecimiento. Se percata de que existen muchos factores que podrían ayudarle a explicar las ventas pero piensa que la inversión en publicidad (en miles de pesos) y el precio (en cientos de pesos) son los principales factores determinantes. Samuel ha reunido los datos que se anexan a continuación.

Observación	Ventas Y	Publicidad X_1	Precio X_2
1	33	3	125
2	61	6	115
3	70	10	140
4	82	13	130
5	17	9	145
6	24	6	140
7	75	11	138
8	80	12	127
9	35	4	128
10	20	8	145

l) Determine los residuales estandarizados para toda la regresión incluyendo el estadístico de Durbin-Watson.

m) Construya la(s) gráfica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.

Cálculo de los residuales estandarizados.

<u>Solución al inciso l.</u>

El residual estandarizado se presenta como la ecuación:

$$SR_i = \frac{\varepsilon_i}{S_{Y.12}\sqrt{1-h_i}}$$

En modelos de regresión h_i mide la distancia de un valor x de observación hasta el promedio de los valores x para todas las observaciones en un conjunto de datos.

Donde,

$$h_i = X_i^{'}(X^{'}X)^{-1}X_i$$

Por tanto para calcular el primer residual haremos lo siguiente:

$\varepsilon_1 = Y_1 - \hat{Y}_1$ donde $\hat{Y}_1 = 223.52438 + 6.56400(3) - 1.70780(125) = \mathbf{29.74138}$

ISBN 978-607-8285-62-4

entonces,

Residuales no estandarizados.

$$\boldsymbol{\varepsilon_1} = 33 - 29.74138 = \mathbf{3.25862}$$

Y así sucesivamente con los siguientes 9 datos.

Para estandarizar estos residuales haremos lo siguiente:

Elementes de la matriz sombrero.

$$h_1 = X_1'(X'X)^{-1}X_1$$
$$= [1 \quad 3 \quad 125]\begin{bmatrix} 17'252{,}036/829{,}796 & 39{,}988/829{,}796 & -131{,}260/829{,}796 \\ 39{,}988/829{,}796 & 8{,}681/829{,}796 & -834/829{,}796 \\ -131{,}260/829{,}796 & -834/829{,}796 & 1{,}036/829{,}796 \end{bmatrix}\begin{bmatrix} 1 \\ 11 \\ 135 \end{bmatrix}$$
$$= [1.16233 \quad -0.04606 \quad -0.00514]\begin{bmatrix} 1 \\ 3 \\ 125 \end{bmatrix} \cong 0.38165$$

Residuales estandarizados.

Entonces

$$\boldsymbol{SR_1} = \frac{\varepsilon_1}{S_{Y:12}\sqrt{1-h_1}} = \frac{3.25862}{12.66383\sqrt{1-0.38165}} = \frac{3.25862}{9.95823} \cong \mathbf{0.327}$$

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Observación	Inversión en publicidad X_1 En miles de pesos	Precio del equipo X_2 En cientos de pesos	Ventas *(Y)* En miles de pesos	$\mathbf{y\ gorro}(\widehat{Y})_i$	Residual $\boldsymbol{\varepsilon_i = Y_i - \widehat{Y}_i}$
1	3	125	33	29.741	3.258
2	6	115	61	66.512	-5.512
3	10	140	70	50.073	19.927
4	13	130	82	86.843	-4.843
5	9	145	17	34.970	-17.970
6	6	140	24	23.817	0.183
7	11	138	75	60.053	14.947
8	12	127	80	85.402	-5.402
9	4	128	35	31.183	3.817
10	8	145	20	28.406	-8.406

ISBN 978-607-8285-62-4

Observación	Inversión en publicidad X_1 En miles de pesos	Precio del equipo X_2 En cientos de pesos	h_i	$S_{Y:X}$	Residual Estandarizado SR_i
1	3	125	0.382	12.664	**0.327**
2	6	115	0.488	12.664	**-0.608**
3	10	140	0.166	12.664	**1.723**
4	13	130	0.386	12.664	**-0.488**
5	9	145	0.259	12.664	**-1.648**
6	6	140	0.236	12.664	**0.017**
7	11	138	0.183	12.664	**1.306**
8	12	127	0.349	12.664	**-0.529**
9	4	128	0.275	12.664	**0.354**
10	8	145	0.276	12.664	**-0.780**

El estadístico de Durbin-Watson determina si la correlación entre los términos de error adyacentes es cero.

Estadístico de Durbin-Watson:

(Y)	(X_1)	(X_2)	$\boxed{\hat{Y}_i}$	$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$	$\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i$	$(\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i)^2$	$\hat{\varepsilon}^2$
33	3	12	29.742	3.258	-8.770	76.91201	10.61511
61	6	115	66.512	-5.512	25.439	647.13767	30.38066
70	10	140	50.073	19.927	-24.770	613.55040	397.08673
82	13	130	86.843	-4.843	-13.127	172.32009	23.45382
17	9	145	34.970	-17.970	18.153	329.53194	322.92050
24	6	140	23.817	0.183	14.764	217.98811	0.03350
75	11	138	60.053	14.947	-20.350	414.11232	223.42615
80	12	127	85.402	-5.402	9.220	85.00429	29.18489
35	4	128	31.183	3.817	-12.223	149.41309	14.57310
20	8	145	28.406	-8.406			
						2705.970	1051.674
					d=	2705.97/ 1051.674 **=2.57301**	

El estadístico de Durbin-Watson es ***d=* 2.57301.** Este valor es mayor a 1.5 por lo que no se puede pensar en que la autocorrelación sea un problema.

Diagnóstico de la regresión.

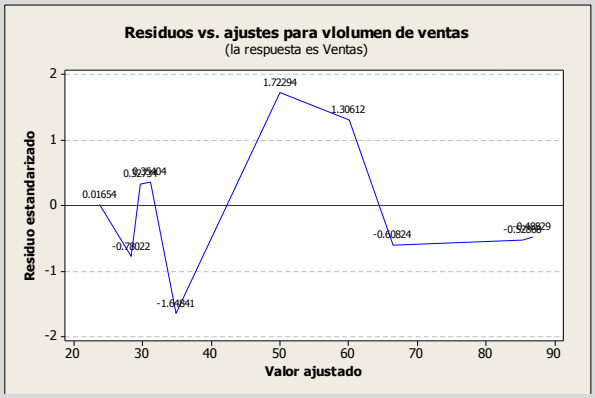
<u>Solución al inciso m.</u>

Linealidad.

Linealidad

Se puede evaluar lo apropiado del modelo de regresión, trazando los "residuales estandarizados" sobre el eje vertical contra los valores $\hat{Y}$ en el eje horizontal. Si el modelo ajustado es apropiado para los datos no habrá un patrón aparente en esta gráfica de los residuales contra $\hat{Y}$. Sin embargo, si el modelo ajustado no es apropiado, habrá relación entre los valores $\hat{Y}$ y los residuales ε_i.

ISBN 978-607-8285-62-4



Así, se puede observar que aunque haya una amplia dispersión en la gráfica residual, no hay patrón ó relación aparente entre los residuales estandarizados y $\hat{Y}$. Los residuales parecen estar distribuidos en forma pareja por encima y por debajo de 0 para diferentes valores de $\hat{Y}$. Por lo tanto se puede concluir que el modelo ajustado parece ser el apropiado.

Homoscedasticidad.

Homoscedasticidad

La suposición de homoscedasticidad se puede evaluar también de la gráfica de residuales estandarizados con $\hat{Y}$. Si parece haber un *"efecto de abanico"* en el cual aumenta ó disminuye la variabilidad de los residuales al aumentar $\hat{Y}$ se demuestra la falta de homogeneidad en las varianzas de Y_i a cada nivel de $\hat{Y}$. Para los datos del volumen de ventas no parece haber diferencias importantes en la variabilidad de SR_i para diferentes valores de $\hat{Y}$. Por lo tanto se puede concluir que para este modelo ajustado no hay violación aparente a la suposición de igual varianza en cada nivel de $\hat{Y}$.

Normalidad.

Normalidad

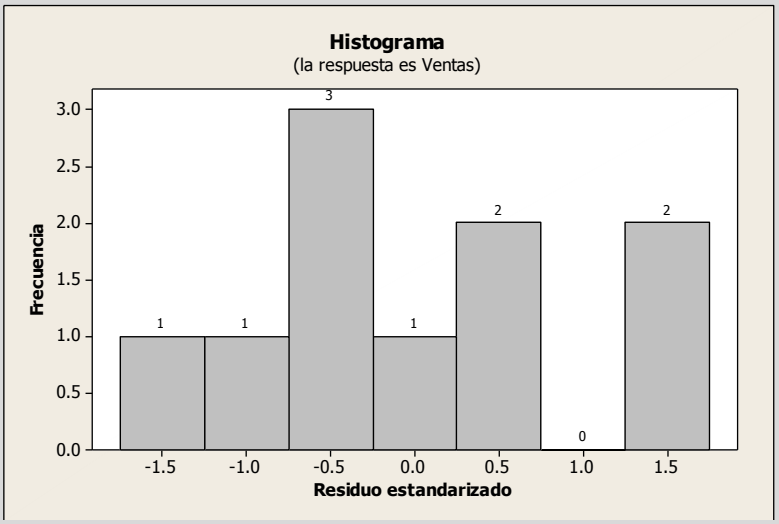
El supuesto de normalidad en la regresión es posible evaluarlo de un análisis residual colocando los residuales estandarizados en una distribución de frecuencias y mostrando los resultados en un histograma. Para los datos del volumen de ventas, los residuales estandarizados se colocaron en la siguiente distribución de frecuencias como se muestra en la siguiente tabla:

Residuales estandarizados	No.
De -1.75 a menos de -1.25	1
De -1.25 a menos de -0.75	1
De -0.75 a menos de -0.25	3
De -0.25 a menos de 0.25	1
De 0.25 a menos de 0.75	2
De 0.75 a menos de 1.25	0
De 1.25 a menos de 1.75	2
Totales	10

ISBN 978-607-8285-62-4

La presencia de largas colas en la gráfica podría indicar sesgo en los datos. Si una o dos barras están lejos de las demás, esos puntos pueden ser valores atípicos. Debido a que el aspecto del histograma cambia según el número de intervalos utilizados para agrupar los datos, utilice la gráfica de probabilidad normal y las pruebas de bondad de ajuste o la prueba de Anderson-Darling para evaluar la normalidad de los residuos.

Los resultados se graficaron en el siguiente histograma:



Es difícil evaluar la suposición de normalidad para una muestra de tan sólo 10 observaciones y los procedimientos de pruebas disponibles quedan fuera del alcance del presente trabajo, sin embargo se puede observar que los datos aunque no parecen tener una "forma de campana" exacta, la mayor parte de los residuales están ubicados cerca del centro de la distribución por lo que parece razonable llegar a la conclusión de que no hay en modo alguno violación a la suposición de normalidad. El histograma indica que los datos podrían tener valores atípicos, lo cual se muestra mediante una barra, en el extremo derecho de la gráfica.

El estadístico de Anderson-Darling mide en este caso si los datos siguen una distribución normal. Mientras mejor se ajuste la distribución a los datos, menor será este estadístico. Utilice el estadístico de Anderson-Darling para comparar el ajuste de varias distribuciones para ver cual es la mejor o probar si una muestra de datos proviene de una población con una distribución normal. Las hipótesis para la prueba de Anderson-Darling son:

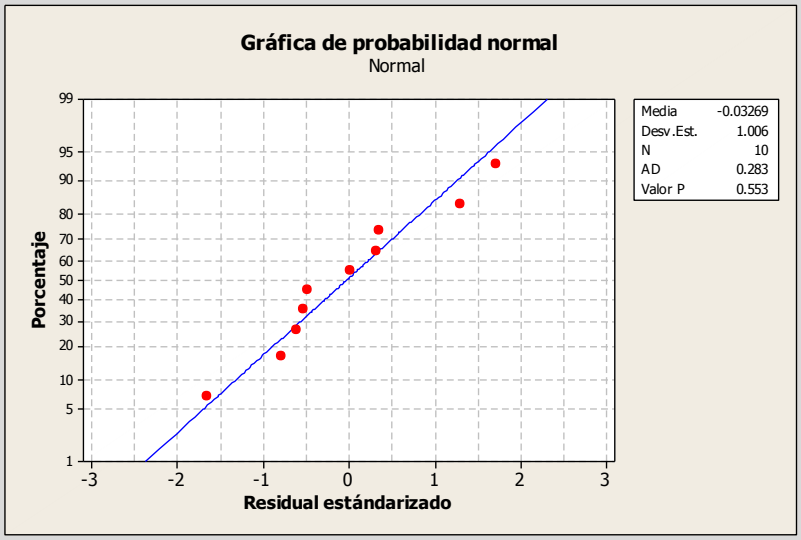
H_0: Los datos siguen una distribución normal

H_1: Los datos no siguen una distribución normal

Si contamos con papel normal o acceso a la computadora, podemos construir una gráfica de probabilidad normal de residuos: Los puntos de esta gráfica deben generalmente formar una línea recta si los residuos se están normalmente distribuidos. Si los puntos en la gráfica salen de una línea recta, el supuesto de normalidad puede ser inválido. Si sus datos tienen menos de 50 observaciones, la gráfica podría mostrar una curvatura en las colas, aun si los residuos están normalmente distribuidos. A medida que el número de observaciones disminuye, la gráfica de probabilidad podría mostrar una variación sustancial no linealidad, aun si los residuos están normalmente distribuidos. Utilice la gráfica de probabilidad y las pruebas de bondad de ajuste, tales como el **estadístico de Anderson-Darling**, para evaluar si los residuos están normalmente distribuidos.

ISBN 978-607-8285-62-4

Si el valor *p* (al estar disponible) para la prueba de Anderson-Darling es inferior al nivel de significación seleccionado (generalmente 0.05 ó 0.10), concluya que los datos no siguen la distribución normal.



La gráfica de probabilidad normal muestra un patrón aproximadamente lineal que concuerda con una distribución normal. Los dos últimos puntos de la esquina superior derecha de la gráfica pueden ser valores atípicos. El destacado de la gráfica identifica estos puntos como 7 y 3, puntos que deberán verificarse como observaciones inusuales ó identificación de valores atípicos. Además el estadístico de Anderson Darling de 0.283 muestra un valor *p* de 0.553, el cual es mayor que 0.05 por lo tanto no rechazamos la hipótesis nula y podemos decir que estadísticamente los residuales estandarizados siguen una distribución normal.

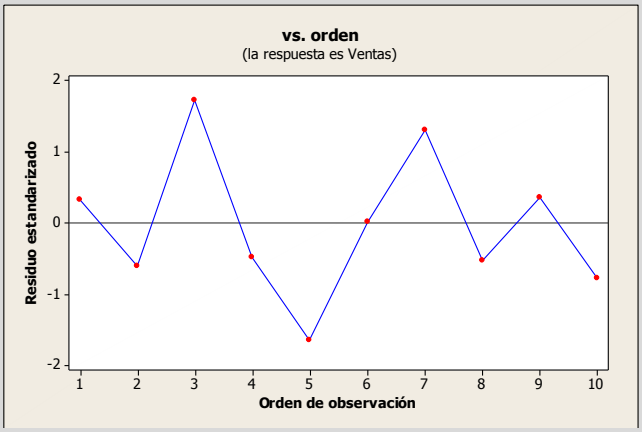
Independencia.

Independencia

La suposición de independencia requiere que el error (diferencia "residual" entre un valor observado y uno predicho de Y) sea independiente para cada valor de $\hat{Y}$. Con frecuencia esta suposición se refiere a datos que se recopilan a lo largo de un periodo. Estos tipos de modelos caen bajo la denominación general de series de tiempo. La suposición de independencia se puede evaluar trazando los residuales en el orden o la sucesión en que se obtuvieron los datos observados.

ISBN 978-607-8285-62-4

Si los verdaderos errores son en realidad independientes, el valor esperado de *d* es alrededor de 2.0. Cualquier valor de *d* menor que 1.5 o 1.6 nos lleva a sospechar que hay autocorrelación.



La gráfica de residuos versus orden no muestra un efecto de "autocorrelación" entre observaciones sucesivas, es decir no hay correlación entre una observación en particular y aquellos valores que la precedieron y la siguieron no afectando la suposición de independencia. Además el estadístico de Durbin-Watson es ***d*= 2.57301.** Este valor es mayor a 1.5 por lo que no se puede pensar en que la autocorrelación sea un problema.

3.4.1.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE **APRENDIZAJE 3.4.1.1 A**NÁLISIS DE **RESIDUALES**

Para el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	11	9	4
2	30	28	14
3	26	17	11
4	20	19	9
5	15	20	8
6	25	24	11
7	35	32	12
8	17	13	21
9	39	36	9
10	45	40	17

ISBN 978-607-8285-62-4

l) Determine los residuales estandarizados para toda la regresión incluyendo el estadístico de Durbin-Watson.
m) Construya la(s) gráfica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Cálculo de los residuales estandarizados.

Solución al inciso l.

El residual estandarizado se presenta como la ecuación:

$$SR_i = \frac{\varepsilon_i}{S_{Y.12}\sqrt{1-h_i}}$$

Donde,

$$h_i = X_i^{'}(X^{'}X)^{-1}X_i$$

Residuales no estandarizados.

Por tanto para calcular el primer residual haremos lo siguiente:

$$\varepsilon_1 = Y_1 - \hat{Y}_1$$

donde $\hat{Y}_1 =$

entonces,

$\boldsymbol{\varepsilon_1} = Y_1 - \hat{Y}_1 =$

Y así sucesivamente con los siguientes 9 datos:

$\boldsymbol{\varepsilon_2} = Y_2 - \hat{Y}_2 =$

ISBN 978-607-8285-62-4

Elementos de la matriz sombrero.

Para estandarizar estos residuales haremos lo siguiente:

$h_1 = X_1^{'}(X^{'}X)^{-1}X_1 =$

Entonces

Residuales estandarizados.

$$\boldsymbol{SR_1} = \frac{\varepsilon_1}{S_{Y:12}\sqrt{1-h_1}} =$$

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Observación	**X_1**	**X_2**	***(Y)***	**y gorro$(\widehat{Y})_i$**	**Residual $\varepsilon_i = Y_i - \widehat{Y}_i$**
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

Observación	**X_1**	**X_2**	h_i	$S_{Y:X}$	**Residual Estandarizado SR_i**
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

ISBN 978-607-8285-62-4

El estadístico de Durbin-Watson.

Estadístico de Durbin-Watson:

(Y)	X_1	X_2	$\boxed{\hat{Y}_i}$	$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$	$\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i$	$(\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i)^2$	$\hat{\varepsilon}^2$
					d=		

Diagnóstico de la regresión.

<u>Solución al inciso m.</u>

Linealidad y Homoscedasticidad.

LINEALIDAD Y HOMOSCEDASTICIDAD. Gráfica de residuales vs Valores ajustados

Interpretación.

INTERPRETACIÓN:

Normalidad.

NORMALIDAD. Gráfica de residuales vs frecuencias

El supuesto de normalidad en la regresión es posible evaluarlo de un análisis residual colocando los residuales estandarizados en una distribución de frecuencias y mostrando los resultados en un histograma. Para los datos del tiempo de entrega de los embarques, los residuales estandarizados se colocaron en la siguiente distribución de frecuencias como se muestra en la siguiente tabla:

ISBN 978-607-8285-62-4

Residuales estandarizados	No.
Totales	

Los resultados se graficaron en el siguiente histograma:

Interpretación.

INTERPRETACIÓN:

Independencia.

INDEPENDENCIA. Gráfica de residuales vs orden de observación:

ISBN 978-607-8285-62-4

Interpretación.

INTERPRETACIÓN:

3.4.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 3.4.1.1
ANÁLISIS DE RESIDUALES

Si tenemos el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	15	7.2	3.0
2	33	4.2	5.0
3	32	1.3	5.0
4	42	1.4	9.5
5	20	6.3	3.7
6	27	3.6	5.5
7	30	2.1	5.2
8	25	5.0	3.1
9	37	1.5	7.8
10	10	8.5	2.5

ISBN 978-607-8285-62-4

l) Determine los residuales estandarizados para toda la regresión incluyendo el estadístico de Durbin-Watson.
m) Construya la(s) gráfica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

Cálculo de los residuales estandarizados.

Solución al inciso l.

El residual estandarizado se presenta como la ecuación:

$$SR_i =$$

Donde,

$$h_i =$$

Por tanto para calcular el primer residual haremos lo siguiente:

Residuales no estandarizados.

$$\varepsilon_1 =$$

Donde

$$\hat{Y}_1 =$$

entonces,

$$\boldsymbol{\varepsilon_1} =$$

Y así sucesivamente con los siguientes 9 datos:

$$\boldsymbol{\varepsilon_2} =$$

ISBN 978-607-8285-62-4

Para estandarizar estos residuales haremos lo siguiente:

$$h_1 =$$

Entonces

Residuales estandarizados.

$$\boldsymbol{SR_1} =$$

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Observación	**X_1**	**X_2**	***(Y)***	$\mathbf{y\ gorro}(\hat{Y})_i$	**Residual** $\boldsymbol{\varepsilon_i = Y_i - \hat{Y}_i}$
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

Observación	**X_1**	**X_2**	$\boldsymbol{h_i}$	$\boldsymbol{S_{Y:X}}$	**Residual Estandarizado** $\boldsymbol{SR_i}$
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

ISBN 978-607-8285-62-4

El estadístico de Durbin-Watson.

Estadístico de Durbin-Watson:

(Y)	X_1	X_2	$\boxed{\hat{Y}_i}$	$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$	$\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i$	$(\hat{\varepsilon}_{i+1} - \hat{\varepsilon}_i)^2$	$\hat{\varepsilon}^2$
					d=		

Diagnóstico de la regresión.

<u>Solución al inciso m.</u>

Linealidad y Homoscedasticidad.

LINEALIDAD Y HOMOSCEDASTICIDAD. Gráfica de residuales vs Valores ajustados

Interpretación.

INTERPRETACIÓN:

ISBN 978-607-8285-62-4

Normalidad.

NORMALIDAD. Gráfica de residuales vs frecuencias

El supuesto de normalidad en la regresión es posible evaluarlo de un análisis residual colocando los residuales estandarizados en una distribución de frecuencias y mostrando los resultados en un histograma. Para los datos del tiempo de entrega de los embarques, los residuales estandarizados se colocaron en la siguiente distribución de frecuencias como se muestra en la siguiente tabla:

Residuales estandarizados	No.
Totales	

Los resultados se graficaron en el siguiente histograma:

Interpretación.

INTERPRETACIÓN:

Independencia

INDEPENDENCIA. Gráfica de residuales vs orden de observación:

Interpretación.

INTERPRETACIÓN:

ISBN 978-607-8285-62-4

3.4.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 3.4.1. ANÁLISIS DE RESIDUALES

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

3.4.1.1 **S**uponga que una compañía de productos de consumo quisiera medir la efectividad de la publicidad en radio y televisión y la publicidad en periódicos en la promoción de sus productos. Se seleccionó una muestra aleatoria de 10 ciudades con poblaciones aproximadamente iguales para el estudio durante un periodo de prueba de un mes. Se registraron las ventas (en miles de pesos) durante el mes de prueba, junto con los niveles de gastos en los medios de publicidad, con los resultados siguientes:

Ciudad	Ventas ($000,000) Y	Publicidad en radio y televisión ($000) X_1	Publicidad en periódicos ($000) X_2
1	9.73	0	400
2	6.25	250	250
3	9.71	300	300
4	9.31	350	350
5	11.77	350	350
6	9.82	400	250
7	16.28	450	450
8	13.29	550	250
9	14.36	600	300
10	17.41	650	350

l) Determine los residuales estandarizados para toda la regresión incluyendo el estadístico de Durbin-Watson.

m) Construya la(s) gráfica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.

3.4.1.2 **E**l gerente distrital de ventas de de un fabricante de automóviles estudia las ventas de éstos. En forma específica, quiere determinar qué factores influyen en el número de automóviles vendidos en una distribuidora. Para investigarlo, seleccionó al azar 10 distribuidoras. De éstas, obtiene el número de automóviles vendidos el mes pasado, los minutos de publicidad en radio comprados el mes pasado y el número de vendedores de tiempo completo contratados. La información es la siguiente:

ISBN 978-607-8285-62-4

Distribuidora	Automóviles vendidos el mes pasado Y	Publicidad (en minutos) X_1	Fuerza de ventas X_2
1	127	18	10
2	159	22	14
3	144	23	12
4	139	17	12
5	128	16	12
6	180	26	17
7	102	15	7
8	163	24	16
9	106	18	10
10	149	30	11

l) Determine los residuales estandarizados para toda la regresión incluyendo el estadístico de Durbin-Watson.

m) Construya la(s) gráfica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.

3.4.1.3 Los siguientes datos representan las calificaciones de estadística para una muestra aleatoria de 10 estudiantes de primer año de determinada institución de enseñanza superior, junto con sus calificaciones en un examen de inteligencia aplicado cuando aún cursaban el último año de secundaria y el número de periodos de clase perdidos por los 10 estudiantes que tomaron el curso de estadística.

Estudiante	Calificación de estadística Y	Calificación del examen X_1	Clases perdidas X_2
1	98	70	5
2	74	50	7
3	87	70	3
4	81	55	4
5	85	65	1
6	90	65	2
7	74	55	4
8	76	55	5
9	91	70	3
10	94	65	2

l) Determine los residuales estandarizados para toda la regresión incluyendo el estadístico de Durbin-Watson.

m) Construya la(s) gráfica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.

ISBN 978-607-8285-62-4

OBJETIVO 3.5. El alumno podrá calcular e interpretar los criterios desarrollados para realizar un análisis de influencias e identificar observaciones influyentes.

ANTECEDENTES

CONCEPTOS DE:

Variable dependiente. Variable independiente. Principio de mínimos cuadrados. Forma general de la ecuación de regresión lineal simple. Punto donde se intercepta la línea de regresión con el eje Y. Pendiente de la línea de regresión. Coeficientes de regresión. Error estándar del estimador. Residual. Residual estandarizado. Distribución ***t*** de Student. Elementos de la matriz sombrero. Valores atípicos.

3.5.1 DIAGNÓSTICO DE LA REGRESIÓN: ANÁLISIS DE INFLUENCIAS

CONCEPTOS BÁSICOS ANÁLISIS DE INFLUENCIA

En modelos de regresión, el apalancamiento *(h_i)* mide la

Las técnicas del **análisis de influencias** se utilizan para determinar si cualquier observación individual tiene una influencia indebida sobre el modelo ajustado.

Se consideran básicamente tres medidas:

1.- Los elementos de la matriz sombrero h_i.
2.- Los residuales de Student eliminados t^*_i .
3.- El estadístico de distancia D_i de Cook.

1.- Uso de los elementos de la matriz sombrero h_i :

$$h_i = X_i'(X'X)^{-1}X_i$$

Donde h_i son los **"elementos diagonales de la matriz sombrero",** que reflejan **la influencia de cada X_i sobre el modelo de regresión ajustado**.

ISBN 978-607-8285-62-4

distancia de un valor X de observación hasta el promedio de los valores X para todas las observaciones en un conjunto de datos.

Las observaciones con valores de apalancamiento grandes pudieran ejercer una influencia desproporcionada sobre el modelo y producir resultados desviados.

Los valores con apalancamiento oscilan entre 0 y 1. Investigue las observaciones con valores con apalancamiento mayores que *2(k + 1)/n*, donde *k* es el número de variables independientes y n es el número de observaciones.

Un residual de Student eliminado es útil en la detección de valores atípicos. Los residuos de Student eliminados de una observación se calculan dividiendo un residuo eliminado de la observación entre un estimado de su error estándar. La observación se omite para ver cómo se comporta el modelo sin este valor atípico potencial. Si una observación tiene un residuo de Student eliminado grande (en general si su valor absoluto es mayor que 2), podría tratarse de un valor atípico en sus datos.

Cada residuo de Student eliminado sigue la distribución t con grados de libertad *(n – k – 2)*, grados de libertad, donde k es igual al número de variables independientes en el modelo de regresión.

Si existen esos puntos de influencia quizá sea necesario **evaluar de nuevo la necesidad de mantenerlos en el modelo.**

Hoaglin y Welsch sugieren la siguiente regla de decisión para un modelo de regresión lineal múltiple con ***k* variables explicatorias**:

$$\text{Si } h_i > 2(k+1)/n$$

entonces X_i es un **punto influyente y removible del modelo**.

2.- Los residuales de Student eliminados t^*_i .

En el estudio del análisis residual se definieron los residuales estandarizados mediante la ecuación:

$$SR_i = \frac{\varepsilon_i}{S_{Y.X}\sqrt{1-h_i}}$$

Para poder **medir mejor** la **repercusión adversa** sobre el modelo de cada caso individual, Hoaglin y Welsch desarrollaron **el residual de Student eliminado t^*_i** que se presenta en la siguiente ecuación:

$$t_i^* = \frac{\varepsilon_i}{S_{(i)}\sqrt{1-h_i}}$$

Donde $\mathbf{S_{(i)}}$ = el error estándar de la estimación para un modelo que incluye todas las observaciones excepto la observación ***i.***

Este **residual de Student eliminado** mide la **diferencia entre cada valor observado Y_i y el valor predicho** obtenidos de un modelo que incluye **todas las demás observaciones excepto *i.*** En el modelo de regresión múltiple Hoaglin y Welsch proponen que si,

$$|\, t^*_i \,| > t._{10.\, n-k-2}$$

entonces los **valores observados y predichos son tan diferentes**, que la observación ***i*** es un **punto influyente que afecta de modo adverso al modelo y puede ser eliminada.**

ISBN 978-607-8285-62-4

La distancia de Cook (*Di)* es una medida de la influencia de una observación sobre el conjunto de coeficientes de regresión en un modelo de regresión. Las observaciones influyentes tienen un impacto desproporcionado sobre el modelo y pueden generar resultados engañosos. Las observaciones influyentes pueden ser puntos de apalancamiento *(h_i),* valores atípicos o ambos.

La distancia de Cook considera tanto el valor de apalancamiento *(h_i)* como el residuo estandarizado de cada observación al determinar su efecto en los coeficientes de regresión. Por lo general, es útil verificar las observaciones en las que D_i es mayor que *F(0.5, k+1, n-k-1*), la mediana de una distribución F, donde k es el número de variables independientes y n es el número de observaciones.

3.- El estadístico de distancia D_i de Cook.

El uso de h_i y t^*_i en la búsqueda de **puntos de datos potencialmente problemáticos** es complementario ya que ninguno de los criterios es suficiente por sí mismo.

Para **decidir** si un punto que ha sido **destacado mediante el criterio h_i ò t^*_i** está **afectando indebidamente al modelo**, Cook y Weisberg sugieren el **uso del estadístico D_i** en la ecuación:

$$D_i = \frac{1}{(k+1)} SR_i^2 \frac{h_i}{(1-h_i)} = \frac{SR_i^2 h_i}{(k+1)(1-h_i)}$$

Donde SR_i es el residual estandarizado.

En el **modelo de regresión múltiple** Cook y Weisberg sugieren que

Si $D_i > F_{.50.\ k+1.\ n-k-1}$

La observación puede **tener una recuperación sobre los resultados** de ajustar un modelo de regresión múltiple.

Para determinar el **grado de influencia**, puede ajustar el modelo **con y sin la observación influyente** y comparar los **coeficientes, valores p, R^2 y otros parámetros del modelo**. Si el modelo **cambia significativamente cuando elimina la observación influyente,** determine primero si la observación es un error de entrada de datos o de medición. De no ser así, examine aun más el modelo para determinar si **omitió una variable**, o ha especificado incorrectamente el modelo. Pudiera necesitar recopilar más datos para resolver el problema.

ISBN 978-607-8285-62-4

3.5.1.1 EJEMPLO ILUSTRATIVO

EJEMPLO ILUSTRATIVO 3.5.1.1 ANÁLISIS DE INFLUENCIA

Samuel Prado, propietario y director general de un establecimiento quiere conocer el comportamiento de las ventas (en miles de pesos) de un equipo de sonido que se expende en el establecimiento. Se percata de que existen muchos factores que podrían ayudarle a explicar las ventas pero piensa que la inversión en publicidad (en miles de pesos) y el precio (en cientos de pesos) son los principales factores determinantes. Samuel ha reunido los datos que se anexan a continuación.

Observación	Ventas Y	Publicidad X_1	Precio X_2
1	33	3	125
2	61	6	115
3	70	10	140
4	82	13	130
5	17	9	145
6	24	6	140
7	75	11	138
8	80	12	127
9	35	4	128
10	20	8	145

n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.

o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.

p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

Elementos de la matriz sombrero h_i.

<u>Solución al inciso n.</u>

Cuando se desarrollo una estimación por intervalo de confianza $\mu_{Y.X}$, se definieron los "elementos diagonales de la matriz sombrero" h_i como:

$$h_i = X_i'(X'X)^{-1}X_i$$

ISBN 978-607-8285-62-4

Así para el primer punto u observación,

Elementos de la matriz sombrero h_i para la observación 1.

$$\boldsymbol{h_1} = X_1'(X'X)^{-1}X_1$$

$$= [1 \quad 3 \quad 125]\begin{bmatrix} {17'252,036}/{829,796} & {39,988}/{829,796} & {-131,260}/{829,796} \\ {39,988}/{829,796} & {8,681}/{829,796} & {-834}/{829,796} \\ {-131,260}/{829,796} & {-834}/{829,796} & {1,036}/{829,796} \end{bmatrix}\begin{bmatrix} 1 \\ 11 \\ 135 \end{bmatrix}$$

$$= [1.16233 \quad -0.04606 \quad -0.00514]\begin{bmatrix} 1 \\ 3 \\ 125 \end{bmatrix} \cong \mathbf{0.38165}$$

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Observación	**1**	**2**	**3**	**4**	**5**
Inversión en Publicidad(X_1) , en miles de pesos.	3	6	10	13	9
Precio del equipo (X_2) en cientos de pesos	125	115	140	130	145
Ventas *(Y), e*n miles de pesos	33	61	70	82	17
Análisis de influencia *hi*	**0.38213**	**0.48782**	**0.16570**	**0.38647**	**0.25879**

Observación	**6**	**7**	**8**	**9**	**10**
Inversión en Publicidad(X_1) , en miles de pesos.	6	11	12	4	8
Precio del equipo (X_2) en cientos de pesos	140	138	127	128	145
Ventas *(Y), e*n miles de pesos	24	75	80	35	20
Análisis de influencia *hi*	**0.23631**	**0.18315**	**0.34874**	**0.27487**	**0.27603**

Los valores con apalancamiento oscilan entre 0 y 1. Investigue las observaciones con valores con apalancamiento mayores que *2(k + 1)/n*, donde *k* es el número de variables independientes y n es el número de observaciones. Minitab identifica, mediante una X en la tabla de observaciones inusuales, las observaciones con apalancamiento superior a *2(k + 1)/n* o .99, el valor que sea menor.

Interpretación: Para los datos del volumen de ventas, puesto que n=10, los criterios deben ser "destacar" cualquier valor $\boldsymbol{h_i}$ superior a ***2(k + 1)/n =0.6*** . Consultando la tabla anterior se puede observar que ninguna observación es candidata potencial para ser removida del modelo del volumen de ventas.

ISBN 978-607-8285-62-4

Residuales de Student eliminados, t_i^*.

Solución al inciso o.

Los residuales de Student eliminados t^*_i .

En el estudio del análisis residual se definieron los residuales estandarizados mediante la ecuación:

$$SR_i = \frac{\varepsilon_i}{S_{Y.X}\sqrt{1-h_i}}$$

Para poder medir mejor la repercusión adversa sobre el modelo de cada caso individual, Hoaglin y Welsch desarrollaron el residual de Student eliminado t^*_i que se presenta en la siguiente ecuación:

$$t_i^* = \frac{\varepsilon_i}{S_{(i)}\sqrt{1-h_i}}$$

Donde $S_{(i)}$ = el error estándar de la estimación para un modelo que incluye todas las observaciones excepto la observación *i*.

Este residual de Student eliminado mide la diferencia entre cada valor observado Y_i y el valor predicho obtenidos de un modelo que incluye todas las demás observaciones excepto *i*. En el modelo de regresión múltiple Hoaglin y Welsch proponen que si,

$$|\ t^*_i\ | > t._{10.\ n-k-2}$$

entonces los valores observados y predichos son tan diferentes, que la observación *i* es un punto influyente que afecta de modo adverso al modelo y puede ser eliminada.

Residuales de Student eliminados, t_i^* para la observación 1.

Así para la observación No. 1 primero debemos calcular la ecuación de regresión considerando sólo las observaciones de la 2 a la 10,

$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_{1i} + \hat{\beta}_{2i} X_{2i}$$

Para encontrar las estimaciones mínimo cuadráticas $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_k$ del término constante y las pendientes parciales en el modelo de regresión múltiple recuerde que el principio de mínimos cuadrados incluye elegir las estimaciones que minimicen las sumas de los cuadrados de los residuos. Las ecuaciones normales que resultan de ello son, en notación matricial,

$$(X'X)\hat{\beta} = X'Y$$

ISBN 978-607-8285-62-4

donde

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

es el vector buscado de de coeficientes estimados. Suponiendo que la matriz $X'X$ tiene una inversa, la solución es

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Observación	Inversión en publicidad X_1 En miles de pesos	Precio del equipo X_2 en cientos de pesos	Ventas *(Y)* En miles de pesos	y gorro($\hat{Y}$	Residual $\varepsilon_i = Y_i - \hat{Y}_i$	h_i	$S_{Y:X}$	Residual Estandarizado SR_i
2	6	115	61	66.512	-5.512	0.448	12.664	**-0.608**
3	10	140	70	50.073	19.927	0.166	12.664	**1.723**
4	13	130	82	86.843	-4.843	0.386	12.664	**-0.488**
5	9	145	17	34.970	-17.970	0.259	12.664	**-1.648**
6	6	140	24	23.817	0.183	0.236	12.664	**0.017**
7	11	138	75	60.053	14.947	0.183	12.664	**1.306**
8	12	127	80	85.402	-5.402	0.349	12.664	**-0.529**
9	4	128	35	31.183	3.817	0.275	12.664	**0.354**
10	8	145	20	28.406	-8.406	0.276	12.664	**-0.780**

Para los datos anteriores,

$$\boldsymbol{Y} = \begin{bmatrix} 61 \\ 70 \\ 82 \\ 17 \\ 24 \\ 75 \\ 80 \\ 35 \\ 20 \end{bmatrix} \quad \text{y} \quad \mathbf{X} = \begin{bmatrix} 1 & 6 & 115 \\ 1 & 10 & 140 \\ 1 & 13 & 130 \\ 1 & 9 & 145 \\ 1 & 6 & 140 \\ 1 & 11 & 138 \\ 1 & 12 & 127 \\ 1 & 4 & 128 \\ 1 & 8 & 145 \end{bmatrix}$$

217.39525
6.80687
-1.68071

ISBN 978-607-8285-62-4

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} 22.97729 & -0.03846 & -0.16785 \\ -0.03846 & 0.01389 & -0.00062 \\ -0.16785 & -0.00062 & 0.00129 \end{bmatrix} \begin{bmatrix} 464 \\ 4{,}514 \\ 61{,}190 \end{bmatrix} = \begin{bmatrix} 217.39525 \\ 6.80687 \\ -1.68071 \end{bmatrix}$$
$$= \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

El modelo ajustado sin considerar la observación No. 1, se puede expresar como:

El modelo ajustado sin considerar la observación No. 1.

$$\boxed{\widehat{\boldsymbol{Y}}_{\boldsymbol{1}} = \mathbf{217.39525} + \mathbf{6.80687}\boldsymbol{X}_{\mathbf{11}} - \mathbf{1.68071}\boldsymbol{X}_{\mathbf{21}}}$$

Ahora se debe calcular el error estándar del estimador para esta ecuación,

$$S_{Y.12} = S_{(1)} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} Y^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y - \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y - \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y}{n-k-1}}$$
$$= \sqrt{\frac{SCE}{g.l.}} = \sqrt{\frac{Y'Y - \hat{\beta}'(X'Y)}{n-k-1}} = \sqrt{CME}$$

Finalmente

$$\hat{\beta}'(X'Y) = [217.39525 \quad 6.80687 \quad -1.68071] \begin{bmatrix} 464 \\ 4{,}514 \\ 61{,}190 \end{bmatrix} \cong 28{,}754.96228$$

Error estándar de la estimación sin considerar la observación 1.

$$SCE = Y'Y - \hat{\beta}'(X'Y) = 29{,}860 - 28{,}754.96228 \cong \mathbf{1{,}105.03772}$$

$$\boldsymbol{S_{Y.12}} = \boldsymbol{S_{(1)}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-k-1}} = \sqrt{\frac{SCE}{n-k-1}} = \sqrt{\frac{1{,}105.03772}{6}} \cong \mathbf{13.571}$$

Residuales de Student eliminado, t_i^* para la observación 1.

Por lo tanto el residual de Student eliminado $\boldsymbol{t_i^*}$

$$\boldsymbol{t_1^*} = \frac{\varepsilon_1}{S_{(1)}\sqrt{1-h_1}} = \frac{3.25862}{13.571\sqrt{1-0.38165}} = \frac{3.25862}{10.67159} \cong \mathbf{0.305}$$

ISBN 978-607-8285-62-4

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Observación	**1**	**2**	**3**	**4**	**5**
Inversión en Publicidad(X_1), en miles de pesos.	3	6	10	13	9
Precio del equipo (X_2) en cientos de pesos	125	115	140	130	145
Ventas *(Y), en* miles de pesos	33	61	70	82	17
Análisis de influencia t_i^*	**0.30541**	**-0.57862**	**2.10190**	**-0.45997**	**-1.95109**

Observación	**6**	**7**	**8**	**9**	**10**
Inversión en Publicidad(X_1), en miles de pesos.	6	11	12	4	8
Precio del equipo (X_2) en cientos de pesos	140	138	127	128	145
Ventas *(Y), en* miles de pesos	24	75	80	35	20
Análisis de influencia t_i^*	**0.01531**	**1.39048**	**-0.49953**	**0.33075**	**-0.75596**

Si una observación tiene un residuo de Student eliminado grande , podría tratarse de un valor atípico en sus datos.

Cada residuo de Student eliminado sigue la distribución t con grados de libertad *(n – k – 2),* grados de libertad, donde k es igual al número de variables independientes en el modelo de regresión.

Interpretación: Para los datos del volumen de ventas, puesto que n=10, los criterios deben ser "destacar" cualquier valor superior **a** $|t_i^*| > t_{0.10.6} = 1.4398$. Consultando la tabla anterior se puede visualizar que $t_3^* = 2.102$ y $t_5^* = -1.951$ Por lo tanto la tercera y la quinta observación pueden tener un efecto adverso sobre el modelo y se pueden considerar candidatos a ser retirados del modelo, sin embargo como de acuerdo al criterio h_i la tienda 3 y 5 no presentaron un efecto adverso, se debe tomar en cuenta otro criterio antes de tomar esa decisión como el criterio D_i de Cook, que se basa tanto en h_i como en el estadístico residual estandarizado t_i^*.

ISBN 978-607-8285-62-4

Estadístico de distancia de Cook, D_i.

Solución al inciso p.

El estadístico de distancia D_i de Cook.

El uso de h_i y t^*_i en la búsqueda de puntos de datos potencialmente problemáticos es complementario ya que ninguno de los criterios es suficiente por sí mismo.

Para decidir si un punto que ha sido destacado mediante el criterio h_i ò t^*_i está afectando indebidamente al modelo, Cook y Weisberg sugieren el uso del estadístico D_i en la ecuación:

$$D_i = \frac{1}{(k+1)} SR_i^2 \frac{h_i}{(1-h_i)} = \frac{SR_i^2 h_i}{(k+1)(1-h_i)}$$

Donde SR_i es el residual estandarizado.

En el modelo de regresión múltiple Cook y Weisberg sugieren que

Si $D_i > F_{.50.\ k+1.\ n-k-1}$

La observación puede tener una recuperación sobre los resultados de ajustar un modelo de regresión múltiple.

Así para el primer punto u observación,

Observación	Inversión en publicidad X_1 En miles de pesos	Precio del equipo X_2 En cientos de pesos	Ventas *(Y)* En miles de pesos	h_i	Residual Estandarizado SR_i
1	3	125	33	0.382	0.327
2	6	115	61	0.488	-0.608
3	10	140	70	0.166	1.723
4	13	130	82	0.386	-0.488
5	9	145	17	0.259	-1.648
6	6	140	24	0.236	0.017
7	11	138	75	0.183	1.306
8	12	127	80	0.349	-0.529
9	4	128	35	0.275	0.354
10	8	145	20	0.276	-0.780

Distancia de Cook para la observación 1.

$$\boldsymbol{D_1} = \frac{1}{(k+1)} SR_1^2 \frac{h_1}{(1-h_1)} = \frac{SR_1^2 h_1}{(k+1)(1-h_1)} = \frac{(0.327)^2(0.382)}{(2+1)(1-0.382)} = \frac{0.04085}{1.854} = \mathbf{0.022}$$

ISBN 978-607-8285-62-4

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Observación	**1**	**2**	**3**	**4**	**5**
Inversión en Publicidad(X_1), en miles de pesos.	3	6	10	13	9
Precio del equipo (X_2) en cientos de pesos	125	115	140	130	145
Ventas *(Y), e*n miles de pesos	33	61	70	82	17
Análisis de influencia D_i	**0.02209**	**0.11745**	**0.19652**	**0.05006**	**0.31623**

La distancia de Cook considera tanto el valor de apalancamiento *(h_i)* como el residuo estandarizado de cada observación al determinar su efecto en los coeficientes de regresión. Por lo general, es útil verificar las observaciones en las que D_i es mayor que $F(0.5, k+1, n-k-1)$, la mediana de una distribución F, donde k es el número de variables independientes y n es el número de observaciones.

Observación	**6**	**7**	**8**	**9**	**10**
Inversión en Publicidad(X_1) , en miles de pesos.	6	11	12	4	8
Precio del equipo (X_2) en cientos de pesos	140	138	127	128	145
Ventas *(Y), e*n miles de pesos	24	75	80	35	20
Análisis de influencia D_i	**0.00003**	**0.12750**	**0.04989**	**0.01584**	**0.07737**

Interpretación: Para el modelo del volumen de ventas (en millones de pesos), puesto que n=10, el criterio sería "destacar" cualquier $\boldsymbol{D_i} > \boldsymbol{F_{0.50.3.7}} = \mathbf{0.871}$. Consultando la tabla anterior se puede observar que ninguna observación es candidata potencial para ser removida del modelo del volumen de ventas. En caso de que alguna observación una vez estudiados los tres criterios fuera necesario eliminar alguna(s) observación(es) se debería estudiar un modelo alterno en el que se hayan eliminado dichas observaciones que no fue el caso en este modelo.

ISBN 978-607-8285-62-4

3.5.1.1 ACTIVIDAD DE **APRENDIZAJE**

ACTIVIDAD DE **APRENDIZAJE 3.5.1.1** ANÁLISIS DE **INFLUENCIA**

Para el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	11	9	4
2	30	28	14
3	26	17	11
4	20	19	9
5	15	20	8
6	25	24	11
7	35	32	12
8	17	13	21
9	39	36	9
10	45	40	17

n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.

o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.

p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ISBN 978-607-8285-62-4

Elementos de la matriz sombrero h_i.

Solución al inciso n.

Cuando se desarrollo una estimación por intervalo de confianza $\mu_{Y.X}$, se definieron los "elementos diagonales de la matriz sombrero" h_i como:

$$h_i = X_i^{'}(X^{'}X)^{-1}X_i$$

Así para el primer punto u observación,

Elementos de la matriz sombrero h_i para la observación 1.

$$\boldsymbol{h_1} = X_1^{'}(X^{'}X)^{-1}X_1 =$$

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Los valores con apalancamiento oscilan entre 0 y 1. Investigue las observaciones con valores con apalancamiento mayores que *2(k + 1)/n*, donde *k* es el número de variables independientes y n es el número de observaciones. Minitab identifica, mediante una X en la tabla de observaciones inusuales, las observaciones con apalancamiento superior a *2(k + 1)/n* o .99, el valor que sea menor.

Observación	**1**	**2**	**3**	**4**	**5**
X_1					
X_2					
Y					
Análisis de influencia *hi*					

Observación	**6**	**7**	**8**	**9**	**10**
X_1					
X_2					
Y					
Análisis de influencia *hi*					

ISBN 978-607-8285-62-4

Interpretación:

Residuales de Student eliminados, t_i^*.

<u>Solución al inciso o.</u>

Así para la observación No. 1 primero debemos calcular la ecuación de regresión considerando sólo las observaciones de la 2 a la 10,

$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_1 X_{1i} + \hat{\beta}_{2i} X_{2i}$$

Para encontrar las estimaciones mínimo cuadráticas $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_k$ del término constante y las pendientes parciales en el modelo de regresión múltiple recuerde que el principio de mínimos cuadrados incluye elegir las estimaciones que minimicen las sumas de los cuadrados de los residuos. Las ecuaciones normales que resultan de ello son, en notación matricial,

$$(X'X)\hat{\beta} = X'Y$$

donde

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

es el vector buscado de de coeficientes estimados. Suponiendo que la matriz $X'X$ tiene una inversa, la solución es

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Observación	X_1	X_2	Y	y gorro($\hat{Y}$	Residual $\varepsilon_i = Y_i - \hat{Y}_i$	h_i	$S_{Y:X}$	Residual Estandarizado SR_i
2								
3								
4								
5								
6								
7								
8								
9								
10								

ISBN 978-607-8285-62-4

Para los datos anteriores,

$\boldsymbol{Y} =$ y $\mathbf{X} =$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} =$$

El modelo ajustado sin considerar la observación No. 1.

El modelo ajustado sin considerar la observación No. 1, se puede expresar como:

$\boxed{\widehat{\boldsymbol{Y}}_{\boldsymbol{i}} =}$

Ahora se debe calcular el error estándar del estimador para esta ecuación,

$$S_{Y.12} = S_{(1)} = \sqrt{\frac{Y'Y - \hat{\beta}'(X'Y)}{n-k-1}}$$

Finalmente

$$\hat{\beta}'(X'Y) =$$

$$SCE = Y'Y - \hat{\beta}'(X'Y) =$$

Error estándar de la estimación sin considerar la observación 1.

$$\boldsymbol{S_{Y.12}} = \boldsymbol{S_{(1)}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-k-1}} = \sqrt{\frac{SCE}{n-k-1}} =$$

ISBN 978-607-8285-62-4

Residuales de Student eliminados, t_i^* para la observación 1.

Por lo tanto el residual de Student eliminado t_i^*

$$t_1^* = \frac{\varepsilon_1}{S_{(1)}\sqrt{1-h_1}} =$$

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Observación	**1**	**2**	**3**	**4**	**5**
X_1					
X_2					
Y					
Análisis de influencia t_i^*					

La observación se omite para ver cómo se comporta el modelo sin este valor atípico potencial. Si una observación tiene un residuo de Student eliminado grande, podría tratarse de un valor atípico en sus datos.

Observación	**6**	**7**	**8**	**9**	**10**
X_1					
X_2					
Y					
Análisis de influencia t_i^*					

Cada residuo de Student eliminado sigue la distribución t con grados de libertad *(n – k – 2),* grados de libertad, donde k es igual al número de variables independientes en el modelo de regresión.

Interpretación:

ISBN 978-607-8285-62-4

Estadístico de distancia de Cook, D_i.

Solución al inciso p.

Así para el primer punto u observación,

Observación	X_1	X_2	Y	h_i	Residual Estandarizado SR_i
1					
2					
3					
4					
5					
7					
8					
9					
10					

Distancia de Cook para la observación 1.

$$\boldsymbol{D_1} = \frac{1}{(k+1)} SR_1^2 \frac{h_1}{(1-h_1)} = \frac{SR_1^2 h_1}{(k+1)(1-h_1)} =$$

$$\boldsymbol{D_2} =$$

ISBN 978-607-8285-62-4

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Observación	**1**	**2**	**3**	**4**	**5**
X_1					
X_2					
Y					
Análisis de influencia D_i					

Observación	**6**	**7**	**8**	**9**	**10**
X_1					
X_2					
Y					
Análisis de influencia D_i					

La distancia de Cook considera tanto el valor de apalancamiento *(h_i)* como el residuo estandarizado de cada observación al determinar su efecto en los coeficientes de regresión. Por lo general, es útil verificar las observaciones en las que D_i es mayor que *F(0.5, k+1, n-k-1)*, la mediana de una distribución F, donde k es el número de variables independientes y n es el número de observaciones.

Interpretación:

ISBN 978-607-8285-62-4

3.5.1.1 EJERCICIO DE AUTOEVALUACIÓN

A continuación se presenta un ejercicio de autoevaluación el cual pone a prueba su comprensión del material anterior. La respuesta a este ejercicio de autoevaluación se encuentra al final del cuaderno de trabajo en el anexo de respuestas. Le recomendamos enfáticamente resolverlo y posteriormente revisar su respuesta como retroalimentación de su aprendizaje

AUTOEVALUACIÓN 3.5.1.1
ANÁLISIS DE INFLUENCIA

Si tenemos el siguiente conjunto de datos:

No. De observación	Y	X_1	X_2
1	15	7.2	3.0
2	33	4.2	5.0
3	32	1.3	5.0
4	42	1.4	9.5
5	20	6.3	3.7
6	27	3.6	5.5
7	30	2.1	5.2
8	25	5.0	3.1
9	37	1.5	7.8
10	10	8.5	2.5

n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.

o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.

p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

NOTA: El uso de un software estadístico como Excel o Minitab, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben comprender primero los pasos del proceso. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y posteriormente utilice un software para comparar sus resultados. Es importante mencionar que pueden existir diferencias en las respuestas debido a la cantidad de dígitos que se utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.

ISBN 978-607-8285-62-4

Elementos de la matriz sombrero h_i.

Elementos de la matriz sombrero h_i para la observación 1.

Solución al inciso n.

Así para el primer punto u observación,

$$\boldsymbol{h_1} =$$

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Observación	**1**	**2**	**3**	**4**	**5**
X_1					
X_2					
Y					
Análisis de influencia *hi*					

Observación	**6**	**7**	**8**	**9**	**10**
X_1					
X_2					
Y					
Análisis de influencia *hi*					

ISBN 978-607-8285-62-4

Interpretación:

Residuales de Student eliminados, t_i^*.

Solución al inciso o.

Observación	X_1	X_2	Y	y gorro($\widehat{Y}$	Residual $\varepsilon_i = Y_i - \widehat{Y}_i$	h_i	$S_{Y:X}$	Residual Estandarizado SR_i
2								
3								
4								
5								
6								
7								
8								
9								
10								

Para los datos anteriores,

$\boldsymbol{Y} =$ y $\mathbf{X} =$

ISBN 978-607-8285-62-4

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} =$$

El modelo ajustado sin considerar la observación No. 1, se puede expresar como:

El modelo ajustado sin considerar la observación No. 1.

$$\boxed{\widehat{\boldsymbol{Y}}_{\boldsymbol{i}} = \qquad\qquad\qquad\qquad}$$

Ahora se debe calcular el error estándar del estimador para esta ecuación,

$S_{Y.12} = S_{(1)} =$

Finalmente

$\hat{\beta}'(X'Y) =$

Error estándar de la estimación sin considerar la observación 1.

$SCE =$

$\boldsymbol{S_{Y.12}} = S_{(1)} =$

Residuales de Student eliminados, t_i^* para la observación 1.

Por lo tanto el residual de Student eliminado $\boldsymbol{t_i^*}$

$\boldsymbol{t_1^*} =$

ISBN 978-607-8285-62-4

Estadístico de distancia de Cook, D_i.

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Observación	**1**	**2**	**3**	**4**	**5**
X_1					
X_2					
Y					
Análisis de influencia t_i^*					

Observación	**6**	**7**	**8**	**9**	**10**
X_1					
X_2					
Y					
Análisis de influencia t_i^*					

Interpretación:

ISBN 978-607-8285-62-4

Distancia de Cook para la observación 1.

Solución al inciso p.

Así para el primer punto u observación,

Observación	X_1	X_2	Y	h_i	Residual Estandarizado SR_i
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

$\boldsymbol{D_1} =$

$\boldsymbol{D_2} =$

ISBN 978-607-8285-62-4

Y así sucesivamente con los siguientes 9 datos con lo que se construye la siguiente tabla resumen:

Observación	**1**	**2**	**3**	**4**	**5**
X_1					
X_2					
Y					
Análisis de influencia D_i					

Observación	**6**	**7**	**8**	**9**	**10**
X_1					
X_2					
Y					
Análisis de influencia D_i					

Interpretación:

ISBN 978-607-8285-62-4

3.5.1 EJERCICIOS DE **REFUERZO**

EJERCICIOS DE REFUERZO 3.5.1 ANÁLISIS DE INFLUENCIA

NOTA:

El uso de un software estadístico como **Excel o Minitab**, entre otros, reduce de gran manera el tiempo de cálculo y la probabilidad de cometer errores en los cálculos aritméticos, sin embargo se deben **comprender primero los pasos del proceso**. Por lo mismo es **muy importante que primero resuelva el ejercicio en forma manual** y **posteriormente utilice un software para comparar sus resultados**. Es importante mencionar que **pueden existir diferencias** en las respuestas debido a la cantidad de dígitos que se **utilizan en los cálculos manuales. Se sugiere utilizar aproximaciones de 5 dígitos.**

3.5.1.1 **S**uponga que una compañía de productos de consumo quisiera medir la efectividad de la publicidad en radio y televisión y la publicidad en periódicos en la promoción de sus productos. Se seleccionó una muestra aleatoria de 10 ciudades con poblaciones aproximadamente iguales para el estudio durante un periodo de prueba de un mes. Se registraron las ventas (en miles de pesos) durante el mes de prueba, junto con los niveles de gastos en los medios de publicidad, con los resultados siguientes:

Ciudad	Ventas ($000,000) Y	Publicidad en radio y televisión ($000) X_1	Publicidad en periódicos ($000) X_2
1	9.73	0	400
2	6.25	250	250
3	9.71	300	300
4	9.31	350	350
5	11.77	350	350
6	9.82	400	250
7	16.28	450	450
8	13.29	550	250
9	14.36	600	300
10	17.41	650	350

n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.

o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.

p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

3.5.1.2 **E**l gerente distrital de ventas de de un fabricante de automóviles estudia las ventas de éstos. En forma específica, quiere determinar qué factores influyen en el número de automóviles vendidos en una distribuidora. Para investigarlo, seleccionó al azar 10 distribuidoras. De éstas, obtiene el número de automóviles vendidos el mes pasado, los minutos de publicidad en radio comprados el mes pasado y el número de vendedores de tiempo completo contratados. La información es la siguiente:

Distribuidora	Automóviles vendidos el mes pasado Y	Publicidad (en minutos) X_1	Fuerza de ventas X_2
1	127	18	10
2	159	22	14
3	144	23	12
4	139	17	12
5	128	16	12
6	180	26	17
7	102	15	7
8	163	24	16
9	106	18	10
10	149	30	11

n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.

o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.

p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

3.5.1.3 Los siguientes datos representan las calificaciones de estadística para una muestra aleatoria de 10 estudiantes de primer año de determinada institución de enseñanza superior, junto con sus calificaciones en un examen de inteligencia aplicado cuando aún cursaban el último año de secundaria y el número de periodos de clase perdidos por los 10 estudiantes que tomaron el curso de estadística.

Estudiante	Calificación de estadística Y	Calificación del examen X_1	Clases perdidas X_2
1	98	70	5
2	74	50	7
3	87	70	3
4	81	55	4
5	85	65	1
6	90	65	2
7	74	55	4
8	76	55	5
9	91	70	3
10	94	65	2

n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.

o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.

p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

R.L.M.

EJEMPLO **ILUSTRATIVO EN EXCEL**

EJEMPLO
ILUSTRATIVO
INTEGRAL EN **EXCEL**

Samuel Prado, propietario y director general de un establecimiento quiere conocer el comportamiento de las ventas (en miles de pesos) de un equipo de sonido que se expende en el establecimiento. Se percata de que existen muchos factores que podrían ayudarle a explicar las ventas pero piensa que la inversión en publicidad (en miles de pesos) y el precio (en cientos de pesos) son los principales factores determinantes. Samuel ha reunido los datos que se anexan a continuación.

Observación	Ventas Y	Publicidad X_1	Precio X_2
1	33	3	125
2	61	6	115
3	70	10	140
4	82	13	130
5	17	9	145
6	24	6	140
7	75	11	138
8	80	12	127
9	35	4	128
10	20	8	145

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$:
c) Determine el error estándar del estimador para toda la regresión lineal múltiple
d) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)
e) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.
f) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados
g) Determine los residuales estandarizados para toda la regresión
h) Construya la(s) grafica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.

ISBN 978-607-8285-62-4

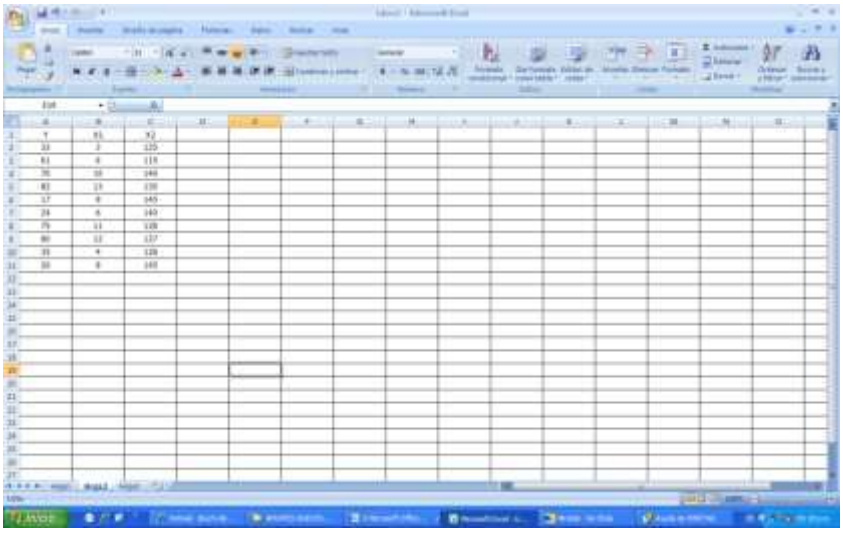
Ecuación de regresión lineal múltiple para los datos anteriores.

Solución al inciso a.

Cuando el número de observaciones en cada variable es extenso, los cálculos manuales son tediosos. Existen muchos paquetes de software que pueden mostrar los resultados entre ellos Excel.

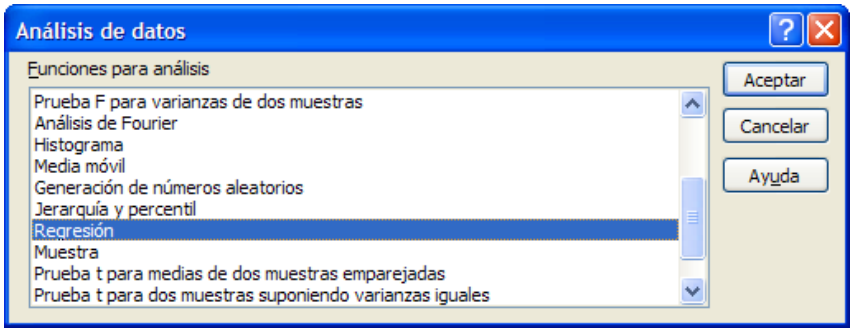
Comenzamos introduciendo los datos en la hoja de Excel, tal y como se muestra a continuación:

HOJA DE TRABAJO DE EXCEL.



Como tenemos un ***modelo de regresión y correlación lineal múltiple*** seleccionamos la opción ***Análisis de datos*** del menú ***Datos***, utilizaremos la opción ***Regresión*** del cuadro ***Análisis de datos*** de la figura siguiente:

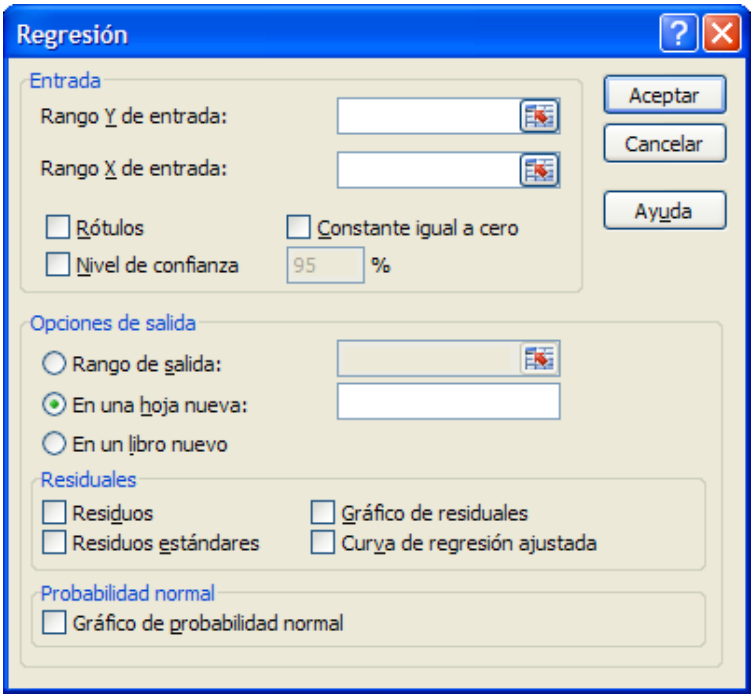
Cuadro de diálogo: Análisis de datos.



En la lista ***Funciones para análisis***, elija la modalidad de ***Regresión*** y oprima el botón ***Aceptar*** para obtener el siguiente cuadro de dialogo rellenando su pantalla de entrada:

ISBN 978-607-8285-62-4

Cuadro de diálogo: Regresión.



Los campos del cuadro de dialogo anterior tienen las siguientes funcionalidades:

En el cuadro ***Rango Y de entrada***: Introduzca la ***referencia*** correspondiente al rango de datos dependientes. El rango debe constar de una única columna o una única columna de datos.

En el cuadro ***Rango X de entrada***: Introduzca la ***referencia*** correspondiente al rango de datos independientes. Microsoft Excel ordenará las variables independientes de este rango en orden ascendente de izquierda a derecha. El número máximo de variables independientes es 16.

En el cuadro ***Rótulos***: Active esta casilla si la primera fila o la primera columna del rango (o rangos) de entrada contiene rótulos. Desactívela si el rango de entrada carece de rótulos; Excel generará los rótulos de datos correspondientes para la tabla de resultados.

En el cuadro ***Nivel de confianza***: Active esta casilla para incluir más niveles en la tabla resumen de resultados. Teclee el nivel de confianza a aplicar además del nivel predeterminado del 95%.

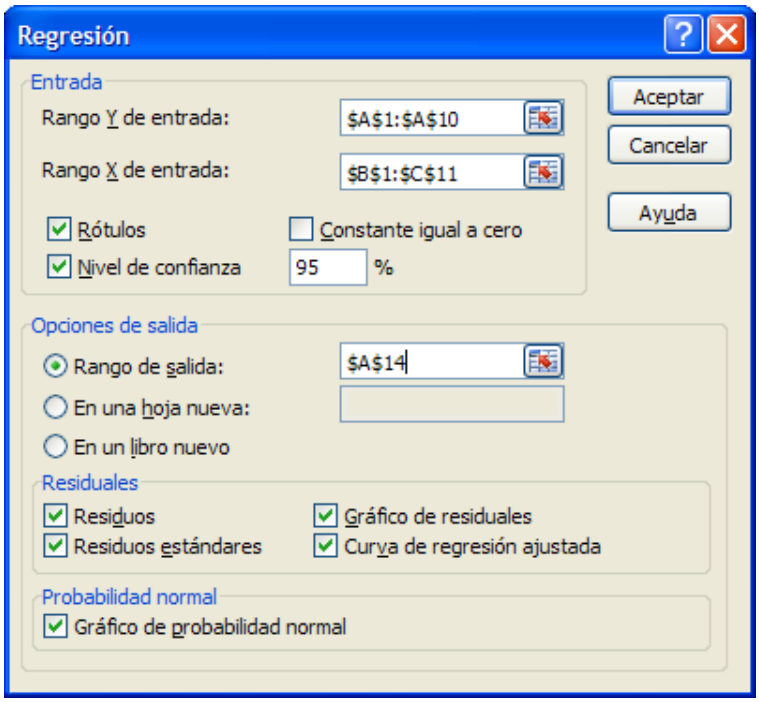
En el cuadro ***Constante igual a cero***: Active esta casilla para que la línea o plano de regresión pase por el origen.

Campos del cuadro de diálogo de regresión.

En el cuadro ***Rango de salida***: Introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados. Deje por lo menos siete columnas disponibles para la tabla de resultados sumarios, que incluirá una tabla de análisis de datos, coeficientes, error típico del pronóstico Y, valores de R^2, número de observaciones y error típico de coeficientes.

ISBN 978-607-8285-62-4

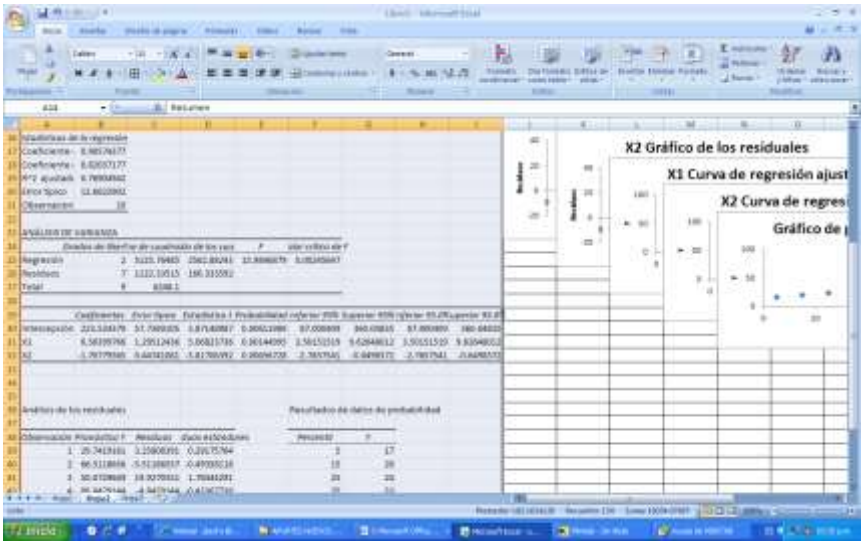
En el cuadro ***En una hoja nueva***: Haga clic en esta opción para insertar una hoja nueva en el libro actual y pegar los resultados, comenzando por la celda A1 de la nueva hoja de cálculo. Para darle un nombre a la nueva hoja de cálculo, escríbalo en el cuadro.
En el cuadro ***En un libro nuevo***: Haga clic en esta opción para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado.
En el cuadro ***Residuos***: Active esta casilla para incluir residuos en la tabla de resultados de residuos.
En el cuadro ***Residuos estándares***: Active esta casilla para incluir residuos estándares en la tabla de resultados de residuos.
En el cuadro ***Gráficos de residuos***: Active esta casilla para generar un gráfico por cada variable independiente frente al residuo.
En el cuadro ***Curva de regresión ajustada***: Active esta casilla para generar un gráfico con los valores pronosticados frente a los observados.
En el cuadro *Trazado de probabilidad normal*: Active esta casilla para generar un gráfico con probabilidad normal.



Cuadro de diálogo: Regresión.

Al pulsar ***Aceptar*** en el cuadro de dialogo anterior, se obtiene la siguiente salida numérica que incluye estadísticos de regresión, cuadro del análisis de varianza del modelo, estimadores, contrastes de significancia de ***F*** y ***T*** con sus *p*-valores asociados, intervalos de confianza para los parámetros y para las predicciones al 95%, y residuos.

ISBN 978-607-8285-62-4



Salida de resultados de Excel.

La ecuación de regresión lineal múltiple se puede expresar como

$$\boxed{\hat{Y}_i = 223.52438 + 6.56400X_{1i} - 1.70780X_{2i}}$$

La ecuación de regresión lineal múltiple.

Respuesta al inciso b.

Interpretación de los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1$ y $\hat{\beta}_2$:

La ordenada al origen $\hat{\beta}_0$, calculada como 223.52438, representa el volumen de ventas (en miles de pesos) que se generaría cuando la inversión en publicidad fuera de $ 0.00 pesos y el precio del equipo de sonido fuera de $ 0.00 pesos.

La pendiente de la inversión en publicidad $\hat{\beta}_1$, calculada como 6.56400, significa que para un equipo de sonido con *determinado* precio fijo (constante), el volumen de ventas se incrementará en $ 6.56400 por cada peso de aumento en la inversión en publicidad.

Asimismo la pendiente del precio del equipo de sonido $\hat{\beta}_2$, calculada como -1.70780, significa que para un equipo de sonido con *determinada* inversión fija en publicidad (constante), el volumen de ventas se disminuirá en $ 1.70780 por cada peso de aumento en el precio del equipo de sonido.

Respuesta al inciso c.

El error estándar del estimador, proporcionado por el símbolo $S_{Y.12}$, se define como

$$S_{Y.12} = 12.66383$$

ISBN 978-607-8285-62-4

El error estándar del estimador para toda la regresión lineal múltiple.

Respuesta al inciso d.

Una vez ajustado un modelo de regresión a un grupo de datos se debe determinar si hay relación significativa entre la variable dependiente y el grupo de variables explicatorias. Las hipótesis se pueden establecer de la siguiente manera:

Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes).

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

$$\boldsymbol{H_o: \beta_1 = \beta_2 = 0\ (no\ existe\ relación)}$$
$$\boldsymbol{H_1: (Por\ lo\ menos\ un\ coeficiente\ de\ regresión\ no\ es\ igual\ a\ cero)}$$

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$F_{calculada} = \frac{CMR}{CME} = \frac{SCR/_{G.L.}}{SCE/_{G.L.}} = 15.98$$

Paso 3.- Establecer la región de rechazo de (H_0).

$$\text{p-level} \leq 0.05$$

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $f_{cal.}$ tiene un p-level ≤ 0.05

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como p-level de 0.00245647 es <0.05 y <0.01 se considera que la prueba es Altamente Significativa (AS) y se rechaza H_o.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación entre el volumen de ventas y al menos una de las variables independientes, ya sea la inversión en publicidad ó el precio del equipo de sonido.

Prueba de hipótesis para cada uno de los coeficientes de regresión y su intervalo de confianza respectivo.

Respuesta al inciso e.

Para la variable independiente "inversión en publicidad"

Se usa el proceso de prueba de hipótesis de cinco pasos.

ISBN 978-607-8285-62-4

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0: β_1 =0 (no existe relación)
$H_{1:}$ $\beta_1 \neq 0$ (existe relación)

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$\boldsymbol{t_{calculada}} = \frac{\hat{\beta}_1}{S_{\beta_1}} = \frac{6.564}{1.29502} \cong 5.06865 \cong \mathbf{5.07}$$

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

p-level ≤ 0.05

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $t_{cal.}$ tiene un p-level ≤ 0.05

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como p-level de 0.00144995 es <0.05 y <0.01 se considera que la prueba es Altamente Significativa (AS) y se rechaza H_o.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación lineal significativa entre el volumen de ventas y la inversión en publicidad, es decir se concluye que el coeficiente de regresión no es cero. La variable independiente "inversión en publicidad" debe incluirse en el análisis.

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de $\hat{\beta}_1$ y determinar si el valor hipotético $(\hat{\beta}_1 = 0)$ está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_1$ se obtendría de la siguiente manera:

$$\boldsymbol{\beta_1} = \boldsymbol{\hat{\beta}_1 \pm t_{n-k-1} S_{\hat{\beta}_1}}$$

Intervalo de confianza de $\hat{\beta}_1$.

$$\boldsymbol{\beta_1} = \begin{cases} LIC = \mathbf{3.50152} \\ LSC = \mathbf{9.62648} \end{cases}$$

Interpretación: Se estima, con una confianza del 95%, que la pendiente real se encuentra entre 3.50 y 9.62. Puesto que estos valores son superiores a cero se puede concluir que hay relación lineal significativa entre el volumen de ventas y la inversión en publicidad. La variable independiente "inversión en publicidad" debe incluirse en el análisis.

ISBN 978-607-8285-62-4

Prueba de hipótesis para cada uno de los coeficientes de regresión y su intervalo de confianza respectivo.

<u>Para la variable independiente "precio del equipo de sonido"</u>

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0: β_2 =0 (no existe relación)
H_1: $\beta_2 \neq 0$ (existe relación)

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$\boldsymbol{t_{calculada}} = \frac{\hat{\beta}_2}{S_{\beta_2}} = \mathbf{-3.81706}$$

Paso 3.- Establecer la región de rechazo de (H_0).

p-level ≤ 0.05

Si el *valor p* es menor que o igual al nivel Alfa (a), rechace la hipótesis nula en favor de la hipótesis alternativa. Si $0.01 \leq p \leq 0.05$ se dice que la prueba es significativa (S). Si $p < 0.01$ se dice que la prueba es altamente significativa (AS).

Si el *valor p* es mayor que el nivel Alfa (a), no rechace la hipótesis nula. Si $p > 0.05$ se dice que la prueba es no significativa (NS).

Los niveles de significancia más utilizado son 0.05 y 0.01

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $t_{cal.}$ tiene un p-level ≤ 0.05

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como p-level de 0.00656728 es <0.05 y <0.01 se considera que la prueba es Altamente Significativa (AS) y se rechaza H_o.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación lineal significativa entre el volumen de ventas y el precio del equipo de sonido, es decir se concluye que el coeficiente de regresión no es cero. La variable independiente "precio del equipo de sonido" debe incluirse en el análisis.

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de $\hat{\beta}_2$ y determinar si el valor hipotético $(\hat{\beta}_2 = 0)$ está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_1$ se obtendría de la siguiente manera:

$$\boldsymbol{\beta_2} = \boldsymbol{\hat{\beta}_2 \mp t_{n-k-1} S_{\hat{\beta}_2}}$$

Intervalo de confianza de $\hat{\beta}_2$.

$$\boldsymbol{\beta_2} = \begin{cases} LIC = \mathbf{-2.76575} \\ LSC = \mathbf{-0.64983} \end{cases}$$

Interpretación: Se estima, con una confianza del 95%, que la pendiente real se encuentra entre -2.76 y -0.65. Puesto que estos valores son inferiores a cero se puede concluir que hay relación lineal significativa entre el volumen de ventas y el precio del equipo de sonido. La variable independiente "precio del equipo de sonido" debe incluirse en el análisis.

ISBN 978-607-8285-62-4

El coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados.

Respuesta al inciso f.

En la regresión múltiple, ya que existen por lo menos dos variables explicatorios, el coeficiente de determinación múltiple representa la proporción de la variación en Y que se explica por el grupo de variables explicativas seleccionadas. En el caso de dos variables explicativas, el coeficiente de determinación múltiple
$(r^2_{Y.12})$ se obtiene de la siguiente manera:

$$\boldsymbol{r^2_{Y.12}} = \frac{SCR}{SCT} = 0.82037x\ 100 = \mathbf{82.03}\%$$

Interpretación: Este coeficiente de determinación múltiple, significa que el 82.03% de la variación del volumen de ventas se puede explicar mediante la variación de la inversión en publicidad y la variación en el precio del equipo de sonido.

No obstante al tratar con modelos de regresión múltiple, algunos investigadores o analistas sugieren que se calcule un R^2 "ajustado" que refleje tanto el número de variables explicatorias en el modelo como el tamaño de la muestra. En la regresión múltiple se puede representar un R^2 ajustado como:

$$r^2_{adj} = 1 - \left[(1 - r^2_{Y.12\ldots k})\frac{n-1}{n-k-1}\right]$$

Por lo tanto

$$\boldsymbol{r^2_{adj}} = 1 - \left[(1 - r^2_{Y.12})\frac{n-1}{n-k-1}\right] = 0.76905x100 = \mathbf{76.90}\%$$

Interpretación: Lo anterior nos dice que el 76.90% de la variación en el volumen de ventas se puede explicar mediante el modelo de regresión lineal múltiple-ajustado por el número de variables de predicción y el tamaño de la muestra.

Por lo general la fuerza de una relación entre dos variables en una población se mide mediante el **coeficiente de correlación**, cuyos valores oscilan entre -1 para la correlación negativa perfecta hasta +1 para la correlación positiva perfecta. Se puede obtener con facilidad el coeficiente de correlación mediante la fórmula:

$$\boldsymbol{r_{Y.12..k}} = \sqrt{\boldsymbol{r^2_{Y.12\ldots k}}}$$

Entonces

Coeficiente de correlación.

$$\boldsymbol{r_{Y.12}} = \sqrt{r^2_{y.12}} = 0.90574\ x\ 100 = \mathbf{90.57}\%$$

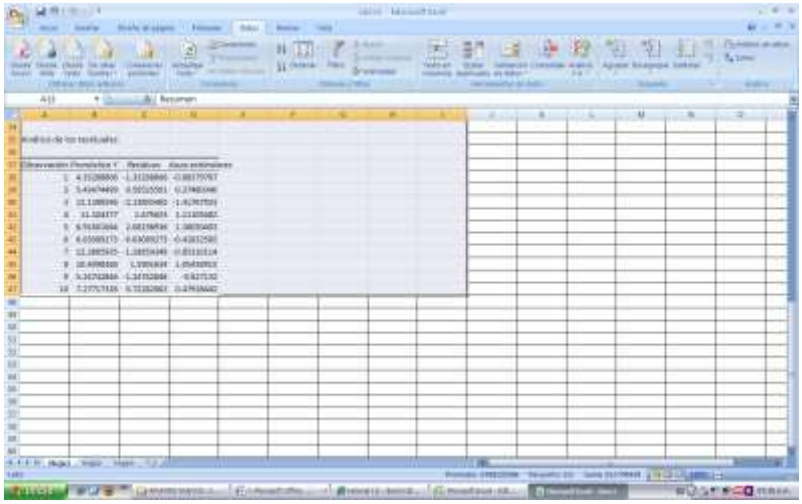
ISBN 978-607-8285-62-4

Interpretación: En este problema del volumen de ventas, puesto que $r^2 = 0.8203$, el coeficiente de correlación se interpreta como ***+0.9057***. La cercanía del coeficiente de correlación con +1.0 implica una fuerte asociación del volumen de ventas(Y) con respecto a la inversión en publicidad(X_1) y el precio del equipo de sonido (X_2).

Residuales estandarizados para toda la regresión.

Respuesta al inciso f.

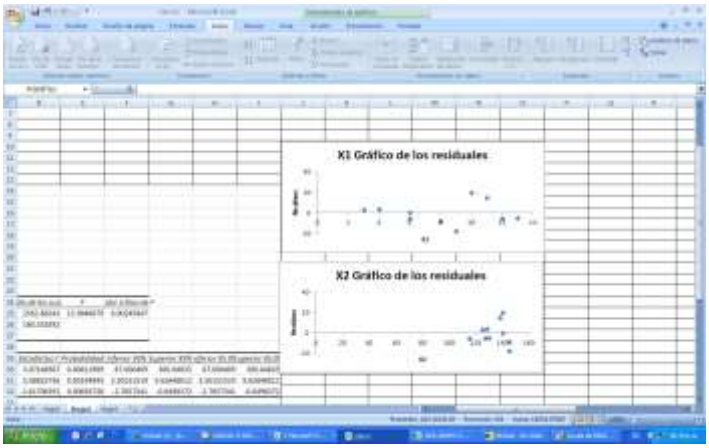
Análisis de residuales



Diagnóstico de la regresión.

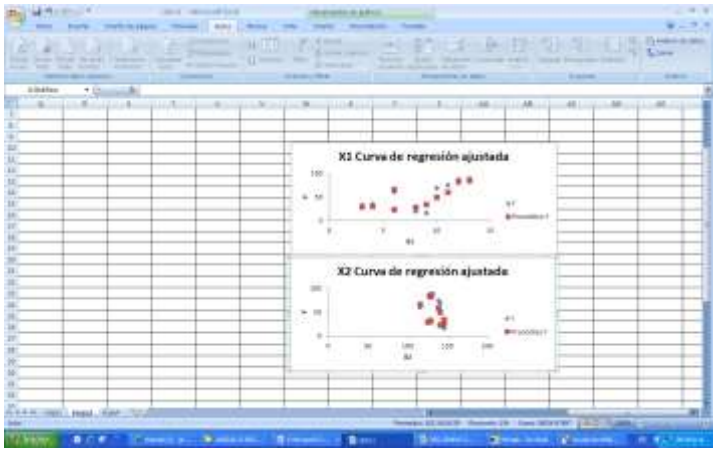
Solución al inciso g.

La figura siguiente presenta el gráfico de cada variable independiente contra los residuales, que sirve para detectar problemas de aleatoridad, linealidad, normalidad, homoscedasticidad y correlación en el modelo de ajuste. **Lo ideal es que todas las gráficas presenten una estructura aleatoria en sus puntos.**



ISBN 978-607-8285-62-4

La figura siguiente presenta el gráfico de cada variable independiente contra los valores predichos, que sirve para detectar problemas de homoscedasticidad. Lo ideal es que todas las gráficas presenten una estructura aleatoria en sus puntos.



Interpretación:

Interpretación:

Linealidad.

Linealidad

Se puede evaluar lo apropiado del modelo de regresión, trazando los "residuales estandarizados" sobre el eje vertical contra los valores $\hat{Y}$ en el eje horizontal. Si el modelo ajustado es apropiado para los datos no habrá un patrón aparente en esta gráfica de los residuales contra $\hat{Y}$. Sin embargo, si el modelo ajustado no es apropiado, habrá relación entre los valores $\hat{Y}$ y los residuales ε_i.

Así, se puede observar que aunque haya una amplia dispersión en la gráfica residual, no hay patrón ó relación aparente entre los residuales estandarizados y $\hat{Y}$. Los residuales parecen estar distribuidos en forma pareja por encima y por debajo de 0 para diferentes valores de $\hat{Y}$. Por lo tanto se puede concluir que el modelo ajustado parece ser el apropiado.

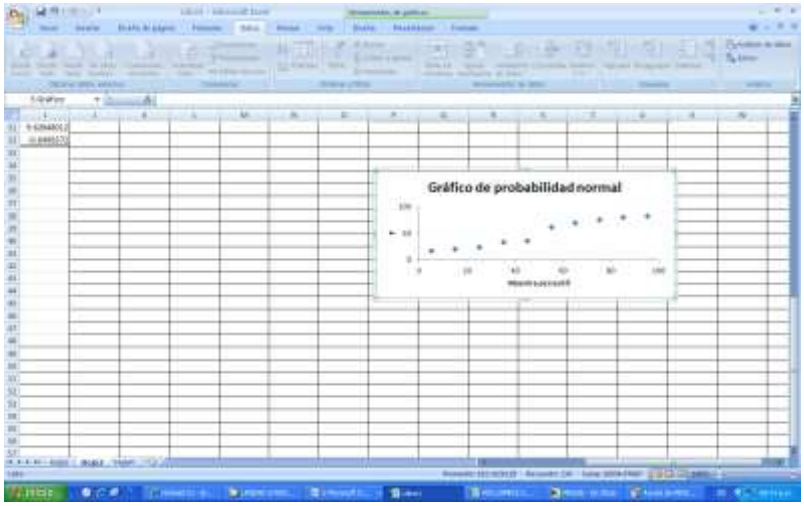
Homoscedasticidad.

Homoscedasticidad

La suposición de homoscedasticidad se puede evaluar también de la gráfica de residuales estandarizados con $\hat{Y}$. Si parece haber un *"efecto de abanico"* en el cual aumenta ó disminuye la variabilidad de los residuales al aumentar $\hat{Y}$ se demuestra la falta de homogeneidad en las varianzas de Y_i a cada nivel de $\hat{Y}$. Para los datos del volumen de ventas no parece haber diferencias importantes en la variabilidad de SR_i para diferentes valores de $\hat{Y}$. Por lo tanto se puede concluir que para este modelo ajustado no hay violación aparente a la suposición de igual varianza en cada nivel de $\hat{Y}$.

ISBN 978-607-8285-62-4

La figura siguiente presenta el gráfico para detectar hipótesis de normalidad en el modelo. La gráfica ideal es la diagonal del primer cuadrante.



Normalidad.

Normalidad

La gráfica de probabilidad normal muestra un patrón aproximadamente lineal que concuerda con una distribución normal. Los dos últimos puntos de la esquina superior derecha de la gráfica pueden ser valores atípicos. El destacado de la gráfica identifica estos puntos como 7 y 3, puntos que deberán verificarse como observaciónes inusuales ó identificación de valores atípicos. Excel no proporciona el estadístico de Anderson-Darling.

NOTA: Excel en este caso no tiene opción para construir intervalos de confianza para los verdaderos valores de Y, ni realizar pruebas para determinar la contribución de las variables explicatorias, ni determinar los coeficientes de determinación parcial, ni verificar multicolinearidad, ni realizar un análisis de influencia.

ISBN 978-607-8285-62-4

R.L.M.

EJEMPLO **ILUSTRATIVO EN MINITAB 15**

EJEMPLO ILUSTRATIVO INTEGRAL EN MINITAB 15

Samuel Prado, propietario y director general de un establecimiento quiere conocer el comportamiento de las ventas (en miles de pesos) de un equipo de sonido que se expende en el establecimiento. Se percata de que existen muchos factores que podrían ayudarle a explicar las ventas pero piensa que la inversión en publicidad (en miles de pesos) y el precio (en cientos de pesos) son los principales factores determinantes. Samuel ha reunido los datos que se anexan a continuación.

Observa-ción	Ventas Y	Publicidad X_1	Precio X_2
1	33	3	125
2	61	6	115
3	70	10	140
4	82	13	130
5	17	9	145
6	24	6	140
7	75	11	138
8	80	12	127
9	35	4	128
10	20	8	145

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$:
c) Determine el error estándar del estimador para toda la regresión lineal múltiple
d) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)
e) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.
f) Construya un intervalo de confianza para las verdaderas ventas cuando se destina una inversión en publicidad de $ 11, 000 y se fija un precio al producto de $ 13,500.00.
g) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.
h) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados.
i) Determine los coeficientes de determinación parcial e intérprete sus resultados
j) Verifique la existencia de multicolinearidad
k) Determine los residuales estandarizados para toda la regresión incluyendo el estadístico de Durbin-Watson.

ISBN 978-607-8285-62-4

l) Construya la(s) grafica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.

m) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.

n) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.

o) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

Ecuación de regresión lineal múltiple para los datos anteriores.

<u>Solución al inciso a</u>

Cuando el número de observaciones en cada variable es extenso, los cálculos manuales son tediosos. Existen muchos paquetes de software que pueden mostrar los resultados entre ellos Minitab (Versión 15)

Comenzamos introduciendo los datos en la hoja de Trabajo 1 de Minitab, tal y como se muestra a continuación:

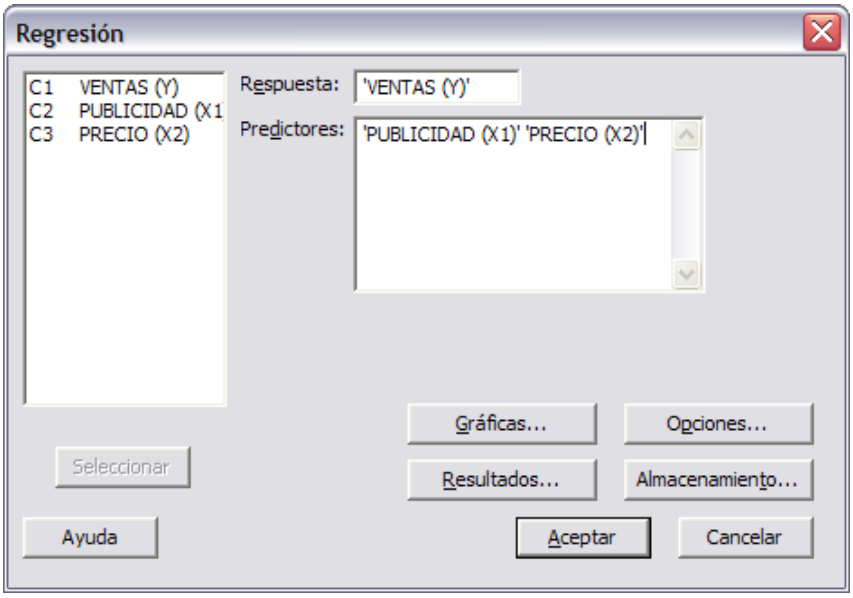
Como tenemos un ***modelo de regresión y correlación lineal múltiple*** seleccionamos la opción ***Regresión y Regresión*** del menú ***Estadísticas***,

Opción regresión del menú estadísticas.

En ***Respuesta***, ingrese C1 VENTAS (Y). En ***Predictores***, ingrese C2 PUBLICIDAD (X1) y C3 PRECIO (X2)

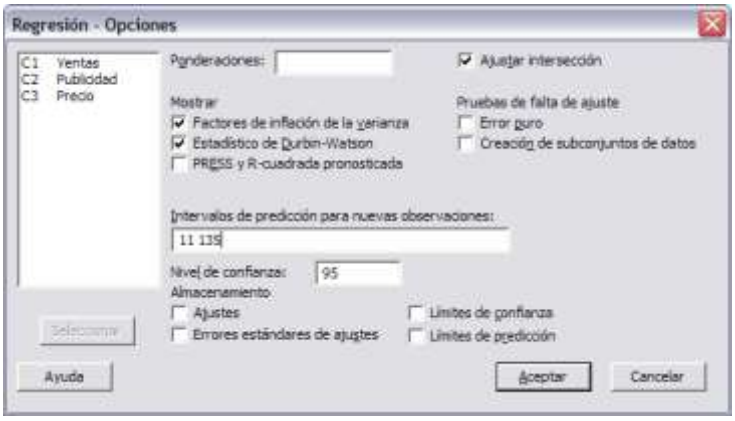
ISBN 978-607-8285-62-4

Cuadro de diálogo: Regresión.



Haga clic en el botón ***Opciones***. Active las casillas que dicen ***Factores de inflación de la varianza*** *y* ***Estadístico de Durbin-Watson***. En la opción ***Intervalos de confianza para nuevas observaciones*** escriba 11 135 (Dejando un espacio en blanco entre el 11 y el 135).

Cuadro de diálogo: Regresión-Opciones.



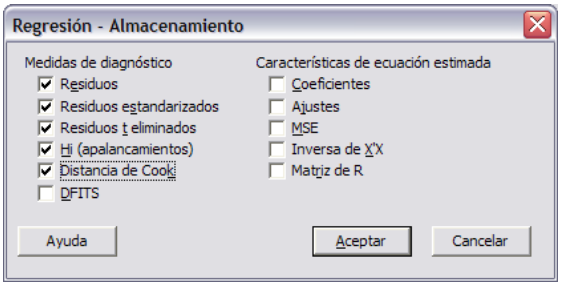
Haga clic en ***Aceptar*** en el cuadro de dialogo para que lo regrese al primer cuadro de dialogo

ISBN 978-607-8285-62-4

Cuadro de diálogo: Regresión.

Haga clic en el botón ***Almacenamiento***. En el campo ***Medidas de diagnóstico***, Active las casillas que dicen ***Residuos, Residuos estandarizados, Residuos t eliminados, H_i (apalancamientos) y Distancia de Cook***

Cuadro de diálogo: Regresión-Almacenamiento.



Haga clic en ***Aceptar*** en el cuadro de dialogo para que lo regrese al primer cuadro de dialogo

Cuadro de diálogo: Regresión.

Haga clic en el botón ***Gráficas***. Active la casilla ***Estandarizado*** en la ***opción Residuos para gráficas.*** En la opción ***Gráficas de residuos*** active las casillas Histograma de residuos, Gráfica normal de residuos, Residuos vs. ajustes y Residuos vs. orden. En la opción ***Residuos vs. las variables*** no coloque nada.

ISBN 978-607-8285-62-4

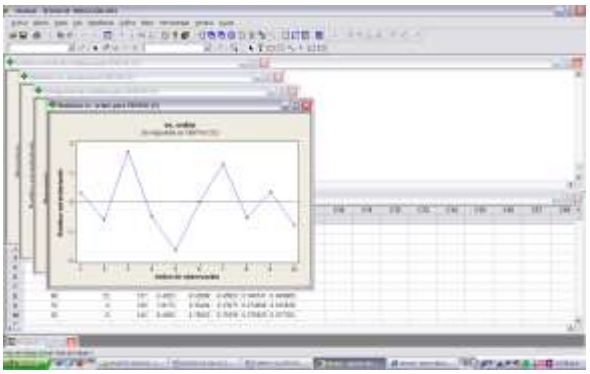
Cuadro de diálogo: Regresión-Gráficas..

Haga clic en ***Aceptar*** en el cuadro de dialogo para que lo regrese al primer cuadro de dialogo

Cuadro de diálogo: Regresión.

Haga clic en ***Aceptar*** en este cuadro de dialogo.

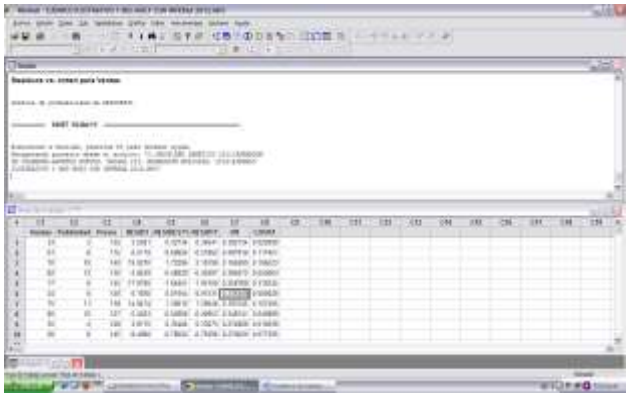
Salida de resultados y gráficas de Minitab 15.



Para probar la hipótesis de normalidad debemos realizar la **prueba de Anderson-Darling** una vez calculamos los residuales estandarizados de la siguiente salida

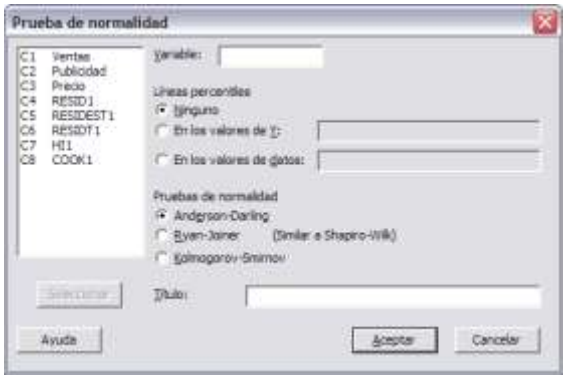
ISBN 978-607-8285-62-4

Salida de resultados de Minitab 15.



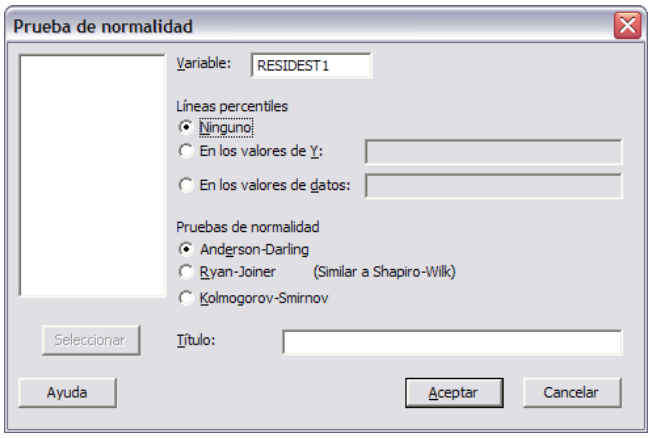
Seleccionamos la opción ***Estadísticas básicas y pruebas de normalidad*** del menú ***Estadísticas*** presentando el siguiente cuadro de dialogo

Cuadro de diálogo: Prueba de Normalidad.



En el campo ***Variable*** selecciona **C5 RESISEST1.** En el campo ***Prueba de normalidad*** ya estará activada la opción **Anderson-Darling**

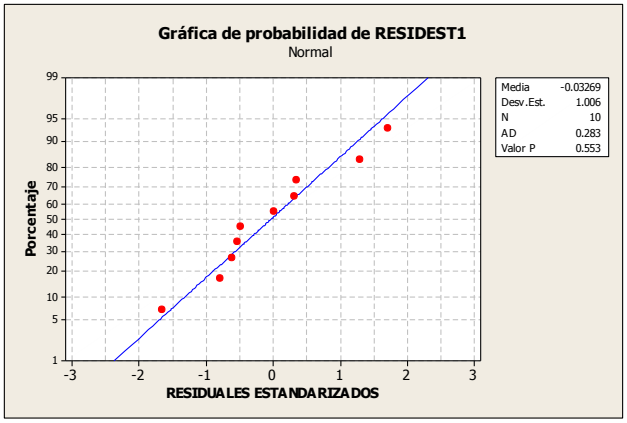
Cuadro de diálogo: Prueba de Normalidad.



ISBN 978-607-8285-62-4

Haga clic en ***Aceptar*** en este cuadro de dialogo para que nos proporcione la siguiente gráfica:

Gráfica de probabilidad Normal en Minitab 15.



Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

Análisis de regresión: VENTAS (Y) vs. PUBLICIDAD (X1), PRECIO (X2)

```
La ecuación de regresión es
VENTAS (Y) = 224 + 6.56 PUBLICIDAD (X1) - 1.71 PRECIO (X2)

                              Coef.
Predictor           Coef      de EE       T       P     VIF
Constante         223.52      57.74    3.87   0.006
PUBLICIDAD (X1)    6.564      1.295    5.07   0.001   1.084
PRECIO (X2)      -1.7078     0.4474   -3.82   0.007   1.084


S = 12.6623   R-cuad. = 82.0%   R-cuad.(ajustado) = 76.9%

Análisis de varianza

Fuente          GL      SC      MC       F       P
Regresión        2  5125.8  2562.9   15.98   0.002
Error residual   7  1122.3   160.3
Total            9  6248.1


Fuente          GL  SC sec.
PUBLICIDAD (X1)   1   2789.7
PRECIO (X2)       1   2336.1

Fuente          GL  SC sec.
PRECIO (X2)       1   1007.3
PUBLICIDAD (X1)   1   4118.5

Estadístico de Durbin-Watson = 2.41102
```


ISBN 978-607-8285-62-4

```
Valores pronosticados para nuevas observaciones

Nueva          Ajuste
  Obs  Ajuste      SE     IC de 95%       PI de 95%
    1   65.18    5.31  (52.61, 77.74)  (32.71, 97.65)

Valores de predictores para nuevas observaciones

Nueva  PUBLICIDAD  PRECIO
  Obs        (X1)    (X2)
    1        11.0     135
```

La ecuación de regresión lineal múltiple se puede expresar como

Ecuación para la regresión lineal múltiple.

$$\boxed{\hat{Y}_i = 224 + 6.56X_{1i} - 1.71X_{2i}}$$

Donde

Y_i = volumen de ventas (en miles de pesos) para la observación *i*.
X_{1i}=Inversión en publicidad (en miles de pesos) para la observación *i*.
X_{2i}= precio del equipo (en cientos de pesos) para la observación *i*.

Salida de la ventana Sesión de Minitab

Salida del programa Minitab 15.

```
La ecuación de regresión es
VENTAS (Y) = 224 + 6.56 PUBLICIDAD (X1) - 1.71 PRECIO (X2)
```

Solución al inciso b.

Interpretación de los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1$ y $\hat{\beta}_2$:

La ordenada al origen $\hat{\beta}_0$, calculada como 224, representa el volumen de ventas (en miles de pesos) que se generaría cuando la inversión en publicidad fuera de $ 0.00 pesos y el precio del equipo de sonido fuera de $ 0.00 pesos.
La pendiente de la inversión en publicidad $\hat{\beta}_1$, calculada como 6.56, significa que para un equipo de sonido con *determinado* precio fijo (constante), el volumen de ventas se incrementará en $ 6.56 por cada peso de aumento en la inversión en publicidad.
Asimismo la pendiente del precio del equipo de sonido $\hat{\beta}_2$, calculada como -1.71, significa que para un equipo de sonido con *determinada* inversión fija en publicidad (constante), el volumen de ventas se disminuirá en $ 1.71 por cada peso de aumento en el precio del equipo de sonido.

Solución al inciso c.

El error estándar del estimador para toda la regresión lineal múltiple.

El error estándar del estimador, proporcionado por el símbolo $S_{Y.X}$, se define como

$$S_{Y.X} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i}{n-2}} = 12.6623$$

ISBN 978-607-8285-62-4

Salida del programa Minitab 1.

Salida de la ventana Sesión de Minitab

```
S = 12.6623   R-cuad. = 82.0%   R-cuad.(ajustado) = 76.9%
```

Nota: en la salida de la ventana Sesión el error estándar del estimador tiene la nomenclatura **(S).**

Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes).

Solución al inciso d.

Una vez ajustado un modelo de regresión a un grupo de datos se debe determinar si hay relación significativa entre la variable dependiente y el grupo de variables explicativas. Las hipótesis se pueden establecer de la siguiente manera:

Juego de hipótesis:

$$\boldsymbol{H_o: \beta_1 = \beta_2 = 0\ (no\ existe\ relación)}$$
$$\boldsymbol{H_1: (Por\ lo\ menos\ un\ coeficiente\ de\ regresión\ no\ es\ igual\ a\ cero)}$$

Se puede probar la hipótesis nula utilizando una prueba *F*. Cuando se prueba la significación de los coeficientes de regresión, a la medida del error aleatorio de le conoce como la **varianza del error**, por lo que la prueba ***F*** es la razón de la varianza debida a la regresión dividida entre la varianza del error

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

Paso 1. Juego de hipótesis.

$$\boldsymbol{H_o: \beta_1 = \beta_2 = 0\ (no\ existe\ relación)}$$
$$\boldsymbol{H_1: (Por\ lo\ menos\ un\ coeficiente\ de\ regresión\ no\ es\ igual\ a\ cero)}$$

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$F_{calculada} = \frac{CMR}{CME} = \frac{SCR/_{G.L.}}{SCE/_{G.L.}} = 15.98$$

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

```
Análisis de varianza

Fuente          GL      SC      MC      F      P
Regresión        2  5125.8  2562.9  15.98  0.002
Error residual   7  1122.3   160.3
Total            9  6248.1
```


ISBN 978-607-8285-62-4

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

p-level ≤ 0.05

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $f_{cal.}$ tiene un p-level ≤ 0.05

Paso 5. Conclusiones

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como p-level de 0.002 es <0.05 y <0.01 se considera que la prueba es Altamente Significativa (AS) y se rechaza H_o.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación entre el volumen de ventas y al menos una de las variables independientes, ya sea la inversión en publicidad ó el precio del equipo de sonido.

Prueba de hipótesis para cada uno de los coeficientes de regresión e intervalo de confianza respectivo.

<u>Solución al inciso e.</u>

Hasta este punto se ha mostrado que alguno, pero no necesariamente todos los coeficientes de regresión, no son iguales a cero y, por tanto, son útiles para las predicciones. El siguiente paso consiste en probar individualmente las variables para determinar cuáles coeficientes de regresión pueden ser 0 y cuáles no. Si una β puede ser cero, ello implica que esta variable independiente en particular no tiene ningún valor para explicar cualquier variación en el valor dependiente. Si hay coeficientes para los cuales no se puede rechazar H_0, se pueden eliminar de la ecuación de regresión.

Ahora se realizarán dos pruebas de hipótesis: para la inversión en publicidad y para el precio del equipo de sonido.

Prueba de hipótesis para β_1.

1. Prueba para la inversión en publicidad.

La distribución t se puede utilizar para realizar pruebas de significancia y construir intervalos de confianza para la verdadera pendiente

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0: $\beta_1 = 0$ (no existe relación)
H_1: $\beta_1 \neq 0$ (existe relación)

ISBN 978-607-8285-62-4

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$\boldsymbol{t_{calculada}} = \frac{\hat{\beta}_1}{S_{\beta_1}} = \mathbf{5.07}$$

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

```
                                Coef.
Predictor            Coef       de EE      T      P     VIF
Constante          223.52       57.74   3.87  0.006
PUBLICIDAD (X1)     6.564       1.295   5.07  0.001
1.084
PRECIO (X2)       -1.7078      0.4474  -3.82  0.007  1.084
```

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

p-level ≤ 0.05

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $t_{cal.}$ tiene un p-level ≤ 0.05

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como p-level de 0.001 es <0.05 y <0.01 se considera que la prueba es Altamente Significativa (AS) y se rechaza H_o.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación lineal significativa entre el volumen de ventas y la inversión en publicidad, es decir se concluye que el coeficiente de regresión no es cero. La variable independiente "inversión en publicidad" debe incluirse en el análisis.

NOTA: Minitab no establece intervalos de confianza para la pendiente de la recta pero pueden construirse fácilmente de la salida de Minitab de la siguiente manera:

Intervalo de confianza para β_1.

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de $\hat{\beta}_1$ y determinar si el valor hipotético $(\hat{\beta}_1 = 0)$ está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_1$ se obtendría de la siguiente manera:

ISBN 978-607-8285-62-4

$$\boldsymbol{\beta_1} = \widehat{\boldsymbol{\beta}}_{\mathbf{1}} \pm \boldsymbol{t_{n-k-1} S_{\widehat{\beta}_1}}$$

$$\beta_1 = 6.564 \mp 2.36(1.295)$$

$$\boldsymbol{\beta_1} = \begin{cases} LIC = 6.564 - 3.0562 = \mathbf{3.50780} \\ LSC = 6.564 + 3.0562 = \mathbf{9.62020} \end{cases}$$

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

```
                                Coef.
Predictor                 Coef   de EE      T      P    VIF
Constante               223.52   57.74   3.87  0.006
PUBLICIDAD (X1)          6.564   1.295   5.07  0.001
1.084
PRECIO (X2)           -1.7078  0.4474  -3.82  0.007  1.084
```

Otra forma de expresar el intervalo de confianza es:

$$\widehat{\boldsymbol{\beta}}_{\mathbf{1}} - \boldsymbol{t_{n-k-1} S_{\widehat{\beta}_1}} \leq \boldsymbol{\beta_1} \leq \widehat{\boldsymbol{\beta}}_{\mathbf{1+}} \boldsymbol{t_{n-k-1} S_{\widehat{\beta}_1}}$$

$$3.50780 \leq \boldsymbol{\beta_1} \leq 9.62020$$

Interpretación: Se estima, con una confianza del 95%, que la pendiente real se encuentra entre 3.50780 y 9.62020. Puesto que estos valores son superiores a cero se puede concluir que hay relación lineal significativa entre el volumen de ventas y la inversión en publicidad. La variable independiente "inversión en publicidad" debe incluirse en el análisis.

Prueba de hipótesis para β_2.

2. Prueba para el precio del equipo de sonido

La distribución t se puede utilizar para realizar pruebas de significancia y construir intervalos de confianza para la verdadera pendiente

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0: β_2 =0 (no existe relación)
H_1: $\beta_2 \neq 0$ (existe relación)

Paso 2. Estadístico de prueba

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$\boldsymbol{t_{calculada}} = \frac{\hat{\beta}_2}{S_{\beta_2}} = \mathbf{-3.82}$$

ISBN 978-607-8285-62-4

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

```
                        Coef.
Predictor              Coef   de EE      T      P    VIF
Constante            223.52   57.74   3.87  0.006
PUBLICIDAD (X1)       6.564   1.295   5.07  0.001  1.084
PRECIO (X2)         -1.7078  0.4474  -3.82  0.007
1.084
```

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

$$\text{p-level} \leq 0.05$$

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si $t_{cal.}$ tiene un p-level ≤ 0.05

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: Como p-level de 0.007 es <0.05 y <0.01 se considera que la prueba es Altamente Significativa (AS) y se rechaza H_o.

Administrativa: Existe evidencia suficiente para decir que estadísticamente existe relación lineal significativa entre el volumen de ventas y el precio del equipo de sonido, es decir se concluye que el coeficiente de regresión no es cero. La variable independiente "precio del equipo de sonido" debe incluirse en el análisis.

NOTA: Minitab no establece intervalos de confianza para la pendiente de la recta pero pueden construirse fácilmente de la salida de Minitab de la siguiente manera:

Un segundo y equivalente método para probar la existencia de una relación lineal entre las variables, es establecer un estimado de intervalo de confianza de $\hat{\beta}_2$ y determinar si el valor hipotético $(\hat{\beta}_2 = 0)$ está incluido en el intervalo. El estimado del intervalo de confianza de $\hat{\beta}_1$ se obtendría de la siguiente manera:

Intervalo de confianza para β_2.

$$\boldsymbol{\beta_2 = \hat{\beta}_2 \mp t_{n-k-1} S_{\hat{\beta}_2}}$$

$$\beta_2 = -1.7078 \mp 2.36(0.4474)$$

$$\boldsymbol{\beta_2} = \begin{cases} LIC = -1.7078 - 1.05586 = \mathbf{-2.76366} \\ LSC = -1.7078 + 1.05586 = \mathbf{-0.65194} \end{cases}$$

ISBN 978-607-8285-62-4

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

```
                Coef.
Predictor          Coef   de EE      T      P    VIF
Constante        223.52   57.74   3.87  0.006
PUBLICIDAD (X1)   6.564   1.295   5.07  0.001  1.084
PRECIO (X2)     -1.7078   0.4474  -3.82  0.007
1.084
```

Otra forma de expresar el intervalo de confianza es:

$$\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{t}_{n-k-1}\boldsymbol{S}_{\widehat{\beta}_2} \leq \boldsymbol{\beta}_2 \leq \widehat{\boldsymbol{\beta}}_{2+}\boldsymbol{t}_{n-k-1}\boldsymbol{S}_{\widehat{\beta}_2}$$

$$-\mathbf{2.76366} \leq \boldsymbol{\beta}_2 \leq -\mathbf{0.65194}$$

Interpretación: Se estima, con una confianza del 95%, que la pendiente real se encuentra entre -2.76366 y -0.65194. Puesto que estos valores son inferiores a cero se puede concluir que hay relación lineal significativa entre el volumen de ventas y el precio del equipo de sonido. La variable independiente "precio del equipo de sonido" debe incluirse en el análisis.

Intervalo de confianza para las verdaderas ventas cuando se destina una inversión en publicidad de $ 11, 000 y se fija un precio al producto de $ 13,500.00.

Solución al inciso f.

Se da otro uso de la notación matricial cuando se utiliza un modelo de regresión múltiple para estimar el valor esperado de $\boldsymbol{Y}$, con nuevos valores de las variables independientes. Las fórmulas en notación matricial incluyen nuevamente a la matriz $(\boldsymbol{X}'\boldsymbol{X})^{-1}$**.**

$$\mu_{Y.12} = \hat{Y}_i \pm t_{\alpha/2.n-k-1} S_{Y:12}\sqrt{h_i}$$

Donde

$$h_i = X_i'(X'X)^{-1}X_i$$

Por lo tanto

$$\mu_{Y.11.135} = \hat{Y}_i \pm t_{0.05,7} S_{Y:12}\sqrt{h_i} = \begin{cases} LIC = 52.61\ (Miles\ de\ pesos) \\ LSC = 77.74 (Miles\ de\ pesos) \end{cases}$$

Otra forma de expresar el intervalo de confianza es:

$$\mathbf{52.61} \leq \mu_{Y.11.135} \leq 77.74$$

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

```
Valores pronosticados para nuevas observaciones

Nueva          Ajuste
  Obs  Ajuste      SE     IC de 95%        PI de 95%
    1   65.18    5.31  (52.61, 77.74)  (32.71, 97.65)
```

Interpretación: En **95** de cada **100** muestras (95% de confianza) de tamaño **10**, el verdadero volumen de ventas promedio cuando se invierte en publicidad **$ 11,000.00 pesos** y se fija un precio al producto de **$13,500.00** oscilará aproximadamente entre **$ 52,610.00** y **$ 77,740.00** pesos.

ISBN 978-607-8285-62-4

Criterio de las "f" parciales para determinar la contribución de las variables explicatorias.

Solución al inciso g

NOTA: Minitab no calcula F parciales sin embargo dependiendo el orden en el que se coloquen las variables independientes proporciona una salida con los valores de SCR(X_2) y la SCR(X_1 | X_2) si se coloca primero C3 PRECIO(X_3) y SCR(X_2) y después C2 PUBLICIDAD (X_1) y SCR(X_1) y la SCR(X_2 | X_1) si se colocó primero C2 PUBLICIDAD (X_1) y después C3 PRECIO(X_3) y. Con estos datos se facilita mucho el cálculo de las F parciales. En este caso:

Contribución de X_1.

1. Contribución de X_1 (inversión en publicidad) una vez incluida X_2 (precio del producto) en el modelo :

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).
H_0: la variable X_1 no mejora en forma significativa el modelo, una vez incluida la variable X_2.
H_1: la variable X_1 mejora en forma significativa el modelo, una vez incluida la variable X_2.

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$F_{1.n-k-1} = F_{1.10-2-1} = F_{1.7} = \frac{SCR(X_1 \mid X_2)}{CME} = \frac{SCR(X_1 y X_2) - SCR(X_2)}{CME}$$
$$= \frac{CMR(X_1 \mid X_2)}{CME} = 25.69$$

SCR(X_2)=1,007.3
SCR(X_1 | X_2)=4,118.5

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

```
Análisis de varianza

Fuente           GL      SC      MC      F      P
Regresión         2  5125.8  2562.9  15.98  0.002
Error residual    7  1122.3   160.3
Total             9  6248.1

Fuente           GL  SC sec.
PRECIO (X2)       1   1007.3
PUBLICIDAD (X1)   1   4118.5
```

Con los datos anteriores se puede calcular:
CMR(X_1 | X_2)= SCR(X_1 | X_2) /G.L.=4,118.5/1=4,118.5

$$F_{1.n-k-1} = F_{1.10-2-1} = F_{1.7} = \frac{CMR(X_1 \mid X_2)}{CME} = \frac{4{,}118.5}{160.3} = 25.69$$

ISBN 978-607-8285-62-4

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

$F_{1.n-k-1} = F_{1.10-2-1} = F_{1.7} = 5.59$

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si fcalc $\geq$ 5.59

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística:

$$Como\ 25.69 > 5.59 > 12.25\ \therefore la\ prueba\ es\ (AS)\ y\ se\ rechaza H_0\ .$$

Administrativa: Existe evidencia suficiente para decir que estadísticamente la variable X_1 (inversión en publicidad) si contribuye significativamente en el modelo, una vez incluida la variable X_2 (precio del producto).

Contribución de X_2.

2. Contribución de X_2 (precio del producto) una vez incluida en el modelo X_1 (inversión en publicidad) :

Se usa el proceso de prueba de hipótesis de cinco pasos.

Paso 1. Juego de hipótesis.

Paso 1.- Establecer la hipótesis nula (H_0) y la hipótesis alternativa (H_1).

H_0: la variable X_2 no mejora en forma significativa el modelo, una vez incluida la variable X_1.

H_1: la variable X_2 mejora en forma significativa el modelo, una vez incluida la variable X_1.

Paso 2. Estadístico de prueba.

Paso 2.- Seleccionar y calcular el valor del estadístico de prueba apropiado.

$$F_{1.n-k-1} = F_{1.10-2-1} = F_{1.7} = \frac{SCR(X_2 \mid X_1)}{CME} = \frac{SCR(X_1 y X_2) - SCR(X_1)}{CME}$$
$$= \frac{CMR(X_2 \mid X_1)}{CME} = 14.57$$

SCR(X_1)=2,789.7

SCR(X_2 | X_1)=2,336.1

ISBN 978-607-8285-62-4

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

```
Análisis de varianza
Fuente          GL      SC      MC      F      P
Regresión        2  5125.8  2562.9  15.98  0.002
Error residual   7  1122.3   160.3
Total            9  6248.

Fuente           GL  SC sec.
PUBLICIDAD (X1)   1   2789.7
PRECIO (X2)       1   2336.1
```

Con los datos anteriores se puede calcular:
CMR(X_2 | X_1)= SCR(X_2 | X_1) /G.L.=2,336.1/1=2,336.1

$F_{1.n-k-1} = F_{1.10-2-1} = F_{1.7} = \frac{CMR(X_2|X_1)}{CME} = \frac{2,336.1}{160.3}$ =14.57

Paso 3. Región de rechazo.

Paso 3.- Establecer la región de rechazo de (H_0).

$F_{1.n-k-1} = F_{1.10-2-1} = F_{1.7} = 5.59$

Paso 4. Regla de decisión.

Paso 4.- Formular una regla de decisión basada en los pasos 1,2 y 3 anteriores.

Se rechaza H_0 si fcalc ≥ 5.59

Paso 5. Conclusiones.

Paso 5.- Tomar una decisión en cuanto a la hipótesis nula con base en la información de la muestra (conclusión estadística). Interpretar los resultados de la prueba (conclusión administrativa).

Estadística: como 14.56
$> 5.59 > 12.25 \therefore la\ prueba\ es\ (AS)\ y\ se\ rechaza H_0$.

Administrativa: Existe evidencia suficiente para decir que estadísticamente la variable X_2 (precio del producto) si contribuye significativamente en el modelo, una vez incluida la variable X_1 (inversión en publicidad).

Por lo tanto, mediante la prueba de la contribución de cada variable explicativa, luego que ha incluida la otra en el modelo, se determinó que cada una de las dos variables explicativas contribuyen a mejorar en forma significativa el modelo. Por consiguiente, el modelo de regresión múltiple debe incluir tanto la inversión en publicidad X_1 como el precio del equipo de sonido X_2.

ISBN 978-607-8285-62-4

El coeficiente de determinación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo.

Solución al inciso h.

En la regresión múltiple, ya que existen por lo menos dos variables explicatorias, el coeficiente de determinación múltiple representa la proporción de la variación en Y que se explica por el grupo de variables explicatorias seleccionadas. En el caso de dos variables explicativas, el coeficiente de determinación múltiple
$(r^2_{Y.12})$ se obtiene de la siguiente manera:

$$\boldsymbol{r^2_{Y.12} = 82.0\%}$$

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

```
S = 12.6623   R-cuad. = 82.0%   R-cuad.(ajustado) = 76.9%
```

Interpretación: Este coeficiente de determinación múltiple, significa que el 82.03% de la variación del volumen de ventas se puede explicar mediante la variación de la inversión en publicidad y la variación en el precio del equipo de sonido.
No obstante al tratar con modelos de regresión múltiple, algunos investigadores o analistas sugieren que se calcule un R^2 "ajustado" que refleje tanto el número de variables explicatorias en el modelo como el tamaño de la muestra. En la regresión múltiple se puede representar un R^2 ajustado como:

El coeficiente de determinación múltiple ajustado.

$$r^2_{adj} = 1 - \left[(1 - r^2_{Y.12\ldots k})\frac{n-1}{n-k-1}\right]$$

Por lo tanto

Salida del programa Minitab 15.

$$\boldsymbol{r^2_{adj}} = 1 - \left[(1 - r^2_{Y.12})\frac{n-1}{n-k-1}\right] = \boldsymbol{76.9\%}$$

Salida de la ventana Sesión de Minitab

```
S = 12.6623   R-cuad. = 82.0%   R-cuad.(ajustado) = 76.9%
```

Interpretación: Lo anterior nos dice que el 76.90% de la variación en el volumen de ventas se puede explicar mediante el modelo de regresión lineal múltiple-ajustado por el número de variables de predicción y el tamaño de la muestra.

ISBN 978-607-8285-62-4

NOTA: Minitab no establece el coeficiente de correlación pero se puede calcular fácilmente en forma manual:

El coeficiente de correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo.

Por lo general la fuerza de una relación entre dos variables en una población se mide mediante el **coeficiente de correlación**, cuyos valores oscilan entre -1 para la correlación negativa perfecta hasta +1 para la correlación positiva perfecta. Se puede obtener con facilidad el coeficiente de correlación mediante la fórmula:

$$\boldsymbol{r_{Y.12..k}} = \sqrt{\boldsymbol{r^2_{Y.12...k}}}$$

Entonces

$$\boldsymbol{r_{Y.12}} = \sqrt{r^2_{y.12}} = \sqrt{0.82} = 0.90554\ x\ 100 = \mathbf{90.55}\%$$

Interpretación: En este problema del volumen de ventas, puesto que $r^2 = 0.82$, el coeficiente de correlación se interpreta como ***+0.9055***. La cercanía del coeficiente de correlación con +1.0 implica una fuerte asociación del volumen de ventas(Y) con respecto a la inversión en publicidad(X_1) y el precio del equipo de sonido (X_2).

Coeficientes de determinación parcial.

<u>Solución al inciso i.</u>

NOTA: Minitab no establece el coeficiente de determinación parcial pero se puede calcular fácilmente en forma manual con datos anteriores:

Los coeficientes de determinación parcial ($r^2_{y1.2}$ y $r^2_{y2.1}$) miden la proporción de la variación en la variable independiente que se explica por cada variable explicatoria, al mismo tiempo que se controlan o se mantienen constantes las otras variables explicatorias.

Para un modelo de regresión múltiple con diversas variables explicatorias (k) resulta que:

$$\boldsymbol{r^2_{Yk}todas\ las\ variables\ exepto\ k} = \frac{\boldsymbol{SCR(X_k todas\ las\ variables\ excepto\ k)}}{\boldsymbol{SCT - SCR(todas\ las\ variables\ incluso\ k + SCR(X_k todas\ las\ variables\ excepto\ k)}}$$

Entonces se puede calcular el coeficiente de determinación parcial de X_1 como,

Coeficiente de determinación parcial de $r^2_{Y1.2}$.

$$\boldsymbol{r^2_{Y1.2}} = \frac{\boldsymbol{SCR(X_1 \mid X_2)}}{\boldsymbol{SCT - SCR(X_1\ y\ X_2) + SCR(X_1 \mid X_2)}} = \frac{\mathbf{4,118.5}}{\mathbf{6,248.1 - 5,125.8 + 4,118.5}} = \mathbf{0.78585x100} = \mathbf{78.58}\%$$

ISBN 978-607-8285-62-4

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

```
Análisis de varianza

Fuente          GL      SC      MC      F      P
Regresión        2  5125.8  2562.9  15.98  0.002
Error residual   7  1122.3   160.3
Total            9  6248.1


PRECIO (X2)       1   1007.3
PUBLICIDAD (X1)   1   4118.5
```

INTERPRETACIÓN: El coeficiente de determinación parcial de la variable dependiente Y, volumen de ventas, cuando se mantiene constante ***X₂***, el precio del equipo de sonido, significa que para un precio fijo (constante) en el equipo de sonido, el **78.58%** de la variación en el volumen de ventas se puede explicar por **la inversión en publicidad *X₁***.

Coeficiente de determinación parcial de $r^2_{Y2.1}$.

Y el coeficiente de determinación parcial de X_2 como,

$$\boldsymbol{r^2_{Y2.1}} = \frac{SCR(X_2 \mid X_1)}{SCT - SCR(X_1\ y\ X_2) + SCR(X_2 \mid X_1)} = \frac{2{,}336.1}{6{,}248.1 - 5{,}125.8 + 2{,}336.1}$$
$$= 0.67549x100 = \mathbf{67.55}\%$$

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

```
Análisis de varianza

Fuente          GL      SC      MC      F      P
Regresión        2  5125.8  2562.9  15.98  0.002
Error residual   7  1122.3   160.3
Total            9  6248.1

Fuente          GL  SC sec.
PUBLICIDAD (X1)   1   2789.7
PRECIO (X2)       1   2336.1
```

INTERPRETACIÓN: El coeficiente de determinación parcial de la variable dependiente Y, volumen de ventas, cuando se mantiene constante ***X₁***, la inversión en publicidad, significa que para una inversión en publicidad fija (constante), el **67.55%** de la variación en el volumen de ventas se puede explicar por **el precio del equipo de sonido.**

ISBN 978-607-8285-62-4

Multicolinearidad.

Solución al inciso j.

Un problema importante en la aplicación del análisis de regresión múltiple incluye la posible multicolinearidad de las variables independientes ó explicatorias. Esta condición se refiere a situaciones en que algunas variables explicatorias estén altamente correlacionadas entre sí. En esas situaciones las variables colineales no proporcionan información nueva y resulta difícil separar el efecto de esas variables sobre la variable dependiente o de respuesta. En esos casos los valores de los coeficientes de regresión para las variables correlacionadas pueden fluctuar en forma importante, dependiendo de qué variables estén incluidas en el modelo.

Un método de medir la colinearidad usa el factor de varianza inflacionaria **(VIF)** para cada variable explicatoria. Este **VIF** se define en la siguiente ecuación:

$$VIF_j = \frac{1}{1 - r_j^2}$$

Donde r_j^2 representa el coeficiente de determinación múltiple de la variable explicatoria X_j con todas las otras variables X.

Cuando sólo hay dos variables explicatorias r_j^2 es el coeficiente de determinación entre X_1 y X_2. Si hubiera tres variables explicatorias, entonces r_j^2 sería el coeficiente de determinación múltiple de X_1 con X_2 y X_3.

Cuando un grupo de variables explicatorias no están correlacionadas, entonces **VIF_j** será **igual a 1**. Si el grupo presenta una alta intercorrelación, entonces **VIF_j** podría **exceder a 10** aunque algunos analistas o investigadores sugieren un criterio más conservador donde se emplearían alternativas a la regresión de mínimos cuadrados si el **VIF_j máximo excediera a 5.**

Puesto que solo hay dos variables explicatorias en el modelo, se puede calcular el **VIF_j** de la siguiente manera

Factor de varianza inflacionaria (VIF).

$$VIF_1 = VIF_2 = \frac{1}{1 - r_{X_1X_2}^2} = 1.084$$

ISBN 978-607-8285-62-4

Salida del programa Minitab 15.

Salida de la ventana Sesión de Minitab

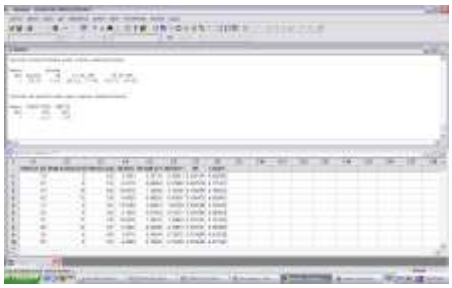
Predictor	Coef	de EE	T	P	VIF
Constante	223.52	57.74	3.87	0.006	
PUBLICIDAD (X1)	6.564	1.295	5.07	0.001	1.084
PRECIO (X2)	-1.7078	0.4474	-3.82	0.007	1.084

INTERPRETACIÓN: como el valor de **VIF** de **1.084 < 5** podemos llegar a la conclusión de que no hay razones para sospechar multicolinearidad alguna entre la variable $\mathbf{X_1}$ (Inversión en publicidad) y $\mathbf{X_2}$ (Precio del producto), por lo que no se debe eliminar ninguna de las dos variables.

Análisis de residuales.

Solución al inciso k.

Cálculo de residuales:



Interpretación: Los residuales normales aparecen en la columna **RESID1** y los residuales estandarizados aparecen en la columna **RESIDEST1.**

Diagnóstico de la regresión.

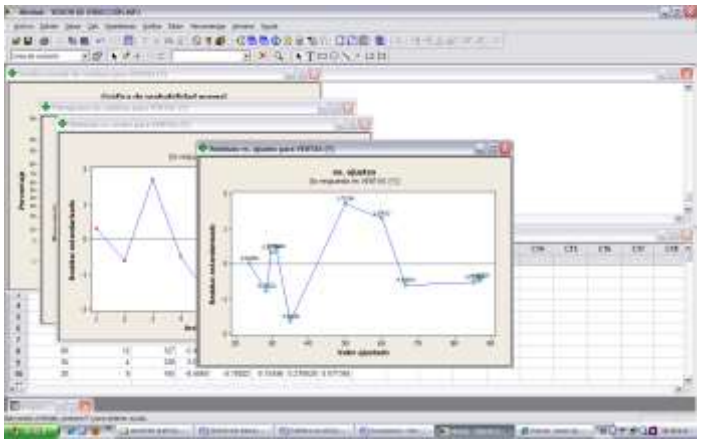
Solución al inciso l.

Se puede evaluar lo apropiado del modelo de regresión, trazando los "residuales estandarizados" sobre el eje vertical contra los valores $\hat{Y}$ en el eje horizontal. Si el modelo ajustado es apropiado para los datos no habrá un patrón aparente en esta gráfica de los residuales contra $\hat{Y}$. Sin embargo, si el modelo ajustado no es apropiado, habrá relación entre los valores $\hat{Y}$ y los residuales ε_i.

Supuestos de Linealidad y Homoscedasticidad;

La figura siguiente presenta el gráfico de valores $\hat{Y}$ en el eje horizontal contra los residuales, que sirve para detectar problemas de linealidad y homoscedasticidad en el modelo de ajuste. **Lo ideal es que todas las gráficas presenten una estructura aleatoria en sus puntos**.

ISBN 978-607-8285-62-4



Interpretación:

Linealidad.

Linealidad

Así, se puede observar que aunque haya una amplia dispersión en la gráfica residual, no hay patrón ó relación aparente entre los residuales estandarizados y $\hat{Y}$. Los residuales parecen estar distribuidos en forma pareja por encima y por debajo de 0 para diferentes valores de $\hat{Y}$. Por lo tanto se puede concluir que el modelo ajustado parece ser el apropiado.

Homoscedasticidad.

Homoscedasticidad

La suposición de homoscedasticidad se puede evaluar también de la gráfica de residuales estandarizados con $\hat{Y}$. Si parece haber un *"efecto de abanico"* en el cual aumenta ó disminuye la variabilidad de los residuales al aumentar $\hat{Y}$ se demuestra la falta de homogeneidad en las varianzas de Y_i a cada nivel de $\hat{Y}$. Para los datos del volumen de ventas no parece haber diferencias importantes en la variabilidad de SR_i para diferentes valores de $\hat{Y}$. Por lo tanto se puede concluir que para este modelo ajustado no hay violación aparente a la suposición de igual varianza en cada nivel de $\hat{Y}$.

Normalidad.

Normalidad:

El supuesto de normalidad en la regresión es posible evaluarlo de un análisis residual colocando los residuales estandarizados en una distribución de frecuencias y mostrando los resultados en un histograma.

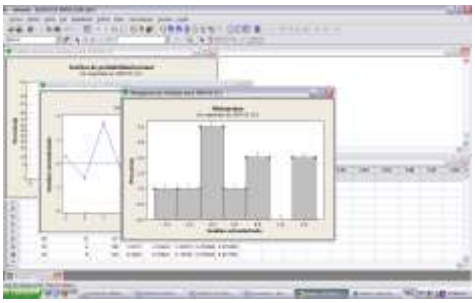
La figura siguiente presenta el Histograma de residuales estandarizados para detectar hipótesis de normalidad en el modelo.

ISBN 978-607-8285-62-4

El estadístico de Anderson-Darling mide en este caso si los residuales estandarizados siguen una distribución normal. Mientras mejor se ajuste la distribución a los residuales estandarizados, menor será este estadístico. Utilice el estadístico de Anderson-Darling para comparar el ajuste de varias distribuciones para ver cual es la mejor o probar si una muestra de datos proviene de una población con una distribución normal. Las hipótesis para la prueba de Anderson-Darling son:

H_0: Los residuales estandarizados siguen una distribución normal

H_1: Los residuales estandarizados no siguen una distribución normal

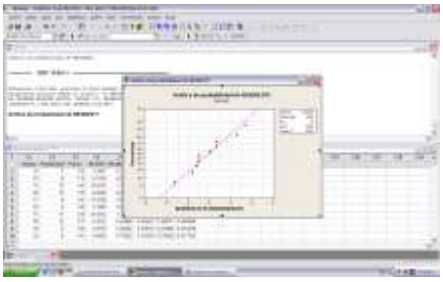


Interpretación:

Es difícil evaluar la suposición de normalidad para una muestra de tan sólo 10 observaciones y los procedimientos de pruebas disponibles quedan fuera del alcance del presente trabajo, sin embargo se puede observar que los datos aunque no parecen tener una "forma de campana" exacta, la mayor parte de los residuales están ubicados cerca del centro de la distribución por lo que parece razonable llegar a la conclusión de que no hay en modo alguno violación a la suposición de normalidad. El histograma indica que los datos podrían tener valores atípicos, lo cual se muestra mediante una barra, en el extremo derecho de la gráfica.

Otra forma de detectar hipótesis de normalidad es mediante la gráfica de probabilidad normal. La gráfica ideal es la diagonal del primer cuadrante.

La figura siguiente presenta el gráfico para detectar hipótesis de normalidad en el modelo.



Si el valor *p* (al estar disponible) para la prueba de Anderson-Darling es inferior al nivel de significación seleccionado (generalmente 0.05 ó 0.10), concluya que los datos no siguen la distribución normal. Minitab no siempre muestra un valor *p* para la prueba de Anderson-Darling, porque ésta no existe matemáticamente para ciertos casos.

Interpretación:

La gráfica de probabilidad normal muestra un patrón aproximadamente lineal que concuerda con una distribución normal. Los dos últimos puntos de la esquina superior derecha de la gráfica pueden ser valores atípicos.

El destacado de la gráfica identifica estos puntos como 7 y 3, puntos que deberán verificarse como observaciones inusuales ó identificación de valores atípicos.

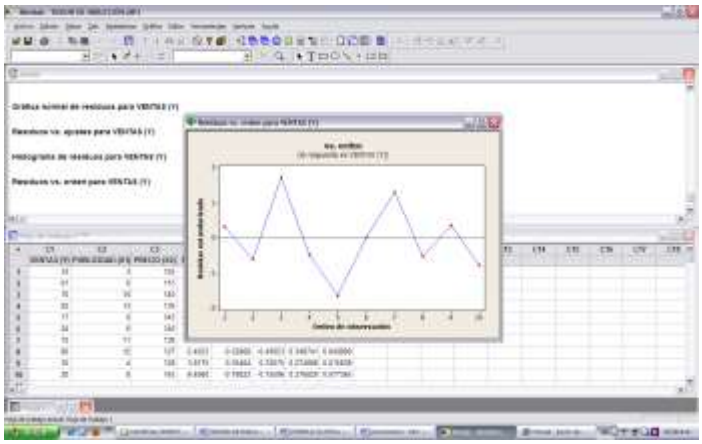
ISBN 978-607-8285-62-4

A la derecha de la gráfica se presenta la prueba de Anderson-Darling que arroja un estadístico de prueba de 0.283 con un nivel p de 0.553 que al se mayor a 0.05 no hace no rechazar la hipótesis nula concluyendo que la distribución de los residuales estandarizados es normal

Independencia.

Independencia

La suposición de independencia requiere que el error (diferencia "residual" entre un valor observado y uno predicho de Y) sea independiente para cada valor de $\hat{Y}$. Con frecuencia esta suposición se refiere a datos que se recopilan a lo largo de un periodo. Estos tipos de modelos caen bajo la denominación general de series de tiempo. La suposición de independencia se puede evaluar trazando los residuales en el orden o la sucesión en que se obtuvieron los datos observados. **Lo ideal es que la gráfica presente una estructura aleatoria en sus puntos**



Interpretación:

La gráfica de residuos versus orden no muestra un efecto de "autocorrelación" entre observaciones sucesivas, es decir no hay correlación entre una observación en particular y aquellos valores que la precedieron y la siguieron no afectando la suposición de independencia. Además el estadístico de Durbin-Watson es ***d*= 2.41102.** Este valor es mayor a 1.5 por lo que se puede pensar en que la autocorrelación no sea un problema.

ISBN 978-607-8285-62-4

Elementos de la matriz sombrero h_i.

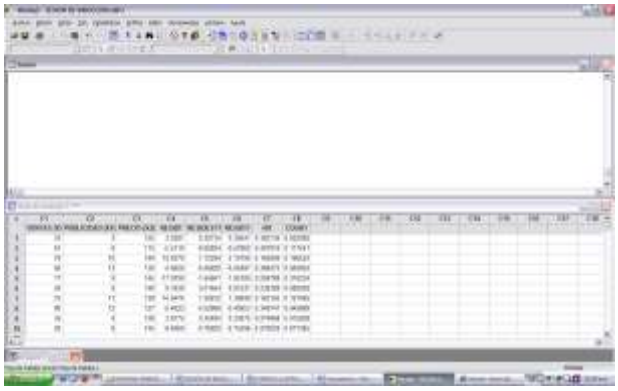
Solución al inciso m.

Las técnicas del análisis de influencias se utilizan para determinar si cualquier observación individual tiene una influencia indebida sobre el modelo ajustado.

Cada h_i refleja la "influencia" de cada X_i sobre el modelo de regresión ajustado. Si existen esos puntos de influencia quizá sea necesario evaluar de nuevo la necesidad de mantenerlos en el modelo. En la regresión lineal múltiple Hoaglin y Welsch sugieren la siguiente regla de decisión: Si $h_i > 2(k+1)/n$ entonces X_i es un punto de influencia y se puede considerar candidato a ser eliminado del modelo.

Cuando se desarrollo una estimación por intervalo de confianza $\mu_{Y.X}$, se definieron los "elementos diagonales de la matriz sombrero" h_i como:

$$h_i = X_i^{'}(X^{'}X)^{-1}X_i$$



Interpretación: (Ver la columna HI1). Para los datos del volumen de ventas, puesto que n=10, los criterios deben ser "destacar" cualquier valor $\boldsymbol{h_i}$ superior a ***2(k + 1)/n =0.6***. Consultando la tabla anterior se puede observar que ninguna observación es candidata potencial para ser removida del modelo del volumen de ventas.

Residuales de Student eliminados, t_i^*.

Solución al inciso n.

Los residuales de Student eliminados t^*_i .

En el estudio del análisis residual se definieron los residuales estandarizados mediante la ecuación:

$$SR_i = \frac{\varepsilon_i}{S_{Y.X}\sqrt{1-h_i}}$$

ISBN 978-607-8285-62-4

Para poder medir mejor la repercusión adversa sobre el modelo de cada caso individual, Hoaglin y Welsch desarrollaron el residual de Student eliminado t^*_i que se presenta en la siguiente ecuación:

$$t_i^* = \frac{\varepsilon_i}{S_{(i)}\sqrt{1-h_i}}$$

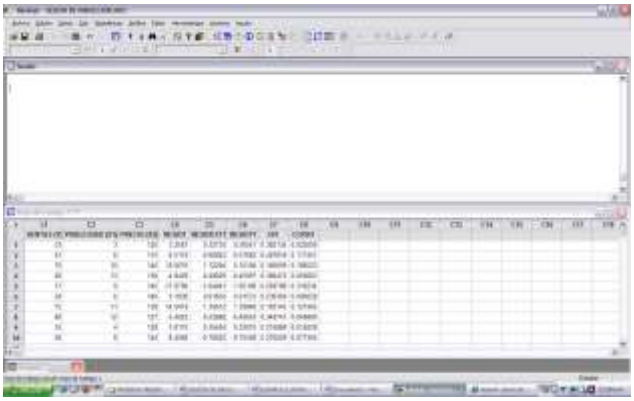
Donde $S_{(i)}$ = el error estándar de la estimación para un modelo que incluye todas las observaciones excepto la observación *i*.

Este residual de Student eliminado mide la diferencia entre cada valor observado Y_i y el valor predicho obtenidos de un modelo que incluye todas las demás observaciones excepto *i*. En el modelo de regresión múltiple Hoaglin y Welsch proponen que si,

$$|\ t^*_i\ | > t_{10.\ n-k-2}$$

entonces los valores observados y predichos son tan diferentes, que la observación *i* es un punto influyente que afecta de modo adverso al modelo y puede ser eliminada.

Salida de resultados de Minitab 15.



Interpretación: (Ver la columna RESIDT1). Para los datos del volumen de ventas, puesto que n=10, los criterios deben ser "destacar" cualquier valor superior **a** $|\boldsymbol{t_i^*}| > \boldsymbol{t_{0.10.6}} = \mathbf{1.4398}$. Consultando la tabla anterior se puede visualizar que $\boldsymbol{t_3^*} = \mathbf{2.102}$ y $\boldsymbol{t_5^*} = \mathbf{-1.951}$ Por lo tanto la tercera y la quinta observación pueden tener un efecto adverso sobre el modelo y se pueden considerar candidatos a ser retirados del modelo, sin embargo como de acuerdo al criterio h_i la tienda 3 y 5 no presentaron un efecto adverso, se debe tomar en cuenta otro criterio antes de tomar esa decisión como el criterio D_i de Cook, que se basa tanto en h_i como en el estadístico residual estandarizado t_i^*.

ISBN 978-607-8285-62-4

Estadístico de distancia de Cook, D_i.

Solución al inciso o.

El estadístico de distancia D_i de Cook.

El uso de h_i y t^*_i en la búsqueda de puntos de datos potencialmente problemáticos es complementario ya que ninguno de los criterios es suficiente por sí mismo.

Para decidir si un punto que ha sido destacado mediante el criterio h_i ò t^*_i está afectando indebidamente al modelo, Cook y Weisberg sugieren el uso del estadístico D_i en la ecuación:

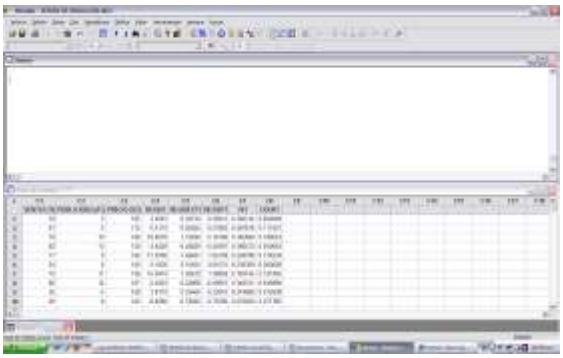
$$D_i = \frac{1}{(k+1)} SR_i^2 \frac{h_i}{(1-h_i)} = \frac{SR_i^2 h_i}{(k+1)(1-h_i)}$$

Donde SR_i es el residual estandarizado.

En el modelo de regresión múltiple Cook y Weisberg sugieren que

$$\text{Si } D_i > F_{.50.\ k+1.\ n-k-1}$$

La observación puede tener una recuperación sobre los resultados de ajustar un modelo de regresión múltiple.



Interpretación: (Ver la columna COOK1). Para el modelo del volumen de ventas (en millones de pesos), puesto que n=10, el criterio sería "destacar" cualquier $\boldsymbol{D_i > F_{0.50.3.7} = 0.871}$. Consultando la tabla anterior se puede observar que ninguna observación es candidata potencial para ser removida del modelo del volumen de ventas. En caso de que alguna observación una vez estudiados los tres criterios fuera necesario eliminar alguna(s) observación(es) se debería estudiar un modelo alterno en el que se hayan eliminado dichas observaciones que no fue el caso en este modelo.

ISBN 978-607-8285-62-4

EJERCICIOS COMPLEMENTARIOS

1 EJERCICIO COMPLEMENTARIO

EJERCICIO COMPLEMENTARIO 1

El departamento de auditoria lleva un registro del número de horas que sus computadoras tardan en detectar los impuestos no pagados. ¿Podríamos combinar esta información con los datos de las horas de trabajo en auditoria de campo y llegar a obtener una ecuación de estimación más precisa para los impuestos no pagados descubiertos al mes?. La tabla siguiente presenta estos datos para los últimos 10 meses

Mes	Impuestos reales Descubiertos no pagados (en millones de pesos) Y	Horas de trabajo en auditoría de campo (en cientos) X_1	Horas de computadora (en cientos) X_2
1	290	44	16
2	240	42	14
3	270	44	15
4	250	45	13
5	260	41	12
6	280	46	14
7	300	44	16
8	280	45	16
9	280	44	15
10	270	43	15

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$:
c) Calcule los impuestos reales $\hat{Y}$ cuando las horas de trabajo X_{1i} son de 4,000 y las horas de computadora X_{2i} de 1,000.
d) Determine el error estándar del estimador para toda la regresión lineal múltiple

ISBN 978-607-8285-62-4

e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)
f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.
g) Construya un intervalo de confianza para los verdaderos impuestos reales cuando las horas de trabajo en auditoria son de 4,000 y las horas de computadora de 1,000.
h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.
i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados
j) Determine los coeficientes de determinación parcial e intérprete sus resultados
k) Verifique la existencia de multicolinearidad
l) Determine los residuales estandarizados para toda la regresión
m) Construya la(s) grafica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.
n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

2 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 2**

El gerente de personal de una empresa intuye que quizá haya relación entre el ausentismo y la edad y el sueldo de un trabajador para desarrollar un modelo de predicción de días de ausencia durante un año laboral. Se seleccionó una muestra aleatoria de 10 trabajadores con los resultados que se presentan a continuación:

Trabajador	Días de ausentismo Y	Edad (en años) X_1	Sueldo anual (miles de pesos) X_2
1	15	27	185
2	6	61	264
3	10	37	243
4	18	23	180
5	9	46	247
6	7	58	240
7	14	29	212
8	11	36	253
9	5	64	269
10	8	40	272

ISBN 978-607-8285-62-4

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$:
c) Calcule los días de ausentismo $\hat{Y}$ cuando la edad X_{1i} es de 30 años y el sueldo anual X_{2i} de $ 200,000 pesos.
d) Determine el error estándar del estimador para toda la regresión lineal múltiple
e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)
f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.
g) Construya un intervalo de confianza para los verdaderos días de ausentismo cuando la edad del trabajador es de 30 años y se fija un sueldo anual de $ 200,000 pesos.
h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.
i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados
j) Determine los coeficientes de determinación parcial e intérprete sus resultados
k) Verifique la existencia de multicolinearidad
l) Determine los residuales estandarizados para toda la regresión
m) Construya la(s) grafica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.
n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

3 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 3**

El departamento de personal de una importante corporación desea desarrollar un modelo para predecir el sueldo semanal de sus empleados administrativos con base en la duración del empleo y su edad. Se seleccionó una muestra aleatoria de 10 empleados administrativos con los resultados siguientes:

Empleado	Sueldo semanal (en miles de pesos) Y	Duración del empleo (meses) X_1	Edad (años) X_2
1	7.46	569	65
2	6.12	343	59
3	5.29	256	61

ISBN 978-607-8285-62-4

4	6.02	215	41
5	5.29	129	37
6	5.92	327	56
7	5.18	113	47
8	4.06	42	28
9	6.74	337	51
10	6.70	375	57

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1\ y\ \hat{\beta}_2$:
c) Calcule el sueldo semana $\hat{Y}$ cuando la duración del empleo X_{1i} es de 300 meses y la edad del empleado X_{2i} es de 50 años..
d) Determine el error estándar del estimador para toda la regresión lineal múltiple
e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)
f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.
g) Construya un intervalo de confianza para el verdadero sueldo semanal cuando la duración del empleo es de 300 meses y se fija la edad en 50 años.
h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.
i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados
j) Determine los coeficientes de determinación parcial e intérprete sus resultados
k) Verifique la existencia de multicolinearidad
l) Determine los residuales estandarizados para toda la regresión
m) Construya la(s) grafica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.
n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

4 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 4**

Un analista financiero querría desarrollar un modelo de regresión para predecir el precio de venta actual (en cientos de pesos) de acciones en la industria de empresas editoriales, sobre la base del valor en libros de la compañía (en pesos por acción) y el rendimiento sobre el capital común (e porcentaje). Se seleccionó una muestra aleatoria de 10 compañías con los resultados siguientes:

ISBN 978-607-8285-62-4

Compañía	Precio de venta actual (en cientos de pesos) Y	Valor en libros (en $ por acción) X_1	Rendimiento sobre el capital común (%) X_2
1	6.2	136.0	20.2
2	3.9	96.1	19.8
3	3.2	101.6	19.5
4	4.4	158.3	17.2
5	6.4	105.6	28.1
6	4.7	241.0	12.7
7	2.5	69.6	11.9
8	2.9	187.1	13.3
9	2.7	61.2	23.3
10	5.9	178.8	20.0

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$:
c) Calcule el precio de venta actual $\hat{Y}$ cuando el valor en libros X_{1i} es de $ 90 pesos y el rendimiento sobre el capital común X_{2i} es de 18%.
d) Determine el error estándar del estimador para toda la regresión lineal múltiple
e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)
f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.
g) Construya un intervalo de confianza para el precio de venta actual de acciones cuando el valor en libros es de $ 90.00 y se fija un rendimiento sobre el capital común del 18%.
h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.
i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados
j) Determine los coeficientes de determinación parcial e intérprete sus resultados
k) Verifique la existencia de multicolinearidad
l) Determine los residuales estandarizados para toda la regresión
m) Construya la(s) grafica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.
n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

5 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 5**

Cierta franquicia se localiza en toda el área metropolitana de cierta Ciudad. Los propietarios, quieren expandirse a otras ciudades. Como parte de su presentación para un banco local, desean comprender mejor los factores que hacen que una tienda de descuento sea productiva. Seleccionaron una muestra aleatoria de 10 tiendas y registraron las ventas diarias promedio (Y), el espacio en el local, es decir el área (X_1) y el ingreso medio de las familias en la región donde está cada tienda. A continuación, presentamos la información de la muestra:

Tienda	Ventas diarias (en miles pesos) Y	Área de la tienda (en m^2) X_1	Ingreso (en miles de $) X_2
1	18.06	508	460
2	18.11	541	490
3	18.03	513	535
4	17.64	499	480
5	17.63	490	480
6	18.25	556	460
7	18.40	532	440
8	18.15	482	430
9	18.46	516	450
10	17.92	514	440

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$:
c) Calcule el número de ventas diarias $\hat{Y}$ cuando el área de la tienda X_{1i} es de 500 m^2 y el ingreso medio de las familias X_{2i} es de $ 495,000.00.
d) Determine el error estándar del estimador para toda la regresión lineal múltiple
e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)
f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.
g) Construya un intervalo de confianza para las verdaderas ventas cuando el área de la tienda es de 500 m^2 y el ingreso medio de las familias es de $ 495,000.00 pesos.
h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.
i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados
j) Determine los coeficientes de determinación parcial e intérprete sus resultados
k) Verifique la existencia de multicolinearidad
l) Determine los residuales estandarizados para toda la regresión

ISBN 978-607-8285-62-4

m) Construya la(s) gráfica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.
n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

6 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 6**

Steve Wilde fue contratado como instructor de administración en una importante empresa de corredores de bolsa. Como su primer proyecto, le pidieron que estudiara los factores que influyen en el ingreso bruto de las empresas de la industria química. Steve seleccionó una muestra aleatoria de 10 empresas y obtuvo la información sobre el número de empleados y el número de dividendos accionarios comunes consecutivos pagados. Sus resultados fueron los siguientes:

Compañia	Ganancias brutas (en millones de $) Y	Número de empleados X_1	Dividendos accionarios X_2
1	57.00	670	64
2	13.00	65	21
3	34.00	480	88
4	28.00	140	12
5	64.40	590	110
6	16.00	115	80
7	6.40	40	14
8	67.00	810	98
9	37.00	810	98
10	38.70	650	60

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$:
c) Calcule las ganancias brutas de una empresa $\hat{Y}$ cuando el número de empleados X_{1i} es de 220 y el número de dividendos accionarios comunes consecutivos pagados X_{2i} fue de 64.
d) Determine el error estándar del estimador para toda la regresión lineal múltiple
e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)

ISBN 978-607-8285-62-4

f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.
g) Construya un intervalo de confianza para las verdaderas ganancias brutas de una empresa de la industria química cuando se emplean a 220 personas y se ha pagado 64 dividendos accionarios comunes consecutivos.
h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.
i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados
j) Determine los coeficientes de determinación parcial e intérprete sus resultados
k) Verifique la existencia de multicolinearidad
l) Determine los residuales estandarizados para toda la regresión
m) Construya la(s) gráfica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.
n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

7 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 7**

El departamento de hipotecas de un banco importante estudia sus préstamos recientes. Quiere saber se que manera factores como los años de escolaridad y la edad del jefe de familia se relacionan con el ingreso familiar anual. Se obtuvo una muestra aleatoria de 10 prestamos recientes:

No. De préstamo	Ingreso (en miles de $) Y	Años de escolaridad X_1	Edad X_2
1	404	15	54
2	396	15	49
3	406	14	50
4	371	14	43
5	395	14	50
6	417	15	49
7	403	14	53
8	404	13	42
9	385	14	45
10	407	14	49

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$:

ISBN 978-607-8285-62-4

c) Calcule el ingreso familiar $\hat{Y}$ cuando los años de escolaridad X_{1i} son 14 y la edad del jefe de familia X_{2i} es de 48.
d) Determine el error estándar del estimador para toda la regresión lineal múltiple
e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)
f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.
g) Construya un intervalo de confianza para el verdadero ingreso familiar con 14 años de escolaridad y 48 años de edad.
h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.
i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados
j) Determine los coeficientes de determinación parcial e intérprete sus resultados
k) Verifique la existencia de multicolinearidad
l) Determine los residuales estandarizados para toda la regresión
m) Construya la(s) gráfica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.
n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

8 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 8**

El gerente de recursos humanos de una corporación debe presentar un análisis de los empleados asalariados como parte de su informe anual para el director ejecutivo. El gerente selecciona una muestra aleatoria de 10 empleados. Para cada empleado, registra el salario mensual; la antigüedad en la empresa y la edad. Los resultados son los siguientes:

Empleado de la muestra	Salario mensual (en miles de pesos) Y	Antigüedad en la empresa (en meses) X_1	Edad (en años) X_2
1	23.67	126	57
2	21.86	129	46
3	25.55	123	59
4	12.00	73	23
5	19.85	90	36

ISBN 978-607-8285-62-4

6	17.49	81	29
7	15.55	104	53
8	16.91	105	32
9	18.19	101	43
10	20.56	106	45

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$:
c) Calcule el salario mensual de un empleado $\hat{Y}$ cuando su antigüedad X_{1i} es de 95 meses y su edad X_{2i} es de 48 años.
d) Determine el error estándar del estimador para toda la regresión lineal múltiple
e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)
f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.
g) Construya un intervalo de confianza para el verdadero salario mensual de un empleado cuando su antigüedad es de 95 meses y su edad es de 48 años.
h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.
i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados
j) Determine los coeficientes de determinación parcial e intérprete sus resultados
k) Verifique la existencia de multicolinearidad
l) Determine los residuales estandarizados para toda la regresión
m) Construya la(s) gráfica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.
n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

9 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 9**

Muchas regiones de la costa del pacífico han experimentado un rápido crecimiento de la población en los últimos 10 años y se espera que el crecimiento continúe en los próximos 10 años. Esto ha tenido influencia en muchas de las cadenas de tiendas de abarrotes que construyen nuevas tiendas en la región. El director de planeación de una de ellas quiere investigar esto para construir más tiendas en una región determinada. Considera que existen dos factores principales que indican la cantidad de dinero que las familias gastan en tiendas de abarrotes. El primero es su ingreso y el segundo es el número de personas en la familia. El director reunió la siguiente información de una muestra de 10 familias:

ISBN 978-607-8285-62-4

Familia	Gasto en abarrotes mensualmente (en miles de pesos) Y	Ingreso familiar mensual (en miles de pesos) X_1	Tamaño de la familia X_2
1	4.0	115	3
2	3.4	46	2
3	3.8	32	3
4	4.2	62	4
5	2.7	105	1
6	4.0	40	5
7	4.2	57	4
8	5.5	68	9
9	2.7	54	1
10	5.0	31	5

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$:
c) Calcule el gasto en abarrotes mensualmente $\hat{Y}$ cuando el ingreso familiar mensual X_{1i} es de $ 90.00 miles de pesos y el tamaño de la familia X_{2i} es de 3 miembros.
d) Determine el error estándar del estimador para toda la regresión lineal múltiple
e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)
f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.
g) Construya un intervalo de confianza para el verdadero gasto en abarrotes cuando el ingreso familiar es de $ 90,000 pesos y el tamaño de la familia es de 3 miembros.
h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.
i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados
j) Determine los coeficientes de determinación parcial e intérprete sus resultados
k) Verifique la existencia de multicolinearidad
l) Determine los residuales estandarizados para toda la regresión
m) Construya la(s) gráfica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.
n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

10 EJERCICIO **COMPLEMENTARIO**

EJERCICIO **COMPLEMENTARIO 10**

Las 50 empresas de servicios al menudeo de cierta revista consisten en empresas al menudeo que están clasificadas del número 1 al 50 en ventas. Se desea estar en posibilidad de predecir la utilidad neta de una empresa con base en las ventas y los activos. Se seleccionó una muestra aleatoria de 10 empresas entre las 50. Los resultados se muestran en la tabla siguiente:

Compañía	Utilidad neta ($ millones) Y	Ventas (miles de millones) X_1	(Activos) Miles de millones) X_2
1	100	5.38	1.27
2	-21	1.90	1.65
3	27	1.90	0.75
4	111	2.03	1.26
5	4	4.10	2.16
6	42	1.53	0.71
7	582	24.25	10.60
8	40	2.11	0.85
9	1351	44.28	65.99
10	103	3.66	1.20

a) Ajuste una ecuación de regresión lineal múltiple para los datos anteriores.
b) Interprete los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1 \; y \; \hat{\beta}_2$:
c) Calcule la utilidad neta $\hat{Y}$ cuando las ventas X_{1i} son de 20,000 millones y los activos X_{2i} de 8,000 millones.
d) Determine el error estándar del estimador para toda la regresión lineal múltiple
e) Prueba la significancia de la relación entre la variable dependiente (y) y las variables explicatorias (independientes)
f) Realice una prueba de hipótesis para cada uno de los coeficientes de regresión y construya su intervalo de confianza respectivo. ¿Se puede eliminar alguna de las variables?.
g) Construya un intervalo de confianza para la verdadera utilidad neta para una empresa con ventas de $ 20 miles de millones y activos por $ 8 miles de millones e interprete los resultados.
h) Utilice el criterio de las "f" parciales para determinar la contribución de las variables explicatorias.
i) Determine el coeficiente de determinación y correlación múltiple para toda la regresión para averiguar con que eficacia los datos observados describen el modelo e intérprete los resultados
j) Determine los coeficientes de determinación parcial e intérprete sus resultados
k) Verifique la existencia de multicolinearidad

ISBN 978-607-8285-62-4

l) Determine los residuales estandarizados para toda la regresión
m) Construya la(s) gráfica(s) correspondiente(s) y determine lo adecuado del ajuste del modelo.
n) Calcule los elementos de la matriz sombrero h_i y determine si existen puntos de influencia.
o) Calcule los residuales de Student eliminados, t_i^* y determine si existen puntos de influencia.
p) Calcule el estadístico de distancia de Cook, D_i y determine si existen puntos de influencia. Si es necesario, analice el nuevo modelo de regresión después de eliminar estas observaciones y compare sus resultados con el modelo original.

ISBN 978-607-8285-62-4

AUTOEVALUACIÓN CON REACTIVOS DE **FALSO Ó VERDADERO**

EN CADA UNO DE LOS REACTIVOS, CONTESTE CON UNA F SI CONSIDERA QUE LA AFIRMACIÓN ES FALSA Y CON UNA V SI CONSIDERA QUE LA AFIRMACIÓN ES VERDADERA.

1. Para determinar si una regresión es significativa como un todo, se calcula y compara un valor calculado de F con un valor tomado de una tabla. **( )**

2. Si una regresión múltiple pudiera incluir a todos los factores explicativos relevantes de la variable dependiente, el coeficiente de correlación sería +1. **( )**

3. Suponga, en la ecuación de la regresión múltiple $\hat{Y}_i = 25 + 5X_1 - 2X_2$ que $\hat{Y}_i$ represente las ventas (en millones de pesos) y X_2, el precio de un producto (en pesos). Entonces para cada peso adicional, puede esperarse que el peso se incremente en 5 pesos. **( )**

4. Según Hoaglin y Welsch, si $|t_i^*| < t_{0.10.n-k-2}$,entonces X_i es un punto influyente y debe ser removido del modelo. **( )**

5. Se desea construir un intervalo de confianza para un valor de *Y* de una ecuación de regresión múltiple. Si hay 10 elementos en la muestra y 2 variables independientes son usadas en la regresión, debería emplear 8 grados de libertad cuando obtiene un valor de la tabla *t*. **( )**

6. Si un grupo de variables explicatorias no están correlacionadas, entonces el factor de varianza inflacionaria (VIF) será igual a 1. **( )**

7. Si una regresión múltiple pudiera incluir a todos los factores explicativos relevantes de la variable dependiente, el coeficiente de determinación sería 1. **( )**

8. Suponga que deseamos probar si los valores de *Y* en una regresión múltiple en realidad dependen de los valores de X_1. La hipótesis nula de nuestra prueba seria: $B_0 = 0$. **( )**

9. Un error estándar pequeño indica que los puntos están lejos al plano de regresión cuando se tienen dos variables independientes. **( )**

10. Según Cook y Weisberg, si $D_i > F_{0.50.k+1.n-k-1}$,entonces X_i es un punto influyente y debe ser removido del modelo. **( )**

11. La principal ventaja de la regresión múltiple sobre la regresión simple es que nos permite estimar la variable dependiente con mayor precisión. **( )**

12. Si se conoce la suma de cuadrados total y la suma de cuadrados de la regresión, la suma de cuadrados del error se puede obtener rápidamente. **( )**

13. Si una regresión múltiple pudiera incluir a todos los factores explicativos relevantes de la variable dependiente, el coeficiente de correlación sería cero. **( )**

14. Se puede calcular el estadístico de Durbin-Watson para detectar y medir la autocorrelación. **( )**

15. La gráfica de los residuales en función del tiempo sirve para investigar patrones en

ISBN 978-607-8285-62-4

los residuales cuando los datos se recolectan en la secuencia del tiempo. ( )

16. Agregando variables adicionales a una regresión múltiple logramos en general incrementar el coeficiente de determinación. ( )

17. Según Hoaglin y Welsch, si $h_i < 2(k+1)/n$,entonces X_i es un punto influyente y debe ser removido del modelo. ( )

18. La presencia de valores atípicos ó aberrantes podría hacer que se violara el supuesto de normalidad en un modelo de regresión. ( )

19. Si una regresión múltiple pudiera incluir a todos los factores explicativos relevantes de la variable dependiente, el error estándar del estimador sería cero.

20. El análisis de los residuales en el modelo de regresión se efectúa para determinar lo adecuado del ajuste del modelo. ( )

21. Suponga, en la ecuación de la regresión múltiple $\hat{Y}_i = 25 + 5X_1 - 2X_2$ que $\hat{Y}_i$ represente las ventas (en millones de pesos) y X_2, el precio de un producto (en pesos). Entonces para cada peso adicional, puede esperarse que las ventas se incrementen en 2 pesos. ( )

22. Aunque es posible hacer inferencias sobre los coeficientes estimados de regresión, no es posible hacerlas sobre la regresión como un todo. ( )

23. Según Hoaglin y Welsch, si $|t_i^*| > t_{0.05.n-k-2}$,entonces X_i es un punto influyente y debe ser removido del modelo. ( )

24. En la regresión lineal múltiple el coeficiente de determinación representa la proporción de la variación en Y que se explica por el grupo de variables explicatorias seleccionadas. ( )

25. Un error estándar pequeño indica que los puntos están cerca al plano de regresión cuando se tienen dos variables independientes. ( )

26. La correlación entre la variable dependiente y las variables independientes se le denomina multicolinearidad. ( )

27. La multicolinearidad es la correlación entre las variables independientes. ( )

28. Según Hoaglin y Welsch, si $h_i > 2(k+1)/n$,entonces X_i es un punto influyente y debe ser removido del modelo. ( )

29. Si un grupo de variables explicatorias están correlacionadas, entonces el factor de varianza inflacionaria (VIF) será igual a 1. ( )

30. Una prueba de significancia global investiga básicamente si es posible que todas las variables independientes tengan coeficientes de regresión neta iguales a cero. ( )

31. Un método para determinar la contribución de una variable explicatoria es conocido como criterio para prueba F parcial. ( )

32. Según Cook y Weisberg, si $D_i > F_{0.05.k+1.n-k-1}$,entonces X_i es un punto influyente y debe ser removido del modelo. ( )

33. La razón F calculada es un estadístico empleado para probar la significancia individual de cada coeficiente de regresión.

34. En la regresión lineal múltiple el coeficiente de determinación representa la proporción de la variación en Y que se explica por al menos una de las variables explicatorias seleccionadas. ( )

35. La razón F parcial calculada es un estadístico empleado para probar la significancia de una regresión como un todo. ( )

36. El coeficiente de determinación parcial mide la proporción de la variación en la variable dependiente que se explica por cada variable explicatoria, si se mantienen constantes las otras variables explicatorias. ( )

37. El estadístico t se emplea para probar la significancia individual de cada coeficiente de regresión.

ISBN 978-607-8285-62-4

38. El coeficiente de determinación múltiple mide la proporción de la variación en la variable dependiente que se explica por cada variable explicatoria, si se mantienen constantes las otras variables explicatorias. **( )**

39. Según Hoaglin y Welsch, si $|t_i^*| > t_{0.10.n-k-2}$,entonces X_i es un punto influyente y debe ser removido del modelo. **( )**

40. La razón F calculada es un estadístico empleado para probar la significancia de una regresión como un todo. **( )**

ISBN 978-607-8285-62-4

AUTOEVALUACIÓN CON REACTIVOS DE **OPCIÓN MÚLTIPLE**

EN CADA UNO DE LOS REACTIVOS SIGUIENTES, SELECCIONE LA OPCIÓN QUE CONSIDERE CORRECTA.

I. Las preguntas 01 a 18 se relacionan con un gerente de mercadotecnia de una corporación desea establecer el salario mensual de los empleados de su departamento (en miles de pesos). Ha utilizado el programa Minitab para desarrollar la regresión del salario de varios empleados sobre la antigüedad en la empresa (x_1) en meses y la edad (x_2) en años. A continuación se muestran sus resultados.

Análisis de regresión: Y vs. X1, X2

```
La ecuación de regresión es
Y = 2.56 + 0.126 X1 + 0.083 X2


Predictor      Coef  Coef. de EE     T      P    VIF
Constante     2.561        5.261  0.49  0.641
X1          0.12619      0.08574  1.47  0.185  3.364
X2           0.0828       0.1326  0.62  0.552  3.364


S = 2.63214   R-cuad. = 66.3%   R-cuad.(ajustado) = 56.7%


Análisis de varianza

Fuente          GL       SC      MC     F      P
Regresión        2   95.491  47.746  6.89  0.022
Error residual   7   48.497   6.928
Total            9  143.988


Fuente  GL  SC sec.
X1       1   92.786
X2       1    2.705

Fuente  GL  SC sec.
X2       1   80.487
X1       1   15.004
```


ISBN 978-607-8285-62-4

1. Si X_1 y X_2 tuvieran valores iguales a cero, entonces podría esperarse que Y tuviera un valor igual a:
a) 0.126
b) -2.56
c) 0.083
d) 2.56

2. La intensidad de la relación del salario con la antigüedad y la edad del empleado es:
a) 66.3 %
b) 26.3%
c) 56.7%
d) 81.4%

3. Se desea comprobar si la edad es una variable explicatoria significativa. Los grados de libertad que se aplicarían en la prueba serían:
a) 10
b) 9
c) 7
d) 2

4. ¿Cuál es el valor de $S_{\hat{\beta}_1}$?
a) 2.63214
b) 0.1326
c) 0.08574
d) 5.261

5. Se desea determinar si la regresión fue significativa como un todo. ¿Cuántos grados de libertad en el numerador se deberían tener en caso de utilizar una prueba F?
a) 1
b) 7
c) 2
d) 9

6. Se desea comprobar si la antigüedad es una variable explicatoria significativa. Los grados de libertad que se aplicarían en la prueba serían:
a) 9
b) 7
c) 2
d) 1

7. Para una edad de 48 años con una antigüedad de 95 meses el salario estimado sería de:
a) $ 18,514.00
b) $ 6,670.00
c) $16,493
d) $8,691

ISBN 978-607-8285-62-4

8. Se desea determinar si la regresión fue significativa como un todo. ¿Cuántos grados de libertad en el denominador se deberían tener en caso de utilizar una prueba F?

a) 1
b) 7
c) 2
d) 9

9. Se desea determinar la contribución de la antigüedad en el modelo. ¿Cuántos grados de libertad en el numerador se deberían tener en caso de utilizar una prueba F parcial?

a) 2
b) 9
c) 7
d) 1

10. ¿Cuál es el valor de $S_{\hat{\beta}_2}$?

a) 2.63214
b) 0.1326
c) 0.08574
d) 5.261

11. Se desea medir la proporción de la variación en la variable independiente que se explica por cada variable explicativa al mismo tiempo que se controlan o se mantienen constantes las otras variables explicativas. ¿Cuál es el valor del coeficiente de determinación parcial $\boldsymbol{r^2_{Y2.1}}$ para la variable independiente edad?

a) 4.26%
b) 23.63%
c) 29.30%
d) 5.28%

12. ¿Cuántas observaciones tomó el director?

a) 9
b) 8
c) 10
d) No puede determinarse con la información disponible

13. Se desea determinar la contribución de la edad en el modelo. ¿Cuántos grados de libertad en el numerador se deberían tener en caso de utilizar una prueba F parcial?

a) 2
b) 9
c) 7
d) 1

ISBN 978-607-8285-62-4

14. Se desea medir la proporción de la variación en la variable independiente que se explica por cada variable explicativa al mismo tiempo que se controlan o se mantienen constantes las otras variables explicativas. ¿Cuál es el valor del coeficiente de determinación parcial $r^2_{Y1.2}$ para la variable independiente antigüedad?

a) 23.63%
b) 4.26%
c) 29.30%
d) 5.28%

15. Se desea determinar la contribución de la edad en el modelo utilizando una prueba F parcial. ¿Cuál es el valor de la $SCR(X_2 | X_1)$

a) 95.491
b) 15.004
c) 48.497
d) 2.705

16. Se desea determinar la contribución de la antigüedad en el modelo utilizando una prueba F parcial. ¿Cuál es el valor de la $SCR(X_1 | X_2)$

a) 92.786
b) 15.004
c) 48.497
d) 2.705

17. Se desea determinar la contribución de la edad en el modelo utilizando una prueba F parcial. ¿Cuál es el valor de la $SC(X_1)$?

a) 92.786
b) 15.004
c) 80.487
d) 2.705

18. Se desea determinar la contribución de la antigüedad en el modelo utilizando una prueba F parcial. ¿Cuál es el valor de la $SC(X_2)$?

a) 92.786
b) 15.004
c) 80.487
d) 2.705

ISBN 978-607-8285-62-4

II. Las preguntas 19 a 36 se relacionan los propietarios de cierta franquicia. Han utilizado el programa Minitab para desarrollar la regresión de las ventas diarias promedio, en miles de pesos) de varias tiendas de descuento registrando el espacio en el local (X_1) en m^2 y el ingreso promedio de las familias en la región (X_2), en miles de pesos. A continuación se muestran sus resultados.

Análisis de regresión: Y vs. X1, X2

```
La ecuación de regresión es
Y = 16.3 + 0.00678 X1 - 0.00367 X2


Predictor       Coef  Coef. de EE      T      P    VIF
Constante     16.284        2.083   7.82  0.000
X1          0.006779     0.003486   1.94  0.093  1.005
X2         -0.003668     0.002554  -1.44  0.194  1.005


S = 0.237600   R-cuad. = 43.9%   R-cuad.(ajustado) = 27.8%


Análisis de varianza

Fuente          GL       SC       MC     F      P
Regresión        2  0.30867  0.15434  2.73  0.133
Error residual   7  0.39518  0.05645
Total            9  0.70385


Fuente  GL  SC sec.
X1       1  0.19223
X2       1  0.11644

Fuente  GL  SC sec.
X2       1  0.09515
X1       1  0.21353
```

19. ¿Cuál es el valor de $S_{\widehat{\beta}_1}$?

a) 0.002554
b) 2.083
c) 0.006779
d) 0.003486

20. Se desea determinar la contribución del ingreso promedio de las familias en el modelo utilizando una prueba F parcial. ¿Cuál es el valor de la SC(X_1) ?

a) 0.30867
b) 0.39518
c) 0.19223
d) 0.21353

ISBN 978-607-8285-62-4

21. Se desea medir la proporción de la variación en la variable independiente que se explica por cada variable explicativa al mismo tiempo que se controlan o se mantienen constantes las otras variables explicativas. ¿Cuál es el valor del coeficiente de determinación parcial $r^2_{Y1.2}$ para la variable independiente espacio en el local?

a) 35.08%
b) 23.28%
c) 40.89%
d) 19.13%

22. Se desea determinar si la regresión fue significativa para la variable independiente espacio en el local (X_1) , entonces el valor de estadístico calculado es:

a) -1.44
b) 2.73
c) 1.94
d) 7.82

23. ¿Cuál es el valor de $S_{\hat{\beta}_2}$?

a) 0.006779
b) 0.003486
c) 0.002554
d) -0.003668

24. Se desea determinar si la regresión fue significativa para la variable independiente espacio en el local (X_1) . ¿Cuántos grados de libertad se deberían tener en caso de utilizar una prueba t?

a) 1
b) 7
c) 2
d) 9

25. Se desea determinar la contribución de la antigüedad en el modelo utilizando una prueba F parcial. ¿Cuál es el valor de la SC(X_2) ?

a) 0.11644
b) 0.19223
c) 0.39518
d) 0.09515

26. Se desea determinar si la regresión fue significativa como un todo. ¿Cuál es el valor del estadístico de prueba calculado usado en esta prueba?

a) 1.94
b) 7.82
c) 2.73
d) -1.44

ISBN 978-607-8285-62-4

27. Se desea determinar si existe autocorrelación entre las variables explicatorias, para ello utilizamos el factor de varianza inflacionaria cuyo valor es:

a) 1.940
b) 1.005
c) 7.820
d) -1.44

28. Se pudo determinar que la regresión como un todo no fue significativa dado que la prueba resultó (NS) no significativa aplicando el criterio de P-level cuyo valor es de:

a) 0.000
b) 0.133
c) 0.194
d) 0.093

29. Se desea medir la proporción de la variación en la variable independiente que se explica por cada variable explicativa al mismo tiempo que se controlan o se mantienen constantes las otras variables explicativas. ¿Cuál es el valor del coeficiente de determinación parcial $r^2_{Y2.1}$ para la variable independiente ingreso promedio de las familias?

a) 14.20%
b) 22.76%
c) 27.39%
d) 35.08%

30. Se desea determinar la contribución del espacio en el local en el modelo utilizando una prueba F parcial. ¿Cuál es el valor de la $SCR(X_1 | X_2)$

a) 0.21353
b) 0.30867
c) 0.11644
d) 0.19223

31. Se desea determinar la contribución del espacio en el local en el modelo. ¿Cuántos grados de libertad en el denominador se deberían tener en caso de utilizar una prueba F parcial?

a) 2
b) 1
c) 7
d) 9

32. La intensidad de la relación de las ventas con el espacio del local y el ingreso promedio de las familias es de:

a) 52.73%
b) 66.26%
c) 43.9%
d) 27.80%

ISBN 978-607-8285-62-4

33. Si se tienen dos variables independientes, puede pensarse en la variación respecto a un plano de regresión. En este caso dicha variación tiene un valor de:

a) 0.003486
b) 2.083
c) 0.002554
d) 0.2376

34. Se desea determinar la contribución del ingreso promedio de las familias en el modelo utilizando una prueba F parcial. ¿Cuál es el valor de la $SCR(X_2 | X_1)$

a) -0.003668
b) 0.09515
c) 0.11644
d) 0.39518

35. ¿Qué tamaño de muestra se obtuvo?

a) 8
b) 10
c) 9
d) 7

36. Se desea determinar la contribución del ingreso promedio de las familias en el modelo. ¿Cuántos grados de libertad en el numerador se deberían tener en caso de utilizar una prueba F parcial?

a) 2
b) 1
c) 7
d) 9

ISBN 978-607-8285-62-4

GLOSARIO DE REGRESIÓN LINEAL MÚLTIPLE. PARTE 1

ANÁLISIS DE CORRELACIÓN. Técnica de con que se determina el grado de relación lineal que hay entre variables.

ANÁLISIS DE RESIDUALES. Respecto a regresión, análisis de las diferencias entre Y y $\hat{Y}$ para valorar las premisas y proporciona guías sobre qué tan bien se ajusta la ecuación a los datos.

ANÁLISIS DE LA VARIANCIA PARA LA REGRESIÓN. Procedimiento con que se calcula la razón F; se emplea para probar la significancia de la regresión como un todo.

COEFICIENTE (r) DE CORRELACIÓN . Una medida de la relación lineal entre dos mediciones numéricas hechas en el mismo conjunto de sujetos. Oscila de -1 a +1, con el cero indicando la ausencia de relación**.** Raíz cuadrada del coeficiente de determinación. Su signo indica la dirección de la relación entre dos variables, directa o inversa.

COEFICIENTE (R^2) DE DETERMINACIÓN . Medida de la proporción de variación de *Y*, la variable dependiente, que se explica con la línea de regresión; esto es, por la relación de las *Y* con la variable independiente. Se interpreta como la cantidad de variación en una variable que puede definirse por el conocimiento de una segunda variable.

COEFICIENTES DE REGRESIÓN. La constante β_1 en la ecuación de regresión lineal simple, $Y = \beta_0 + \beta_1 X$. ,se interpreta como la pendiente de la línea de regresión y β_0 como la ordenada al origen..

ECUACIÓN DE ESTIMACIÓN. Fórmula matemática que relaciona la variable desconocida con las variables conocidas en el análisis de regresión.

ECUACIÓN DE REGRESIÓN MÚLTIPLE. Es una ecuación que define la relación lineal entre dos ó más variables.

ERROR ESTÁNDAR DE ESTIMACIÓN MÚLTIPLE. Medida de al confiabilidad de la ecuación de estimación, que indica la variabilidad de los puntos observados alrededor de un plano de regresión (en el caso de dos variables independientes); es decir, hasta qué punto los valores observados difieren de los predichos en el plano de regresión.

ERROR ESTÁNDAR DEL COEFICIENTE DE REGRESIÓN. Medida de la variabilidad de los coeficientes de regresión de la muestra alrededor del verdadero coeficiente de regresión de la población.

F CALCULADA. Estadístico que se usa como prueba de la significancia de una variable explicatoria individual.

ISBN 978-607-8285-62-4

GLOSARIO DE REGRESIÓN LINEAL MÚLTIPLE. PARTE 2

INTERSECCIÓN EN *Y*. Constante de cualquier recta, cuyo valor representa el valor de la variable *Y* cuando la variable *X* tiene un valor de 0.

MULTICOLINEALIDAD. Problema estadístico que en ocasiones se presenta en el análisis de regresión múltiple; en él se reduce la confiabilidad de los coeficientes de regresión, a causa de un alto nivel de correlación entre las variables independientes.

PENDIENTE. Constante de cualquier recta, cuyo valor representa en qué medida el cambio de cada unidad de la variable independiente modifica la variable dependiente.

PRINCIPIO DE MÍNIMOS CUADRADOS. Técnica empleada para obtener la ecuación de regresión, minimizando la suma de los cuadrados de las distancias verticales entre los valores verdaderos de ***Y*** y los valores pronosticados de ***Y.***

RAZÓN F CALCULADA. Estadístico que se usa para probar la significancia de la regresión como un todo.

REGRESIÓN. (de **Y** en **X**) proceso por el cual se determina una ecuación para predecir **Y** a partir de **X**. Proceso general de predecir una variable a partir de otra con medios estadísticos, usando datos anteriores.

REGRESIÓN MÚLTIPLE. Proceso estadístico por medio del cual se utilizan, algunas variables para predecir otra variable.

RELACIÓN DIRECTA. Relación entre dos variables en la cual, al aumentar un valor de la variable independiente, también aumenta el de la variable dependiente.

RELACIÓN INVERSA: Relación entre dos variables en la cual, al aumentar la variable independiente, disminuye la variable dependiente.

RELACIÓN LINEAL. Tipo particular de asociación entre dos variables, que puede ser descrita matemáticamente con una recta.

RESIDUAL. Diferencia entre el valor probable (predicción) y el valor real de la variable dependiente (resultado o respuesta) en regresión.

TÉCNICAS DE MODELADO. Métodos con que se decide que variables incluir en un modelo de regresión y las diferentes maneras de incluirlas.

TRANSFORMACIONES. Manipulaciones matemáticas para darle a una variable una forma diferente, de modo que se ajuste a las curvas y también a las líneas por regresión.

VARIABLE DEPENDIENTE. Aquella que estamos tratando de predecir en el análisis de regresión.

VARIABLE INDEPENDIENTE. La variable, o variables, conocidas en el análisis de regresión.

VARIABLE FICTICIA. Aquella que toma el valor 0 o 1, permitiéndonos incluir en un modelo de regresión factores cualitativos tales como sexo, estado civil y nivel de escolaridad.

ISBN 978-607-8285-62-4

αA

SIMBOLOGÍA

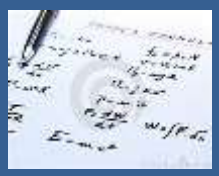
$=$	Igual	Σ	Letra griega mayúscula signa; símbolo que indica una suma
$\neq$	Desigual	$g.l.$	Grados de libertad
$<$	Menor que	ε	Letra griega épsilon; usada para simbolizar el error experimental
$\leq$	Menor que o igual a	F	Símbolo para la prueba y la distribución F
$>$	Mayor que	n	Tamaño de la muestra
$\geq$	Mayor que o igual que	r	Correlación de la muestra
H_o	Hipótesis nula	r^2	Correlación al cuadrado, llamado coeficiente de determinación
H_1	Hipótesis alterna	S	Desviación estándar de la muestra
α	Letra griega alfa; probabilidad de un error tipo I	SE	Error estándar de la muestra
β	Letra griega beta; probabilidad de un error tipo II	$S_{Y.X}$	Error estándar de la estimación en regresión
β_0	Valor poblacional de la ordenada al origen de la línea de regresión	t	Símbolo para la razón t (la razón crítica que sigue a una distribución t)
β_1	Valor poblacional de la pendiente de la línea de regresión.	X	Variable independiente (explicatoria, predictora) en regresión
$\hat{\beta}_0$	Valor estimado de la ordenada al origen de la línea de regresión	$\overline{X}$	Media de la muestra; X con barra
$\hat{\beta}_1$	Valor estimado de la pendiente de la línea de regresión	Y	Variable dependiente (resultado, respuesta, criterio) en regresión
μ	Letra griega mu; media de la población	$\hat{y}$	Valor probable (predicción) de Y en regresión
ρ	Letra griega rho, correlación de la población	SCT	Suma de Cuadrados Total
σ	Legra griega minúscula sigma; desviación estándar de población	SCR	Suma de Cuadrados de la Regresión
τ	Legra griega tau; usada para simbolizar términos en el modelo ANOVA	SCE	Suma de Cuadrados del Error
CMR	Cuadrado Medio de la Regresión	CME	Cuadrado Medio del Error

ISBN 978-607-8285-62-4

FÓRMULAS CLAVE. PARTE 1

• **Ecuaciones normales** $$\sum_{i=1}^{n} Y = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_2$$ $$\sum_{i=1}^{n} X_1 Y = \hat{\beta}_0 \sum_{i=1}^{n} X_1 + \hat{\beta}_1 \sum_{i=1}^{n} X_1^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_1 X_2$$ $$\sum_{i=1}^{n} X_2 Y = \hat{\beta}_0 \sum_{i=1}^{n} X_2 + \hat{\beta}_1 \sum_{i=1}^{n} X_1 X_2 + \hat{\beta}_2 \sum_{i=1}^{n} X_2^2$$	**(1)**	• **Vector de observaciones de *Y*, y sea la matriz *X* de tamaño *n* x (k+1)** $$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}$$	**(2)**
• **Vector columna *Y* de tamaño (*n* x 1)** $$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$	**(3)**	• **Coeficientes de regresión** $$\hat{\beta} = (X'X)^{-1} X'Y$$ $$X = A^{-1}K$$ $$(X'X)^{-1} = \frac{1}{\lvert X'X \rvert} \alpha (X'X)'$$	**(4)**
• **Varianza de la población** $$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$	**(5)**	• **Desviación estándar de la población** $$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$	**(6)**
• **Varianza muestral** $$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$	**(7)**	• **Desviación estándar muestral** $$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$	**(8)**
• **Error estándar del estimador** $$S_{Y.12} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-k-1}} = \sqrt{\frac{Y'Y - \hat{\beta}'(X'Y)}{n-k-1}} = \sqrt{CME}$$	**(9)**	• **Suma de cuadrados de la Regresión** $$SCR = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y + \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y - n\bar{Y}^2$$	**(10)**

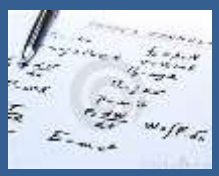
ISBN 978-607-8285-62-4



FÓRMULAS CLAVE. PARTE 2

• **Suma de cuadrados del Error** $$SCE = \sum_{i=1}^{n} Y^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y - \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y - \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y$$	**(11)**	• **Suma de Cuadrados Total** $$SCT = SCR + SCE = \sum_{i=1}^{n} Y^2 - n\bar{Y}^2$$	**(12)**
• **Suma de cuadrados de la Regresión** $$SCR = \hat{\beta}'(X'Y) - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$	**(13)**	• **Suma de cuadrados del Error** $$\boldsymbol{SCE = Y'Y - \widehat{\beta}'(X'Y)}$$	**(14)**
• **Suma de Cuadrados Total** $$SCT = Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$	**(15)**	• **Razón *F* calculada** $$\boldsymbol{F_{calculada} = \frac{CMR}{CME} = \frac{SCR/_{g.l.}}{SCE/_{g.l.}}}$$	**(16)**
• **Estadístico *t* para los coeficientes** $$t_{\alpha/2.n-k-1} = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}}$$	**(17)**	• **Estadístico *t* para el coeficiente $\widehat{\boldsymbol{\beta}}_1$** $$t_{\alpha/2.n-k-1} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$	**(18)**
• **Error estándar del coeficiente $\widehat{\boldsymbol{\beta}}_1$** $$\boldsymbol{S_{\widehat{\beta}_1}} = S_{Y.12}\sqrt{v_{11}}$$ $$(X'X)^{-1} = \begin{bmatrix} v_{00} & & \\ & v_{11} & \\ & & v_{22} \end{bmatrix}$$	**(19)**	• **Estadístico *t* para el coeficiente $\widehat{\boldsymbol{\beta}}_2$** $$t_{\alpha/2.n-k-1} = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}}$$	**(20)**
• **Error estándar del coeficiente $\widehat{\boldsymbol{\beta}}_2$** $$\boldsymbol{S_{\widehat{\beta}_2}} = S_{Y.12}\sqrt{v_{22}}$$ $$(X'X)^{-1} = \begin{bmatrix} v_{00} & & \\ & v_{11} & \\ & & v_{22} \end{bmatrix}$$	**(21)**	• **Intervalo de confianza para $\widehat{\boldsymbol{\beta}}_1$** $$\boldsymbol{\beta_1 = \widehat{\beta}_1 \pm t_{n-k-1} S_{\widehat{\beta}_1}}$$	**(22)**

ISBN 978-607-8285-62-4



FÓRMULAS CLAVE. PARTE 3

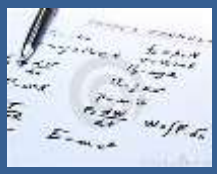
Intervalo de confianza para $\widehat{\boldsymbol{\beta}}_2$ $$\boldsymbol{\beta_2} = \widehat{\boldsymbol{\beta}}_2 \mp \boldsymbol{t_{n-k-1} S_{\widehat{\beta}_2}}$$	**(23)**	• **Intervalo de confianza** $$\mu_{Y.X} = Y = \hat{Y} \mp t_{\alpha/2,n-k-1} S_{Y.12..k}\sqrt{h_i}$$ donde: $$h_i = X_i'(X'X)^{-1}X_i$$	**(24)**
• ***F* parcial para X_1** $$F_{1.n-k-1} = \frac{SCR(X_1 \mid X_2)}{CME} = \frac{SCR(X_1 y X_2) - SCR(X_2)}{CME} = \frac{CMR(X_1 \mid X_2)}{CME}$$ $$SCR(X_2) = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y - n(\bar{Y})^2$$ $$SCR(X_1 \mid X_2) = \mathrm{SCR}(X_1 \text{ y } X_2) - \mathrm{SCR}(X_2)$$ **OPCIONAL :** $$\boxed{SCR(X_1 \mid X_2) = \frac{\hat{\beta}_1^2 \mathrm{CME}}{S^2_{\hat{\beta}_1}}}$$	**(25)**	• ***F* parcial para X_2** $$F_{1.n-k-1} = \frac{SCR(X_2 \mid X_1)}{CME} = \frac{SCR(X_1 y X_2) - SCR(X_1)}{CME} = \frac{CMR(X_2 \mid X_1)}{CME}$$ $$SCR(X_2) = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y - n(\bar{Y})^2$$ $$SCR(X_2 \mid X_1) = \mathrm{SCR}(X_1 \text{ y } X_2) - \mathrm{SCR}(X_1)$$ **OPCIONAL :** $$\boxed{SCR(X_2 \mid X_1) = \frac{\hat{\beta}_2^2 \mathrm{CME}}{S^2_{\hat{\beta}_2}}}$$	**(26)**
• **Coeficiente de correlación** $$\boldsymbol{r_{Y.12..k} = \sqrt{r^2_{Y.12...k}}}$$	**(27)**	• **Coeficiente de determinación parcial $r^2_{y1.2}$** $$r^2_{Y1.2} = \frac{SCR(X_1 \mid X_2)}{SCT - SCR(X_1 \text{ y } X_2) + SCR(X_1 \mid X_2)}$$	**(28)**

ISBN 978-607-8285-62-4

FÓRMULAS CLAVE. PARTE 4

• **Coeficiente de Determinación** $r^2_{Y.12} = \frac{SCR}{SCT}$ $SCR = \hat{\beta}_0 \sum_{i=1}^{n} Y + \hat{\beta}_1 \sum_{i=1}^{n} X_1 Y + \hat{\beta}_2 \sum_{i=1}^{n} X_2 Y - n\bar{Y}^2 = \hat{\beta}'(X'Y) - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$ $SCT = SCR + SCE = \sum_{i=1}^{n} Y^2 - n\bar{Y}^2 = Y'Y - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$	**(29)**	• **Multicolinearidad. Factor de varianza inflacionaria (VIF)** $\boldsymbol{VIF_1 = VIF_2 = \frac{1}{1 - r^2_{X_1X_2}}}$ $\boldsymbol{r^2_{X_1X_2} = \left[\frac{cov(X_1X_2)}{S_{X_1}S_{X_2}}\right]^2}$ $\boldsymbol{cov(X_1X_2) = \frac{\sum[(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)]}{n-1}}$ $\boldsymbol{S_{X_1} = \sqrt{\frac{\sum_{i=1}^{n} X_1^2 - n\bar{X}_1^2}{n-1}}}$ $\boldsymbol{S_{X_2} = \sqrt{\frac{\sum_{i=1}^{n} X_2^2 - n\bar{X}_2^2}{n-1}}}$	**(30)**
• **Coeficiente de determinación parcial $r^2_{y2.1}$** $r^2_{Y2.1} = \frac{SCR(X_2 \mid X_1)}{SCT - SCR(X_1 \, y \, X_2) + SCR(X_2 \mid X_1)}$	**(31)**	• **Coeficiente de Determinación ajustado** $r^2_{adj} = 1 - \left[(1 - r^2_{Y.12\ldots k})\frac{n-1}{n-k-1}\right]$	**(32)**
• **Error ó residual** $\boldsymbol{\varepsilon_i = Y_i - \hat{Y}_i}$	**(33)**	• **Residual estándarizado** $SR_i = \frac{\varepsilon_i}{S_{Y.12}\sqrt{1 - h_i}}$ $h_i = X_i'(X'X)^{-1}X_i$	**(34)**

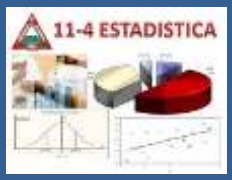
ISBN 978-607-8285-62-4



FÓRMULAS **CLAVE. PARTE 5**

• **Estadístico Durbin-Watson** $$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$	**(35)**	• **Elementos matriz sombrero** $$h_i = X_i'(X'X)^{-1}X_i$$	**(36)**
• **Residual de Student eliminado** $$SR_i = \frac{\varepsilon_i}{S_{Y.X}\sqrt{1-h_i}}$$ $$t_i^* = \frac{\varepsilon_i}{S_{(i)}\sqrt{1-h_i}}$$	**(37)**	• **Estadístico D_i** $$D_i = \frac{1}{(k+1)} SR_i^2 \frac{h_i}{(1-h_i)} = \frac{SR_i^2 h_i}{(k+1)(1-h_i)}$$	**(38)**

ISBN 978-607-8285-62-4



APÉNDICE TABLAS. SECCIÓN 1

TABLA A1 PUNTOS PORCENTUALES DEL RANGO STUDENTIZADO. PARTE 1

g.l.del error	t= número de niveles del tratamiento										
	α	2	3	4	5	6	7	8	9	10	11
2	.05	6.08	8.33	9.80	10.88	11.74	12.44	13.03	13.54	13.99	14.40
	.01	13.90	19.02	22.56	25.37	27.76	29.86	31.73	33.41	34.93	36.29
3	.05	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	9.72
	.01	8.26	10.62	12.17	13.32	14.24	15.00	15.65	16.21	16.71	17.16
4	.05	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.03
	.01	6.51	8.12	9.17	9.96	10.58	11.10	11.54	11.92	12.26	12.57
5	.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17
	.01	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48
6	.05	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65
	.01	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30
7	.05	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30
	.01	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55
8	.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05
	.01	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86	8.03
9	.05	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87
	.01	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49	7.65
10	.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72
	.01	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36
11	.05	3.11	3.82	4.26	4.57	4.82	5.03	5.30	5.35	5.49	5.61
	.01	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13
12	.05	3.08	3.77	4.20	4.52	4.75	4.95	5.12	5.27	5.39	5.51
	.01	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94
13	.05	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43
	.01	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79
14	.05	3,03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36
	.01	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66

ISBN 978-607-8285-62-4

TABLA A1	**PUNTOS PORCENTUALES DEL RANGO STUDENTIZADO. PARTE 2**

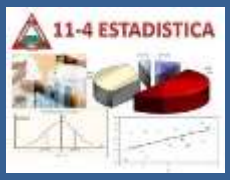
g.l.del error	t= número de niveles del tratamiento										
	α	2	3	4	5	6	7	8	9	10	11
15	.05	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31
	.01	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55
16	.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26
	.01	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46
17	.05	2.97	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21
	.01	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38
18	.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17
	.01	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31
19	.05	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14
	.01	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25
20	.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	4.01	5.11
	.01	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19
21	.05	2.94	3.56	4.21	4.42	4.60	4.74	4.87	4.98	5.08	5.17
	.01	4.00	4.61	4.99	5.26	5.47	5.65	5.79	5.92	6.04	6.14
22	.05	2.93	3.55	3.93	4.20	4.41	4.58	4.72	4.85	4.96	5.06
	.01	3.99	4.59	4.96	5.22	5.43	5.61	5.75	5.88	5.99	6.10
23	.05	2.93	3.54	3.91	4.18	4.39	4.56	4.70	4.63	4.94	5.03
	.01	3.97	4.57	4.93	5.20	5.40	5.57	5.72	5.84	5.95	6.05
24	.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01
	.01	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02
25	.05	2.91	3.52	3.89	4.15	4.36	4.53	4.67	4.79	4.90	4.99
	.01	3.94	4.53	4.89	5.14	5.35	5.51	5.65	5.78	5.89	5.98
26	.05	2.91	3.51	3.88	4.14	4.35	4.51	4.65	4.77	4.88	4.98
	.01	3.94	4.53	4.89	5.14	5.35	5.51	5.65	5.78	5.89	5.98
27	.05	2.90	3.51	3.87	4.13	4.33	4.50	4.64	4.76	4.86	4.96
	.01	3.92	4.49	4.85	5.10	5.30	5.46	5.60	5.72	5.83	5.92
28	.05	2.90	3.50	3.86	4.12	4.32	4.49	4.62	4.74	4.85	4.94
	.01	3.90	4.48	4.83	5.08	5.28	5.44	5.58	5.70	5.80	5.90
29	.05	2.89	3.49	3.85	4.11	4.31	4.47	4.61	4.73	4.84	4.93
	.01	3.90	4.47	4.81	5.06	5.26	5.42	5.56	5.67	5.78	5.87
30	.05	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92
	.01	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85
31	.05	2.88	3.48	3.84	4.09	4.29	4.45	4.59	4.71	4.81	4.90
	.01	3.88	4.44	4.79	5.03	5.23	5.38	5.52	5.63	5.74	5.83
32	.05	2.88	3.48	3.83	4.09	4.28	4.45	4.58	4.70	4.80	4.89
	.01	3.87	4.43	4.77	5.02	5.21	5.37	5.50	5.61	5.72	5.81
33	.05	2.88	3.47	3.83	4.08	4.28	4.44	4.57	4.69	4.79	4.88
	.01	3.87	4.42	4.76	5.00	5.20	5.35	5.48	5.60	5.70	5.79

ISBN 978-607-8285-62-4

TABLA A1	**PUNTOS PORCENTUALES DEL RANGO STUDENTIZADO. PARTE 3**

g.l.del error	t= número de niveles del tratamiento										
	α	2	3	4	5	6	7	8	9	10	11
34	.05	2.87	3.47	3.82	4.07	4.27	4.43	4.56	4.68	4.78	4.87
	.01	3.86	4.41	4.75	4.99	5.18	5.34	5.47	5.58	5.68	5.77
35	.05	2.87	3.46	3.81	4.07	4.26	4.42	4.56	4.67	4.77	4.86
	.01	3.85	4.40	4.74	4.98	5.17	5.32	5.45	5.57	5.67	5.75
36	.05	2.87	3.46	3.81	4.06	4.25	4.41	4.55	4.66	4.76	4.85
	.01	3.85	4.40	4.73	4.97	5.16	5.31	5.44	5.55	5.65	5.74
37	.05	2.87	3.45	3.80	4.05	4.25	4.41	4.54	4.66	4.76	4.85
	.01	3.84	4.39	4.72	4.96	5.15	5.30	5.43	5.54	5.64	5.72
38	.05	2.86	3.45	3.80	4.05	4.24	4.40	4.53	4.65	4.75	4.84
	.01	3.83	4.38	4.71	4.95	5.13	5.29	5.41	5.53	5.62	5.71
39	.05	2.86	3.45	3.79	4.04	4.24	4.39	4.53	4.64	4.74	4.83
	.01	3.83	4.37	4.70	4.94	5.12	5.28	5.40	5.51	5.61	5.70
40	.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82
	.01	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60	5.69
41	.05	2.86	3.44	3.79	4.03	4.23	4.38	4.51	4.63	4.73	4.82
	.01	3.82	4.36	4.69	4.92	5.11	5.26	5.38	5.49	5.59	5.67
42	.05	2.85	3.44	3.78	4.03	4.22	4.38	4.51	4.62	4.72	4.81
	.01	3.82	4.35	4.68	4.91	5.10	5.25	5.37	5.48	5.58	5.66
43	.05	2.85	3.43	3.78	4.03	4.22	4.37	4.50	4.62	4.72	4.80
	.01	3.81	4.35	4.67	4.91	5.09	5.24	5.36	5.47	5.57	5.65
44	.05	2.85	3.43	3.78	4.02	4.21	4.37	4.50	4.61	4.71	4.80
	.01	3.81	4.34	4.67	4.90	5.08	5.23	5.35	5.46	5.56	5.64
45	.05	2.85	3.43	3.77	4.02	4.21	4.36	4.49	4.61	4.70	4.79
	.01	3.80	4.34	4.66	4.89	5.07	5.22	5.34	5.45	5.55	5.63
50	.05	2.84	3.42	3.76	4.00	4.19	4.34	4.47	4.58	4.68	4.77
	.01	3.79	4.32	4.63	4.86	5.04	5.19	5.31	5.41	5.51	5.59
60	.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73
	.01	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45	5.53
120	.05	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64
	.01	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.37
∞	.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55
	.01	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23

ISBN 978-607-8285-62-4



APÉNDICE TABLAS. SECCIÓN 2

TABLA A2 VALORES CRÍTICOS DE t. PARTE 1 EXCEL

Grados de libertad	Áreas de la cola superior					
	0.25	0.10	0.05	0.025	0.01	0.005
1	1.00000	3.07768	6.31375	12.70620	31.82052	63.65674
2	0.81650	1.88562	2.91999	4.30265	6.96456	9.92484
3	0.76489	1.63774	2.35336	3.18245	4.54070	5.84091
4	0.74070	1.53321	2.13185	2.77645	3.74695	4.60409
5	0.72669	1.47588	2.01505	2.57058	3.36493	4.03214
6	0.71756	1.43976	1.94318	2.44691	3.14267	3.70743
7	0.71114	1.41492	1.89458	2.36462	2.99795	3.49948
8	0.70639	1.39682	1.85955	2.30600	2.89646	3.35539
9	0.70272	1.38303	1.83311	2.26216	2.82144	3.24984
10	0.69981	1.37218	1.81246	2.22814	2.76377	3.16927
11	0.69745	1.36343	1.79588	2.20099	2.71808	3.10581
12	0.69548	1.35622	1.78229	2.17881	2.68100	3.05454
13	0.69383	1.35017	1.77093	2.16037	2.65031	3.01228
14	0.69242	1.34503	1.76131	2.14479	2.62449	2.97684
15	0.69120	1.34061	1.75305	2.13145	2.60248	2.94671
16	0.69013	1.33676	1.74588	2.11991	2.58349	2.92078
17	0.68920	1.33338	1.73961	2.10982	2.56693	2.89823
18	0.68836	1.33039	1.73406	2.10092	2.55238	2.87844
19	0.68762	1.32773	1.72913	2.09302	2.53948	2.86093
20	0.68695	1.32534	1.72472	2.08596	2.52798	2.84534
21	0.68635	1.32319	1.72074	2.07961	2.51765	2.83136
22	0.68581	1.32124	1.71714	2.07387	2.50832	2.81876
23	0.68531	1.31946	1.71387	2.06866	2.49987	2.80734
24	0.68485	1.31784	1.71088	2.06390	2.49216	2.79694
25	0.68443	1.31635	1.70814	2.05954	2.48511	2.78744
26	0.68404	1.31497	1.70562	2.05553	2.47863	2.77871
27	0.68368	1.31370	1.70329	2.05183	2.47266	2.77068
28	0.68335	1.31253	1.70113	2.04841	2.46714	2.76326
29	0.68304	1.31143	1.69913	2.04523	2.46202	2.75639
30	0.68276	1.31042	1.69726	2.04227	2.45726	2.75000
31	0.68249	1.30946	1.69552	2.03951	2.45282	2.74404
32	0.68223	1.30857	1.69389	2.03693	2.44868	2.73848
33	0.68200	1.30774	1.69236	2.03452	2.44479	2.73328
34	0.68177	1.30695	1.69092	2.03224	2.44115	2.72839
35	0.68156	1.30621	1.68957	2.03011	2.43772	2.72381

Fuente: Valores generados con Microsoft Excel.

ISBN 978-607-8285-62-4

TABLA A2 VALORES CRÍTICOS DE t. PARTE 2 EXCEL

Grados de libertad	Áreas de la cola superior					
	0.25	0.10	0.05	0.025	0.01	0.005
36	0.68137	1.30551	1.68830	2.02809	2.43449	2.71948
37	0.68118	1.30485	1.68709	2.02619	2.43145	2.71541
38	0.68100	1.30423	1.68595	2.02439	2.42857	2.71156
39	0.68083	1.30364	1.68488	2.02269	2.42584	2.70791
40	0.68067	1.30308	1.68385	2.02108	2.42326	2.70446
41	0.68052	1.30254	1.68288	2.01954	2.42080	2.70118
42	0.68038	1.30204	1.68195	2.01808	2.41847	2.69807
43	0.68024	1.30155	1.68107	2.01669	2.41625	2.69510
44	0.68011	1.30109	1.68023	2.01537	2.41413	2.69228
45	0.67998	1.30065	1.67943	2.01410	2.41212	2.68959
46	0.67986	1.30023	1.67866	2.01290	2.41019	2.68701
47	0.67975	1.29982	1.67793	2.01174	2.40835	2.68456
48	0.67964	1.29944	1.67722	2.01063	2.40658	2.68220
49	0.67953	1.29907	1.67655	2.00958	2.40489	2.67995
50	0.67943	1.29871	1.67591	2.00856	2.40327	2.67779
51	0.67933	1.29837	1.67528	2.00758	2.40172	2.67572
52	0.67924	1.29805	1.67469	2.00665	2.40022	2.67373
53	0.67915	1.29773	1.67412	2.00575	2.39879	2.67182
54	0.67906	1.29743	1.67356	2.00488	2.39741	2.66998
55	0.67898	1.29713	1.67303	2.00404	2.39608	2.66822
56	0.67890	1.29685	1.67252	2.00324	2.39480	2.66651
57	0.67882	1.29658	1.67203	2.00247	2.39357	2.66487
58	0.67874	1.29632	1.67155	2.00172	2.39238	2.66329
59	0.67867	1.29607	1.67109	2.00100	2.39123	2.66176
60	0.67860	1.29582	1.67065	2.00030	2.39012	2.66028
61	0.67853	1.29558	1.67022	1.99962	2.38905	2.65886
62	0.67847	1.29536	1.66980	1.99897	2.38801	2.65748
63	0.67840	1.29513	1.66940	1.99834	2.38701	2.65615
64	0.67834	1.29492	1.66901	1.99773	2.38604	2.65485
65	0.67828	1.29471	1.66864	1.99714	2.38510	2.65360
66	0.67823	1.29451	1.66827	1.99656	2.38419	2.65239
67	0.67817	1.29432	1.66792	1.99601	2.38330	2.65122
68	0.67811	1.29413	1.66757	1.99547	2.38245	2.65008
69	0.67806	1.29394	1.66724	1.99495	2.38161	2.64898
70	0.67801	1.29376	1.66691	1.99444	2.38081	2.64790
71	0.67796	1.29359	1.66660	1.99394	2.38002	2.64686
72	0.67791	1.29342	1.66629	1.99346	2.37926	2.64585
73	0.67787	1.29326	1.66600	1.99300	2.37852	2.64487
74	0.67782	1.29310	1.66571	1.99254	2.37780	2.64391
75	0.67778	1.29294	1.66543	1.99210	2.37710	2.64298
76	0.67773	1.29279	1.66515	1.99167	2.37642	2.64208
77	0.67769	1.29264	1.66488	1.99125	2.37576	2.64120
78	0.67765	1.29250	1.66462	1.99085	2.37511	2.64034
79	0.67761	1.29236	1.66437	1.99045	2.37448	2.63950
80	0.67757	1.29222	1.66412	1.99006	2.37387	2.63869

Fuente: Valores generados con Microsoft Excel.

ISBN 978-607-8285-62-4

TABLA A2 VALORES CRÍTICOS DE t. PARTE 3 EXCEL

Grados de libertad	Áreas de la cola superior					
	0.25	0.10	0.05	0.025	0.01	0.005
81	0.67753	1.29209	1.66388	1.98969	2.37327	2.63790
82	0.67749	1.29196	1.66365	1.98932	2.37269	2.63712
83	0.67746	1.29183	1.66342	1.98896	2.37212	2.63637
84	0.67742	1.29171	1.66320	1.98861	2.37156	2.63563
85	0.67739	1.29159	1.66298	1.98827	2.37102	2.63491
86	0.67735	1.29147	1.66277	1.98793	2.37049	2.63421
87	0.67732	1.29136	1.66256	1.98761	2.36998	2.63353
88	0.67729	1.29125	1.66235	1.98729	2.36947	2.63286
89	0.67726	1.29114	1.66216	1.98698	2.36898	2.63220
90	0.67723	1.29103	1.66196	1.98667	2.36850	2.63157
91	0.67720	1.29092	1.66177	1.98638	2.36803	2.63094
92	0.67717	1.29082	1.66159	1.98609	2.36757	2.63033
93	0.67714	1.29072	1.66140	1.98580	2.36712	2.62973
94	0.67711	1.29062	1.66123	1.98552	2.36667	2.62915
95	0.67708	1.29053	1.66105	1.98525	2.36624	2.62858
96	0.67705	1.29043	1.66088	1.98498	2.36582	2.62802
97	0.67703	1.29034	1.66071	1.98472	2.36541	2.62747
98	0.67700	1.29025	1.66055	1.98447	2.36500	2.62693
99	0.67698	1.29016	1.66039	1.98422	2.36461	2.62641
100	0.67695	1.29007	1.66023	1.98397	2.36422	2.62589
101	0.67693	1.28999	1.66008	1.98373	2.36384	2.62539
102	0.67690	1.28991	1.65993	1.98350	2.36346	2.62489
103	0.67688	1.28982	1.65978	1.98326	2.36310	2.62441
104	0.67686	1.28974	1.65964	1.98304	2.36274	2.62393
105	0.67683	1.28967	1.65950	1.98282	2.36239	2.62347
106	0.67681	1.28959	1.65936	1.98260	2.36204	2.62301
107	0.67679	1.28951	1.65922	1.98238	2.36170	2.62256
108	0.67677	1.28944	1.65909	1.98217	2.36137	2.62212
109	0.67675	1.28937	1.65895	1.98197	2.36105	2.62169
110	0.67673	1.28930	1.65882	1.98177	2.36073	2.62126
115	0.67663	1.28896	1.65821	1.98081	2.35921	2.61926
120	0.67654	1.28865	1.65765	1.97993	2.35782	2.61742
125	0.67646	1.28836	1.65714	1.97912	2.35655	2.61573
130	0.67638	1.28810	1.65666	1.97838	2.35537	2.61418
135	0.67631	1.28785	1.65622	1.97769	2.35429	2.61274
140	0.67625	1.28763	1.65581	1.97705	2.35328	2.61140
145	0.67619	1.28742	1.65543	1.97646	2.35234	2.61016
150	0.67613	1.28722	1.65508	1.97591	2.35146	2.60900
170	0.67594	1.28655	1.65387	1.97402	2.34848	2.60506
200	0.67572	1.28580	1.65251	1.97190	2.34514	2.60063
∞	0.67449	1.28155	1.64486	1.95997	2.32635	2.57583

Fuente: Valores generados con Microsoft Excel.

ISBN 978-607-8285-62-4

TABLA A3 VALORES CRÍTICOS DE F $\alpha = 0.05$, PARTE 1 EXCEL

Denominador gl_2	Grados de libertad en el numerador gl_1									
	1	2	3	4	5	6	7	8	9	10
1	161.4476	199.5000	215.7073	224.5832	230.1618	233.9860	236.7684	238.8826	240.5432	241.8817
2	18.51282	19.00000	19.16429	19.24679	19.29641	19.32953	19.35322	19.37099	19.38483	19.39590
3	10.12796	9.55209	9.27663	9.11718	9.01346	8.94065	8.88674	8.84524	8.81230	8.78552
4	7.70865	6.94427	6.59138	6.38823	6.25606	6.16313	6.09421	6.04104	5.99878	5.96437
5	6.60789	5.78614	5.40945	5.19217	5.05033	4.95029	4.87587	4.81832	4.77247	4.73506
6	5.98738	5.14325	4.75706	4.53368	4.38737	4.28387	4.20666	4.14680	4.09902	4.05996
7	5.59145	4.73741	4.34683	4.12031	3.97152	3.86597	3.78704	3.72573	3.67667	3.63652
8	5.31766	4.45897	4.06618	3.83785	3.68750	3.58058	3.50046	3.43810	3.38813	3.34716
9	5.11736	4.25649	3.86255	3.63309	3.48166	3.37375	3.29275	3.22958	3.17889	3.13728
10	4.96460	4.10282	3.70826	3.47805	3.32583	3.21717	3.13546	3.07166	3.02038	2.97824
11	4.84434	3.98230	3.58743	3.35669	3.20387	3.09461	3.01233	2.94799	2.89622	2.85362
12	4.74723	3.88529	3.49029	3.25917	3.10588	2.99612	2.91336	2.84857	2.79638	2.75339
13	4.66719	3.80557	3.41053	3.17912	3.02544	2.91527	2.83210	2.76691	2.71436	2.67102
14	4.60011	3.73889	3.34389	3.11225	2.95825	2.84773	2.76420	2.69867	2.64579	2.60216
15	4.54308	3.68232	3.28738	3.05557	2.90129	2.79046	2.70663	2.64080	2.58763	2.54372
16	4.49400	3.63372	3.23887	3.00692	2.85241	2.74131	2.65720	2.59110	2.53767	2.49351
17	4.45132	3.59153	3.19678	2.96471	2.81000	2.69866	2.61430	2.54796	2.49429	2.44992
18	4.41387	3.55456	3.15991	2.92774	2.77285	2.66130	2.57672	2.51016	2.45628	2.41170
19	4.38075	3.52189	3.12735	2.89511	2.74006	2.62832	2.54353	2.47677	2.42270	2.37793
20	4.35124	3.49283	3.09839	2.86608	2.71089	2.59898	2.51401	2.44706	2.39281	2.34788
21	4.32479	3.46680	3.07247	2.84010	2.68478	2.57271	2.48758	2.42046	2.36605	2.32095
22	4.30095	3.44336	3.04913	2.81671	2.66127	2.54906	2.46377	2.39650	2.34194	2.29670
23	4.27934	3.42213	3.02800	2.79554	2.64000	2.52766	2.44223	2.37481	2.32011	2.27473
24	4.25968	3.40283	3.00879	2.77629	2.62065	2.50819	2.42263	2.35508	2.30024	2.25474
25	4.24170	3.38519	2.99124	2.75871	2.60299	2.49041	2.40473	2.33706	2.28210	2.23647
26	4.22520	3.36902	2.97515	2.74259	2.58679	2.47411	2.38831	2.32053	2.26545	2.21972
27	4.21001	3.35413	2.96035	2.72777	2.57189	2.45911	2.37321	2.30531	2.25013	2.20429
28	4.19597	3.34039	2.94669	2.71408	2.55813	2.44526	2.35926	2.29126	2.23598	2.19004
29	4.18296	3.32765	2.93403	2.70140	2.54539	2.43243	2.34634	2.27825	2.22287	2.17684
30	4.17088	3.31583	2.92228	2.68963	2.53355	2.42052	2.33434	2.26616	2.21070	2.16458
31	4.15962	3.30482	2.91133	2.67867	2.52254	2.40943	2.32317	2.25491	2.19936	2.15316
32	4.14910	3.29454	2.90112	2.66844	2.51225	2.39908	2.31274	2.24440	2.18877	2.14249
33	4.13925	3.28492	2.89156	2.65887	2.50264	2.38939	2.30298	2.23456	2.17886	2.13250
34	4.13002	3.27590	2.88260	2.64989	2.49362	2.38031	2.29383	2.22534	2.16956	2.12314
35	4.12134	3.26742	2.87419	2.64147	2.48514	2.37178	2.28524	2.21668	2.16083	2.11434
36	4.11317	3.25945	2.86627	2.63353	2.47717	2.36375	2.27714	2.20852	2.15261	2.10605
37	4.10546	3.25192	2.85880	2.62605	2.46965	2.35618	2.26951	2.20083	2.14485	2.09824
38	4.09817	3.24482	2.85174	2.61899	2.46255	2.34903	2.26230	2.19356	2.13753	2.09086
39	4.09128	3.23810	2.84507	2.61231	2.45583	2.34226	2.25549	2.18668	2.13060	2.08387
40	4.08475	3.23173	2.83875	2.60597	2.44947	2.33585	2.24902	2.18017	2.12403	2.07725
50	4.03431	3.18261	2.79001	2.55718	2.40041	2.28644	2.19920	2.12992	2.07335	2.02614
60	4.00119	3.15041	2.75808	2.52522	2.36827	2.25405	2.16654	2.09697	2.04010	1.99259
120	3.92012	3.07178	2.68017	2.44724	2.28985	2.17501	2.08677	2.01643	1.95876	1.91046
∞	3.84155	2.99582	2.60500	2.37202	2.21419	2.09869	2.00968	1.93851	1.87998	1.83080

Fuente: Valores generados con Microsoft Excel.

ISBN 978-607-8285-62-4

TABLA A3 VALORES CRÍTICOS DE F $\alpha = 0.05$, PARTE 2 EXCEL

Denominador gl_2	Grados de libertad en el numerador gl_1									
	11	12	13	14	15	16	17	18	19	20
1	242.9834	243.9060	244.6898	245.3639	245.9499	246.4639	246.9184	247.3232	247.6860	248.0130
2	19.40496	19.41251	19.41890	19.42438	19.42914	19.43329	19.43696	19.44022	19.44314	19.44577
3	8.76333	8.74464	8.72868	8.71490	8.70287	8.69229	8.68290	8.67452	8.66699	8.66019
4	5.93581	5.91173	5.89114	5.87335	5.85781	5.84412	5.83197	5.82112	5.81136	5.80254
5	4.70397	4.67770	4.65523	4.63577	4.61876	4.60376	4.59044	4.57853	4.56782	4.55813
6	4.02744	3.99994	3.97636	3.95593	3.93806	3.92228	3.90826	3.89571	3.88441	3.87419
7	3.60304	3.57468	3.55034	3.52923	3.51074	3.49441	3.47988	3.46686	3.45514	3.44452
8	3.31295	3.28394	3.25902	3.23738	3.21841	3.20163	3.18670	3.17332	3.16125	3.15032
9	3.10249	3.07295	3.04755	3.02547	3.00610	2.98897	2.97370	2.96000	2.94765	2.93646
10	2.94296	2.91298	2.88717	2.86473	2.84502	2.82757	2.81201	2.79805	2.78545	2.77402
11	2.81793	2.78757	2.76142	2.73865	2.71864	2.70091	2.68510	2.67090	2.65808	2.64645
12	2.71733	2.68664	2.66018	2.63712	2.61685	2.59888	2.58284	2.56843	2.55541	2.54359
13	2.63465	2.60366	2.57693	2.55362	2.53311	2.51492	2.49867	2.48407	2.47087	2.45888
14	2.56550	2.53424	2.50726	2.48373	2.46300	2.44461	2.42818	2.41340	2.40004	2.38790
15	2.50681	2.47531	2.44811	2.42436	2.40345	2.38488	2.36827	2.35333	2.33982	2.32754
16	2.45637	2.42466	2.39725	2.37332	2.35222	2.33348	2.31672	2.30164	2.28798	2.27557
17	2.41256	2.38065	2.35306	2.32895	2.30769	2.28880	2.27189	2.25667	2.24289	2.23035
18	2.37416	2.34207	2.31430	2.29003	2.26862	2.24959	2.23255	2.21720	2.20330	2.19065
19	2.34021	2.30795	2.28003	2.25561	2.23406	2.21490	2.19773	2.18226	2.16825	2.15550
20	2.30999	2.27758	2.24951	2.22496	2.20327	2.18398	2.16670	2.15112	2.13701	2.12416
21	2.28292	2.25036	2.22216	2.19747	2.17567	2.15626	2.13887	2.12319	2.10898	2.09603
22	2.25852	2.22583	2.19750	2.17269	2.15078	2.13126	2.11377	2.09799	2.08369	2.07066
23	2.23642	2.20361	2.17516	2.15024	2.12822	2.10860	2.09101	2.07515	2.06075	2.04764
24	2.21631	2.18338	2.15482	2.12980	2.10767	2.08796	2.07028	2.05433	2.03986	2.02666
25	2.19793	2.16489	2.13623	2.11111	2.08889	2.06909	2.05132	2.03529	2.02074	2.00747
26	2.18107	2.14793	2.11917	2.09395	2.07164	2.05176	2.03391	2.01780	2.00318	1.98984
27	2.16554	2.13230	2.10345	2.07815	2.05575	2.03579	2.01787	2.00169	1.98699	1.97359
28	2.15120	2.11787	2.08893	2.06354	2.04107	2.02103	2.00304	1.98678	1.97203	1.95856
29	2.13791	2.10449	2.07547	2.05000	2.02746	2.00735	1.98928	1.97297	1.95815	1.94462
30	2.12556	2.09206	2.06296	2.03742	2.01480	1.99462	1.97650	1.96012	1.94524	1.93165
31	2.11405	2.08048	2.05131	2.02569	2.00301	1.98276	1.96457	1.94813	1.93320	1.91956
32	2.10331	2.06966	2.04042	2.01474	1.99199	1.97168	1.95343	1.93694	1.92195	1.90826
33	2.09325	2.05954	2.03023	2.00448	1.98167	1.96130	1.94300	1.92645	1.91141	1.89767
34	2.08382	2.05004	2.02066	1.99486	1.97199	1.95157	1.93321	1.91660	1.90151	1.88773
35	2.07496	2.04111	2.01167	1.98581	1.96288	1.94241	1.92400	1.90735	1.89221	1.87838
36	2.06661	2.03270	2.00321	1.97729	1.95431	1.93378	1.91532	1.89862	1.88344	1.86956
37	2.05873	2.02477	1.99522	1.96925	1.94622	1.92564	1.90713	1.89039	1.87516	1.86124
38	2.05129	2.01728	1.98767	1.96165	1.93857	1.91794	1.89939	1.88260	1.86733	1.85338
39	2.04425	2.01018	1.98053	1.95445	1.93133	1.91066	1.89206	1.87523	1.85992	1.84593
40	2.03758	2.00346	1.97376	1.94764	1.92446	1.90375	1.88511	1.86824	1.85289	1.83886
50	1.98606	1.95153	1.92143	1.89493	1.87138	1.85031	1.83133	1.81413	1.79846	1.78412
60	1.95221	1.91740	1.88702	1.86024	1.83644	1.81511	1.79589	1.77845	1.76255	1.74798
120	1.86929	1.83370	1.80255	1.77503	1.75050	1.72846	1.70854	1.69043	1.67388	1.65868
∞	1.78874	1.75227	1.72025	1.69187	1.66649	1.64362	1.62287	1.60395	1.58661	1.57063

Fuente: Valores generados con Microsoft Excel.

ISBN 978-607-8285-62-4

TABLA A3 VALORES CRÍTICOS DE F $\alpha = 0.05$, PARTE 3 EXCEL

Denominador gl_2	Grados de libertad en el numerador gl_1									
	21	22	23	24	25	30	40	60	120	∞
1	248.3093	248.5790	248.8255	249.0517	249.2600	250.0951	251.1431	252.1957	253.2528	254.3143
2	19.44815	19.45031	19.45228	19.45409	19.45575	19.46241	19.47074	19.47906	19.48739	19.49572
3	8.65402	8.64839	8.64324	8.63850	8.63414	8.61658	8.59441	8.57200	8.54935	8.52645
4	5.79453	5.78723	5.78054	5.77439	5.76872	5.74588	5.71700	5.68774	5.65811	5.62808
5	4.54933	4.54129	4.53393	4.52715	4.52090	4.49571	4.46379	4.43138	4.39845	4.36500
6	3.86489	3.85640	3.84862	3.84146	3.83484	3.80816	3.77429	3.73980	3.70467	3.66887
7	3.43487	3.42604	3.41795	3.41049	3.40361	3.37581	3.34043	3.30432	3.26745	3.22976
8	3.14037	3.13128	3.12293	3.11524	3.10813	3.07941	3.04278	3.00530	2.96692	2.92758
9	2.92626	2.91693	2.90837	2.90047	2.89318	2.86365	2.82593	2.78725	2.74752	2.70668
10	2.76360	2.75407	2.74532	2.73725	2.72978	2.69955	2.66086	2.62108	2.58012	2.53788
11	2.63584	2.62613	2.61720	2.60897	2.60136	2.57049	2.53091	2.49012	2.44802	2.40448
12	2.53281	2.52293	2.51386	2.50548	2.49773	2.46628	2.42588	2.38417	2.34099	2.29620
13	2.44794	2.43792	2.42870	2.42020	2.41232	2.38033	2.33918	2.29660	2.25241	2.20644
14	2.37681	2.36665	2.35731	2.34868	2.34069	2.30821	2.26635	2.22295	2.17781	2.13070
15	2.31632	2.30603	2.29657	2.28783	2.27973	2.24679	2.20428	2.16011	2.11406	2.06585
16	2.26423	2.25383	2.24425	2.23541	2.22721	2.19384	2.15071	2.10581	2.05890	2.00964
17	2.21890	2.20839	2.19871	2.18977	2.18148	2.14771	2.10400	2.05841	2.01066	1.96039
18	2.17909	2.16847	2.15870	2.14966	2.14129	2.10714	2.06289	2.01664	1.96810	1.91685
19	2.14383	2.13313	2.12326	2.11414	2.10569	2.07119	2.02641	1.97954	1.93024	1.87803
20	2.11240	2.10160	2.09165	2.08245	2.07392	2.03909	1.99382	1.94636	1.89632	1.84319
21	2.08419	2.07331	2.06328	2.05400	2.04540	2.01025	1.96452	1.91649	1.86574	1.81171
22	2.05873	2.04777	2.03767	2.02832	2.01964	1.98420	1.93802	1.88945	1.83802	1.78311
23	2.03563	2.02460	2.01442	2.00501	1.99627	1.96054	1.91394	1.86484	1.81276	1.75700
24	2.01458	2.00348	1.99324	1.98376	1.97496	1.93896	1.89195	1.84236	1.78964	1.73306
25	1.99532	1.98415	1.97385	1.96431	1.95545	1.91919	1.87180	1.82173	1.76840	1.71100
26	1.97763	1.96639	1.95603	1.94643	1.93751	1.90101	1.85325	1.80272	1.74879	1.69061
27	1.96131	1.95002	1.93959	1.92994	1.92097	1.88424	1.83613	1.78515	1.73065	1.67169
28	1.94622	1.93487	1.92439	1.91469	1.90567	1.86871	1.82026	1.76886	1.71380	1.65408
29	1.93222	1.92082	1.91029	1.90053	1.89147	1.85429	1.80552	1.75370	1.69811	1.63765
30	1.91920	1.90775	1.89716	1.88736	1.87825	1.84087	1.79179	1.73957	1.68345	1.62227
31	1.90706	1.89555	1.88492	1.87507	1.86592	1.82834	1.77896	1.72636	1.66973	1.60784
32	1.89571	1.88415	1.87348	1.86358	1.85439	1.81662	1.76696	1.71398	1.65686	1.59428
33	1.88507	1.87347	1.86275	1.85281	1.84358	1.80564	1.75569	1.70236	1.64475	1.58149
34	1.87508	1.86344	1.85268	1.84270	1.83342	1.79531	1.74510	1.69142	1.63334	1.56941
35	1.86569	1.85400	1.84320	1.83318	1.82387	1.78559	1.73512	1.68111	1.62258	1.55799
36	1.85683	1.84510	1.83427	1.82421	1.81486	1.77642	1.72570	1.67136	1.61239	1.54716
37	1.84847	1.83671	1.82583	1.81574	1.80636	1.76776	1.71680	1.66215	1.60274	1.53688
38	1.84057	1.82876	1.81785	1.80773	1.79831	1.75957	1.70838	1.65342	1.59359	1.52710
39	1.83308	1.82124	1.81029	1.80014	1.79069	1.75180	1.70039	1.64513	1.58489	1.51779
40	1.82598	1.81410	1.80312	1.79294	1.78346	1.74443	1.69280	1.63725	1.57661	1.50891
50	1.77095	1.75879	1.74753	1.73708	1.72734	1.68716	1.63368	1.57565	1.51147	1.43827
60	1.73459	1.72222	1.71077	1.70012	1.69019	1.64914	1.59427	1.53431	1.46727	1.38929
120	1.64467	1.63170	1.61966	1.60844	1.59796	1.55434	1.49520	1.42901	1.35189	1.25387
∞	1.55585	1.54213	1.52935	1.51740	1.50621	1.45921	1.39409	1.31817	1.22157	1.00776

Fuente: Valores generados con Microsoft Excel.

ISBN 978-607-8285-62-4

TABLA A4 VALORES CRÍTICOS DE F $\alpha = 0.01$. PARTE 1

EXCEL

Denominador gl_2	Grados de libertad en el numerador gl_1									
	1	2	3	4	5	6	7	8	9	10
1	4052.180	4999.500	5403.352	5624.583	5763.649	5858.986	5928.355	5981.070	6022.473	6055.846
2	98.50251	99.00000	99.16620	99.24937	99.29930	99.33259	99.35637	99.37421	99.38809	99.39920
3	34.11622	30.81652	29.45670	28.70990	28.23708	27.91066	27.67170	27.48918	27.34521	27.22873
4	21.19769	18.00000	16.69437	15.97702	15.52186	15.20686	14.97576	14.79889	14.65913	14.54590
5	16.25818	13.27393	12.05995	11.39193	10.96702	10.67225	10.45551	10.28931	10.15776	10.05102
6	13.74502	10.92477	9.77954	9.14830	8.74590	8.46613	8.26000	8.10165	7.97612	7.87412
7	12.24638	9.54658	8.45129	7.84665	7.46044	7.19140	6.99283	6.84005	6.71875	6.62006
8	11.25862	8.64911	7.59099	7.00608	6.63183	6.37068	6.17762	6.02887	5.91062	5.81429
9	10.56143	8.02152	6.99192	6.42209	6.05694	5.80177	5.61287	5.46712	5.35113	5.25654
10	10.04429	7.55943	6.55231	5.99434	5.63633	5.38581	5.20012	5.05669	4.94242	4.84915
11	9.64603	7.20571	6.21673	5.66830	5.31601	5.06921	4.88607	4.74447	4.63154	4.53928
12	9.33021	6.92661	5.95254	5.41195	5.06434	4.82057	4.63950	4.49937	4.38751	4.29605
13	9.07381	6.70096	5.73938	5.20533	4.86162	4.62036	4.44100	4.30206	4.19108	4.10027
14	8.86159	6.51488	5.56389	5.03538	4.69496	4.45582	4.27788	4.13995	4.02968	3.93940
15	8.68312	6.35887	5.41696	4.89321	4.55561	4.31827	4.14155	4.00445	3.89479	3.80494
16	8.53097	6.22624	5.29221	4.77258	4.43742	4.20163	4.02595	3.88957	3.78042	3.69093
17	8.39974	6.11211	5.18500	4.66897	4.33594	4.10151	3.92672	3.79096	3.68224	3.59307
18	8.28542	6.01290	5.09189	4.57904	4.24788	4.01464	3.84064	3.70542	3.59707	3.50816
19	8.18495	5.92588	5.01029	4.50026	4.17077	3.93857	3.76527	3.63052	3.52250	3.43382
20	8.09596	5.84893	4.93819	4.43069	4.10268	3.87143	3.69874	3.56441	3.45668	3.36819
21	8.01660	5.78042	4.87405	4.36882	4.04214	3.81173	3.63959	3.50563	3.39815	3.30983
22	7.94539	5.71902	4.81661	4.31343	3.98796	3.75830	3.58666	3.45303	3.34577	3.25761
23	7.88113	5.66370	4.76488	4.26357	3.93919	3.71022	3.53902	3.40569	3.29863	3.21060
24	7.82287	5.61359	4.71805	4.21845	3.89507	3.66672	3.49593	3.36287	3.25599	3.16807
25	7.76980	5.56800	4.67546	4.17742	3.85496	3.62717	3.45675	3.32394	3.21722	3.12941
26	7.72125	5.52633	4.63657	4.13996	3.81834	3.59108	3.42099	3.28840	3.18182	3.09411
27	7.67668	5.48812	4.60091	4.10562	3.78477	3.55799	3.38822	3.25583	3.14939	3.06175
28	7.63562	5.45294	4.56809	4.07403	3.75389	3.52756	3.35807	3.22587	3.11955	3.03199
29	7.59766	5.42045	4.53779	4.04487	3.72540	3.49947	3.33025	3.19822	3.09201	3.00452
30	7.56248	5.39035	4.50974	4.01788	3.69902	3.47348	3.30450	3.17262	3.06652	2.97909
31	7.52977	5.36239	4.48369	3.99281	3.67453	3.44934	3.28059	3.14886	3.04285	2.95548
32	7.49928	5.33634	4.45943	3.96948	3.65173	3.42688	3.25834	3.12675	3.02082	2.93351
33	7.47080	5.31203	4.43679	3.94770	3.63046	3.40591	3.23757	3.10611	3.00026	2.91300
34	7.44414	5.28928	4.41561	3.92733	3.61056	3.38631	3.21815	3.08681	2.98103	2.89381
35	7.41912	5.26794	4.39575	3.90824	3.59191	3.36793	3.19995	3.06872	2.96301	2.87583
36	7.39560	5.24789	4.37710	3.89031	3.57440	3.35068	3.18286	3.05173	2.94609	2.85895
37	7.37344	5.22902	4.35954	3.87343	3.55792	3.33444	3.16677	3.03574	2.93016	2.84305
38	7.35254	5.21122	4.34299	3.85752	3.54238	3.31913	3.15161	3.02067	2.91515	2.82807
39	7.33279	5.19441	4.32736	3.84250	3.52771	3.30468	3.13730	3.00644	2.90097	2.81392
40	7.31410	5.17851	4.31257	3.82829	3.51384	3.29101	3.12376	2.99298	2.88756	2.80055
50	7.17058	5.05661	4.19934	3.71955	3.40768	3.18643	3.02017	2.89001	2.78496	2.69814
60	7.07711	4.97743	4.12589	3.64905	3.33888	3.11867	2.95305	2.82328	2.71845	2.63175
120	6.85089	4.78651	3.94910	3.47953	3.17355	2.95585	2.79176	2.66291	2.55857	2.47208
∞	6.63515	4.60538	3.78182	3.31936	3.01744	2.80216	2.63951	2.51146	2.40751	2.32110

Fuente: Valores generados con Microsoft Excel.

ISBN 978-607-8285-62-4

TABLA A4 VALORES CRÍTICOS DE F $\alpha = 0.01$. PARTE 2

EXCEL

Denominador gl_2	Grados de libertad en el numerador gl_1									
	11	12	13	14	15	16	17	18	19	20
1	6083.316	6106.320	6125.864	6142.673	6157.284	6170.101	6181.434	6191.528	6200.575	6208.730
2	99.40828	99.41585	99.42226	99.42775	99.43251	99.43668	99.44035	99.44362	99.44654	99.44917
3	27.13257	27.05182	26.98306	26.92380	26.87219	26.82686	26.78671	26.75091	26.71878	26.68979
4	14.45228	14.37359	14.30650	14.24863	14.19820	14.15386	14.11457	14.07951	14.04803	14.01961
5	9.96265	9.88828	9.82481	9.77001	9.72222	9.68016	9.64287	9.60957	9.57966	9.55265
6	7.78957	7.71833	7.65748	7.60490	7.55899	7.51857	7.48271	7.45066	7.42186	7.39583
7	6.53817	6.46909	6.41003	6.35895	6.31433	6.27501	6.24010	6.20889	6.18082	6.15544
8	5.73427	5.66672	5.60891	5.55887	5.51512	5.47655	5.44228	5.41163	5.38405	5.35909
9	5.17789	5.11143	5.05451	5.00521	4.96208	4.92402	4.89019	4.85992	4.83266	4.80800
10	4.77152	4.70587	4.64961	4.60083	4.55814	4.52045	4.48692	4.45691	4.42987	4.40539
11	4.46244	4.39740	4.34162	4.29324	4.25087	4.21344	4.18013	4.15029	4.12340	4.09905
12	4.21982	4.15526	4.09985	4.05176	4.00962	3.97237	3.93921	3.90950	3.88271	3.85843
13	4.02452	3.96033	3.90520	3.85734	3.81537	3.77825	3.74520	3.71556	3.68884	3.66461
14	3.86404	3.80014	3.74524	3.69754	3.65570	3.61868	3.58570	3.55611	3.52942	3.50522
15	3.72990	3.66624	3.61151	3.56394	3.52219	3.48525	3.45231	3.42275	3.39608	3.37189
16	3.61616	3.55269	3.49810	3.45063	3.40895	3.37205	3.33914	3.30960	3.28293	3.25874
17	3.51851	3.45520	3.40072	3.35333	3.31169	3.27482	3.24193	3.21240	3.18573	3.16152
18	3.43379	3.37061	3.31622	3.26888	3.22729	3.19043	3.15754	3.12801	3.10132	3.07710
19	3.35960	3.29653	3.24221	3.19491	3.15334	3.11650	3.08361	3.05406	3.02736	3.00311
20	3.29411	3.23112	3.17686	3.12960	3.08804	3.05120	3.01830	2.98873	2.96201	2.93774
21	3.23587	3.17295	3.11874	3.07150	3.02995	2.99311	2.96019	2.93061	2.90386	2.87956
22	3.18374	3.12089	3.06671	3.01949	2.97795	2.94109	2.90816	2.87855	2.85178	2.82745
23	3.13682	3.07402	3.01987	2.97267	2.93112	2.89425	2.86130	2.83167	2.80487	2.78050
24	3.09437	3.03161	2.97749	2.93029	2.88873	2.85185	2.81888	2.78923	2.76239	2.73800
25	3.05577	2.99306	2.93895	2.89175	2.85019	2.81329	2.78030	2.75061	2.72375	2.69932
26	3.02053	2.95785	2.90375	2.85655	2.81498	2.77807	2.74505	2.71534	2.68845	2.66399
27	2.98823	2.92557	2.87149	2.82429	2.78270	2.74577	2.71273	2.68299	2.65607	2.63158
28	2.95851	2.89588	2.84180	2.79460	2.75300	2.71605	2.68299	2.65322	2.62627	2.60174
29	2.93108	2.86847	2.81440	2.76719	2.72558	2.68860	2.65552	2.62573	2.59874	2.57419
30	2.90569	2.84310	2.78902	2.74181	2.70018	2.66319	2.63008	2.60026	2.57325	2.54866
31	2.88211	2.81953	2.76546	2.71823	2.67659	2.63958	2.60645	2.57660	2.54956	2.52494
32	2.86016	2.79759	2.74353	2.69629	2.65463	2.61760	2.58444	2.55457	2.52750	2.50285
33	2.83968	2.77712	2.72305	2.67580	2.63413	2.59708	2.56390	2.53399	2.50690	2.48222
34	2.82052	2.75797	2.70390	2.65664	2.61495	2.57788	2.54467	2.51474	2.48762	2.46292
35	2.80256	2.74002	2.68594	2.63867	2.59697	2.55987	2.52665	2.49669	2.46954	2.44481
36	2.78569	2.72315	2.66907	2.62179	2.58007	2.54296	2.50971	2.47973	2.45255	2.42779
37	2.76982	2.70728	2.65320	2.60591	2.56417	2.52704	2.49376	2.46376	2.43656	2.41177
38	2.75485	2.69232	2.63823	2.59093	2.54918	2.51202	2.47873	2.44870	2.42147	2.39666
39	2.74072	2.67819	2.62410	2.57678	2.53501	2.49784	2.46452	2.43447	2.40722	2.38239
40	2.72735	2.66483	2.61073	2.56340	2.52162	2.48442	2.45108	2.42101	2.39374	2.36888
50	2.62503	2.56250	2.50833	2.46089	2.41896	2.38160	2.34807	2.31780	2.29032	2.26524
60	2.55867	2.49612	2.44188	2.39435	2.35230	2.31480	2.28113	2.25070	2.22305	2.19781
120	2.39900	2.33630	2.28181	2.23395	2.19150	2.15357	2.11943	2.08851	2.06035	2.03459
∞	2.24790	2.18492	2.13004	2.08170	2.03871	2.00018	1.96540	1.93381	1.90497	1.87850

Fuente: Valores generados con Microsoft Excel.

ISBN 978-607-8285-62-4

TABLA A4 VALORES CRÍTICOS DE F $\alpha = 0.01$. PARTE 3

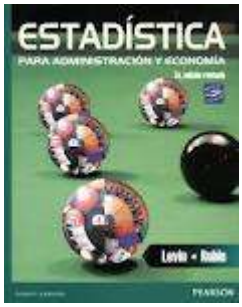
EXCEL

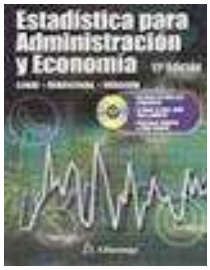
Denominador gl_2	Grados de libertad en el numerador gl_1									
	21	22	23	24	25	30	40	60	120	∞
1	6216.118	6222.843	6228.990	6234.630	6239.825	6260.648	6286.782	6313.030	6339.391	6365.861
2	99.45155	99.45371	99.45569	99.45750	99.45917	99.46583	99.47416	99.48250	99.49083	99.49916
3	26.66350	26.63955	26.61764	26.59752	26.57898	26.50453	26.41081	26.31635	26.22114	26.12518
4	13.99383	13.97033	13.94882	13.92906	13.91085	13.83766	13.74538	13.65220	13.55810	13.46306
5	9.52812	9.50576	9.48529	9.46647	9.44912	9.37933	9.29119	9.20201	9.11177	9.02043
6	7.37219	7.35063	7.33088	7.31272	7.29597	7.22853	7.14322	7.05674	6.96902	6.88003
7	6.13238	6.11134	6.09205	6.07432	6.05795	5.99201	5.90845	5.82357	5.73729	5.64954
8	5.33641	5.31571	5.29673	5.27926	5.26314	5.19813	5.11561	5.03162	4.94605	4.85881
9	4.78556	4.76508	4.74629	4.72900	4.71303	4.64858	4.56665	4.48309	4.39777	4.31056
10	4.38313	4.36278	4.34411	4.32693	4.31106	4.24693	4.16529	4.08186	3.99648	3.90899
11	4.07688	4.05662	4.03803	4.02091	4.00509	3.94113	3.85957	3.77607	3.69044	3.60245
12	3.83633	3.81612	3.79757	3.78049	3.76469	3.70079	3.61918	3.53547	3.44944	3.36082
13	3.64254	3.62236	3.60383	3.58675	3.57096	3.50704	3.42529	3.34129	3.25476	3.16540
14	3.48317	3.46300	3.44447	3.42739	3.41159	3.34760	3.26564	3.18127	3.09419	3.00403
15	3.34984	3.32966	3.31112	3.29403	3.27822	3.21411	3.13191	3.04713	2.95945	2.86844
16	3.23668	3.21649	3.19793	3.18081	3.16497	3.10073	3.01825	2.93305	2.84474	2.75284
17	3.13944	3.11923	3.10064	3.08350	3.06764	3.00324	2.92046	2.83481	2.74585	2.65304
18	3.05500	3.03476	3.01615	2.99897	2.98308	2.91852	2.83542	2.74931	2.65970	2.56597
19	2.98098	2.96071	2.94207	2.92487	2.90894	2.84420	2.76079	2.67421	2.58394	2.48929
20	2.91558	2.89528	2.87660	2.85936	2.84340	2.77848	2.69475	2.60771	2.51678	2.42120
21	2.85737	2.83704	2.81832	2.80105	2.78505	2.71995	2.63590	2.54839	2.45681	2.36031
22	2.80523	2.78486	2.76611	2.74880	2.73276	2.66749	2.58311	2.49515	2.40292	2.30549
23	2.75825	2.73785	2.71907	2.70172	2.68565	2.62019	2.53550	2.44708	2.35421	2.25586
24	2.71571	2.69527	2.67646	2.65907	2.64296	2.57733	2.49232	2.40346	2.30996	2.21070
25	2.67701	2.65653	2.63768	2.62026	2.60411	2.53831	2.45299	2.36369	2.26956	2.16940
26	2.64164	2.62113	2.60224	2.58479	2.56860	2.50262	2.41701	2.32728	2.23254	2.13148
27	2.60919	2.58865	2.56973	2.55224	2.53602	2.46987	2.38396	2.29381	2.19846	2.09653
28	2.57933	2.55875	2.53979	2.52227	2.50602	2.43970	2.35350	2.26294	2.16700	2.06419
29	2.55174	2.53113	2.51214	2.49458	2.47830	2.41182	2.32534	2.23437	2.13785	2.03418
30	2.52618	2.50553	2.48651	2.46892	2.45260	2.38597	2.29921	2.20785	2.11076	2.00624
31	2.50243	2.48175	2.46270	2.44508	2.42873	2.36194	2.27491	2.18317	2.08552	1.98015
32	2.48031	2.45960	2.44052	2.42286	2.40648	2.33954	2.25225	2.16014	2.06194	1.95574
33	2.45965	2.43891	2.41980	2.40211	2.38570	2.31861	2.23107	2.13859	2.03985	1.93283
34	2.44031	2.41955	2.40040	2.38269	2.36625	2.29902	2.21123	2.11838	2.01912	1.91129
35	2.42218	2.40138	2.38221	2.36447	2.34800	2.28063	2.19260	2.09940	1.99962	1.89099
36	2.40513	2.38431	2.36511	2.34734	2.33084	2.26333	2.17507	2.08153	1.98124	1.87183
37	2.38909	2.36824	2.34901	2.33121	2.31468	2.24704	2.15855	2.06468	1.96389	1.85370
38	2.37395	2.35307	2.33381	2.31599	2.29944	2.23167	2.14295	2.04876	1.94748	1.83652
39	2.35965	2.33874	2.31946	2.30161	2.28503	2.21714	2.12820	2.03369	1.93194	1.82022
40	2.34611	2.32518	2.30588	2.28800	2.27140	2.20338	2.11423	2.01941	1.91719	1.80472
50	2.24226	2.22111	2.20158	2.18349	2.16666	2.09759	2.00659	1.90903	1.80260	1.68314
60	2.17465	2.15334	2.13363	2.11536	2.09837	2.02848	1.93602	1.83626	1.72632	1.60066
120	2.01091	1.98906	1.96882	1.95002	1.93249	1.86001	1.76285	1.65569	1.53299	1.38055
∞	1.85410	1.83152	1.81055	1.79101	1.77275	1.69660	1.59247	1.47321	1.32486	1.01099

Fuente: Valores generados con Microsoft Excel.

ISBN 978-607-8285-62-4

BIBLIOGRAFÍA





1. **B**erenson, Mark L. y Levine, David M. **ESTADÍSTICA PARA ADMINISTRACIÓN.** Editorial Pearson. Cuarta Edición, 2006. Formato: Rústico. Idioma: español. País: México. ISBN: 9702608023. No. de páginas: 648.
2. Carlberg, Conrad. **ANÁLISIS ESTADÍSTICO CON EXCEL**. Editorial Anaya multimedia-Anaya interactiva. Edición, 2011. Formato: Rústico. Idioma: Español. País: España. ISBN: 9788441530263. No. de páginas: 528.
3. Carrascal Arranz, Ursicino. **ESTADÍSTICA DESCRIPTIVA CON MS MICROSIFT EXCEL 2010: VERSIONES 97 A 2010**. Editorial Alfaomega grupo editor. Edición, 2012. Formato: Rústico. Idioma: español. País: México. ISBN: 9786077071969. No. de páginas: 288.
4. Dawson-Saunders, Beth y G. Trapp, Robert. **BIOESTADÍSTICA MÉDICA**. Editorial Manual modern. Segunda edición, 1997. Formato: Rústico. Idioma: español. País: México. ISBN: 9684267517. No. de páginas: 403.
5. Gutiérrez Pulido, Humberto y De la Mara Salazar, Román. **ANÁLISIS Y DISEÑO DE EXPERIMENTOS**. Editorial Mc Graw Hill (México). Edición, 2008. Formato: Libro electrónico. Idioma: español. País: México. Código producto: 9786071501394. Tamaño: 12.20 MB.
6. Hildebrand, David K y Ott R., Lyman. **ESTADISTICA APLICADA** a la administración y a la economía. Editorial Addison–Wesley Iberoamericana, 1997. Formato: Rústico. Idioma: español. País: México. ISBN: 0201625520. No. de páginas: 943.
7. Johnson, Robert. **ESTADÍSTICA ELEMENTAL**. Editorial Cengage Learnin. Décima edición, 2008. Formato: Rústico. Idioma: español. País: México. ISBN: 9789706868350. No. de páginas: 725.
8. Kazmier, Leonard y Díaz Mata, Alfredo. **ESTADÍSTICA APLICADA** a la administración y a la economía. Editorial Mc Graw-Hill Interamericana. Edición, 2006. Formato: Rústico. Idioma: español. País: México. ISBN: 9701059182. No. de páginas: 406.
9. Levin, Richard I. **ESTADÍSTICA PARA ADMINISTRADORES**. Editorial Prentice- Hall hispanoamericana, S.A. Segunda edición, 1988. Formato: Rústico. Idioma: español. País: México. ISBN: 9688801526. No. de páginas: 940.
10. Levin, Richard I. **ESTADÍSTICA PARA ADMINISTRACIÓN Y ECONOMIA**. Editorial Pearson. Séptima edición, 2010. Formato: Rústico. Idioma: español. País: México. ISBN: 9786074429053. No. de páginas: 952
11. Levine, David M, Berenson, Mark I. **ESTADÍSTICA PARA ADMINISTRACIÓN**. Editorial Pearson. Cuarta Edición, 2008. Formato: Rústico. Idioma: español. País: México. ISBN: 9702608023. No. de páginas: 648.
12. Lind, Douglas A., Marchal, William G. y Wathen, Samuel A. Wathen.

ISBN 978-607-8285-62-4

ESTADÍSTICA APLICADA a los Negocios y a la Economía. Editorial Mc Graw –Hill Interamericana. 12ª. Edición, 2008. Formato: Rústico. Idioma: español. País: México. ISBN: 9701048342. No. de páginas: 800.

13. Lind, Douglas A., Marchal, William G. y Mason, Robert D. **ESTADÍSTICA PARA ADMINISTRACIÓN Y ECONOMÍA**. Editorial Alfaomega. Onceava edición, 2004. Formato: Rústico. Idioma: español. País: México. ISBN: 9701509749. No. de páginas: 830.
14. **MANUAL DE MINITAB 15**. Versión en español para Windows. Edición, 2007. Formato: electrónico. Idioma: español. País: México.
15. Márquez, Felicidad. **ESTADÍSTICA DESCRIPTIVA** a través de Excel. Editorial Alfaomega grupo editor. Edición, 2009. Formato: Rústico. Idioma: español. País: México. ISBN: 9786077686989. No. de páginas: 288.
16. Montgomery, Douglas C. **DISEÑO Y ANÁLISIS DE EXPERIMENTOS**. Editorial Limusa. Primera edición, 2008. Formato: Rústico. Idioma: español. País: México. ISBN: 9681861566. No. de páginas:
17. Pérez López, Cesar. **ESTADÍSTICA APLICADA** a través de Excel. Editorial Pearson-Prentice Hall. Edición, 2002. Formato: Rústico. Idioma: español. País: México. ISBN: 8420535362. No. de páginas: 596.
18. Velasco Sotomayor, Gabriel. **ESTADÍSTICA CON EXCEL**. Editorial Trillas. Primera edición, 2005. Formato: Rústico. Idioma: español. País: México. ISBN: 9682406269. No. de páginas:
19. Walpole, Ronald E. **PROBABILIDAD Y ESTADÍSTICA**. Editorial Mc Graw-Hill Interamericana. Edición, 1992. Formato: Rústico. Idioma: español. País: México. ISBN: 9684229925. No. de páginas:

ISBN 978-607-8285-62-4