ISBN 978-607-8617-98-2

Apuntes de estadística inferencial básica para ingeniería: regresión, series de tiempo y diseños experimentales

M.C. ISAAC SÁNCHEZ ANASTACIO
M.A.F.O. CLAUDIA VELÁSQUEZ CORTÉS
M.C. JUAN CARLOS ROJAS MARTÍNEZ

RED IBEROAMERICANA DE ACADEMIAS DE INVESTIGACIÓN A.C.
RED IBAI
RED TEMÁTICA IBEROAMERICANA DE ACADEMIAS DE INVESTIGACIÓN SOBRE MIGRACIÓN Y DESARROLLO
MIGRACIÓN Y DESARROLLO
REDIBAI-MYD
CONACYT
Consejo Nacional de Ciencia y Tecnología
Redes Temáticas
ITSZ
INGENIEROS
INSTITUTO TECNOLÓGICO SUPERIOR DE ZONGOLICA

APUNTES DE ESTADÍSTICA INFERENCIAL BÁSICA PARA INGENIERÍA: REGRESIÓN, SERIES DE TIEMPO Y DISEÑOS EXPERIMENTALES

AUTORES
ISAAC SÁNCHEZ ANASTACIO
CLAUDIA VELÁSQUEZ CORTÉS
JUAN CARLOS ROJAS MARTÍNEZ

EDITORIAL

EDITA: RED IBEROAMERICANA DE ACADEMIAS DE INVESTIGACIÓN A.C
DUBLÍN 34, FRACCIONAMIENTO MONTE MAGNO
C.P. 91190. XALAPA, VERACRUZ, MÉXICO.
CEL 2282386072
PONCIANO ARRIAGA 15, DESPACHO 101.
COLONIA TABACALERA
DELEGACIÓN CUAUHTÉMOC
C.P. 06030. MÉXICO, D.F. TEL. (55) 55660965
www.redibai.org
redibai@hotmail.com

Sello editorial: Red Iberoamericana de Academias de Investigación, A.C. (607-8617)
Primera Edición, Xalapa, Veracruz, México.
No. de ejemplares: 200
Presentación en medio electrónico digital: Cd-Rom
formato PDF 10 MB
Fecha de aparición 11/12/2020
ISBN 978-607-8617-98-2

RED IBEROAMERICANA DE ACADEMIAS DE INVESTIGACIÓN A.C.

SELLO EDITORIAL INDAUTOR/ISBN 607-8617

Dublín 34
Fracc. Monte Magno
Xalapa, Ver.
C.P. 91193

CERTIFICACIÓN EDITORIAL DEL LIBRO ELECTRÓNICO *APUNTES DE ESTADÍSTICA INFERENCIAL BÁSICA PARA INGENIERÍA: REGRESIÓN, SERIES DE TIEMPO Y DISEÑOS EXPERIMENTALES*
(ISBN 978-607-8617-98-2)

La Red Iberoamericana de Academias de Investigación A.C. con el sello editorial Nº 607-8617 otorgado por la agencia mexicana de ISBN, hace constar que el libro electrónico **APUNTES DE ESTADÍSTICA INFERENCIAL BÁSICA PARA INGENIERÍA: REGRESIÓN, SERIES DE TIEMPO Y DISEÑOS EXPERIMENTALES** con ISBN 978-607-8617-98-2; es publicado por nuestro sello con fecha del 11 de diciembre de 2020 cumpliendo con todos los requisitos de calidad científica y normalización que exige nuestra política editorial.

Apuntes de estadística inferencial básica para ingeniería: regresión, series de tiempo y diseños experimentales fue arbitrado bajo el sistema de administración y publicación de libros electrónicos OJS versión 3.2.0.3. del Public Knowled Project cuyo desarrollo promueve las tecnologías para el uso de la investigación académica. El proceso de arbitraje constó de dos etapas.

La primera revisión fue realizada por parte de la Secretaría Técnica de la REDIBAI. AC, en conjunto con el Instituto Tecnológico Superior de Zongolica, quien verificó que la propuesta cumpliera con los requisitos básicos establecidos: enfoque temático, extensión, apego a las normas de citación, estructura, formato, entre otros. Posteriormente el trabajo pasó a una primera lectura a cargo del Editor en Jefe que forma parte del Comité Editorial del sello editorial, quien determinó la pertinencia de la propuesta y decidió que cumplía con los requisitos de calidad académica. Esta fase se desarrolló en un tiempo de 15 días.

En la segunda etapa el trabajo se sometió al proceso de evaluación de pares académicos a través del procedimiento doble ciego, a cargo de árbitros anónimos especialistas en el tema pertenecientes a instituciones educativas a nivel nacional e internacional, lo que busca garantizar la calidad de las revisiones. Ningún veredicto de los dictaminadores fue contradictorio, por lo que no se recurrió a un tercer árbitro para tomar la decisión final de publicarlo, el resultado de este esfuerzo académico y científico fue aprobado. Este proceso comprendió de dos meses.

RED IBEROAMERICANA DE ACADEMIAS DE INVESTIGACIÓN A.C.

SELLO EDITORIAL INDAUTOR/ISBN 607-8617

Dublín 34
Fracc. Monte Magno
Xalapa, Ver.
C.P. 91193

El proceso de evaluación de las dos etapas se desarrolló en un tiempo promedio de 2 meses y medio, iniciado desde el momento de su recepción el 2 de septiembre de 2020, hasta la terminación del arbitraje el 29 de noviembre de 2020 y se publicó el 11 de diciembre de 2020 tomando en cuenta los criterios de originalidad, pertinencia, relevancia de los hallazgos, manejo de la teoría especializada, rigor metodológico, congruencia, claridad argumentativa y calidad de la redacción.

El cuerpo de arbitraje estuvo integrado por los cuerpos académicos pertenecientes al comité científico de la REDIBAI MyD y al comité científico del Instituto Tecnológico Superior de Zongolica

Todos los soportes concernientes a los procesos editoriales y de evaluación reposan en Editorial REDIBAI, las cuales ponemos a disposición de la comunidad académica interna y externa en el momento que se requiera.

Atentamente

Xalapa Enríquez, Veracruz, a 11 de diciembre de 2020

MTRO. DANIEL ARMANDO OLIVERA GÓMEZ
Editor
Secretario Ejecutivo de la REDIBAI A.C.

APUNTES DE ESTADÍSTICA INFERENCIAL BÁSICA PARA INGENIERÍA: REGRESIÓN, SERIES DE TIEMPO Y DISEÑOS EXPERIMENTALES

AUTORES
ISAAC SÁNCHEZ ANASTACIO
CLAUDIA VELÁSQUEZ CORTÉS
JUAN CARLOS ROJAS MARTÍNEZ

Contenido

Introducción

El presente libro fue realizado con el objetivo de poder explicar los temas de regresión lineal simple, regresión lineal múltiple, diseños factoriales de un factor y diseños factoriales de bloques completamente aleatorizados, de una manera simple, describiendo los pasos que los libros de estadística omiten, permitiendo al alumno, un entendimiento adecuado de estos temas.

1. Regresión lineal simple y correlación

1.1 ¿Cómo determinar la relación entre variables?

- En el análisis de regresión, desarrollaremos una ecuación de estimación, esto es, una fórmula matemática que relaciona las variables conocidas con la variable desconocida.
- Después de conocer el patrón de esta relación, podremos aplicar el análisis de correlación para determinar el grado en el que las variables se relacionan. El análisis de correlación, entonces, nos indica qué tan bien la ecuación de estimación describe realmente la relación.

A continuación, describiremos un ejemplo de la aplicación de ambos conceptos.

1.2 Análisis de regresión lineal

El análisis de regresión es una herramienta muy valiosa para el gerente actual. La regresión se ha utilizado para modelar cuestiones como la relación entre el nivel de educación y el ingreso, el precio de una casa y los pies cuadrados de construcción, así como el volumen de ventas para una compañía en relación con el dinero gastado en publicidad. Cuando un negocio intenta decidir cuál lugar es mejor para abrir una nueva tienda o sucursal, los modelos de regresión se utilizan con frecuencia. Los modelos de estimación de costos muchas veces son modelos de regresión. Las posibilidades de aplicación del análisis de regresión son prácticamente ilimitadas.

En general, hay dos propósitos en el análisis de regresión. **El primero es entender la relación entre las variables como gastos en publicidad y ventas. El segundo es predecir el valor de una de las variables con base en el valor de la otra.** Por ello, la regresión es una técnica muy importante para realizar predicciones, siendo la materia de Gestión de la Producción I, donde la aplicarán para realizar pronósticos. En cualquier modelo de regresión, la variable que se quiere **predecir** se llama **variable dependiente o variable de respuesta**. Se dice que su valor es **dependiente** del valor de una **variable independiente**, que algunas veces se **llama variable explicativa o variable predictiva**.

A la ecuación con que se describe cómo se relaciona ***y*** con ***x,*** y en la que se da un término para el error, se le llama **modelo de regresión**. El siguiente es el modelo que se emplea en la regresión lineal simple.

Modelo de regresión lineal simple

$$y = \beta_0 + \beta_1 x + \in \tag{6.1}$$

$\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$ se conocen como los parámetros del modelo, y $\in$ (la letra griega épsilon) es una variable aleatoria que se conoce como término del error. El término del error da cuenta de la variabilidad de y que no puede ser explicada por la relación lineal entre ***x*** y ***y***.

A la ecuación que describe la relación entre el valor esperado de ***y***, que se denota ***E(x)*** y ***x*** se le llama, ecuación de regresión. La siguiente es la ecuación de regresión para la regresión lineal simple.

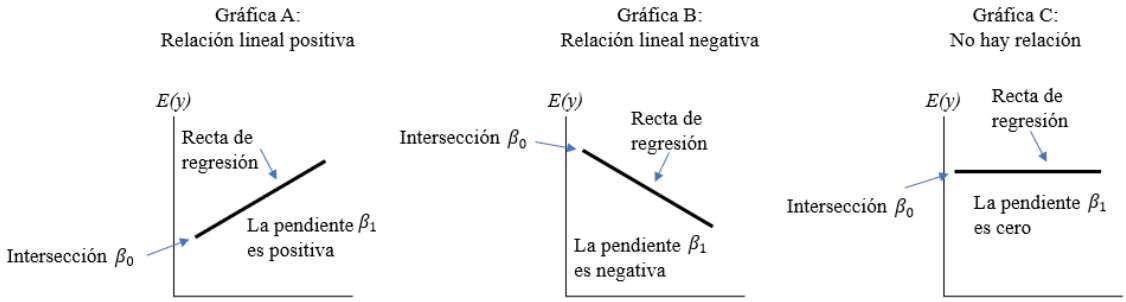
Ecuación de regresión lineal simple

$$E(y) = \beta_0 + \beta_1 x \tag{6.2}$$

La gráfica de la ecuación de regresión lineal simple es una línea recta; $\boldsymbol{\beta_0}$ es la intersección de la recta de regresión con el eje ***y***, $\boldsymbol{\beta_1}$ es la pendiente y ***E(y)*** es la media o valor esperado de ***y*** para un valor dado de ***x.***

En la figura 1.1 se presentan ejemplos de posibles rectas de regresión. La recta de regresión de la gráfica A indica que el valor medio de ***y*** está relacionado **positivamente** con ***x***. La recta de regresión de la gráfica B indica que el valor medio de ***y*** está relacionado negativamente con ***x***, valores menores de ***E(y)*** corresponden a valores mayores de ***x***. La recta de regresión de la gráfica C muestra el caso en el que el valor medio de ***y*** no está relacionado con ***x***; es decir, el valor medio de ***y*** es el mismo para todos los valores de ***x***.

Figura 1.1 Ejemplos de posibles rectas de regresión



1.3 El método de mínimos cuadrados

El método de mínimos cuadrados es un método en el que se usan los datos muestrales para hallar la ecuación de regresión estimada. Las estimaciones de la pendiente y la intersección se encuentran a partir de los datos muestrales. La mejor recta de regresión se define como la que tiene la suma mínima de los cuadrados de los errores. Por tal razón, algunas veces el análisis de regresión se conoce como regresión de mínimos cuadrados. Los estadísticos han desarrollado fórmulas para encontrar la ecuación de una recta que minimiza la suma de los cuadrados de los errores. La ecuación de regresión lineal simple es:

$$\hat{y} = a + bx \tag{6.3}$$

donde:

$\hat{y}$ = valor pronosticado de y, la variable dependiente o variable de respuesta

a = estimación de β_0, la intersección de la recta de regresión con el eje y, según los resultados de la muestra, cuando x vale 0

b = estimación de β_1, la pendiente de la recta de estimación, según los resultados de la muestra.

x = variable independiente (variable predictiva o variable explicativa)

Las siguientes fórmulas sirven para calcular la intersección y la pendiente:

$$\bar{x} = \frac{\sum x}{n} \quad \text{media de los valores de } x \tag{6.4}$$

$$\bar{y} = \frac{\sum y}{n} \quad \text{media de los valores de } y \tag{6.5}$$

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{6.6}$$

$$a = \bar{y} - b\bar{x} \tag{6.7}$$

1.4 Análisis de correlación

El análisis de correlación es la herramienta estadística que podemos usar para describir el grado en el que una variable está linealmente relacionada con otra. Con frecuencia, el análisis de correlación se utiliza junto con el de regresión para medir qué tan bien la línea de regresión explica los cambios de la variable dependiente, Y. Sin embargo, la correlación también se puede usar sola para medir el grado de asociación entre dos variables. Los estadísticos han desarrollado dos medidas para describir la correlación entre dos variables: **el coeficiente de determinación y el coeficiente de correlación**.

El coeficiente de determinación (r^2)

El coeficiente de determinación es la principal forma en que podemos medir el grado, o fuerza, de la asociación que existe entre dos variables, *x* y *y*. Debido a que usamos una muestra de puntos para desarrollar rectas de regresión, nos referimos a esta medida como el coeficiente de determinación muestral. Un punto que debemos resaltar es que r^2 mide sólo la fuerza de una relación lineal entre dos variables. La fórmula es la siguiente:

$$r^2 = \frac{a\sum y + b\sum xy - n\bar{y}^2}{\sum y^2 - n\bar{y}^2} \tag{6.8}$$

Si cada punto de la muestra estuviera sobre la recta de regresión (es decir, si todos los errores fueran 0), entonces, la ecuación de regresión explicaría 100% de la variabilidad en *y*, de manera que r^2=1. El menor valor posible de r^2 es 0 e indica que *x* explica 0% de la variabilidad en *y*. Así, r^2 puede tener valores desde 0, el más bajo, hasta 1, el más alto. Al desarrollar las ecuaciones de regresión, un buen modelo tendrá un valor de r^2 cercano a 1.

El coeficiente de correlación r

Otra medida relacionada con el coeficiente de determinación es el coeficiente de correlación. Esta medida también expresa el grado o fuerza de la relación lineal. En general, se expresa como *r* y puede ser cualquier número entre +1 y –1, incluyendo ambos valores. La figura 1.2 ilustra los diagramas de dispersión posibles para diferentes valores de *r*. El valor de *r* es la raíz cuadrada de r^2. **Es negativo si la pendiente es negativa y es positivo si la pendiente es positiva**. La fórmula para obtener r es:

$$r = \pm\sqrt{r^2} \tag{6.9}$$

1.5 Error estándar de la media

Una vez obtenida la ecuación de estimación de la recta de regresión, se requiere medir la dispersión de los datos muestrales en torno a la función de regresión ajustada. Se puede desarrollar una medida de dispersión análoga a la desviación estándar muestral. Esta medid, llamada el **error estándar de la estimación**, mide la dispersión de los datos respecto a la línea ajustada en la dirección *y*, se puede determinar mediante la siguiente fórmula:

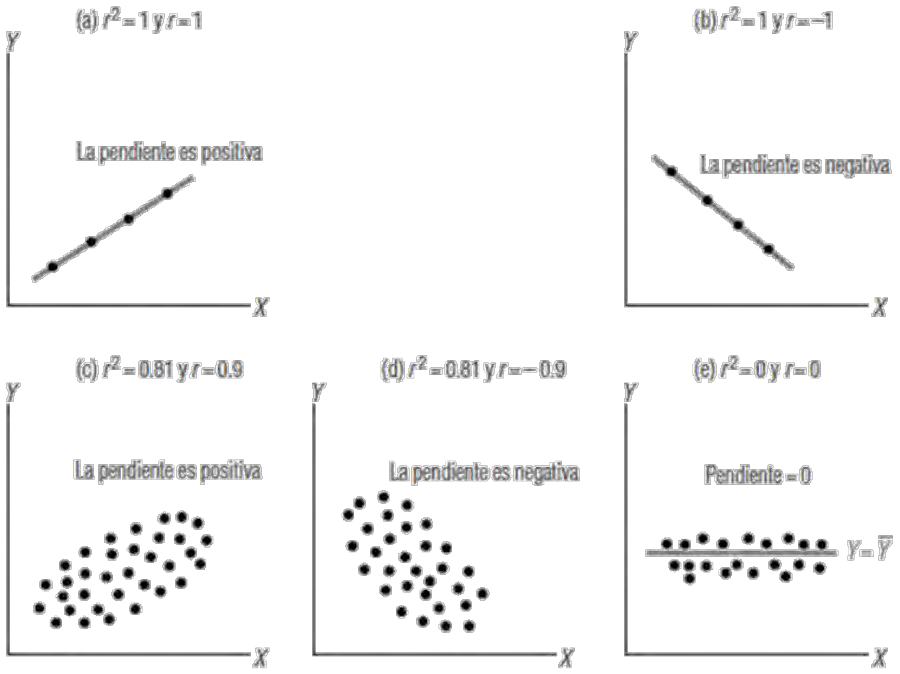
$$S_e = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}} \tag{6.10}$$

El **error estándar de la estimación** mide la cantidad por la cual los valores verdaderos *y*, difieren de los valores estimados $\hat{y}$. Para muestras relativamente grandes, se esperaría que

alrededor del 67% de las diferencias $(y - \hat{y})$ estuvieran a una distancia inferior a S_e del 0 y aproximadamente el 95% de estas diferencias estarían a $2S_e$ del 0.

Un análisis de regresión con un pequeño error estándar de la estimación significativa que todos los datos descansan muy cera de la línea de regresión ajustada. Si el error estándar de la estimación es grande, los datos se dispersan considerablemente respecto de la línea ajustada.

Figura 1.2 Valores posibles del coeficiente de correlación *r*



1

1.6 Pasos para realizar el análisis de regresión lineal simple y correlación

Para describir los pasos de la aplicación de un análisis de regresión lineal simple y correlación, pondremos como ejemplo el caso del vicepresidente de investigación y desarrollo (I+D) de una gran compañía química y de fabricación de fibras, él cree que las ganancias anuales de la empresa dependen de la cantidad gastada en I+D. El nuevo presidente de la compañía no está de acuerdo y ha solicitado pruebas. Los datos de seis años se presentan en el cuadro 6.1

Cuadro 6.1 Concentrado de datos de gastos y ganancias
del 2014 al 2019.

Año	Gastos en Investigación y Desarrollo I+D (Millones de pesos)	Ganancias anuales (Millones de pesos)
2014	5	31
2015	7	35
2016	4	30
2017	5	34
2018	3	25
2019	2	20

El vicepresidente de I+D desea una ecuación para pronosticar los beneficios anuales derivados de la cantidad presupuestada para I+D y ocupar dicha ecuación para pronosticar las ganancias del 2020, considerando una inversión en I+D de 6 millones para el año 2020.

Paso 1. Realizar el diagrama de dispersión

El primer paso para determinar si existe una relación entre dos variables es examinar la gráfica de los datos observados (o conocidos). Esta gráfica, o dibujo, se llama diagrama de dispersión. **Un diagrama de dispersión** nos puede dar dos tipos de información. Visualmente, podemos identificar patrones que indiquen que las variables están relacionadas. Si esto sucede, podemos ver qué tipo de línea, o ecuación de estimación, describe esta relación.

En la figura 6.3 se muestra el gráfico de dispersión, para realizar este tipo de gráfico se deben realizar los siguientes pasos:

a) Identificar del problema cuál es la variable independiente (x) y la variable dependiente (y). De acuerdo al problema planteado, el vicepresidente de I+D desea establecer una ecuación que pronostique las ganancias (estas sería la variable dependiente) y el gasto en I+D (sería la variable independiente).
b) Una vez definida quien sería la variable x y y, se procede a graficar, de acuerdo a los datos el eje x tendría valores entre 2 y 7 (valores mínimo y máximo del gasto en I+D). Para el eje y se tendría valores entre 20 y 35 (valores mínimo y máximo del gasto en I+D). Proceder a trazar los ejes x y y con los rangos que se definieron.
c) Ir agregando los puntos de intersección, ejemplo: en la figura 6.3, el punto (A), se obtuvo de la intersección del valor 5 en el eje x y del valor 31 en el eje y. Agregar los otros 5 puntos restantes que resulten de la intersección de ambas coordenadas.

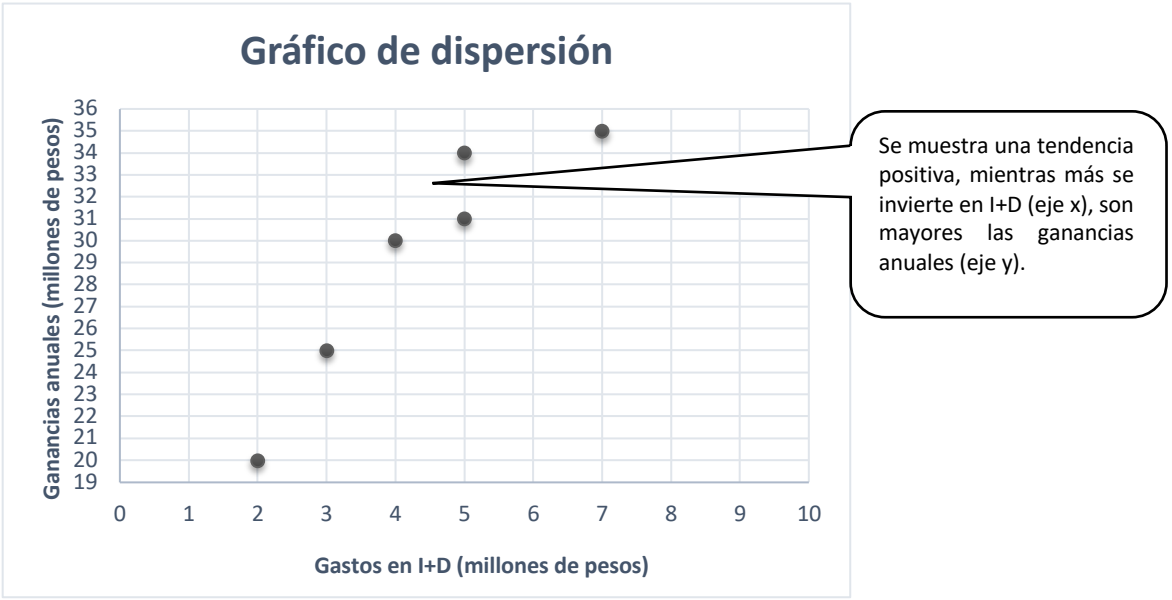


Figura 6.3. Gráfico de dispersión de los datos de la compañía química

(A)

De acuerdo al diagrama de dispersión de la figura 6.3, los puntos muestran una tendencia positiva, es decir, que sí se aumenta la inversión de I+D, aumenta las ganancias anuales.

Nota: Como hay una tendencia, sí es factible realizar un análisis de regresión lineal y correlación, por otra parte, como la tendencia es positiva, el valor que se obtenga de "r" (coeficiente de correlación) será positivo, su valor y significado lo veremos más adelante.

Paso 2. Realizar la sumatoria de los valores (de acuerdo a lo indicado por las fórmulas de mínimos cuadrados.

En el cuadro 6.1, se muestra la suma de los valores de la variable *x* (el gasto en I+D) y los valores de *y* (las ganancias anuales).

Cuadro 6.2 Cálculos del análisis de regresión lineal y correlacional

	x	y	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	xy	y^2
	5	31						
	7	35						
	4	30						
	5	34						
	3	25						
	2	20						
∑=	**26**	**175**						

Paso 3. Calcular la media de *x* (fórmula 6.4) y de *y* (fórmula 6.5)

Ocupando los valores de las sumatorias de *x* y *y*, se obtienen los valores de las medias de la variable dependiente e independiente.

$$\bar{x} = \frac{\sum x}{n} = \frac{26}{6} = 4.333 \qquad \bar{y} = \frac{\sum y}{n} = \frac{175}{6} = 29.167$$

Los valores de $\bar{x}$ y $\bar{y}$ se ocupan en los cálculos del paso 4.

Paso 4. Realizar los cálculos de lo que indican las columnas 3 a la 8, del cuadro 6.2

	x	y	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	xy	y^2
	5	31	0.667	0.445	1.833	1.223	155	961
	7	35	2.667	7.113	5.833	15.557	245	1225
	4	30	-0.333	0.111	0.833	-0.277	120	900
	5	34	0.667	0.445	4.833	3.224	170	1156
	3	25	-1.333	1.777	-4.167	5.555	75	625
	2	20	-2.333	5.443	-9.167	21.387	40	400
∑=	**26**	**175**		**15.334**		**46.669**	**805**	**5267**

Se describen los cálculos de la primera fila:

$x_i - \bar{x}$ = 5-4.33= 0.667

$(x_i - \bar{x})^2$ = 0.667*0.667=0.445

$y_i - \bar{y}$ = 31-29.167= 1.833

$(x_i - \bar{x})(y_i - \bar{y})$ = 0.667*1.833=1.223

xy = 5*31=155

y^2 = 31*31 = 961

Paso 5. Con los valores obtenidos en los pasos 2 al 4, sustituirlos en las fórmulas (6.6) y (6.7), para obtener respectivamente, b (la pendiente de la recta de estimación) y el valor de a (la ordenada).

$$(6.6)\ b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{46.669}{15.334} = 3.043$$

$$(6.7)\ a = \bar{y} - b\bar{x} =$$

29.167- (3.043*4.333) = 15.982

Paso 6. Sustituir los valores de a y b en la fórmula (6.3) para obtener la ecuación de estimación de la recta de este problema.

Esta es la ecuación de estimación para pronosticas las ganancias, dado el gasto destinado a I+D

$$(6.3)\ \hat{y} = a + bx = {}^{15.982 + (3.043\ x\ )}$$

$$\hat{y} = \ 15.982+3.043x$$

Sustituir en la ecuación de estimación de la recta obtenida en el paso 6, el valor de "*x*"; de acuerdo al problema, se desea estimar las ganancias anuales, considerando un gasto en I+D de 6 millones de pesos.

$$\hat{y} = \ 15.982+3.043x$$

$$\hat{y} = \ 15.982+3.043(6)$$

$$\hat{y} = \ 34.24$$

Sí en el 2020 se realiza un gasto de 6 millones de pesos en I+D, las ganancias anuales estimadas serían de 34.24 millones de pesos.

Paso 7. Obtener el valor del coeficiente de determinación r^2 (fórmula 6.8) y coeficiente de correlación r (fórmula 6.9)

Datos obtenidos de los pasos 3 y 4

$\sum y = 175$ $\sum xy = 805$ $\sum y^2 = 5267$ $n = 6$ $\bar{y} = 29.167$

$\sum (x_i - \bar{x})(y_i - \bar{y}) = 46.669$ $\sum (x_i - \bar{x})^2 = 15.334$

Datos obtenidos del paso 5

$a = 15.982$ $b = 3.043$

$$(6.8)\ r^2 = \frac{a\sum y + b\sum xy - n\bar{y}^2}{\sum y^2 - n\bar{y}^2} = \frac{((15.982 * 175) + (3.043 * 805) - (6 * 29.167^{\wedge} 2))}{(5267 - (6 * 29.167^{\wedge}2))}$$

$$r^2 = \frac{a\sum y + b\sum xy - n\bar{y}^2}{\sum y^2 - n\bar{y}^2} = \frac{142.182}{162.717}$$

$r^2 = 0.874$

¿Qué significa el valor obtenido de r^2= 87.4%?

Podemos concluir que la variación en los gastos de I+D (la variable independiente *x*) explica el 87.4% de la variación en las ganancias anuales (la variable dependiente *y*).

$$(6.9) r = \pm\sqrt{r^2} = \sqrt{0.874}$$

$r = 0.935$

¿Qué significa el valor obtenido de r = 93.5%?

La relación entre las dos variables **es directa y la pendiente es positiva**; por tanto, el signo de r es positivo. El valor de r=0.935 significa que el 93.5% de los datos se relacionan entre sí.

Nota: Si la relación fuera negativa, el valor de r sería -0.935

Nota: para propósitos descriptivos

$$r \geq 0.8 \qquad Relación\ fuerte$$

$$0.5 < r < 0.8 \qquad Relación\ moderada$$

$$r \leq 0.5 \qquad Relación\ débil$$

De acuerdo a los datos mostrados, el valor de r=0.935 = 93.5% indica una relación fuerte entre el gasto en I+D y las ganancias anuales obtenidas.

Paso 8. Obtener el error estándar de la regresión con la fórmula 6.10.

$$(6.10)\ S_e = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}} = \sqrt{\frac{(5267 - (15.982 * 175) - (3.043 * 805))}{(6-2)}}$$

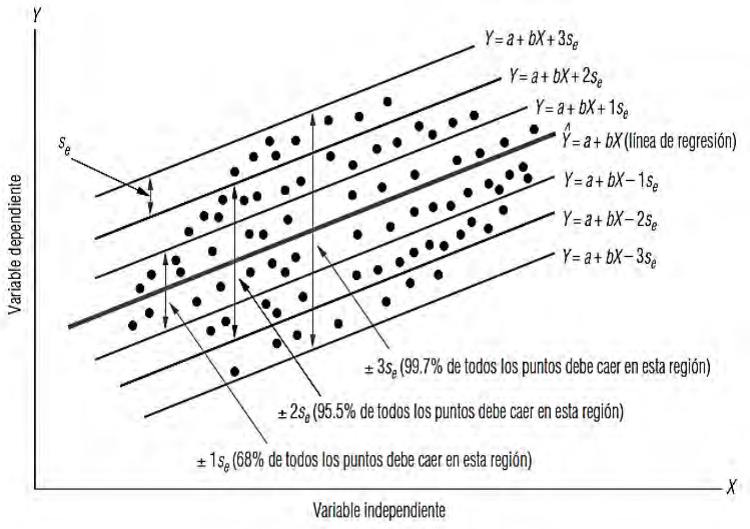
$$= \sqrt{\frac{20.535}{4}}$$

$$S_e = 2.266$$

S_e=2.266 millones de pesos
El 68% de los datos caerán dentro de ±1S_e, el 95% de los datos caerían dentro de ±2S_e y el 99.7% de los datos caerían dentro de ±3S_e, respeto a la distancia recta sobre el eje *y*, con respecto a la línea de estimación de la recta obtenida en el paso 6.

De una manera visual, el error estándar de la ecuación de regresión se muestra en la figura 6.4.

Figura 6.4. Límites alrededor de la línea de regresión



Intervalos de confianza para la estimación (o el valor esperado)

Podemos concebir al error estándar de la estimación como la herramienta estadística que podemos usar para hacer afirmaciones de probabilidad acerca del intervalo alrededor del valor estimado de $\hat{y}$ dentro del cual cae el valor real de y.

En la figura 6.4 podemos ver, por ejemplo, que hay una seguridad del 95.5% de que el valor real de y caerá dentro de dos errores estándar del valor estimado de $\hat{y}$. Llamamos a estos intervalos alrededor de la $\hat{y}$ estimada, intervalos de confianza para la estimación.

Ahora, aplicando el concepto de intervalos de confianza para la estimación al problema del vicepresidente de I+D, sabemos que la ecuación de estimación usada para predecir las ganancias anuales respecto a los gastos en I+D es:

$$\hat{y} = 15.982+3.043x$$

Y sabemos qué, si la empresa decide considerar un gasto en I+D de 6 millones de pesos, predecimos que tendrá una ganancia anual de $34.24 millones de pesos:

$$\hat{y} = 15.982+3.043x$$

$$\hat{y} = 15.982+3.043 * 6$$

$$\hat{y} = 34.24$$

Por último, recordar que calculamos el error estándar de la estimación como S_e = 2.266 ($2,266,00). Ahora podemos combinar estas dos piezas de información y decir que estamos seguros aproximadamente el 68% de las veces, la ganancia anual estará dentro de ± 1 error estándar de la estimación de $\hat{y}$. Podemos calcular los límites superior e inferior de este intervalo de confianza de las ganancias anuales.

Nota: Recuerde que los estadísticos aplican los intervalos de confianza para la estimación basados en la distribución normal (el 68% para 1S_e, el 95.5% para 2S_e y el 99.7% para 3S_e) sólo para muestras grandes, esto es, cuando n > 30 (de acuerdo al Teorema del límite central). Para el caso de este problema, nuestro tamaño de muestra es demasiado pequeño (n=6). Por tanto, nuestras conclusiones son inexactas. Si deseamos evitar inexactitudes ocasionadas por el tamaño de la muestra, necesitamos usar **la distribución t.** Recuerde que esta distribución t es apropiada cuando n es menor que 30 y la desviación estándar de la población no se conoce. Estas dos condiciones, se cumplen puesto que n=6, y S_e es una estimación y no la desviación estándar conocida de la población.

Si el vicepresidente de I+D desea tener una confianza del 90% de que las ganancias anuales caerán en el intervalo de estimación, ¿Cómo calculamos este intervalo? Como la distribución *t* que se muestra en el Apéndice 2, se concentra en la probabilidad de que el parámetro que estamos estimando caerá fuera del intervalo de predicción, necesitamos consultar en el Apéndice 2, en la columna de 100% - 90% = 10% (0.1). Una vez localizada la columna, buscamos el renglón para 4 grados de libertad; porque n= 6 y sabemos que perdemos 2 grados de libertad (al estimar los valores de a y b), entonces 6 - 2 = 4. Encontraremos que el valor apropiado de *t* es 2.132.

Ahora, usando este valor de t, podemos hacer un cálculo más exacto de los límites del intervalo de la estimación, de la siguiente manera:

$\hat{y} = 34.24$ $S_e = 2.266$ $t = 2.132$

$Límite\ superior \rightarrow \hat{y} + \mathrm{t}\,(S_e) = 34.24 + (2.132 * 2.266)$
$= 34.24 + 4.831$
$= 39.071$ → 39,071,000

$Límite\ inferior \rightarrow \hat{y} - \mathrm{t}\,(S_e) = 34.24 - (2.132 * 2.266)$
$= 34.24 - 4.831$
$= 29.409$ → 29,409,000

Así, el vicepresidente de I+D, puede estar 90% seguro de que las ganancias anuales, estarán entre $39,071,000 y $29, 409,000.

1.7 Ejercicios

Ejercicio 1.1

En cierto producto de prueba metálico se sabe que la tensión normal sobre un espécimen se relaciona funcionalmente con la resistencia al corte. El siguiente es un conjunto de datos experimentales codificados para esas dos variables.

a) Traza el diagrama de dispersión y haz una conclusión del mismo.
b) Determina la ecuación de regresión lineal
c) Si se tiene una tensión norma de 24 ¿cuál será el valor de su resistencia al corte?
d) Estima el valor de la resistencia al corte si la tensión normal es igual a 30.
e) Calcula el coeficiente de determinación y menciona si el modelo es confiable para hacer predicciones
f) Calcula el coeficiente de correlación y determina el tipo de relación entre las variables.
g) Calcula el error estándar y obtén el intervalo de confianza de la resistencia al corte si se somete a una tensión normal de 23 kg/cm^2, considerando un nivel de confianza del 90%.

Tensión normal (kg/cm^2)	Resistencia al corte (kg/cm^2)
26.8	26.5
25.4	27.3
28.9	24.2
23.6	27.1
27.7	23.6
23.9	25.9
24.7	26.3
28.1	22.5
26.9	21.7
27.4	21.4
22.6	25.8
25.6	25.6

Ejercicio 1.2

Una compañía administra a sus vendedores en capacitación una prueba de ventas antes de salir a trabajar. La administración de la compañía está interesada en determinar la relación entre las calificaciones de la prueba y las ventas logradas por esos vendedores al final de un año de trabajo. Se recolectaron los siguientes datos de 10 agentes de ventas que han estado en el campo un año.

Número de vendedor	Calificación de la prueba (T)	Número de unidades vendidas (S)
1	2.6	95
2	3.7	140
3	2.4	85
4	4.5	180
5	2.6	100
6	5	195
7	2.8	115
8	3	136
9	4	175
10	3.4	150

a) Traza el diagrama de dispersión y haz una conclusión del mismo.
b) Determina la ecuación de regresión lineal
c) ¿Cuál sería el número estimado de unidades vendidas de un vendedor considerando que obtuvo una calificación de 4.2?
d) Calcula el coeficiente de determinación y menciona si el modelo es confiable para hacer predicciones.
e) Calcula el coeficiente de correlación y determina el tipo de relación entre las variables.
f) Calcula el error estándar y obtén el intervalo de confianza de las unidades vendidas por un vendedor cuya calificación obtenida fue de 4.2, considerando un nivel de confianza del 95%.

2. Regresión lineal múltiple y correlación

2.1 Cuándo utilizar la regresión lineal múltiple

En el tema seis de regresión lineal simple, recordemos que la ecuación que se obtuvo para estimar las ganancias anuales de acuerdo al gasto de I+D, tuvo un valor bueno del coeficiente de determinación (r^2) de 87.4%, pero, si se hubiera obtenido un valor por ejemplo de 56%, la ecuación de estimación que se obtuvo no sería adecuada para estimar las ganancias anuales, ¿que se podría hacer en ese caso?.

Una opción es utilizar otra variable independiente para estimar la variable dependiente e intentar, de esta manera, aumentar la precisión de la estimación, por ejemplo, agregar la variable de gasto en publicidad de la empresa. Este proceso se conoce como análisis de regresión múltiple y correlación. Está basado en las mismas suposiciones y procedimientos que encontramos al utilizar la regresión simple.

Ciertamente, podemos encontrar una ecuación de estimación sencilla que relacione a estas dos variables. ¿Podemos también hacer más precisa nuestra ecuación incluyendo en el proceso de estimación el gasto en I+D y el gasto en mercadotecnia? Probablemente la respuesta sea sí. Y ahora, como deseamos utilizar esas dos variables independientes para predecir las ganancias anuales, debemos utilizar regresión múltiple, no simple, para determinar la relación.

La principal ventaja de la regresión múltiple es que nos permite utilizar más información disponible para estimar la variable dependiente. En algunas ocasiones, la correlación entre dos variables puede resultar insuficiente para determinar una ecuación de estimación confiable; sin embargo, si agregamos los datos de más variables independientes, podemos determinar una ecuación de estimación que describa la relación con mayor precisión.

La regresión múltiple y el análisis de correlación implican un proceso de tres etapas, que son:

1. Describimos la ecuación de regresión múltiple;
2. Utilizamos el análisis de correlación múltiple para determinar qué tan bien la ecuación de regresión describe los datos observados.
3. Examinamos el error estándar de regresión múltiple de la estimación

Además, en la regresión múltiple podemos observar cada una de las variables independientes y probar si contribuyen de manera significativa a la forma en que la regresión describe los

datos. En este capítulo, veremos cómo encontrar la ecuación de regresión de mejor ajuste para un conjunto dado de datos, y cómo analizar la ecuación obtenida.

2.2 Análisis de regresión múltiple y correlación

En el siguiente ejemplo, se describirán las tres etapas y los pasos que deben realizarse para desarrollar un análisis de regresión múltiple y correlación.

Etapa 1. Describir la ecuación de regresión múltiple

Con el siguiente ejemplo, se mostrará como calcular la ecuación de regresión múltiple. Por conveniencia, utilizaremos sólo dos variables independientes en el problema que trabajaremos. Sin embargo, se debe tener en mente que, en principio, la misma clase de técnica se aplica a cualquier número de variables independientes.

El Servicio de Administración Tributaria (SAT) está tratando de estimar la cantidad mensual de impuestos no pagados descubiertos por su departamento de auditorías. En el pasado, el SAT estimaba esta cantidad con base en el número esperado de horas de trabajo de auditorías de campo. En los últimos años, sin embargo, las horas de trabajo de auditorías de campo se han convertido en un pronosticador errático de los impuestos no pagados reales. Como resultado, la dependencia está buscando otro factor para mejorar la ecuación de estimación.

El departamento de auditorías tiene un registro del número de horas que usa sus computadoras para detectar impuestos no pagados. ¿Podríamos combinar esta información con los datos referentes las horas de trabajo de auditorías de campo y obtener una ecuación de estimación más precisa para los impuestos no pagados descubiertos cada mes? En el cuadro 7.1 se presentan esos datos para los últimos 10 meses.

Cuadro 7.1. Datos recolectados en 10 meses por el SAT

Mes	x_1 Horas de trabajo de auditorías de campo (se omiten 2 ceros)	x_2 Horas en computadora (se omiten 2 ceros)	y Impuestos reales no pagados descubiertos (millones de pesos)
Enero	45	16	29
Febrero	42	14	24
Marzo	44	15	27
Abril	45	13	25
Mayo	43	13	26
Junio	46	14	28
Julio	44	16	30
Agosto	45	16	28
Septiembre	44	15	28
Octubre	43	15	27

En este problema, x_1 representa el número de horas de trabajo de auditoría de campo y x_2 el número de horas en computadora. La variable dependiente, *y*, será los impuestos reales no pagados descubiertos. La fórmula que se usa cuando tenemos dos variables independientes es la siguiente:

$$\hat{y} = \mathrm{a} + b_1 x_1 + b_2 x_2 \tag{7.1}$$

Donde:

$\hat{y}$ = valor estimado correspondiente a la variable dependiente

a =ordenada de *y*

x_1, x_2 = valores de las dos variables independientes

b_1, b_2 = pendientes asociadas con x_1, x_2 respectivamente.

Para hallar ahora los valores de a, b_1, b_2 se ocupará nuevamente el Método de mínimos cuadrados, ocupando las siguientes tres ecuaciones para determinar los valores de dichas constantes numéricas.

$$\sum y = na \quad + b_1 \sum x_1 \quad + \ b_2 \sum x_2 \tag{7.2}$$

$$\sum x_1 y = a \sum x_1 \ + b_1 \sum x_1^2 \quad + \ b_2 \sum x_1 x_2 \tag{7.3}$$

$$\sum x_2 y = a \sum x_2 + b_1 \sum x_1 x_2 + \ b_2 \sum x_2^2 \tag{7.4}$$

Con las ecuaciones 7.2 a 7.4 se procederá a utilizarlas para obtener la ecuación de estimación de regresión lineal múltiple para dos variables independientes. Los pasos para obtener dicha ecuación son los siguientes.

Paso 1. Realizar los cálculos del cuadro 7.2

Realizar los cálculos indicados en los encabezados del cuadro 7.2

Se describen los cálculos de la primera fila:

$x_1{}^2 =$ 45*45= 2025 $x_2{}^2$ = 16*16=256

$(x_1 x_2)$ = 45*16= 720 $y x_1$ = 29*45=1305

$y x_2$ = 29*16= 464

Realizar el resto de cálculos y las sumatorias de cada columna; verificar con los resultados del cuadro 7.2

Cuadro 7.2 Cálculos de valores para el análisis de regresión lineal múltiple

	y	x_1	x_2	x_1^2	x_2^2	$x_1 x_2$	$y x_1$	$y x_2$
	29	45	16	2025	256	720	1305	464
	24	42	14	1764	196	588	1008	336
	27	44	15	1936	225	660	1188	405
	25	45	13	2025	169	585	1125	325
	26	43	13	1849	169	559	1118	338
	28	46	14	2116	196	644	1288	392
	30	44	16	1936	256	704	1320	480
	28	45	16	2025	256	720	1260	448
	28	44	15	1936	225	660	1232	420
	27	43	15	1849	225	645	1161	405
Σ=	**272**	**441**	**147**	**19461**	**2173**	**6485**	**12005**	**4013**
	$\sum y$	$\sum x_1$	$\sum x_2$	$\sum x_1^2$	$\sum x_2^2$	$\sum x_1 x_2$	$\sum y x_1$	$\sum y x_2$

Paso 2. Sustituir los valores obtenidos de las sumatorias del cuadro 7.2, en las ecuaciones normales (7.2, 7.3 y 7.4) que contienen las constantes numéricas *a, b_1 y b_2*.

$$\sum y = na + b_1 \sum x_1 + b_2 \sum x_2 \rightarrow 272 = 10\ a + 441\ b_1 + 147\ b_2$$

$$\sum x_1 y = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 \rightarrow 12005 = 441\ a + 19461\ b_1 + 6485\ b_2$$

$$\sum x_2 y = a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 \rightarrow 4013 = 147\ a + 6485\ b_1 + 2173\ b_2$$

Se obtiene un sistema de ecuaciones lineales con tres constantes numéricas desconocidas, *a, b_1 y b_2.*

Paso 3. Resolver el sistema de ecuaciones lineales que contienen las constantes numéricas *a, b_1 y b_2*, obtenido en el paso 2.

Para resolver el sistema de ecuaciones lineales, ocuparemos el Método de Cramer o Regla de Cramer para resolver un sistema de ecuaciones lineales de: 3 ecuaciones lineales por 3 variables (3x3). Para hallar los valores de *a, b_1 y b_2,* ocuparemos las ecuaciones 7.5, 7.6 y 7.7 respectivamente.

$$a = \frac{D_1}{D} \qquad (7.5) \qquad\qquad b_1 = \frac{D_2}{D} \qquad (7.6) \qquad\qquad b_2 = \frac{D_3}{D} \qquad (7.7)$$

donde:

D= Determinante principal

D_1= Determinante uno

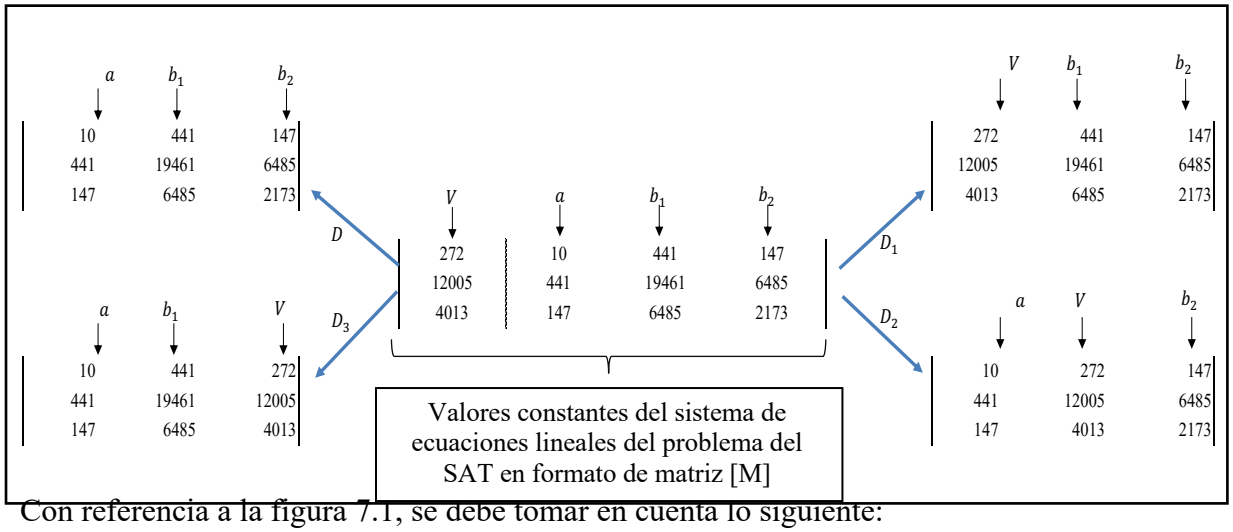
D_2= Determinante dos

D_3= Determinante tres

Realizaremos los siguientes pasos para hallar los valores de cada determinante y posteriormente, hallar los valores de las constantes numéricas *a, b_1 y b_2.*

Paso 3.1 Del sistema de ecuaciones lineales obtenido del paso 2, obtener los componentes de cada determinante, para su posterior cálculo. En la figura 7.1 se muestra cómo queda cada determinante *D, D_1, D_2 y D_3.*

Figura 7.1 Obtención de los componentes de los 4 determinantes, con base en el sistema de ecuaciones de 3x3.



Con referencia a la figura 7.1, se debe tomar en cuenta lo siguiente:

a) El determinante *D*, se forma agregando en las columnas 1, 2 y 3, los valores de las columnas *a, b_1 y b_2* obtenidos de la matriz [M].

b) El determinante D_1 se forma agregando en la columna 1, los valores de la columna *V* de [M]; los valores de las columnas 2 y 3, se toman de las columnas b_1 *y* b_2 de [M]

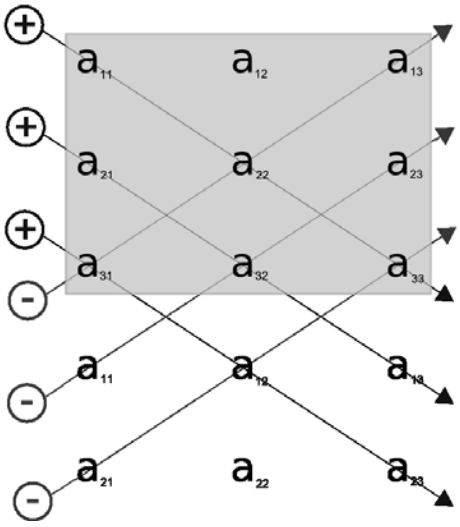
c) El determinante D_2, se forma agregando en la columna 1, los valores de la columna *a* de [M]; los valores de la columna 2 de toman de los valores de la columna *V* de [M]; los valores de la columna *3,* se toman de las columnas b_2 de [M].

d) El determinante D_3, se forma agregando en la columna 1 y 2, los valores de las columnas *a* y b_1 de [M] respectivamente; los valores de la columna *3,* se toman de la columna *V* de [M].

Paso 3.2 Se procede a calcular el valor de cada determinante: *D, D_1, D_2 y D_3.*

Para obtener los valores de los 4 determinantes, se ocupará la Regla de Sarrus para resolver un determinante de 3x3 (3 filas x 3 columnas). La Regla de Sarrus se muestra en la figura 7.2

Figura 7.2 Regla de Sarrus para resolver un determinante de 3x3



De acuerdo a la figura 7.2, para resolver un determinante de 3x3, se repiten en la parte inferior las dos primeras filas, en la parte inferior de los componentes del determinante. Realizando una multiplicación de izquierda a derecha entre los valores en diagonal, considerando lo siguiente:

a) Los valores que se multiplican de arriba hacia abajo, ejemplo: valores de las posiciones a11, a22 y a33 se multiplican por (+) al inicio de la multiplicación.

b) Los valores que se multiplican de abajo hacia arriba, ejemplo: valores de las posiciones a31, a22 y a13 se multiplican por (-) al inicio de la multiplicación. Puede ser al final de la multiplicación también por el signo (-).

Siguiendo la Regla de Sarrus, se procede a realizar las multiplicaciones de cada determinante *D, D_1, D_2 y D_3* para hallar el valor de cada uno. Se muestran a continuación los cálculos hechos para cada determinante.

$$D = \begin{vmatrix} 10 & 441 & 147 \\ 441 & 19461 & 6485 \\ 147 & 6485 & 2173 \\ 10 & 441 & 147 \\ 441 & 19461 & 6485 \end{vmatrix}$$

a) 10 * 19461 * 2173 = 422887530
b) 441 * 6485 * 147 = 420403095
c) 147 * 441 * 6485 = 420403095
d) (147 * 19461 * 147) (-1) = -420532749
e) (10 * 6485 * 6485) (-1) = -420552250
f) (441 * 441 * 2173) (-1) = -422607213

$$D = 1508$$

$$D_1 = \begin{vmatrix} 272 & 441 & 147 \\ 12005 & 19461 & 6485 \\ 4013 & 6485 & 2173 \\ 272 & 441 & 147 \\ 12005 & 19461 & 6485 \end{vmatrix}$$

a) 272 * 19461 * 2173 = 11502540816
b) 12005 * 6485 * 147 = 11444306475
c) 4013 * 441 * 6485 = 11476718505
d) (4013 * 19461 * 147) (-1) = -11480257971
e) (272 * 6485 * 6485) (-1) = -11439021200
f) (12005 * 441 * 2173) (-1) = -11504307465

$$D_1 = -20840$$

$$D_2 = \begin{vmatrix} 10 & 272 & 147 \\ 441 & 12005 & 6485 \\ 147 & 4013 & 2173 \\ 10 & 272 & 147 \\ 441 & 12005 & 6485 \end{vmatrix}$$

a) 10 * 12005 * 2173 = 260868650
b) 441 * 4013 * 147 = 260150751
c) 147 * 272 * 6485 = 259296240
d) (147 * 12005 * 147) (-1) = -259416045
e) (10 * 4013 * 6485) (-1) = -260243050
f) (441 * 272 * 2173) (-1) = -260655696

$$D_2 = 850$$

$$D_3 = \begin{vmatrix} 10 & 441 & 272 \\ 441 & 19461 & 12005 \\ 147 & 6485 & 4013 \\ 10 & 441 & 272 \\ 441 & 19461 & 12005 \end{vmatrix} = \begin{array}{lll} a)\ 10 * 19461 * 4013 & = & 780969930 \\ b)\ 441 * 6485 * 272 & = & 777888720 \\ c)\ 147 * 441 * 12005 & = & 778248135 \\ d)\ (147 * 19461 * 272)(-1) & = & -778128624 \\ e)\ (10 * 6485 * 12005)(-1) & = & -778524250 \\ f)\ (441 * 441 * 4013)(-1) & = & -780452253 \end{array} = 1658$$

Paso 3.3 Obtener los valores de las constantes numéricas *a, b_1 y b_2.*

Ocuparemos las ecuaciones 7.5, 7.6 y 7.7 para obtener los valores:

$$a = \frac{D_1}{D} = \frac{-20840}{1508} = -13.82$$

$$b_1 = \frac{D_2}{D} = \frac{850}{1508} = 0.564$$

$$b_2 = \frac{D_3}{D} = \frac{1658}{1508} = 1.099$$

Paso 4. Obtener la ecuación de estimación de regresión lineal múltiple.

Se sustituyen los valores obtenidos de *a, b_1 y b_2* en la ecuación 7.1

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

$$\hat{y} = -13.82 + 0.564\ x_1 + 1.099\ x_2$$

Se obtiene la ecuación de regresión lineal múltiple para poder estimar los impuestos reales no pagados descubiertos (variable dependiente *y*), de acuerdo a las horas de trabajo en campo B_1 y las horas de trabajo en computadora B_2.

Suponga que, en noviembre, el SAT intenta dejar las horas de trabajo en auditorías de campo y las horas en computadora en sus niveles de octubre (4,300 y 1,500). ¿Cuánto de impuestos no pagados esperan descubrir en noviembre? Sustituyendo estos valores en la ecuación de regresión estimada, obtenemos:

$$\hat{y} = -13.82 + 0.564\ x_1 + 1.099\ x_2$$

$$x_1 = 43$$

$$x_2 = 15$$

$$\hat{y} = -13.82 + 0.564 * 43 + 1.099 * 15$$

$$\hat{y} = 26.917$$

Impuestos no pagados estimados descubiertos 26,917,000

De modo que el departamento de auditorías espera descubrir aproximadamente $27 millones de evasión de impuestos en noviembre, considerando 4,300 horas de trabajo en auditorías de campo (43) y 1500 horas en computadora (15).

Etapa 2. Realizar el análisis de correlación múltiple para determinar qué tan bien la ecuación de regresión describe los datos observados.

Al estudiar el análisis de correlación simple, medimos la fuerza de la relación entre dos variables, utilizando el coeficiente de determinación de la muestra, r^2. Este coeficiente de determinación es la fracción de la variación total de la variable dependiente *y* que se explica con la ecuación de estimación.

Similarmente, en la correlación múltiple mediremos la fuerza de la relación entre tres variables utilizando el coeficiente de determinación múltiple, $\mathbf{r^2}$, o su raíz cuadrada, **r** (el coeficiente de correlación múltiple). Este coeficiente de determinación múltiple es también la fracción que representa la porción de la variación total de ***y*** que "explica" el plano de regresión. Para calcular el coeficiente de determinación múltiple $\mathbf{r^2}$ y el coeficiente de correlación múltiple **r**, se ocuparán las siguientes ecuaciones.

$$r^2 = \frac{SCR}{SCT} = \frac{\sum\left(\hat{y}_i - \bar{y}\right)^2}{\sum(y - \bar{y})^2} \tag{7.8}$$

$$r = \pm\sqrt{r^2} = \tag{7.9}$$

Los pasos para hallar ambos coeficientes son los siguientes:

Paso 5. Realizar los cálculos para obtener los valores que se requieren para sustituir en las ecuaciones 7.8 y 7.9, y hallar el valor del coeficiente de determinación y correlación múltiple.

En el cuadro 7.3 se muestran las operaciones de cada una de las columnas que se requieren calcular para obtener el coeficiente de determinación y coeficiente de correlación.

Cuadro 7.3 Cálculos de valores para el análisis de correlación y error estándar de la regresión

	y	x_1	x_2	$\hat{y}_i$	$\sum(\hat{y}_i - \bar{y})^2$	$\sum(y - \bar{y})^2$	$\sum(y - \hat{y})^2$
	29	45	16	-13.82 + 0.564 * 45 + 1.099 * 16 = 29.144	3.779	3.24	0.021
	24	42	14	-13.82 + 0.564 * 42 + 1.099 * 14 = 25.254	3.787	10.24	1.573
	27	44	15	-13.82 + 0.564 * 44 + 1.099 * 15 = 27.481	0.079	0.04	0.231
	25	45	13	-13.82 + 0.564 * 45 + 1.099 * 13 = 25.847	1.831	4.84	0.717
	26	43	13	-13.82 + 0.564 * 43 + 1.099 * 13 = 24.719	6.155	1.44	1.641
	28	46	14	-13.82 + 0.564 * 46 + 1.099 * 14 = 27.51	0.096	0.64	0.24
	30	44	16	-13.82 + 0.564 * 44 + 1.099 * 16 = 28.58	1.904	7.84	2.016
	28	45	16	-13.82 + 0.564 * 45 + 1.099 * 16 = 29.144	3.779	0.64	1.309
	28	44	15	-13.82 + 0.564 * 44 + 1.099 * 15 = 27.481	0.079	0.64	0.269
	27	43	15	-13.82 + 0.564 * 43 + 1.099 * 15 = 26.917	0.08	0.04	0.007
∑=	**272**				**21.569**	**29.6**	**8.024**

$$\bar{y} = \frac{\sum y}{n} = \quad 27.2$$

Con los valores obtenidos del cuadro 7.3, sustituirlos en las ecuaciones 7.8 y 7.9 como a continuación se muestra.

$$r^2 = \frac{SCR}{SCT} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{21.569}{29.6} = 0.729$$

$$r = \pm\sqrt{r^2} = \sqrt{0.729} = 0.854$$

El valor del coeficiente de determinación r^2= 0.729 nos indica que las dos variables independientes (horas de trabajo en auditoría de campo y las horas en computadora), explican el 72.9% de la variación total de impuestos no pagados descubiertos. Por otra parte, el coeficiente de correlación r= 0.854, indica que las horas de trabajo en auditoría de campo y las horas en computadora, junto con él total de impuestos no pagados descubiertos, se relacionan un 85.4%. Como el coeficiente de determinación r^2 es menor a 80%, se sugiere agregar otra variable o variables que permitan una ecuación de regresión múltiple que explique un porcentaje mayor de la variación del total de impuestos no pagados descubiertos.

Etapa 3. Examinar el error estándar de la estimación de la regresión múltiple

Ya que hemos determinado la ecuación que relaciona a nuestras tres variables, y sabemos el porcentaje al que responde dicha ecuación de regresión múltiple a la variabilidad de la variable dependiente, necesitamos una medida de la dispersión alrededor de este plano de regresión múltiple. En la regresión simple, la estimación es más precisa conforme el grado de dispersión alrededor de la regresión es menor. Lo mismo es cierto para los puntos de la muestra que se encuentran alrededor del plano de regresión múltiple. Para medir esta variación, debemos utilizar de nuevo la medida conocida como error estándar de la estimación, la ecuación para obtenerla es la siguiente:

$$S_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - k - 1}} \tag{7.10}$$

donde

y = valores muestrales de la variable dependiente
$\hat{y}$ = valores correspondientes estimados con la ecuación de regresión
n= número de puntos de la muestra
k= número de variables independientes (3 en el ejemplo que estamos desarrollando)

Paso 6. Obtener el error estándar de la estimación de la regresión múltiple

Para calcular S_e, observamos los errores individuales en el plano de regresión ajustado $(y - \hat{y})$, los cuales se pueden visualizar en el cuadro 2.3, los elevamos al cuadrado, calculamos su media (dividiendo entre n - k - 1 en lugar de n) y tomamos la raíz cuadrada del resultado. Debido a la forma en que se calcula, S_e se conoce a veces como raíz del error cuadrático medio [o raíz de mse (mean-square error)]. A continuación, se muestran las operaciones realizadas para obtener S_e.

k= 2

$$S_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - k - 1}} = \sqrt{37} =$$

$$\frac{8.024}{10 - 2 - 1} \qquad 1.071$$

El valor obtenido del error estándar de la estimación fue de 1.071, es decir, la ecuación de estimación obtenida en el paso 4, tiene un error de estimación de 1,071,000 (un millón setenta y un mil pesos), recordar que la columna de impuestos no pagados en el cuadro 7.1 del problema indica que son millones de pesos.

¿Cuál es la utilidad del valor S_e?

Con el valor obtenido de S_e y la distribución *t*, podemos obtener el intervalo de confianza alrededor del valor estimado de $\hat{y}$. En el problema se estimó qué, para 4,300 horas de trabajo en auditorías de campo y 1,500 horas en computadora, los impuestos no pagados descubiertos ($\hat{y}$) se calcularon en $26,917,000, además el valor de S_e fue de $1,071,000.

Si deseamos construir un intervalo de confianza del 95% alrededor de esta estimación de $26,917,000, miramos en el Apéndice 2, en la columna del 5% (0.005), localizamos ahí el renglón correspondiente a $n- k-1$ = 10-2-1= 7 grados de libertad. El valor apropiado de t para nuestra estimación del intervalo es de 2.365. En consecuencia, podemos calcular los límites de nuestro intervalo de confianza como sigue:

$\hat{y} =$ 26.917 $\qquad S_e =$ 1.071 $\qquad t =$ 2.365

$$Límite\ superior \rightarrow \hat{y} + \mathrm{t}\,(S_e) = 26.917 + (2.365 * 1.071)$$
$$= 26.917 + 2.533$$
$$= 29.45$$

29,450,000

$$Límite\ inferior \rightarrow \hat{y} - \mathrm{t}\,(S_e) = 26.917 - (2.365 * 1.071)$$
$$= 26.917 - 2.533$$
$$= 24.384$$

24,384,000

Con un nivel de confianza del 95%, el departamento de auditorías del SAT, puede sentirse seguro de que los descubrimientos reales estarán en este intervalo, que va de $24,384,000 a $29,450,000.

Si el SAT desea usar un nivel de confianza menor, como 90%, puede reducir el intervalo de valores para la estimación de descubrimientos de impuestos no pagados. Igual que con la regresión simple, podemos utilizar la distribución normal estándar (Apéndice 1) para aproximar la distribución t siempre que los grados de libertad (n menos el número de coeficientes de regresión estimados) sea un número mayor que 30.

2.3 Ejercicios y/o actividades para evaluar con fechas de entrega

Ejercicio 2.1

Se piensa que la energía eléctrica consumida mensualmente por una empresa se relaciona con la temperatura ambiente promedio y las toneladas del producto producido. Los datos del último año se muestran en el cuadro de abajo.

Temperatura ambiente promedio (°C)	Toneladas producidas	Energía Eléctrica Consumida (kW*h)
12.5	16	29
15	14	24
22.5	15	27
30	13	25
32.5	15	27

a) Determina la ecuación de regresión lineal múltiple.

b) Calcula los coeficientes de correlación de las variables independientes (r^2 y r), escribe una conclusión de ambos coeficientes.

e) Calcula el error de estimación.

f) Estima el valor de la energía consumida, si la temperatura es 25°C y las toneladas producidas son 20.

g) Determina el intervalo de confianza de la energía consumida, tomando los datos del inciso f y considerando un nivel de confianza de 90%.

Ejercicio 2.2

La Reserva Federal de Estados Unidos realiza un estudio preliminar para determinar la relación entre ciertos indicadores económicos y el cambio porcentual anual en el producto interno bruto (PIB). Dos de los indicadores examinados son el monto del déficit del gobierno

federal (en miles de millones de dólares) y el promedio industrial Dow Jones (el valor medio del año). Los datos correspondientes a seis años son:

Cambio en el PIB (y) (%)	Déficit Federal (x_1) (miles de millones de dólares)	Down Jones (x_2) (miles)
2.5	100	2.85
-1.0	400	2.10
4.0	120	3.30
1.0	200	2.40
1.5	180	2.55
3.0	80	2.70

a) Determina la ecuación de regresión lineal múltiple.

b) Calcula los coeficientes de correlación de las variables independientes (r^2 y r), escribe una conclusión de ambos coeficientes.

e) Calcula el error de estimación.

f) ¿Qué porcentaje de cambio en el PIB se esperaría en un año en el cual el déficit federal fue 240,000 millones de dólares y el promedio Dow Jones fue 3,000?

g) Determina el intervalo de confianza del cambio porcentual en el PIB, tomando los datos del inciso f y considerando un nivel de confianza de 95%.

3. Análisis de series de tiempo

3.1 Modelos de series de tiempo

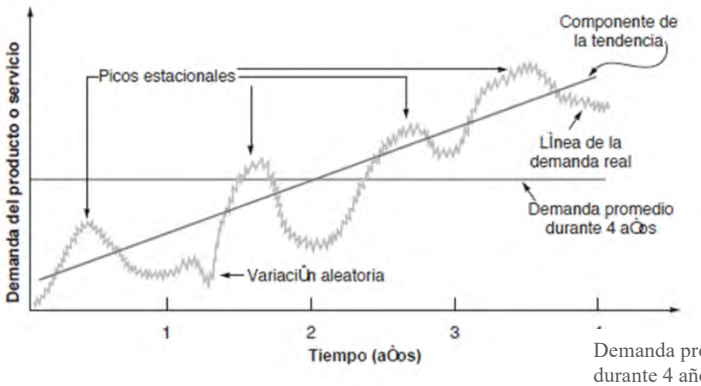
Los modelos de series de tiempo predicen bajo el supuesto de que el futuro es una función del pasado. En otras palabras, observan lo que ha ocurrido durante un periodo y usan una serie de datos históricos para hacer un pronóstico. Si estamos pronosticando las ventas semanales de cortadoras de césped, utilizamos datos de las ventas pasadas de cortadoras de césped para hacer el pronóstico.

Una serie de tiempo se basa en una secuencia de datos puntuales igualmente espaciados (semanales, mensuales, trimestrales, etc.). Los ejemplos incluyen las ventas semanales de Nike Air Jordans, los informes de ingresos trimestrales en Microsoft, los embarques diarios de cerveza Coors, y los índices anuales de precios al consumidor.

Los datos para pronósticos de series de tiempo implican que los valores futuros se predicen solamente a partir de los valores pasados y que se pueden ignorar otras variables, sin importar qué tan potencialmente valiosas sean.

En la figura 8.1 se ilustra una demanda en un periodo de 4 años. Se muestra el promedio, la tendencia, las componentes estacionales y las variaciones aleatorias alrededor de la curva de demanda. La demanda promedio es la suma de la demanda medida en cada periodo y dividida entre el número de periodos con datos.

Figura 8.1 Ejemplo de una gráfica de demanda



Demanda promedio durante 4 años

Variación aleatoria

Tiempo (años)

3.2 Método de Promedios Móviles

El pronóstico de promedios móviles usa un número de valores de datos históricos reales para generar un pronóstico. Los promedios móviles son útiles si podemos suponer que la demanda del mercado permanecerá relativamente estable en el tiempo. Un promedio móvil de 4 meses se encuentra simplemente al sumar la demanda medida durante los últimos 4 meses y dividiéndola entre cuatro. Al concluir cada mes, los datos del mes más reciente se agregan a la suma de los 3 meses previos y se elimina el dato del mes más antiguo. Esta práctica tiende a suavizar las irregularidades del corto plazo en las series de datos.

Matemáticamente, el promedio móvil simple (que sirve como estimación de la demanda del siguiente periodo) se expresa como:

$$Promedio\ móvil = \frac{\sum Demanda\ de\ los\ n\ periodos\ previos}{n} \qquad (8.1)$$

donde n es el número de periodos incluidos en el promedio móvil —por ejemplo, 4, 5 o 6 meses, respectivamente, para un promedio móvil de 4, 5 o 6 periodos. En el ejemplo siguiente se muestra cómo calcular los promedios móviles.

Ejemplo 1

La tienda de suministros para jardín Ferresol quiere hacer un pronóstico con el promedio móvil de 3 meses, incluyendo un pronóstico para las ventas de cobertizos el próximo enero.

Método: Las ventas de cobertizos para almacenamiento se muestran en la columna media de la tabla que se encuentra en la parte superior de la próxima página. A la derecha se presenta un promedio móvil de 3 meses. Calcular el resto de pronósticos mediante promedios móviles para los siguientes meses y para enero del siguiente año.

Mes	Ventas reales de cobertizos	Promedio móvil de tres meses
Enero	10	
Febrero	12	
Marzo	13	
Abril	16	$(10 + 12 + 13)/3 = 11\frac{2}{3}$
Mayo	19	$(12 + 13 + 16)/3 = 13\frac{2}{3}$
Junio	23	$(13 + 16 + 19)/3 = 16$
Julio	26	$(16 + 19 + 23)/3 = 19\frac{1}{3}$
Agosto	30	$(19 + 23 + 26)/3 = 22\frac{2}{3}$
Septiembre	28	$(23 + 26 + 30)/3 = 26\frac{1}{3}$
Octubre	18	$(26 + 30 + 28)/3 = 28$
Noviembre	16	$(30 + 28 + 18)/3 = 25\frac{1}{3}$
Diciembre	14	$(28 + 18 + 16)/3 = 20\frac{2}{3}$

Para enero el resultado sería: $\frac{18 + 16 + 14}{3} = 16$

Razonamiento: Ahora la administración tiene un pronóstico que promedia las ventas para los últimos 3 meses. Es fácil de usar y entender.

Cambios en los pronósticos: Si las ventas reales en diciembre fueran de 18 (en vez de 14), ¿cuál es el nuevo pronóstico para enero?

Respuesta: $17.33 \ \ (17\frac{1}{3})$

3.3 Promedios Móviles Ponderados

Cuando se presenta una tendencia o un patrón localizable, pueden utilizarse ponderaciones para dar más énfasis a los valores recientes. Esta práctica permite que las técnicas de pronóstico respondan más rápido a los cambios, puesto que puede darse mayor peso a los periodos más recientes. La elección de las ponderaciones es un tanto arbitraria porque no existe una fórmula establecida para determinarlas. Por lo tanto, decidir qué ponderaciones emplear requiere cierta experiencia. Por ejemplo, si el último mes o periodo se pondera demasiado alto, el pronóstico puede reflejar un cambio grande inusual, demasiado rápido en el patrón de demanda o de ventas.

Un promedio móvil ponderado puede expresarse matemáticamente como:

$$Promedio\ móvil\ ponderado = \frac{\sum(Ponderación\ para\ el\ periodod\ n)(Demanda\ de\ periodo\ n)}{\sum Ponderaciones} \qquad (8.2)$$

El ejemplo 2 muestra cómo calcular un promedio móvil ponderado.

La tienda de suministros para jardín de Donna quiere pronosticar las ventas de cobertizos ponderando los últimos 3 meses, dando más peso a los datos recientes para hacerlos más significativos.

Método: Se asigna más ponderación a los datos recientes, de la siguiente manera:

Ponderación Aplicada	Periodo
3	Último mes
2	Hace dos meses
1	Hace tres meses
6	Suma de ponderaciones

Pronósticos para este mes:

$$\frac{3\ xVentas\ del\ último\ mes + 2\ xVentas\ de\ hace\ 2\ meses + 1\ xVentas\ de\ hace\ 3\ meses}{\sum Ponderaciones}$$

Mes	Ventas reales de cobertizos	Promedio móvil ponderado de tres meses
Enero	10	
Febrero	12	
Marzo	13	
Abril	16	$[(3\times13)+(2\times12)+(10)]/6=12\frac{1}{6}$
Mayo	19	$[(3\times16)+(2\times13)+(12)]/6=14\frac{1}{3}$
Junio	23	$[(3\times19)+(2\times16)+(13)]/6=17$
Julio	26	$[(3\times23)+(2\times19)+(16)]/6=20\frac{1}{2}$
Agosto	30	$[(3\times26)+(2\times23)+(19)]/6=23\frac{5}{6}$
Septiembre	28	$[(3\times30)+(2\times26)+(23)]/6=27\frac{1}{2}$
Octubre	18	$[(3\times28)+(2\times30)+(26)]/6=28\frac{1}{3}$
Noviembre	16	$[(3\times18)+(2\times28)+(30)]/6=23\frac{1}{3}$
Diciembre	14	$[(3\times16)+(2\times18)+(28)]/6=18\frac{2}{3}$

Para enero el resultado sería: $\frac{(3*14)+(2*16)+18}{6}=15.33$

Razonamiento: En esta situación particular de pronóstico, se observa que cuanto más se pondera el último mes, la proyección que se obtiene es mucho más precisa.

Cambios en los pronósticos: Si las ponderaciones asignadas fueran 4, 2 y 1 (en lugar de 3, 2 y 1), ¿cuál es el pronóstico para enero con el promedio móvil ponderado?

Respuesta: $(15\frac{1}{7})$

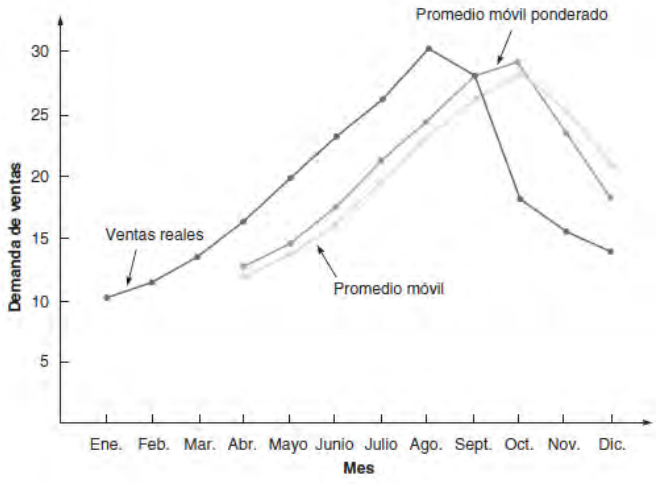
Tanto los promedios móviles simples como los ponderados son efectivos para suavizar las fluctuaciones repentinas en el patrón de la demanda con el fin de obtener estimaciones estables. Sin embargo, los promedios móviles presentan tres problemas:

1. Aumentar el tamaño de n (el número de periodos promediados) suaviza de mejor manera las fluctuaciones, pero resta sensibilidad al método ante cambios reales en los datos.
2. Los promedios móviles no reflejan muy bien las tendencias. Porque son promedios, siempre se quedarán en niveles pasados, no predicen los cambios hacia niveles más altos ni más bajos. Es decir, retrasan los valores reales.

3. Los promedios móviles requieren amplios registros de datos históricos.

En la figura 8.2, una gráfica de los datos de los ejemplos 1 y 2, se ilustra el efecto de retraso de los modelos de promedios móviles. Observe que tanto las líneas de los promedios móviles simples como las de promedios móviles ponderados retrasan la demanda real. Sin embargo, los promedios móviles ponderados usualmente reaccionan más rápido ante los cambios detectados en la demanda. Incluso en periodos a la baja (vea noviembre y diciembre), siguen la demanda de manera más cercana.

Figura 8.2 Comparativo entre promedio móvil y promedio móvil ponderado



Elección de ponderaciones

La experiencia y las pruebas son las formas más sencillas de elegir las ponderaciones. Por regla general, el pasado más reciente es el indicador más importante de lo que se espera en el futuro y, por lo tanto, debe tener una ponderación más alta. Los ingresos o la capacidad de la planta del mes pasado, por ejemplo, serían un mejor estimado para el mes próximo que los ingresos o la capacidad de la planta de hace varios meses.

No obstante, si los datos son estacionales, por ejemplo, las ponderaciones se deben establecer en forma correspondiente. Las ventas de trajes de baño en julio del año pasado deben tener una ponderación más alta que las ventas de trajes de baño en diciembre (en el hemisferio norte).

3.4 Método de Suavizamiento Exponencial

El suavizamiento exponencial es un sofisticado método de pronóstico de promedios móviles ponderado que sigue siendo bastante fácil de usar. Implica mantener muy pocos registros de datos históricos. La fórmula básica para el suavizamiento exponencial se expresa como sigue:

Nuevo pronóstico = Pronóstico del periodo anterior + α (Demanda real del mes anterior – Pronóstico del periodo anterior)

donde α es la ponderación, o constante de suavizamiento, elegida por quien pronostica, que tiene un valor de entre 0 y 1.

En los métodos de pronósticos anteriores (promedios móviles simple y ponderado), la principal desventaja es la necesidad de manejar en forma continua gran cantidad de datos históricos (esto también sucede con las técnicas de análisis de regresión, que se estudiarán en breve). En estos métodos, al agregar cada nueva pieza de datos, se elimina la observación anterior y se calcula el nuevo pronóstico. En muchas aplicaciones (quizás en la mayor parte), las ocurrencias más recientes son más indicativas del futuro que aquellas en el pasado más distante. Si esta premisa es válida (que la importancia de los datos disminuye conforme el pasado se vuelve más distante), es probable que el método más lógico y fácil sea la suavización exponencial.

La razón por la que se llama suavización exponencial es que cada incremento en el pasado se reduce (1 − α). Por ejemplo, si α es 0.05, las ponderaciones para los distintos periodos serían las siguientes (α se define a continuación):

	PESO EN $\alpha = 0.05$
Peso más reciente = $(1-\alpha)^0$	0.0500
Datos de un periodo anterior = $\alpha(1-\alpha)^1$	0.0457
Datos de dos periodos anteriores = $\alpha(1-\alpha)^2$	0.0451
Datos de tres periodos anteriores = $\alpha(1-\alpha)^3$	0.0429

Por lo tanto, los exponentes 0, 1, 2, 3,…, etc. le dan su nombre.

La suavización exponencial es la más utilizada de las técnicas de pronóstico. Es parte integral de casi todos los programas de pronóstico por computadora, y se usa con mucha frecuencia al ordenar el inventario en las empresas minoristas, las compañías mayoristas y las agencias de servicios.

Las técnicas de suavización exponencial se han aceptado en forma generalizada por seis razones principales:

- Los modelos exponenciales son sorprendentemente precisos.
- Formular un modelo exponencial es relativamente fácil
- El usuario puede entender cómo funciona el modelo.
- Se requieren muy pocos cálculos para utilizar el modelo.
- Los requerimientos de almacenamiento en la computadora son bajos debido al uso limitado de datos históricos.
- Es fácil calcular las pruebas de precisión relacionadas con el desempeño del modelo.

En el método de suavización exponencial, sólo se necesitan tres piezas de datos para pronosticar el futuro: el pronóstico más reciente, la demanda real que ocurrió durante el periodo de pronóstico y una constante de uniformidad alfa (α). Esta constante de suavización determina el nivel de uniformidad y la velocidad de reacción a las diferencias entre los pronósticos y las ocurrencias reales.

La constante de suavizamiento, α, se encuentra generalmente en un intervalo de .05 a .50 para aplicaciones de negocios. Puede cambiarse para dar más peso a datos recientes (cuando α es alta) o más peso a datos anteriores (si α es baja). Cuando α llega al extremo de 1.0, entonces en la ecuación $Ft = 1.0\ At\text{-}1$. Todos los valores anteriores se desechan y el pronóstico se vuelve idéntico al modelo intuitivo.

- En ocasiones, los usuarios del promedio móvil simple cambian a la suavización exponencial pero conservan las proyecciones similares a las del promedio móvil simple. En este caso, α se calcula 2 ÷ (*n* + 1), donde *n* es el número de periodos.

- La ecuación anterior también puede escribirse matemáticamente como:
  - $\boldsymbol{F_t = F_{t-1} + \propto (A_{t-1} - F_{t-1})}$

- donde F_t = nuevo pronóstico
- F_{t-1} = pronóstico del periodo anterior
- α = constante de suavizamiento (o ponderación) (0 ≤ α ≤ 1)
- A_{t-1} = demanda real en el periodo anterior.

El concepto no es complicado. La última estimación de la demanda es igual a la estimación anterior ajustada por una fracción de la diferencia entre la demanda real del último periodo y la estimación anterior. En el ejemplo 3 se muestra cómo usar el suavizamiento exponencial para obtener un pronóstico.

En enero, un vendedor de automóviles predijo que la demanda para febrero sería de 142 Ford Mustang. La demanda real en febrero fue de 153 automóviles. Usando la constante de suavizamiento que eligió la administración de α=0.20, el vendedor quiere pronosticar la demanda para marzo usando el modelo de suavizamiento exponencial.

$$\boldsymbol{F_t = F_{t-1} + \propto (A_{t-1} - F_{t-1})}$$

donde:

F_t= nuevo pronóstico

F_{t-1}= pronóstico del periodo anterior

α = constante de suavizamiento (o ponderación) (0 ≤ α ≤ 1)

A_{t-1}= demanda real en el periodo anterior

Solución: Al sustituir en la fórmula los datos de la muestra, se obtiene:

$$Nuevo\ pronóstico\ (marzo) = 142 + 0.2(153 - 142)$$
$$= 142 + 2.2 = 144.2$$

Así, el pronóstico de la demanda de marzo para los Ford Mustang se redondea a 144.
Razonamiento: Usando sólo dos elementos de datos, el pronóstico y la demanda real, más una constante de suavizamiento, se desarrolló un pronóstico de 144 Ford Mustang para marzo. Ejercicio de aprendizaje: Si la constante de suavizamiento se cambia a 0.30, ¿cuál es el nuevo pronóstico? Respuesta: 145.3

Ejemplo 4
Durante los últimos 8 trimestres, en el puerto de Baltimore se han descargado de los barcos grandes cantidades de grano. El administrador de operaciones del puerto quiere probar el uso de suavizamiento exponencial para ver qué tan bien funciona la técnica para predecir el tonelaje descargado. Supone que el pronóstico de grano descargado durante el primer trimestre fue de 175 toneladas. Se examinan dos valores de α = .10 y α= .50. En el cuadro siguiente se muestran los datos del problema:

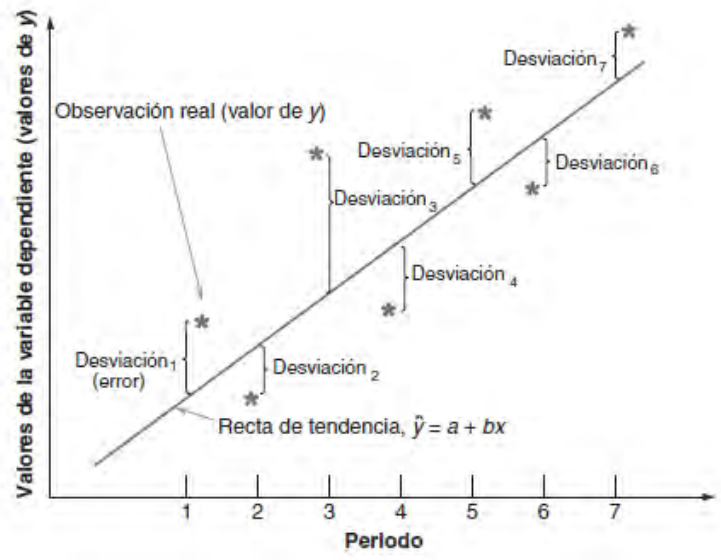
Trimestre	Tonelaje real descargado	Pronóstico con α = .10	Pronóstico con α = .50
1	180	175	175
2	168	175.50 = 175.00 + .10(180 − 175)	177.50
3	159	174.75 = 175.50 + .10(168 − 175.50)	172.75
4	175	173.18 = 174.75 + .10(159 − 174.75)	165.88
5	190	173.36 = 173.18 + .10(175 − 173.18)	170.44
6	205	175.02 = 173.36 + .10(190 − 173.36)	180.22
7	180	178.02 = 175.02 + .10(205 − 175.02)	192.61
8	182	178.22 = 178.02 + .10(180 − 178.02)	186.30
9	?	178.59 = 178.22 + .10(182 − 178.22)	184.15

3.5 Proyecciones de tendencia

El último método de pronósticos de series de tiempo que analizaremos es la proyección de la tendencia. Esta técnica ajusta una recta de tendencia a una serie de datos puntuales históricos, y después proyecta dicha recta al futuro para obtener pronósticos de mediano y largo plazos. Se pueden desarrollar varias ecuaciones matemáticas (por ejemplo, exponencial y cuadrática), pero en esta sección veremos sólo tendencias lineales (en línea recta).

Si decidimos desarrollar una recta de tendencia lineal mediante un método estadístico preciso, podemos aplicar el método de mínimos cuadrados. Este enfoque resulta en una línea recta que minimiza la suma de los cuadrados de las diferencias verticales o desviaciones de la recta hacia cada una de las observaciones reales. En la figura 3.3 se ilustra el método de mínimos cuadrados.

Figura 3.3 Método de mínimos cuadrados para encontrar la recta que mejor se ajuste



- Una recta de mínimos cuadrados se describe en términos de su intersección con el eje *y* (la altura a la cual cruza al eje *y*) y su pendiente (el ángulo de la recta). Si podemos calcular la intersección con el eje *y* y la pendiente, podremos expresar la recta con la siguiente ecuación:

$$\hat{y} = a + bx$$

donde:

($\hat{y}$ que se lee "*y* gorro") = valor calculado de la variable que debe predecirse (llamada *variable dependiente*)
a = intersección con el eje *y*
b = pendiente de la recta de regresión (o la tasa de cambio en *y* para los cambios dados en *x*)
x = variable independiente (que en este caso es el *tiempo*)

Los estadísticos han desarrollado ecuaciones que se utilizan para encontrar los valores de *a* y *b* para cualquier recta de regresión. La pendiente *b* se encuentra mediante:

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

donde
b = pendiente de la recta de regresión
Σ = signo de sumatoria
x = valores conocidos de la variable independiente
y = valores conocidos de la variable dependiente
$\bar{x}$= promedio de los valores de *x*
$\bar{y}$= promedio de los valores de *y*
n = número de puntos de datos u observaciones

La intersección con el eje *y*, *a*, puede calcularse como sigue:

$$a = \bar{y} - b\bar{x}$$

Ejemplo:

En la tabla siguiente se muestra la demanda de energía eléctrica en N. Y. Edison durante el periodo 2001 a 2007, en megawatts. La empresa quiere pronosticar la demanda para 2008 ajustando una recta de tendencia a estos datos.

Año	Demanda de energía eléctrica	Año	Demanda de energía eléctrica
2001	74	2005	105
2002	79	2006	142
2003	80	2007	122
2004	90		

Método: Con una serie de datos en función del tiempo, podemos minimizar los cálculos transformando los valores de x (tiempo) en números más simples. En este caso podemos designar el año 2001 como año 1, 2002 como año 2, etc. Después pueden usarse las ecuaciones para crear el modelo de proyección de la tendencia.

Año	Periodo (x)	Demanda de energía eléctrica (y)	x^2	xy
2001	1	74	1	74
2002	2	79	4	158
2003	3	80	9	240
2004	4	90	16	360
2005	5	105	25	525
2006	6	142	36	852
2007	7	122	49	854
	$\Sigma x = 28$	$\Sigma y = 692$	$\Sigma x^2 = 140$	$\Sigma xy = 3{,}063$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{28}{7} = 4 \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{692}{7} = 98.86$$

$$b = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2} = \frac{3{,}063 - (7)(4)(98.86)}{140 - (7)(4^2)} = \frac{295}{28} = 10.54$$

$$a = \bar{y} - b\bar{x} = 98.86 - 10.54(4) = 56.70$$

Así, la ecuación de mínimos cuadrados para la tendencia es $\hat{y}$= 56.70 + 10.54x. Para proyectar la demanda en 2008, primero denotamos el año 2008 en nuestro nuevo sistema de código como x = 8.

Demanda en 2008 = 56.70 + 10.54(8)

= 141.02, o 141 megawatts

3.6 Ejercicios

Ejercicio 3.1

Las ventas mensuales en AC Delco Bateries, Inc., fueron como sigue:

Mes	Ventas
Enero	20
Febrero	21
Marzo	15
Abril	14
Mayo	13
Junio	16
Julio	17
Agosto	18
Septiembre	20
Octubre	20
Noviembre	21
Diciembre	23

Pronostique las ventas para enero usando cada una de las técnicas siguientes:

a) Un promedio móvil de 3 meses.

b) Un promedio móvil ponderado de 6 meses empleando .1, .1, .1, .2, .2 y .3, con las ponderaciones más altas a los meses más recientes.

c) Suavizamiento exponencial con $\alpha = .3$ y un pronóstico para septiembre de 18.

Con los datos proporcionados, ¿qué método le permitiría elaborar el pronóstico de ventas para el próximo mes de marzo?

Ejercicio 3.2

Considere los siguientes niveles de demanda real y pronosticada para las hamburguesas Big Mac en un restaurante McDonald's local.

Día	Demanda real	Demanda pronosticada
Lunes	88	88
Martes	72	88
Miércoles	68	84
Jueves	48	80
Viernes		

El pronóstico para el lunes se obtuvo observando el nivel de demanda para lunes y estableciendo el nivel pronosticado a este mismo nivel real. Los pronósticos subsecuentes se obtuvieron usando suavizamiento exponencial con una constante de suavizamiento de 0.25. Usando este método de suavizamiento exponencial, ¿cuál es el pronóstico para la demanda de Big Mac el viernes?

Ejercicio 3.3

En la tabla siguiente se muestra el número de transistores (en millones) fabricados en una planta de Japón durante los últimos 5 años.

Año	Transistores
1	140
2	160
3	190
4	200
5	210

Usando regresión lineal, pronostique el número de transistores que se fabricará el próximo año (año 6).

4. Diseño experimental para un factor

4.1 Introducción, conceptualización, importancia y alcances del diseño experimental en el ámbito empresarial

4.2 Clasificación de los diseños experimentales

4.3 Nomenclatura y simbología en el diseño experimental

4.4 Identificación de los efectos de los diseños experimentales

4.5 La importancia de la aleatorización de los especímenes de prueba

4.6 Supuestos estadísticos en las pruebas experimentales

4.7 Prueba de Duncan

4.8 Aplicaciones industriales

5. Metodología del diseño experimental de bloques al azar

5.1 Metodología del diseño experimental de bloques al azar

5.2 Diseño de bloques completos al azar

Cuando se quieren comparar ciertos tratamientos o estudiar el efecto de un factor, es deseable que las posibles diferencias se deban principalmente al factor de interés y no a otros factores que no se consideran en el estudio. Cuando esto no ocurre y existen otros factores que no se controlan o nulifican para hacer la comparación, las conclusiones podrían ser afectadas sensiblemente. Por ejemplo, supongamos que se quieren comprar varias máquinas, si cada máquina es manejada por un operador diferente y se sabe que éste tiene una influencia en el resultado, entonces es claro que el factor operador debe tomarse en cuenta si se quiere comparar a las máquinas de manera justa.

Un operador más hábil puede hacer ver a su máquina (aunque ésta sea la peor) como la que tiene el mejor desempeño, lo cual impide hacer una comparación adecuada de los equipos. Para evitar este sesgo hay dos maneras de anular el posible efecto del factor operador: la manera lógica es utilizar el mismo operador en las cuatro máquinas; sin embargo, tal estrategia no siempre es aconsejable, ya que utilizar al mismo sujeto elimina el efecto del factor operador, pero restringe la validez de la comparación con dicho operador, y es posible que el resultado no se mantenga al utilizar a otros operadores.

La otra forma de anular el efecto operador en la comparación consiste en que cada operador trabaje durante el experimento con cada una de las máquinas. Esta estrategia es la más recomendable, ya que utilizar a todos los operadores con todas las máquinas permite tener resultados de la comparación que son válidos para todos los operadores. Esta última forma de nulificar el efecto de operadores, recibe el nombre de bloqueo.

5.2.1 Factores de bloque

A los factores adicionales al factor de interés que se incorporan de manera explícita en un experimento comparativo se les llama factores de bloque. Éstos tienen la particularidad de

que no se incluyen en el experimento porque interese analizar su efecto, sino como un medio para estudiar de manera adecuada y eficaz al factor de interés.

Los factores de bloque entran al estudio en un nivel de importancia secundaria con respecto al factor de interés y, en este sentido, se puede afirmar que se estudia un solo factor, porque es uno el factor de interés. Por ejemplo, en el caso de comparar cuatro máquinas que son manejadas por cuatro operadores, es pertinente incluir explícitamente al factor operadores (bloques) para lograr el propósito del estudio, pero esta inclusión no es con el fin de estudiar el efecto del factor operador (o comparar a los operadores).

Más bien, la inclusión de los operadores es un medio y no un fin para lograr una comparación adecuada y eficaz de las máquinas. Puede ser que además de los operadores existan otros factores de bloque que deban controlarse durante el experimento para lograr una comparación adecuada de las máquinas. También se podrían controlar: el tipo de material, lotes, tipo de producto, día, turno, etc., pero no se trata de caer en el extremo de querer controlarlo todo, sino básicamente aquellos factores qué por conocimiento del proceso o experiencia previa, se sabe que afectan en forma considerable el resultado de la comparación.

En un diseño en bloques completos al azar (DBCA) se consideran tres fuentes de variabilidad: el factor de tratamientos, el factor de bloque y el error aleatorio, es decir, se tienen tres posibles "culpables" de la variabilidad presente en los datos. La palabra completo en el nombre del diseño se debe a que en cada bloque se prueban todos los tratamientos, o sea, los bloques están completos. La aleatorización se hace dentro de cada bloque; por lo tanto, no se realiza de manera total como en el diseño completamente al azar. El hecho de que existan bloques hace que no sea práctico o que incluso sea imposible aleatorizar en su totalidad.

Los factores de bloqueo que aparecen en la práctica son: turno, lote, día, tipo de material, línea de producción, operador, máquina, método, etc. La imposibilidad de aleatorizar de bloque a bloque se aprecia clara mente cuando se bloquean factores como día o turno, ya que no tiene sentido pensar en seleccionar al azar el orden de los días o los turnos porque es imposible regresar el tiempo.

Supongamos una situación experimental con k tratamientos y b bloques. El aspecto de los datos para este caso se muestra en cuadro 5.1, y considera una repetición en cada combinación de tratamiento y bloque.

Cuadro 5.1 Arreglo de los datos en un diseño en bloques completos al azar.

Tratamiento	Bloque					
		1	2	3	...	*b*
	1	Y_{11}	Y_{12}	Y_{13}	...	Y_{1b}
	2	Y_{21}	Y_{22}	Y_{23}	...	Y_{2b}
	3	Y_{31}	Y_{32}	Y_{33}	...	Y_{3b}
	⋮	⋮	⋮	⋮	⋮	⋮
	k	Y_{k1}	Y_{k2}	Y_{k3}	...	Y_{kb}

5.2.2 Modelo estadístico

Cuando se decide utilizar un DBCA, el experimentador piensa que cada medición será el resultado del efecto del tratamiento donde se encuentre, del efecto del bloque al que pertenece y de cierto error que se espera sea aleatorio. El modelo estadístico para este diseño está dado por:

$$Y_{ij} = \mu + \tau_i + \gamma_j + \varepsilon_{ij}; \begin{Bmatrix} i = 1,2,\ldots,k \\ j = 1,2,\ldots,b \end{Bmatrix} \qquad (5.1)$$

donde Y_{ij} es la medición que corresponde al tratamiento i y al bloque j (ver cuadro 5.1); m es la media global poblacional; τ_i es el efecto debido al tratamiento i, γ_j es el efecto debido al bloque j, y ε_{ij} es el error aleatorio atribuible a la medición Y_{ij}. Se supone que los errores se distribuyen de manera normal con media cero y varianza constante σ^2 [$N(0, \sigma^2)$], y que son independientes entre sí.

5.2.3 Hipótesis a probar

La hipótesis de interés es la misma para todos los diseños comparativos, y está dada por:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k = \mu$$
$$H_A : \mu_i \neq \mu_j \text{ para algún } i \neq j \qquad (5.2)$$

que también se puede expresar como

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \ldots = \tau_k = 0$$
$$H_A : \tau_i \neq 0 \text{ para algún } i \qquad (5.3)$$

En cualquiera de estas hipótesis la afirmación a probar es que la respuesta media poblacional lograda con cada tratamiento es la misma para los k tratamientos y que, por lo tanto, cada respuesta media μ_i es igual a la media global poblacional, m. De manera alternativa, es posible afirmar que todos los efectos de tratamiento sobre la variable de respuesta son nulos, porque cuando el efecto $\tau_i = \mu_i - \mu = 0$, entonces necesariamente la respuesta media del tratamiento es igual a la media global ($\mu_i = \mu$).

5.2.4 Análisis de varianza

La hipótesis dada por (5.2 o 5.3) se prueba con un análisis de varianza con dos criterios de clasificación, porque se controlan dos fuentes de variación: el factor de tratamientos y el factor de bloque. En el cuadro 5.2 se muestra el aspecto del ANOVA para diseño DBCA.

Los cálculos necesarios pueden ser manuales, pero siempre es más práctico hacerlos con un software estadístico, porque además proporciona muchas otras opciones gráficas y tabulares útiles (no sólo el ANOVA).

Cuadro 5.2 ANOVA para un diseño en bloques completos al azar.

Tabla ANOVA					
Fuente de Variación (FV)	Suma de Cuadrados (SC)	Grados de Libertad (GL)	Cuadrado Medio (CM)	Fcalculada (F_0)	Ftablas (F)
Tratamientos	$SC_{TRAT} = \sum_{j=1}^{k} \frac{Y_{i.}^2}{b} - \frac{Y_{..}^2}{N}$	$k-1$	$CM_{TRAT} = \frac{SC_{TRAT}}{k-1}$	$\frac{CM_{TRAT}}{CM_E}$	$F = (\propto, k-1, b-1)$
Bloques	$SC_B = \sum_{j=1}^{b} \frac{Y_j^2}{k} - \frac{Y_{..}^2}{N}$	$b-1$	$CM_B = \frac{SC_B}{b-1}$	$\frac{CM_B}{CM_E}$	$F = (\propto, k-1, b-1)$
Error	$SC_E = SC_T - SC_{TRAT} - SC_B$	$N-k$	$CM_E = \frac{SC_E}{N-k}$		
Total	$SC_T = \sum_{j=1}^{b}\sum_{i=1}^{k} Y_{ij}^2 - \frac{Y_{..}^2}{N}$	$N-1$			

Utilizando la notación de puntos, las fórmulas más prácticas para calcular las sumas de cuadrados son:

$$SC_T = \sum_{j=1}^{b}\sum_{i=1}^{k} Y_{ij}^2 - \frac{Y_{..}^2}{N}$$
$$SC_{TRAT} = \sum_{i=1}^{k} \frac{Y_{i\bullet}^2}{b} - \frac{Y_{..}^2}{N} \qquad (5.4)$$
$$SC_B = \sum_{j=1}^{b} \frac{Y_{\bullet j}^2}{k} - \frac{Y_{..}^2}{N}$$

y la del error se obtiene por sustracción como:

$$SC_E = SC_T - SC_{TRAT} - SC_B \qquad (5.5)$$

5.2.5 Ejemplo de aplicación

Comparación de cuatro métodos de ensamble. Un equipo de mejora investiga el efecto de cuatro métodos de ensamble A, B, C y D, sobre el tiempo de ensamble en minutos. Se va a controlar activamente en el experimento a los operadores que realizarán el ensamble (cuatro operadores), lo que da lugar al siguiente diseño en bloques completos al azar.

Método	Operador			
	1	2	3	4
A	6	9	7	8
B	7	10	11	8
C	10	16	11	14
D	10	13	11	9

Recordemos que la variable de respuesta son los minutos en que se realiza el ensamble. Para comparar los cuatro métodos se plantea la hipótesis:

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu$$
$$H_A : \mu_i \neq \mu_j \text{ para algún } i \neq j = A, B, C, D$$

la cual se prueba mediante el análisis de varianza dado en la Cuadro 5.5. Para obtener dicho cuadro de ANOVA, se realizaron los siguientes pasos:

Paso 1. En el cuadro 5.3, se calcularon las sumatorias de las observaciones por filas $(Y_{i.})$ que son por cada método (tratamientos), las sumatorias de las observaciones por columnas $(Y_{.j})$, que son por cada operador (bloques), y la sumatoria del total de las observaciones $(Y_{..})$.

Cuadro 5.3 Cálculo de sumatorias de las observaciones

Método (k)	Operador (b)				Total por tratamiento
	1	2	3	4	
A	6	9	7	8	30
B	7	10	11	8	36
C	10	16	11	14	51
D	10	13	11	9	43
Total	33	48	40	39	160

$Y_{i.}$ $Y_{..}$ $Y_{.j}$

Paso 2. En el cuadro 5.4 se muestran los cálculos para obtener los valores de las siguientes sumatorias:

a) la suma de las observaciones por tratamiento elevadas al cuadrado $\sum_{j=1}^{k} Y_{i.}^2$

b) la suma de las observaciones por bloque (por operario) elevadas al cuadrado $\sum_{j=1}^{b} Y_{.j}^2$

$\sum_{j=1}^{b}\sum_{i=1}^{k} Y_{ij}^2$

c) cada una de las observaciones elevadas al cuadrado

Cuadro 5.4 Cálculo de sumatorias de las observaciones al cuadrado

Y_{ij}^2				Total	$Y_{.j}^2$	$Y_{i.}^2$
6^2=36	9^2=81	7^2=49	8^2=64	230	33^2 = 1089	30^2 = 900
7^2=49	10^2=100	11^2=121	8^2=64	334	48^2 = 2304	36^2 = 1296
10^2=100	16^2=256	11^2=121	14^2=196	673	40^2 = 1600	51^2 = 2601
10^2=100	13^2=169	11^2=121	9^2=81	471	39^2 = 1521	43^2 = 1849
				1708	**6514**	**6646**

$\sum_{j=1}^{b}\sum_{i=1}^{k} Y_{ij}^2$ $\quad$ $\sum_{j=1}^{b} Y_{.j}^2$ $\quad$ $\sum_{j=1}^{k} Y_{i.}^2$

Paso 3. Realizar los cálculos de las sumas de cuadrados ocupando las fórmulas 5.4 y 5.5, como se muestra a continuación:

$$SC_T = \sum_{j=1}^{b}\sum_{i=1}^{k} Y_{ij}^2 - \frac{Y_{..}^2}{N}$$

$SC_T =$ 1708 - $\frac{160^2}{16}$

$SC_T =$ 1708 - 1600

$SC_T =$ **108**

$$SC_{TRAT} = \sum_{j=1}^{k} \frac{Y_{i.}^2}{b} - \frac{Y_{..}^2}{N}$$

$SC_{TRAT} = \frac{6646}{4} - \frac{160^2}{16}$

$SC_{TRAT} =$1661.5 - 1600

$SC_{TRAT} =$ **61.5**

$$SC_B = \sum_{j=1}^{b} \frac{Y_{.j}^2}{k} - \frac{Y_{..}^2}{N}$$

$SC_B = \frac{6514}{4} - \frac{160^2}{16}$

$SC_B =$ 1628.5 - 1600

$SC_B =$ **28.5**

$SC_E = SC_T - SC_{TRAT} - SC_B$

$SC_E =$ 108 - 61.5 - 28.5

$SC_E =$ **18**

Paso 4. Con los resultados obtenidos en el paso 3, de las sumas de los cuadrados y, con las fórmulas del cuadro 5.2 del ANOVA, se procede a calcular todos los elementos que requiere el cuadro de ANOVA (Cuadro 5.5), para poder realizar una conclusión del mismo.

Cuadro 5.5 ANOVA obtenido de los datos del ejemplo 5.1

Cuadro ANOVA					
Fuente de Variación (FV)	Suma de Cuadrados (SC)	Grados de Libertad (GL)	Cuadrado Medio (CM)	Fcalculada (F_0)	F tablas (F)
Métodos	61.5	3	20.5	10.25	3.86
Operadores	28.5	3	9.5	4.75	3.86
Error	18	9	2		
Total	108	15			

Del cuadro 5.5. el valor F de tablas, se obtuvo considerando un nivel de significancia del 5% (tomando en cuenta que 1-Nivel de confianza es el nivel de significancia, si se requiere un nivel de confianza del 95%=0.95, el nivel de significancia α=1-0.95= 0.05 = 5%).

Para hallar el valor de F de tablas para la fila "métodos", se considera lo siguiente: con el nivel de significancia α=0.05, y con los grados de libertad (GL) de los métodos (numerador) =3 y los grados de libertad (GL) de error (denominador)=9, obtenidos en el ANOVA (cuadro 5.5), se procede a buscar esa información en el Apéndice 5 (Distribución de Fischer con α= 5%). En la figura 5.1 se muestra el valor obtenido de F(tablas) = 3.86.

Para el caso del del valor F de tablas de la fila de "operadores", se considera lo siguiente: α=0.05, GL de los operadores=3 (grados de libertad del numerador) y GL del error=9 (grados de libertad del denominador). Como ambos valores son iguales, se muestra el valor obtenido en la figura 5.1.

Analizando los resultados del ANOVA (cuadro 5.5), se observa que para los métodos se obtuvo un F_{tablas} = 3.86, y considerando que si F_0= 10.25 es mayor que F_{tablas} = 3.86, por tanto, se rechaza la hipótesis H_0 de que el tiempo medio poblacional de los métodos de ensamble son iguales, y se acepta que al menos dos de los métodos son diferentes en cuanto al tiempo promedio que requieren.

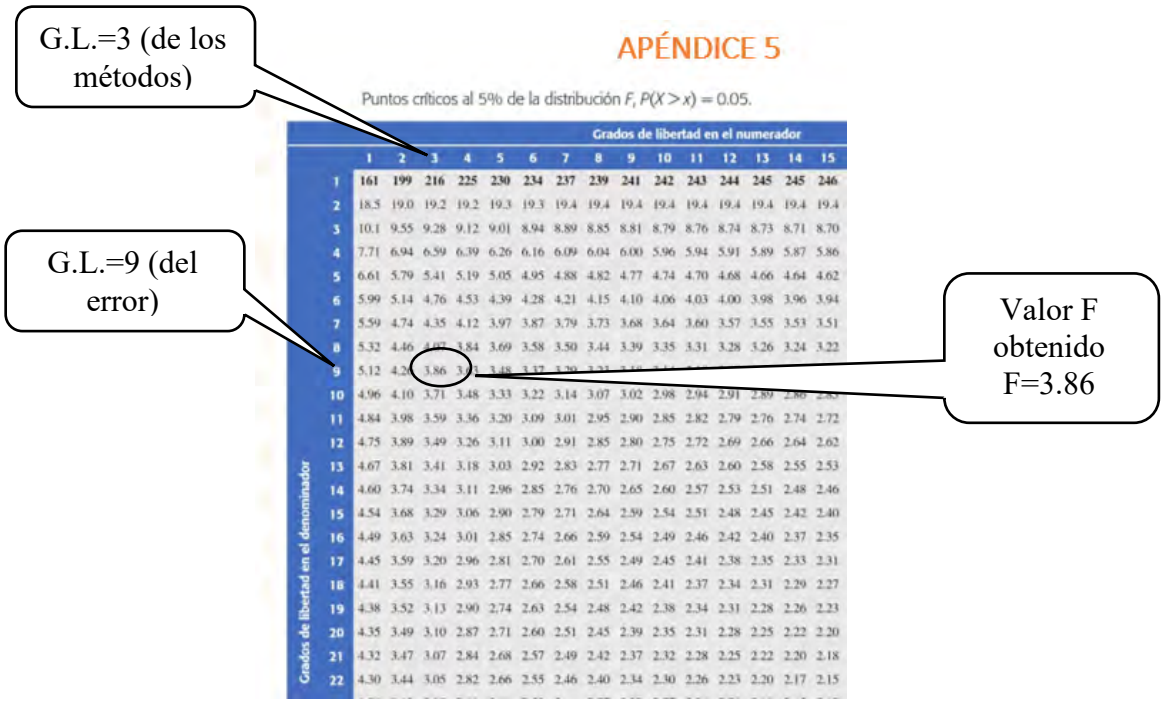
De la misma manera para operadores, se obtuvo un F_{tablas} = 3.86, y considerando que si F_0= 10.25 es mayor que F_{tablas} = 3.86, por tanto, el factor de bloques (operadores) también afecta, es decir, existen diferencias entre los operadores en cuanto al tiempo promedio.

Sin embargo, recordemos que no es objetivo del experimento comparar a los operadores, y su control en el estudio se utiliza para lograr una comparación más justa y precisa de los métodos de ensamble. En otras palabras, mientras que los métodos de ensamble se comparan con el objetivo final de elegir el más eficiente en términos de tiempo, con los operadores no

se trata de elegir uno; en todo caso, quizá como información extra se pueda tomar alguna decisión sobre los operadores, como por ejemplo dar mayor entrenamiento a quien lo requiera por salirse en forma significativa del comportamiento del resto.

Cuando mediante un diseño de bloques se concluye que los tratamientos son diferentes, es probable que no se haya llegado a esa conclusión, sino que se haya considerado el factor de bloque. Por ejemplo, si en el ANOVA cuadro 5.5 no se considera el efecto de bloque (operador), entonces la variabilidad y los grados de libertad atribuibles a operadores se irían al error, lo cual puede modificar las conclusiones sobre los tratamientos (métodos).

Figura 5.1 Obtención del valor F de tablas de los métodos (Tratamientos)



APÉNDICE 5

Puntos críticos al 5% de la distribución F, $P(X > x) = 0.05$.

Grados de libertad en el denominador	Grados de libertad en el numerador														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	161	199	216	225	230	234	237	239	241	242	243	244	245	245	246
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20	2.18
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15

5.2.5 Ejercicios

Ejercicio 5.2.1

Deberá realizar el siguiente ejercicio a mano en hojas blancas o de libreta. Deben realizar sus cálculos redondeando a 3 decimales todos sus resultados, realizarlo de manera ordenada.

Ejercicio.

Se diseñó un experimento para estudiar el rendimiento de cuatro detergentes. Las siguientes lecturas de “blancura” se obtuvieron con un equipo especial diseñado para 12 cargas de lavado, distribuidas en tres modelos de lavadoras:

Detergente	Lavadora 1	Lavadora 2	Lavadora 3
A	45	43	51
B	47	44	52
C	50	49	57
D	42	37	49

a) Plantea la hipótesis que se requiere probar para este experimento

b) Realice el ANOVA y redacte las conclusiones.

5.3 Diseño factorial 2K

5.4 Diseño de cuadrados latinos

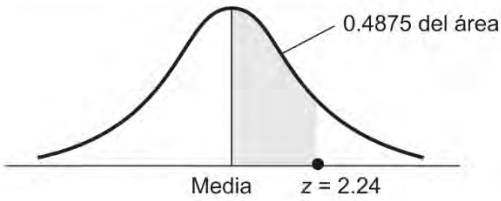
5.5 Diseño de cuadrados grecolatinos

5.6 Aplicaciones

APÉNDICE 1

DISTRIBUCIÓN DE PROBABILIDAD NORMAL ESTÁNDAR

Áreas bajo la distribución de probabilidad normal estándar entre la media y valores positivos de *z



EJEMPLO: Para encontrar el área bajo la curva que se encuentra entre la media y un punto situado a 2.24 desviaciones estándar a la derecha de la media, busque el valor en el renglón correspondiente a 2.2 bajo la columna 0.04 de la tabla; 0.4875 del área bajo la curva se encuentra entre la media y un valor z de 2.24

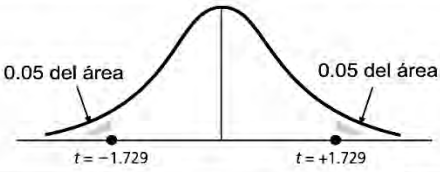
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0319
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3816	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4788	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4986	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

*Tomado de Robert D. Mason, *Essentials of Statistics.* © 1976, pág. 307. Impreso con licencia de Prentice-Hall, Inc., Engiewood Cliffs, NJ.

APÉNDICE 2

DISTRIBUCIÓN *t*

Áreas en los dos extremos combinados para la distribución *t* de Student.*



EJEMPLO: Para encontrar el valor de t que corresponde a un área de 0.10 en los dos extremos combinados de la distribución, cuando existen 19 grados de libertad, busque en la columna del 0.10 hacia abajo hasta el renglón correspondiente a 19 grados de libertad, el valor *t* apropiado es 1.729

Grados de libertad	Área en los dos extremos combinados			
	0.10	0.05	0.02	0.01
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.291
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
39	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
60	1.671	2.000	2.390	2.660
120	1.658	1.980	2.358	2.617
Distribución normal	1.645	1.960	2.326	2.576

*Tomado de la Tabla III de Fisher y Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, publicado por Longman Group, Ltd., Londres (publicada anteriormente por Olivier & Boyd, Edimburgo) y con licencia de los autores y los editores.

APÉNDICE 5

Puntos críticos al 5% de la distribución F, $P(X > x) = 0.05$.

Grados de libertad en el denominador \ Grados de libertad en el numerador	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30	40	50	75	100	∞
1	**161**	**199**	**216**	**225**	**230**	**234**	**237**	**239**	**241**	**242**	**243**	**244**	**245**	**245**	**246**	**248**	**249**	**250**	**251**	**252**	**253**	**253**	**254**
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	91.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.66	8.63	8.62	8.59	8.58	8.56	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.80	5.77	5.75	5.72	5.70	5.68	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.56	4.52	4.50	4.46	4.44	4.42	4.41	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.87	3.83	3.81	3.77	3.75	3.73	3.71	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.44	3.40	3.38	3.34	3.32	3.29	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.15	3.11	3.08	3.04	3.02	2.99	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.94	2.89	2.86	2.83	2.80	2.77	2.76	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.77	2.73	2.70	2.66	2.64	2.60	2.59	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.65	2.60	2.57	2.53	2.51	2.47	2.46	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.54	2.50	2.47	2.43	2.40	2.37	2.35	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.46	2.41	2.38	2.34	2.31	2.28	2.26	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.39	2.34	2.31	2.27	2.24	2.21	2.19	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.33	2.28	2.25	2.20	2.18	2.14	2.12	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.28	2.23	2.19	2.15	2.12	2.09	2.07	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.23	2.18	2.15	2.10	2.08	2.04	2.02	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.19	2.14	2.11	2.06	2.04	2.00	1.98	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.16	2.11	2.07	2.03	2.00	1.96	1.94	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.12	2.07	2.04	1.99	1.97	1.93	1.91	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20	2.18	2.10	2.05	2.01	1.96	1.94	1.90	1.88	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15	2.07	2.02	1.98	1.94	1.91	1.87	1.85	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.18	2.15	2.13	2.05	2.00	1.96	1.91	1.88	1.84	1.82	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13	2.11	2.03	1.97	1.94	1.89	1.86	1.82	1.80	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11	2.09	2.01	1.96	1.92	1.87	1.84	1.80	1.78	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.07	1.99	1.94	1.90	1.85	1.82	1.78	1.76	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.10	2.08	2.06	1.97	1.92	1.88	1.84	1.81	1.76	1.74	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06	2.04	1.96	1.91	1.87	1.82	1.79	1.75	1.73	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.08	2.05	2.03	1.94	1.89	1.85	1.81	1.77	1.73	1.71	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01	1.93	1.88	1.84	1.79	1.76	1.72	1.70	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.92	1.84	1.78	1.74	1.69	1.66	1.61	1.59	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.89	1.86	1.84	1.75	1.69	1.65	1.59	1.56	1.51	1.48	1.39
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85	1.82	1.79	1.77	1.68	1.62	1.57	1.52	1.48	1.42	1.39	1.28
∞	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75	1.72	1.69	1.67	1.57	1.51	1.46	1.40	1.35	1.28	1.25	1.03

APÉNDICE 6

Puntos críticos al 10% de la distribución F, $P(X > x) = 0.10$.

Grados de libertad en el numerador	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30	40	50	75	100	∞
1	40	50	54	56	57	58	59	59	60	60	60	61	61	61	61	62	62	62	63	63	63	63	63

EDITA: RED IBEROAMERICANA DE ACADEMIAS DE INVESTIGACIÓN A.C
DUBLÍN 34, FRACCIONAMIENTO MONTE MAGNO
C.P. 91190. XALAPA, VERACRUZ, MÉXICO.
CEL 2282386072
PONCIANO ARRIAGA 15, DESPACHO 101.
COLONIA TABACALERA
DELEGACIÓN CUAUHTÉMOC
C.P. 06030. MÉXICO, D.F. TEL. (55) 55660965
www.redibai.org
redibai@hotmail.com

Sello editorial: Red Iberoamericana de Academias de Investigación, A.C. (607-8617)
Primera Edición, Xalapa, Veracruz, México.
No. de ejemplares: 200
Presentación en medio electrónico digital: Cd-Rom formato PDF 10MB
Fecha de aparición 11/12/2020
ISBN 978-607-8617-98-2

ISBN: 978-607-8617-98-2