Estadística aplicada

Emma Barreno • Jorge Chue • Rosa Millones
Félix Vásquez • Carlos Castillo

UNIVERSIDAD DE LIMA

FONDO EDITORIAL

Estadística aplicada

Emma Barreno Vereau • Jorge Chue Gallardo
Rosa Millones Rivalles • Félix Vásquez Urbano
• Carlos Castillo Crespo

UNIVERSIDAD
DE LIMA

FONDO EDITORIAL

Índice

Presentación

La ciencia estadística, cuyos orígenes se fusionan con la historia de la humanidad, es fundamental en el desarrollo alcanzado en todas las disciplinas científicas. La estadística es parte integrante del proceso de toma de decisiones y su importancia se incrementa en una sociedad que requiere determinaciones acertadas, rápidas y eficientes. Las posibilidades de incrementar nuestra productividad individual o colectiva serían casi nulas sin un adecuado conocimiento de la estadística.

En este contexto se presenta *Estadística aplicada*, libro que condensa una extensa e intensa experiencia en la docencia universitaria y en el ámbito profesional, y que está dirigido a los interesados en aplicaciones de la estadística en el campo de la ingeniería, las ciencias administrativas y económicas, pero que ha sido diseñado para que pueda ser consultado con provecho por cualquier profesional. El objetivo principal es explicar los principios de la estadística inferencial, la regresión y el diseño de experimentos, temas determinantes en la toma de decisiones con herramientas cuantitativas.

Los más de trescientos ejercicios incluidos proporcionan el enlace necesario entre los conceptos teóricos y las aplicaciones reales. Este adecuado énfasis en explorar las articulaciones de la teoría y la práctica constituye uno de los aportes más significativos de este texto.

En la solución de los ejemplos y los problemas desarrollados se han empleado el software estadístico Minitab for Windows y el MS Excel.

Dividido en cinco capítulos, que establecen una secuencia gradual de aprendizaje en el que cada apartado conserva cierta independencia y autonomía, este texto tiene la ventaja de adecuarse a las necesidades del lector, quien podrá revisarlo en forma secuencial o selectiva.

Las distribuciones de 'probabilidad' y el proceso de 'estimación puntual' e 'interválica' son los temas tratados en los dos primeros capítulos del libro; mientras que en el tercero se desarrolla el importante punto referido a la 'prueba de hipótesis' y en el siguiente capítulo se expone una sección de aplicación de las herramientas propuestas como la de 'regresión lineal simple' y "múltiple". El quinto y último capítulo está dedicado a la descripción y el análisis del método de diseño de experimentos completamente aleatorizado.

Esperamos que este texto sea de utilidad para nuestros lectores, a quienes agradecemos de antemano sus comentarios y sugerencias.

Los autores

Capítulo

1 Distribuciones muestrales

En este capítulo trataremos los siguientes temas:

- Conceptos básicos
- Muestra aleatoria
- Tipos de muestreo
- Principales estimadores
- Distribución de la media muestral
- Teorema central del límite
- Distribuciones de probabilidad
- Distribuciones muestrales de un estimador obtenido de una muestra
- Distribuciones muestrales para estimadores obtenidos de dos muestras

Uno de los temas más importantes de la estadística es el teorema central del límite, el cual señala que la distribución de la media muestral de cualquier variable discreta o continua se aproximará a una distribución normal para tamaños de muestra lo suficientemente grande, y cuyo estudio en este capítulo es precedido por el desarrollo de puntos como la muestra aleatoria, los tipos de muestreo, los principales estimadores y la distribución de la media muestral. A continuación se aborda lo relacionado con las distribuciones de muestras pequeñas, las distribuciones muestrales de un estimador de una y de dos muestras. Al finalizar este capítulo, y con la ayuda de la variedad de ejercicios resueltos y de los propuestos se estará familiarizado con estos conceptos de tal forma que se conocerá el comportamiento de los diferentes tipos de distribuciones.

Las poblaciones suelen ser demasiado grandes para estudiarlas en su totalidad; por ello, se investigan temas particulares como el consumo promedio *per cápita* en una región del país o el porcentaje de consumidores que prefieren un determinado producto. En estos casos, es preferible elegir una muestra representativa que tenga un tamaño más manejable y que permita obtener conclusiones válidas sobre la población objetivo que interesa conocer. Para el ejemplo anterior, se puede calcular la media aritmética $\overline{x}$ de la muestra de consumidores y utilizarla como una estimación de la media aritmética de la población μ. Para trabajar con una muestra y obtener conclusiones sobre un tema poblacional se deben aplicar las técnicas de la estadística inferencial.

En la estadística inferencial se desarrollan dos puntos importantes: el problema de estimación de los parámetros y el de la dócima o prueba de hipótesis, que serán desarrollados en los capítulos posteriores.

1. Conceptos básicos

a. *Unidad de análisis*.- Se define como el elemento que se observa y del que se busca información de características o variables de interés.
b. *Población.*- Se entiende por población o universo a la totalidad de elementos, ya sean empresas, personas, objetos, etcétera, que presentan una o más características observables.
c. *Población objetivo*.- Es la población completamente caracterizada. Por ejemplo, en una encuesta sobre aceptación de un nuevo producto de belleza de una empresa que produce cosméticos, la población objetivo estará dada por todas las mujeres que son usuarias de los productos de la empresa con edades entre 20 y 39 años, pertenecientes al nivel socioeconómico medio alto, a partir de la cual se selecciona una muestra de mujeres para la investigación.
d. *Marco muestral*.- Se define como el listado de elementos desde los que se seleccionará la muestra.

e. *Unidad de muestreo*.- Es aquella que contiene las unidades de análisis de la población que se utilizarán para confeccionar la muestra. En general, es la selección de los conjuntos de unidades de análisis que serán tomados en cuenta para conformar la muestra final en la investigación.

2. Muestra aleatoria

Tanto la estimación de parámetros como las pruebas de hipótesis se basan generalmente en la información proporcionada por las unidades de observación o de análisis sobre una variable o característica de estudio X a través de sus valores x_1, $x_2,\ldots, x_n$.

Estas unidades de observación se eligen de manera independiente y deben tener la misma probabilidad de ser seleccionadas. El conjunto de estas unidades seleccionadas reciben el nombre de *muestra aleatoria*.

Cuando se trata de poblaciones finitas de N elementos se seleccionarán $k = C_n^N$ muestras diferentes sin reemplazamiento donde $C_n^N = \dfrac{N!}{n!(N-n)!}$; si el muestreo es con reemplazamiento se seleccionarán $k = N^n$ muestras diferentes.

2.1 Definición

Se dice que los valores $x_1, x_2,\ldots, x_n$ de la variable de interés X con función de probabilidad $f(x)$ constituyen una muestra aleatoria de tamaño n si son variables aleatorias independientes y distribuidas de forma idéntica. Es decir, si se verifica que la ley de probabilidad es la misma para cada una de las observaciones, esto es:

$$f(x_1) = f(x_2) = \ldots = f(x_n)$$

La función de probabilidad de las observaciones muestrales está dada por:

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) \qquad (1)$$

La expresión (1) es conocida como función de probabilidad conjunta.

3. Tipos de muestreo

Se dispone de dos métodos para seleccionar las muestras de poblaciones:

- Muestreo probabilístico.
- Muestreo no probabilístico.

3.1 Muestreo probabilístico

En este tipo de muestreo se cuenta con información de las probabilidades de las unidades de análisis seleccionadas en la muestra. El muestreo probabilístico per-

mite calcular el grado hasta el cual el valor obtenido de la muestra puede diferir del valor de la población de interés. Esta diferencia recibe el nombre de error muestral.

Existen varios tipos de muestreo probabilístico, los cuales se mencionan a continuación:

3.1.1 Muestreo aleatorio simple

En este tipo de muestreo cada unidad de la población tiene igual probabilidad de ser seleccionada, se recomienda cuando la variable en estudio es homogénea. Se facilita su aplicación mediante el uso de una tabla de números aleatorios, que se construye sobre la base de un proceso de aleatorización de los dígitos del 0 al 9 o a través de un software de aplicación. Su uso se ilustra en el siguiente ejemplo.

Ejemplo 1:

Suponga que se desea seleccionar una muestra aleatoria simple de 20 personas de una población total de 100. A cada persona se le asigna un número del 001 al 100 y se hace uso de la Tabla de Números Aleatorios (véase el anexo 1) para seleccionar la muestra, eligiendo primeramente al alzar un número que será el punto de partida, en este caso la quinta columna y la cuarta fila del primer bloque de la tabla y que corresponde al número 18. A partir de dicho número y siguiendo un camino aleatorio, desde arriba hasta el final de la columna se van seleccionando los números de la muestra hasta completar los 20 números asociados a cada persona. Los 20 números seleccionados constituyen la muestra:

18	23	26	39	82	62	90	48	82	68
42	56	59	86	27	38	14	04	29	64

CON MINITAB. Para la obtención de la muestra aleatoria haciendo uso del Minitab se sigue el siguiente procedimiento:

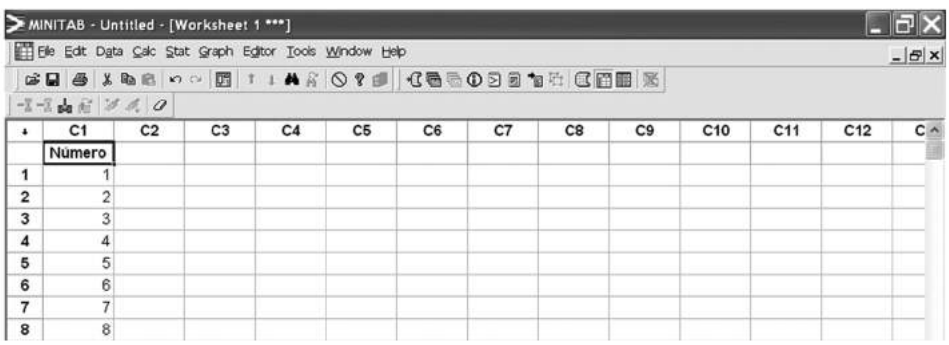


Figura 1. Ingreso de datos en el Minitab.

- Se disponen en una columna los 100 números, un número asignado a cada persona, como se muestra en la figura 1.
- En la barra de herramientas, se selecciona la opción Calc / Random Data / Sample From Columns.
- Se coloca el tamaño de la muestra que se desea extraer:
  Sample: 20 rows from column(s).
- Se selecciona el marco muestral, es decir la columna donde se encuentra la numeración asignada a las personas. En este caso la columna C1 Número.
- Se indica la columna donde se desea almacenar los resultados del muestreo, por ejemplo en la columna C3.
  Store samples in: C3.
- Se verifica que no se encuentre seleccionada la opción de muestreo con reemplazo.
- Finalmente, se pulsa el botón Ok.

Lo anteriormente expuesto se aprecia en la figura 2.

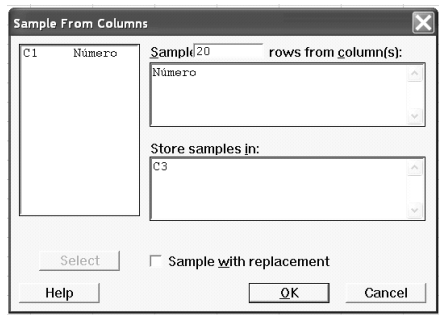
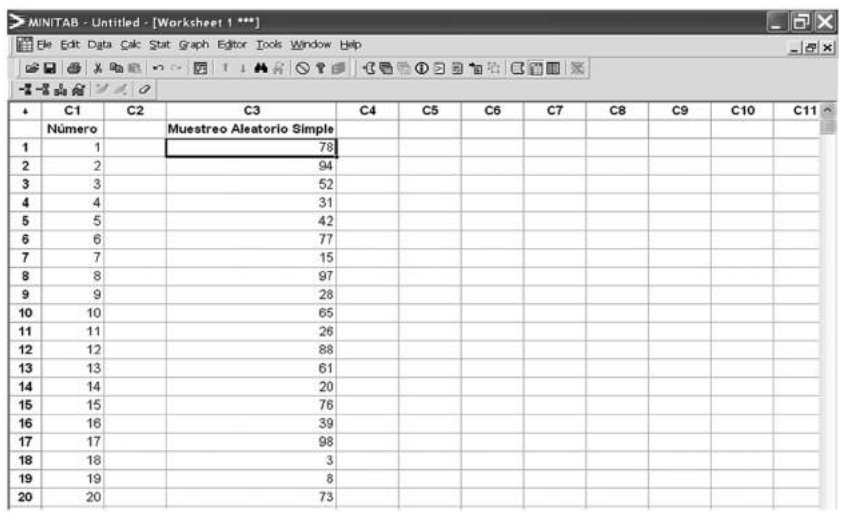


Figura 2. Cuadro de diálogo - Sample From Column.

Los resultados se almacenaron en la columna C3, tal como se indicó, entonces se procede a etiquetar la columna; por ejemplo, Muestreo Aleatorio Simple. De acuerdo con el resultado (figura 3), la muestra estará conformada por personas cuyos números asignados sean: 78, 94, 52, …, 73.



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
	Número		Muestreo Aleatorio Simple								
1	1		78								
2	2		94								
3	3		52								
4	4		31								
5	5		42								
6	6		77								
7	7		15								
8	8		97								
9	9		28								
10	10		65								
11	11		26								
12	12		88								
13	13		61								
14	14		20								
15	15		76								
16	16		39								
17	17		98								
18	18		3								
19	19		8								
20	20		73								

Figura 3. Salida de resultados - Sample From Column.

Nota: Cada vez que se realice el muestreo se obtendrán resultados diferentes, ya que son resultados aleatorios.

Si se desea, los resultados del muestreo se pueden ordenar para su mejor visualización mediante el siguiente procedimiento:

- En la barra de herramientas, seleccione la opción <Data / Sort>.
- Seleccione la columna donde se encuentra la muestra seleccionada.
  <Sort column(s)>: Muestreo Aleatorio Simple.
- Indique la forma de ordenamiento.
  <By column>: Muestreo Aleatorio Simple.
- La opción <Descending> debe permanecer desactivada, a menos que se desee realizar el ordenamiento en forma descendente.
- Marque la opción <Original Column(s)> (si se desea que el ordenamiento se realice en la misma columna), o marcar la opción <Column(s) of Current Work>; luego indicar otra columna de trabajo, por ejemplo, C4 (si se desea que el ordenamiento se realice en otra columna).
- Pulse el botón <Ok>.

Lo anteriormente expuesto se aprecia en la figura 4.

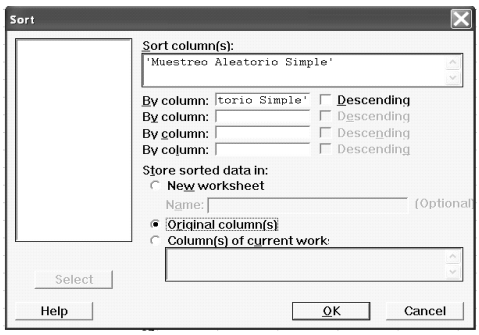
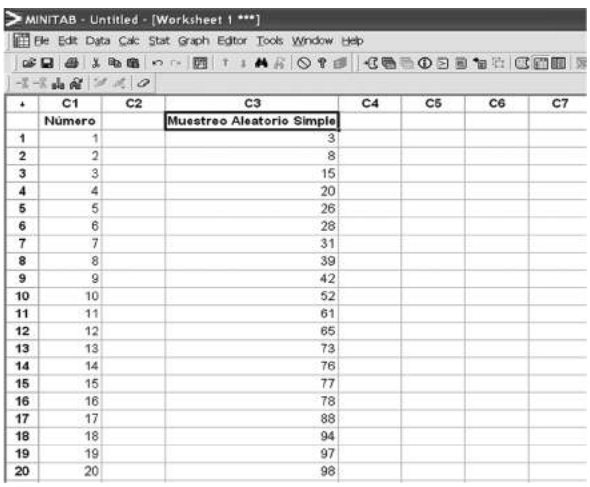


Figura 4. Cuadro de diálogo – Sort.

Después de haber realizado estos pasos la muestra se debe encontrar ordenada en forma ascendente (véase la figura 5).



	C1	C2	C3	C4	C5	C6	C7
	Número		Muestreo Aleatorio Simple				
1	1		3				
2	2		8				
3	3		15				
4	4		20				
5	5		26				
6	6		28				
7	7		31				
8	8		39				
9	9		42				
10	10		52				
11	11		61				
12	12		65				
13	13		73				
14	14		76				
15	15		77				
16	16		78				
17	17		88				
18	18		94				
19	19		97				
20	20		98				

Figura 5. Ordenamiento de los resultados - Sort.

CON SOFTWARE MS EXCEL. A continuación se presenta la metodología para obtener una muestra aleatoria haciendo uso del MS Excel.

- Se deben disponer en una columna los 100 números asignados a cada persona, como se muestra a continuación:

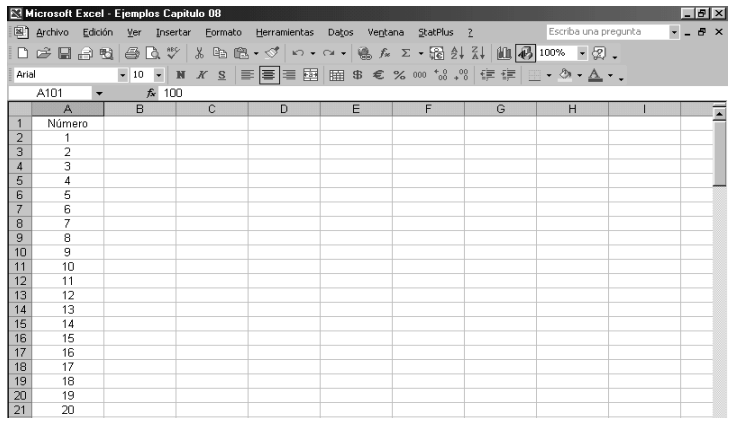


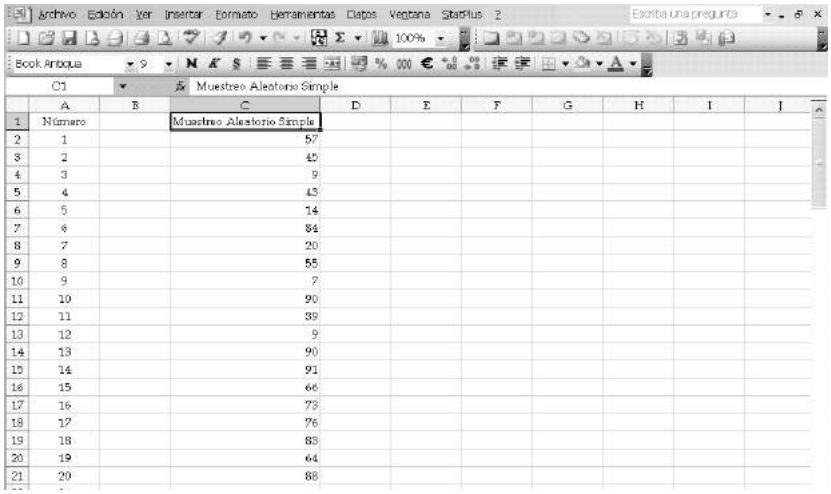
Figura 6. Ingreso de datos en el MS Excel.

- En la barra de herramientas, elija la opción Herramientas / Análisis de datos.
- Seleccione la opción: Muestra.
- Pulse en el botón Aceptar.
- Seleccione el marco muestral.
  Rango de entrada: $A$1:$A$101. (Incluye el rótulo.)
- Marque la opción Rótulos. (Desactivar en el caso de no haber incluido el rótulo al momento de seleccionar el marco muestral.)
- Seleccione la opción Aleatorio.
- Indique el tamaño de la muestra. Número de muestras: 20.
- Seleccione la opción de salida de resultados. Seleccionar la opción Rango de salida e indicar la celda donde se desea que se empiecen a grabar los resultados.
  Rango de salida: C2.
- Pulse el botón Aceptar.

Lo anteriormente expuesto se aprecia en la figura 7.

Figura 7. Cuadro de diálogo – Muestreo aleatorio.

Los resultados se almacenaron a partir de la celda C2, tal como se indicó, entonces se procede a etiquetar la columna; por ejemplo, Muestreo Aleatorio Simple. De acuerdo con el resultado (figura 8) la muestra estará conformada por personas cuyos números asignados son: 57, 45, 9, …, 88.



	A	B	C
1	Número		Muestreo Aleatorio Simple
2	1		57
3	2		45
4	3		9
5	4		43
6	5		14
7	6		84
8	7		20
9	8		55
10	9		7
11	10		90
12	11		39
13	12		9
14	13		90
15	14		91
16	15		66
17	16		73
18	17		76
19	18		83
20	19		64
21	20		88

Figura 8. Salida de resultados – Muestreo aleatorio.

Nota: Al igual que en el Minitab, cada vez que se realice el muestreo se obtendrán resultados diferentes, ya que son resultados aleatorios.
Si se desea, se pueden ordenar los resultados del muestreo para su mejor visualización mediante el siguiente procedimiento:

- Marque toda la muestra obtenida. (Incluyendo el rótulo.)
- En la barra de herramientas, seleccione la opción Datos / Ordenar.
- Ya se encuentra la muestra seleccionada. Marque la opción Ascendente.
- Verifique que se encuentre activada la opción que indica que el rango de datos tiene fila de encabezamiento. Pulsar el botón Aceptar.

Lo anteriormente expuesto se aprecia en la figura 9.

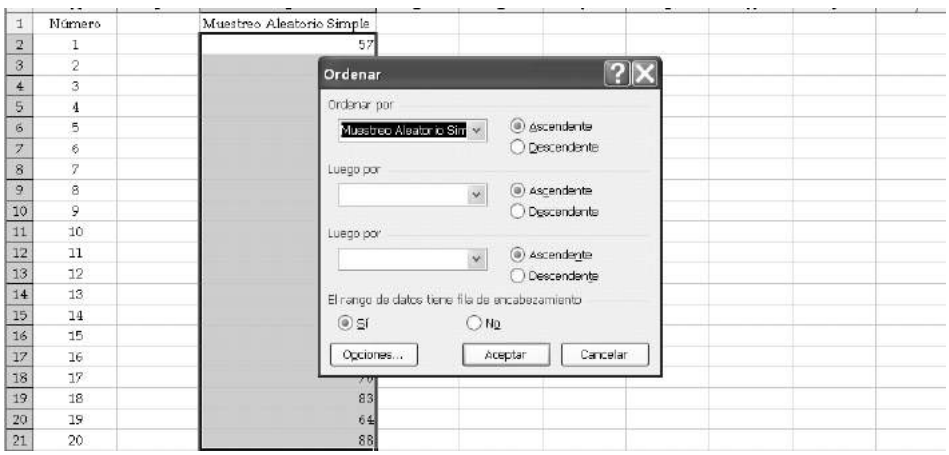
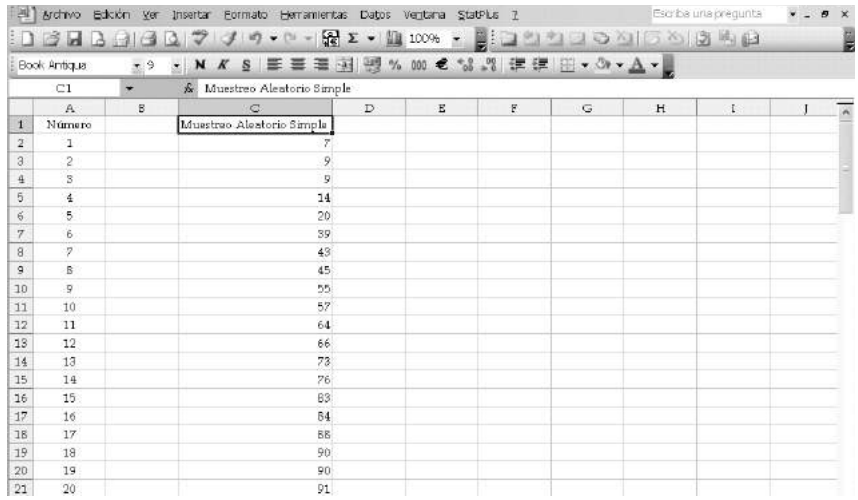


Figura 9. Cuadro de diálogo – Ordenar.

Luego de realizados estos pasos la muestra se debe encontrar ordenada en forma ascendente (véase la figura 10).



	A	B	C	D	E	F	G	H	I	J
1	Número		Muestreo Aleatorio Simple							
2	1		7							
3	2		9							
4	3		9							
5	4		14							
6	5		20							
7	6		39							
8	7		43							
9	8		45							
10	9		55							
11	10		57							
12	11		64							
13	12		66							
14	13		73							
15	14		76							
16	15		83							
17	16		84							
18	17		88							
19	18		90							
20	19		90							
21	20		91							

Figura 10. Ordenamiento de los resultados - Ordenar.

Nota: En la mayoría de los casos se pueden presentar valores repetidos asociados a las unidades de observación; en este caso, la persona asignada con el número 90. Esto se debe a que el MS Excel realiza un muestreo con reemplazo, en este caso se debe seleccionar en forma aleatoria otro número para completar la muestra, por ejemplo 96.

3.1.2 Muestreo sistemático

Es un tipo de muestreo que simplifica el proceso de selección de las unidades de análisis, las cuales se seleccionan en un intervalo constante (salto), que se mide en el tiempo, en el orden o en el espacio.

El método consiste en determinar el tamaño de salto sistemático (k) y elegir el arranque aleatorio (A).

Determinación del salto sistemático: $k = \frac{N}{n}$, donde N es el tamaño de la población y n es el tamaño de la muestra.

Elección del arranque aleatorio: se debe elegir un número aleatorio A entre 1 y k, es decir que A se encuentra acotado de la siguiente manera $1 \leq A \leq k$.

Ejemplo 2:

De acuerdo con el ejemplo anterior:

Sean $N = 100$ y $n = 20$, entonces $k = \frac{100}{20} = 5$.

El arranque aleatorio se selecciona entre las cinco primeras observaciones ($1 \leq A \leq 5$), por ejemplo, si $A = 2$. Las otras unidades seleccionadas con un intervalo o salto sistemático de $k = 5$ son:
2, 7, 12, 17, 22, 27, 32, 37, 42, 47, 52, 57, 62, 67, 72, 77, 82, 87, 92, 97; números que corresponden a la numeración asignada a las personas.

Con Minitab. Para la obtención de la muestra sistemática haciendo uso del Minitab se debe seguir el siguiente procedimiento:

- En la barra de herramientas, seleccione la opción Calc / Make Patterned Data / Sample Set of numbers.
- Indique la columna donde se desea que se almacenen los resultados del muestreo sistemático, <Store patterned data:>, por ejemplo en la columna C5.
- Indique el arranque aleatorio, <From first:>, coloque el valor 2.
- Indique el último valor de la numeración asignada en el marco muestral, <To last:>, colocar valor 100.
- Indique el tamaño del salto sistemático, <In steps of:>, coloque valor 5.
- Pulsar el botón <Ok>.

Lo anteriormente expuesto se aprecia en la figura 11.

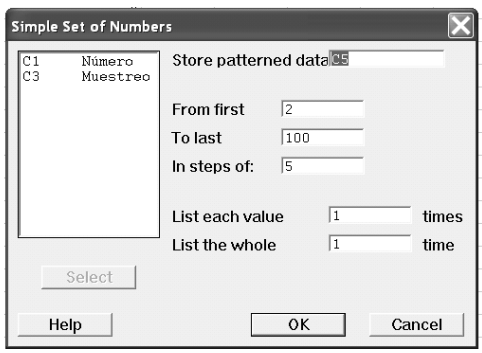
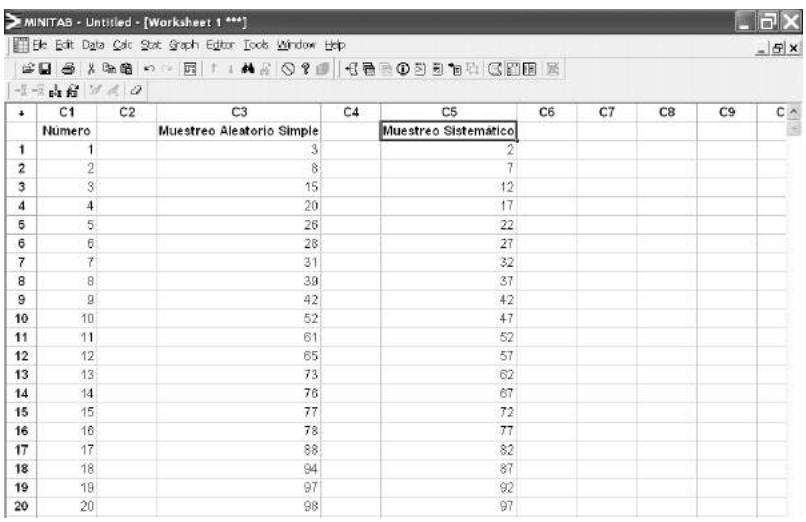


Figura 11. Cuadro de diálogo – Sample Set of Numbers.

Nota: No es necesario trabajar directamente con el marco muestral, pero los números obtenidos deben relacionarse con las posiciones asignadas en el marco muestral.

Los resultados se almacenaron en la columna C5, tal como se indicó, entonces, se procede a etiquetar la columna; por ejemplo, Muestreo Sistemático. De acuerdo con el resultado (figura 12) la muestra estará conformada por personas cuyos números asignados sean: 2, 7, 12, …, 97.



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C
	Número		Muestreo Aleatorio Simple		Muestreo Sistemático					
1	1		3		2					
2	2		8		7					
3	3		15		12					
4	4		20		17					
5	5		26		22					
6	6		28		27					
7	7		31		32					
8	8		39		37					
9	9		42		42					
10	10		52		47					
11	11		61		52					
12	12		65		57					
13	13		73		62					
14	14		76		67					
15	15		77		72					
16	16		78		77					
17	17		88		82					
18	18		94		87					
19	19		97		92					
20	20		98		97					

Figura 12. Salida de resultados – Sample Set of Numbers.

Con software MS Excel. A continuación se presenta la metodología para obtener una muestra sistemática haciendo uso del MS Excel:

- Se debe disponer de una columna con los números asignados a cada unidad muestral.
- En la barra de herramientas, seleccione la opción Herramientas / Análisis de datos.
- Seleccione la opción <Muestra>.
- Pulse en el botón <Aceptar>.
- Seleccione el marco muestral.
  <Rango de entrada:> $A$1:$A$101. (Incluye el rótulo.)
- Marque la opción Rótulos. (Desactive en el caso de no haber incluido el rótulo al momento de seleccionar el marco muestral.)
- Seleccione la opción <Periódico>.
- Indique el tamaño del salto sistemático.
  <Período:> 5.
- Seleccione la opción de salida de resultados. Seleccione la opción Rango de salida e indique la celda donde se desea iniciar la grabación de los resultados.
  <Rango de salida:> E2.
- Pulse el botón <Aceptar>.

Lo anteriormente expuesto se aprecia en la figura 13.

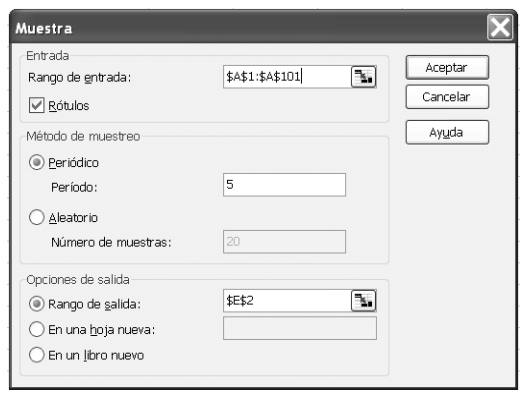
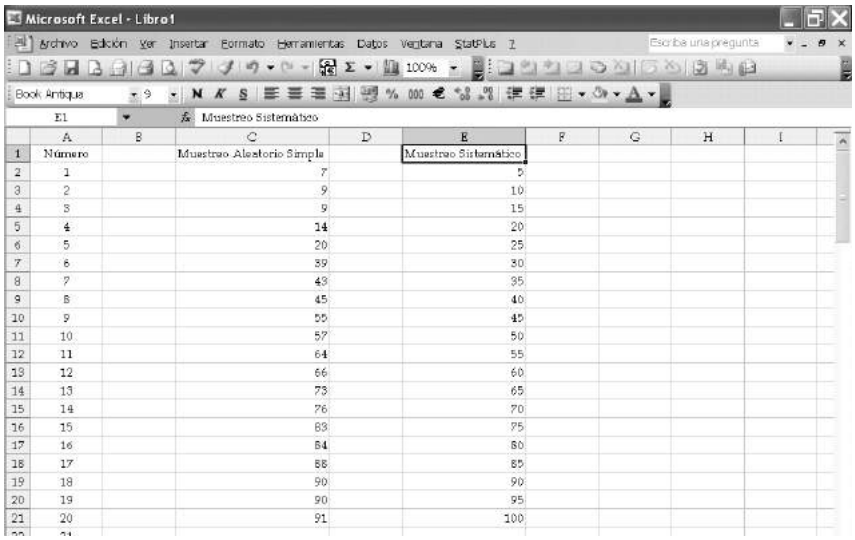


Figura 13. Cuadro de diálogo – Muestreo sistemático

Nota: El MS Excel realiza un muestreo sistemático tomando un arranque aleatorio igual al salto sistemático.
Los resultados se almacenaron a partir de la celda E2, tal como se indicó, entonces se procede a etiquetar la columna; por ejemplo, Muestreo Sistemático. De acuerdo con el resultado (figura 14) la muestra estará conformada por personas cuyos números asignados sean: 5, 10, 15, …, 100.



Microsoft Excel - Libro1

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana StatPlus ?

Escriba una pregunta

Book Antiqua 9 N K S

E1 Muestreo Sistemático

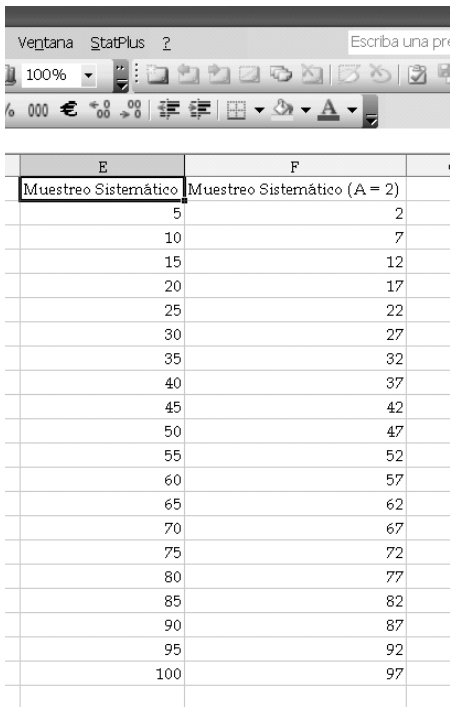
	A	B	C	D	E	F	G	H	I
1	Número		Muestreo Aleatorio Simple		Muestreo Sistemático				
2	1		7		5				
3	2		9		10				
4	3		9		15				
5	4		14		20				
6	5		20		25				
7	6		39		30				
8	7		43		35				
9	8		45		40				
10	9		55		45				
11	10		57		50				
12	11		64		55				
13	12		66		60				
14	13		73		65				
15	14		76		70				
16	15		83		75				
17	16		84		80				
18	17		88		85				
19	18		90		90				
20	19		90		95				
21	20		91		100				

Figura 14. Salida de resultados – Muestreo sistemático

Para poder obtener en el MS Excel un muestreo sistemático con un determinado arranque aleatorio diferente del salto sistemático, se deben tomar como referencia los resultados obtenidos y aplicar fórmulas sencillas de acuerdo con el arranque aleatorio deseado. Por ejemplo, si se deseara una muestra sistemática considerando un arranque aleatorio $A = 2$, se debería seguir el siguiente procedimiento:

- Coloque el cursor en la celda F2 (celda contigua al primer valor de la muestra sistemática hallada anteriormente).
- Introduzca la siguiente fórmula: = E2 – 3, ya que a 5 hay que restarle 3 para que sea igual al arranque aleatorio ($A = 2$).
- Copie la fórmula para las demás celdas contiguas a los demás datos de la muestra sistemática hallada con anterioridad.
- Etiquete la nueva columna; por ejemplo: Muestreo Sistemático ($A = 2$).

La nueva muestra obtenida se puede apreciar en la figura 15.



Ventana StatPlus ?

Escriba una pr

100%

E	F
Muestreo Sistemático	Muestreo Sistemático (A = 2)
5	2
10	7
15	12
20	17
25	22
30	27
35	32
40	37
45	42
50	47
55	52
60	57
65	62
70	67
75	72
80	77
85	82
90	87
95	92
100	97

Figura 15. Cálculos y salida de resultados – Muestreo sistemático (A = 2).

3.1.3 Muestreo estratificado

En este tipo de muestreo la población se divide en grupos o estratos. El principio básico del muestreo estratificado es que los estratos tengan una gran homogeneidad o similitud dentro de cada estrato y heterogeneidad de estrato a estrato.

Cuando ya se ha determinado el número de estratos L y las unidades pertenecientes a cada uno de ellos, el siguiente paso es definir el número de las unidades muestrales a seleccionar dentro de cada estrato. Este proceso es conocido como Asignación o Afijación de la muestra.

Asignación proporcional de la muestra: Es un tipo de asignación que consiste en la distribución de la muestra entre los L estratos de tal manera que el tamaño de cada muestra sea proporcional al tamaño de cada estrato que la origina.

Si N_h es el tamaño del estrato h y n_h es el tamaño de la muestra en dicho estrato, se verifica que:

$$\frac{n_h}{N_h} = \frac{n}{N}$$

Lo que lleva a $n_h = nW_h$, siendo $W_h = \frac{N_h}{N}$ llamado también ponderación del estrato h.

Ejemplo 3:

La prueba de producto es una forma de investigación comercial que tiene como objetivo conocer las bondades del producto en base a la opinión de una muestra de consumidores.

IMA S.A. es una empresa de investigación comercial que realizará una prueba de producto sobre un detergente de ropa.

Para la investigación se seleccionará una muestra de hogares de los 47 distritos de Lima y Callao.

a. Indique la población objetivo del estudio.
b. Para la información necesitada proponga un tipo de muestreo probabilístico. Sustente.
c. Si se planea utilizar un muestreo estratificado, indique cómo distribuiría el tamaño de muestra en los estratos. Explique.

Solución:

a. La población objetivo está constituida por todos los hogares de los 47 distritos de Lima y Callao.
b. Es adecuado utilizar el muestreo estratificado debido a que la selección se realizará de acuerdo con el número de viviendas que tiene cada distrito, el cual es agrupado en estratos homogéneos (distritos).
c. La distribución del tamaño de muestra se realizará mediante la asignación proporcional de acuerdo con el número de viviendas de cada uno de los 47 distritos de Lima y Callao.
   Así por ejemplo:
   – El número total de viviendas en Lima y Callao lo constituyen las viviendas de los 47 distritos.

– Dividiendo el número de viviendas de cada uno de los distritos entre el total se obtiene el porcentaje de viviendas para cada distrito.
– De acuerdo con el porcentaje de viviendas por distrito se distribuirá proporcionalmente la muestra en cada estrato.

3.1.4 Muestreo por conglomerados

A diferencia de las otras técnicas donde se seleccionaban unidades de muestreo, en el muestreo por conglomerados se divide la población en grupos o conglomerados y luego se selecciona una muestra aleatoria de ellos. Por ejemplo: si la unidad de muestreo es la vivienda, el conglomerado puede ser la manzana constituida por viviendas.

La característica del muestreo por conglomerados es que estos son internamente heterogéneos y homogéneos de conglomerado a conglomerado. Por ejemplo, si se desea encuestar a los empleados de una gran empresa con el propósito de averiguar su percepción con respecto a las condiciones de implementación de ciertos préstamos personales que ofrece la empresa, un primer paso será seleccionar una muestra de los diversos departamentos de la empresa; posteriormente se realizaría una selección aleatoria de los empleados dentro de cada uno de los departamentos que resultaran elegidos.

3.2 Muestreo no probabilístico

Los métodos de muestreo no probabilísticos, a diferencia de los probabilísticos, no permiten determinar el error de muestreo; no es posible determinar el nivel de confianza sobre la representatividad de la muestra; además, no permiten realizar inferencias sobre la población.

Existen varios tipos de muestreo no probabilístico, pero los más usados son los siguientes:

3.2.1 Muestreo por cuotas

Esta es una técnica de uso frecuente en la investigación de mercados, sobre todo en las encuestas de opinión. Se basa en el conocimiento de los estratos de una población y de los individuos más representativos de esta para los fines del estudio que se está realizando; en este tipo de muestreo se seleccionan unas cuotas de individuos que reúnen ciertas condiciones; por ejemplo: 50 clientes de telefonía fija que reciben facturación a través de mensajería. Una vez especificada la cuota se eligen los primeros que cumplan estas características.

3.2.2 Muestreo por conveniencia

En este caso, como su nombre lo indica, las unidades reportantes en la muestra se seleccionan de acuerdo con la conveniencia del investigador. Por ejemplo, se puede solicitar a algunas personas que colaboren voluntariamente para probar los productos y después realizar un proceso de monitoreo con ellas. También se puede solicitar la opinión de personas que transitan en un centro comercial. En cada caso, la unidad de muestreo se selecciona sobre la base de su fácil disponibilidad.

3.2.3 *Muestreo de juicio*

Este tipo de muestreo consiste en seleccionar las unidades reportantes a juicio del investigador, el que considera quiénes representan a la población. Una importante diferencia radica en que la muestra no es típica, sino que el investigador la considera como tal. Como se observa, entonces, la eficacia de la muestra de juicio depende de la opinión del investigador o experto que selecciona las unidades por entrevistar.

4. PRINCIPALES ESTIMADORES

La media y la varianza muestral son estimadores y se caracterizan porque varían de muestra a muestra; mientras que la media y la varianza poblacional son valores fijos y en general desconocidos.

Se define la media muestral como: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Y la varianza muestral como: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\bar{x}^2$

Si se tiene una población conformada por N unidades con parámetros μ y σ^2, la representación esquemática de la obtención de k muestras de tamaño n con su propia media y varianza está dada por:

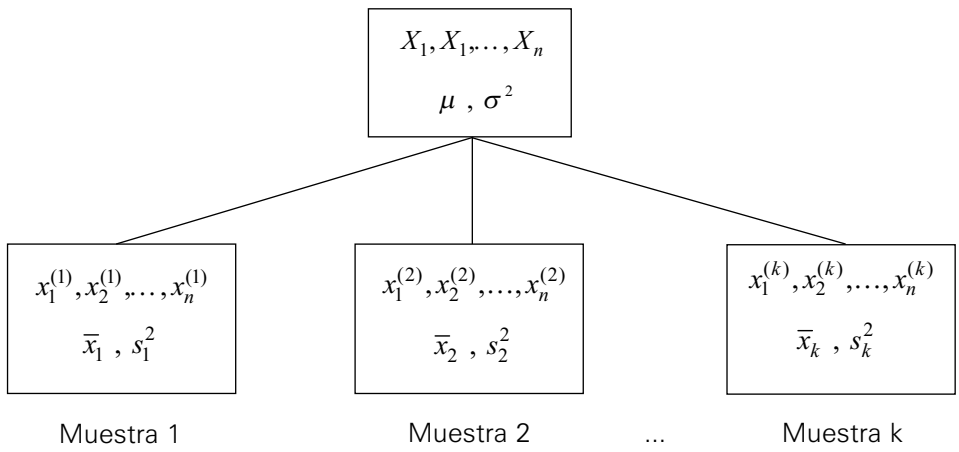


Figura 16. Esquema de la obtención de *k* muestras.

5. Distribución de la media muestral

La distribución de la variable X es: $X \sim (\mu, \sigma^2)$. La distribución de la media muestral se refiere a las características de esperanza y varianza de la media muestral.

Se verifica que la esperanza de la media muestral es igual a la media poblacional.

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{n}{n}\mu = \mu$$

También que la varianza de la media muestral es igual a la varianza poblacional dividida entre el tamaño de la muestra.

$$V(\bar{x}) = V\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}V\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(x_i) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

Se observa, entonces, que mientras mayor sea el tamaño de la muestra, menor será la variabilidad de la media.

Por consiguiente:

$$\bar{x} \rightarrow \left(\mu, \frac{\sigma^2}{n}\right)$$

Ejemplo 4:

Una población está constituida por cuatro productos, los cuales reportan sus precios en nuevos soles:

2, 4, 8, 9

a. Calcule la media y la varianza poblacional de los precios.
b. Determine la distribución muestral de la media de las muestras de tamaño 2 en un muestreo con reposición.

Solución:

Sea X: Precios asociados a los productos.

a. La media y la varianza poblacional son, respectivamente:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

Luego, la media y varianza poblacional que se obtienen son:

$$\mu = \frac{1}{4}\sum_{i=1}^{4} x_i = \frac{2+4+8+9}{4} = 5.75$$

$$\sigma^2 = \frac{1}{4}\sum_{i=1}^{4} x_i^2 - \mu^2 = \frac{2^2+4^2+8^2+9^2}{4} - 5.75^2 = 41.25 - 33.0625 = 8.1875$$

b. Se tiene que $N = 4$ y $n = 2$, como el muestreo es con reemplazamiento, el número de muestras de tamaño 2 es $N^n = 4^2 = 16$.

Muestras de tamaño 2			
(2,2)	(4,2)	(8,2)	(9,2)
(2,4)	(4,4)	(8,4)	(9,4)
(2,8)	(4,8)	(8,8)	(9,8)
(2,9)	(4,9)	(8,9)	(9,9)

Medias muestrales			
2.0	3.0	5.0	5.5
3.0	4.0	6.0	6.5
5.0	6.0	8.0	8.5
5.5	6.5	8.5	9.0

La distribución de probabilidad de las medias muestrales se presenta a continuación:

$\bar{x}$	2.0	3.0	4.0	5.0	5.5	6.0	6.5	8.0	8.5	9.0
$p(\bar{x})$	$1/16$	$2/16$	$1/16$	$2/16$	$2/16$	$2/16$	$2/16$	$1/16$	$2/16$	$1/16$

Obteniéndose:

$$E(\bar{x}) = \sum_{i=1}^{16} \bar{x}_i p(\bar{x}_i) = \left[\left(2\frac{1}{16}\right) + \left(3\frac{2}{16}\right) + \cdots + \left(9\frac{1}{16}\right) \right] = 5.75$$

$$V(\bar{x}) = \sum_{i=1}^{16} \bar{x}_i^2 p(\bar{x}_i) - \mu^2 = \left[\left(2^2\frac{1}{16}\right) + \left(3^2\frac{2}{16}\right) + \cdots + \left(9^2\frac{1}{16}\right) \right] - 5.75^2$$

$$V(\bar{x}) = 37.15625 - 33.0625 = 4.09375$$

Se verifica entonces que:

$E(\bar{x}) = \mu = 5.75$ nuevos soles

$$\sigma_{\bar{x}}^2 = V(\bar{x}) = \frac{\sigma^2}{n} = \frac{8.1875}{2} = 4.0937 \quad \sigma_{\bar{x}} = \sqrt{4.09375} = 2.023302 \text{ nuevos soles}$$

6. Teorema central del límite

Sea X una variable aleatoria con cualquier tipo de distribución, con media μ y varianza finita. Si se toma una muestra aleatoria de tamaño n entonces:

$$\lim_{n\to\infty} \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \lim_{n\to\infty} \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma} \to N(0,1)$$

Lo que implica que cuando el tamaño de muestra crece, la media muestral luego de ser estandarizada converge a una distribución Normal estándar o reducida con parámetros $\mu = 0$ y $\sigma^2 = 1$.

Ejemplo 5:

Los ingresos mensuales, en cientos de dólares, de una empresa consultora tienen distribución Normal con media y desviación estándar, en cientos de dólares, de 100 y 10, respectivamente. Si el promedio mensual de los ingresos varía entre 95 y 105 cientos de dólares, entonces se considera que la situación financiera de la empresa es estable.

a. Calcule la probabilidad de que la empresa consultora mantenga una situación financiera estable en un periodo de 16 meses.
b. ¿Cuál es la probabilidad de que el promedio muestral mensual de los ingresos difiera de su respectiva media poblacional en más de trecientos dólares?
c. ¿De cuántos meses se deberían registrar datos para tener una probabilidad de 0.97 de que el promedio muestral de ingresos se encuentre en el intervalo de 95 a 105 cientos de dólares?

Solución:

a. Si la empresa tuviera una situación financiera estable, los ingresos deberían mantenerse en promedio entre 95 y 105 cientos de dólares. Se define la variable aleatoria X como los ingresos mensuales y $\bar{x}$ como su media en los 16 meses, la distribución de probabilidad es $X \sim N(100,10^2)$, mientras que la media de la muestra se distribuye normalmente con μ $\mu = 100$ y $\sigma_{(\bar{x})} = \sqrt{100/16} = 2.5$. Entonces $\bar{x} \to N\left(100, 2.5^2\right)$.
   Por consiguiente: $P(95 < \bar{x} < 105) = P(\bar{x} < 105) - P(\bar{x} \le 95)$

   Haciendo uso del Minitab

   Cumulative Distribution Function

   ```
   Normal with mean = 100 and standard deviation = 2.5
   x     P( X <= x )
   95    0.022750
   105   0.977250
   ```

   $P(95 < \bar{x} < 105) = 0.977250 - 0.022750 = 0.9545$

   Interpretación:
   Hay una probabilidad de 0.9545 de que la empresa mantenga una situación financiera estable en un período de 16 meses.
b. Se pide que:

   $$P(|\bar{x} - \mu| > 3) = 1 - P(|\bar{x} - \mu| \le 3) = 1 - P(-3 \le \bar{x} - \mu \le 3)$$

   pero se sabe que:

   $$\bar{x} \to N\left(100, 2.5^2\right)$$

   entonces:

   $$P(|\bar{x} - \mu| > 3) = 1 - P(|\bar{x} - \mu| \le 3) = 1 - P(-3 \le \bar{x} - \mu \le 3) = 1 - P(97 \le \bar{x} \le 103)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 100 and standard deviation = 2.5
x     P( X <= x )
97    0.115070
103   0.884930
```

Luego:

$$P\left(\left|\bar{x}-\mu\right|>3\right)=1-\left[P\left(\bar{x}\le 103\right)-P\left(\bar{x}<97\right)\right]=1-\left(0.884930-0.115070\right)=0.23014$$

Interpretación:

La probabilidad de que el promedio muestral difiera de su promedio poblacional en más de trecientos dólares es de 0.23014.

c. En este caso, se desea hallar el valor del tamaño de muestra *n* tal que:

$$P\left(95<\bar{x}<105\right)=0.97$$

Realizando un proceso de estandarización:

$$P\left(95<\bar{x}<105\right)=P\left(\frac{95-100}{10/\sqrt{n}}<\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}<\frac{105-100}{10/\sqrt{n}}\right)=P\left(\frac{-\sqrt{n}}{2}<Z<\frac{\sqrt{n}}{2}\right)$$

$$P\left(95<\bar{x}<105\right)=\left[P\left(Z<\frac{\sqrt{n}}{2}\right)-P\left(Z\le\frac{-\sqrt{n}}{2}\right)\right]=P\left(z<\frac{\sqrt{n}}{2}\right)-\left[1-P\left(Z\le\frac{\sqrt{n}}{2}\right)\right]$$

$$P\left(95<\bar{x}<105\right)=0.97=2P\left(Z\le\frac{\sqrt{n}}{2}\right)-1$$

Por consiguiente:

$$P\left(Z<\frac{\sqrt{n}}{2}\right)=\frac{1+0.97}{2}=0.985$$

Haciendo uso del Minitab

Inverse Cumulative Distribution Function

```
Normal with mean = 0 and standard deviation = 1
P( X <= x )      x
0.985            2.17009
```

Entonces:

$$\frac{\sqrt{n}}{2}=2.17009 \qquad \Rightarrow n=\left((2)2.17009\right)^2=18.837162\approx 19 \text{ meses.}$$

Interpretación:

Se deben registrar datos de aproximadamente 19 meses.

6.1 Aplicación del Teorema Central del Límite a diferentes distribuciones

El Teorema Central del Límite es útil para aproximar la distribución de la media muestral ($\overline{x}$) cuando la muestra aleatoria es obtenida de diferentes distribuciones de probabilidad para valores grandes del tamaño n de la muestra.

6.1.1 Distribución de Poisson

Sea $X \sim P(\lambda)$

Con $E(X)=\lambda$ y $V(X)=\lambda$.

Si se toman muestras de tamaño n grande, la distribución de la media muestral es:

$$E(\overline{x})=\lambda \quad , \qquad V(\overline{x})=\frac{\lambda}{n}$$

Es decir: $\overline{x} \to N\left(\lambda, \frac{\lambda}{n}\right)$

Y por el Teorema Central del Límite: $\underset{n\to\infty}{Lim} \frac{\overline{x}-\lambda}{\sqrt{\lambda/n}} \to N(0,1)$

Ejemplo 6:

Según las autoridades de una entidad bancaria capitalina, la variable número de cuentas cerradas diariamente en el sistema financiero tiene una distribución de Poisson con media 80. Suponga que se seleccionan al azar 50 días y se registra el número de cuentas cerradas por día, ¿cuál es la probabilidad de que el promedio de cuentas cerradas en la muestra sea mayor que 82?
Nota: Si X tiene distribución de Poisson, entonces $E(X) = V(X) = \lambda$.

Solución:

Sea: $X =$ Número de cuentas cerradas diariamente en el sistema financiero.
$X \sim (\lambda = 80)$

Por el Teorema Central del Límite: $\overline{x} \to N(80, 1.2649^2)$

Donde: $\sigma_{(\overline{x})} = \sqrt{80/50} = \sqrt{1.6} = 1.2649$
Luego: $P(\overline{x} > 82) = 1 - P(\overline{x} \le 82)$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 80 and standard deviation = 1.2649

x       P( X <= x )
82      0.943078
```

$P(\overline{x} > 82) = 1 - 0.943078 = 0.056922$

Es decir, la probabilidad de que el promedio de la muestra sea mayor que 82 es de 0.056922.

6.1.2 Distribución uniforme

Sea $X \sim U(\alpha\ ,\ \beta)$

Con $E(X)=\dfrac{\alpha+\beta}{2}$ y $V(X)=\dfrac{(\beta-\alpha)^2}{12}$.

Si se toma una muestra de tamaño n, la distribución de la media muestral es:

$$E(\bar{x})=\frac{\alpha+\beta}{2} \text{ y } V(\bar{x})=\frac{(\beta-\alpha)^2}{12n}$$

Es decir: $\bar{x} \to N\left(\dfrac{\alpha+\beta}{2}, \dfrac{(\beta-\alpha)^2}{12n}\right)$

Y por el Teorema Central del Límite: $\underset{n\to\infty}{Lim} \dfrac{\bar{x}-\left[(\alpha+\beta)/2\right]}{(\beta-\alpha)/\sqrt{12n}} \to N(0,1)$

Ejemplo 7:

Se sabe que el peso de ciertos caramelos es una variable aleatoria con distribución uniforme con parámetros de 10 y 12 gramos. ¿Cuál es la probabilidad de que una caja con 100 caramelos pese más de 1.1 kg? Considere que el peso de la caja vacía es despreciable.

Solución:

Sea X: Peso de un caramelo, cuya distribución es: $X \sim U(10,12)$ entonces:

$$E(X)=\frac{10+12}{2}=11 \qquad \text{y} \quad V(X)=\frac{(12-10)^2}{12}=\frac{4}{12}=0.33$$

El peso de una caja de 100 caramelos está dada por $\sum_{i=1}^{100} x_i$ y su promedio $\bar{x}$.

Por el Teorema Central del Límite, la distribución de $\bar{x}$ está dada por:

$E(\bar{x})=11$ y $\sigma_{(\bar{x})}=\sqrt{0.33/100}=\sqrt{0.0033}=0.057446$

Entonces $\bar{x} \to N(11, 0.057446^2)$.

Como el valor de la variable está en gramos y la probabilidad pedida está en kilos, se realiza la conversión de kilos a gramos, por lo tanto:

$$P\left(\sum_{i=1}^{100} x_i > 1100\right)=P\left(\sum_{i=1}^{100} x_i \Big/ n > 1100/100\right)=P(\bar{x}>11)=1-P(\bar{x}\le 11)=0.5$$

Haciendo uso del Minitab

```
Cumulative Distribution Function
Normal with mean = 11 and standard deviation = 0.0574446
x       P( X <= x )
11      0.5
```

7. Distribuciones de probabilidad

En esta sección se estudiarán las distribuciones Ji Cuadrado, t de Student y F de Fisher.

7.1 Distribución Ji Cuadrado

Definición

Sea $X \sim N(\mu, \sigma^2)$

Se sabe que: $\frac{x_1 - \mu}{\sigma} \sim N(0,1)$

Entonces: $\sum_{i=1}^{n} \left(\frac{x_1 - \mu}{\sigma} \right)^2 \sim \chi^2_{(n)}$

Propiedad: La variable $v = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$

Esto es, la variable aleatoria v tiene una distribución X^2 con (n–1) grados de libertad.

Características. Si X es una variable aleatoria con distribución Ji Cuadrado, entonces:

$$E(X) = n \quad \text{y} \quad V(X) = 2n$$

El parámetro n de la distribución se conoce con el nombre de grados de libertad y es considerado como el número de valores que la variable puede tomar libremente con la condición de que la suma debe ser igual a un valor fijo, este valor es asociado con el tamaño de la muestra.

Ejemplo 8:

En la elaboración de una prueba de aptitud para diferentes puestos de trabajo en una empresa, el departamento de Recursos Humanos planificó una dispersión bastante elevada en las calificaciones para identificar con facilidad a los mejores candidatos para los diferentes puestos. Si se supone que las calificaciones se distribuyen normalmente con un promedio de 80 puntos y una desviación estándar de 10 puntos y la empresa cuenta con 12 postulantes, a los cuales se les aplicará la prueba de aptitud, calcule la probabilidad aproximada de que la varianza muestral de las calificaciones de dichos postulantes sea mayor que 15 puntos^2.

Solución:

Sea X: Calificación de los postulantes.
De acuerdo con el enunciado, se tienen los siguientes datos:

$$\mu = 80 \text{ puntos} \quad \sigma = 10 \text{ puntos} \quad n = 12$$

Se sabe también que: $\dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$

Se desea calcular:

$$P(s^2 > 15) = P\left[\frac{(n-1)s^2}{\sigma^2} > \frac{(12-1)15}{10^2}\right] = P(\chi^2_{(11)} > 1.65) = 1 - P(\chi^2_{(11)} \leq 1.65)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Chi-Square with 11 DF
    x     P( X <= x )
 1.65     0.0006037
```

$$P\left(s^2 > 15\right) = 1 - 0.0006037 = 0.9993963$$

Interpretación:
La probabilidad de que la varianza muestral de las calificaciones de dichos postulantes sea mayor que 15 puntos^2 es 0.9993963.

7.2 Distribución *t* de Student

Definición. Sean $Z \sim N(0, 1)$ y $Y \sim \chi^2_{(k)}$ variables aleatorias independientes. Sea la variable aleatoria T definida como:

$$T = \frac{Z}{\sqrt{\dfrac{Y}{k}}} \sim t_{(k)}$$

Entonces, la variable T tiene distribución t con k grados de libertad.

Características:

$X \sim t_{(k)}$. Entonces:

$E(X) = 0$; para $k > 1$

$V(X) = \dfrac{k}{k-2}$; para $k > 2$

La distribución t de Student es muy similar a la distribución Normal, ya que ambas varían en el conjunto de los números reales, aunque la distribución t tiene mayor dispersión. Sin embargo, la varianza de la distribución t se aproxima a 1 cuando k tiende a valores muy grandes.

Propiedad. Sea $X \sim N(\mu, \sigma^2)$. Si se obtienen muestras de tamaño n, esto es x_1, $x_2,\ldots, x_n$ y se calcula su media y varianza muestral. Entonces se define la variable:

$$\frac{\bar{x}-\mu}{s/\sqrt{n}}$$

Esta variable tiene entonces una distribución t con $(n-1)$ grados de libertad.

Ejemplo 9:

El equipo de soporte técnico de una compañía aseguradora está en un proceso de reemplazo de 22 unidades antiguas de computadoras por igual número de última generación. El tiempo de instalación de las 22 nuevas máquinas para que operen a nivel satisfactorio tiene una desviación estándar de 3.90 horas.

El equipo, por trabajos previos, manifiesta que el tiempo de instalación promedio antes de operar satisfactoriamente es de 8.1 horas. Para el conjunto de las nuevas máquinas, ¿cuál es la probabilidad de que el tiempo promedio de instalación sea menor a 7 horas?

Solución:

Sea X: Tiempo de instalación de las nuevas computadoras
Además: $\mu = 8.1 \qquad n = 22 \qquad s = 3.90$
Se desea hallar: $P(\bar{x} < 7)$

$$P(\bar{x}<7) = P\left(\frac{\bar{x}-\mu}{s/\sqrt{n}} < \frac{7-8.1}{3.9/\sqrt{22}}\right) = P(t_{(21)} < -1.32)$$

Haciendo uso del Minitab

```
Cumulative Distribution Function
Student's t distribution with 21 DF
    x       P( X <= x )
-1.32      0.100522
```

Entonces: $P(\bar{x} < 7) = 0.100522$

7.3 Distribución *F* de Fisher

Definición. Sean $X \sim \chi^2_{(m)}$ y $Y \sim \chi^2_{(n)}$ variables aleatorias independientes, entonces la variable definida como:

$$W = \frac{\frac{X}{m}}{\frac{Y}{n}} = \frac{nX}{mY}$$

Tiene una distribución F con parámetros m y n, y se denota $F_{(m,n)}$.
Para la variable $W \sim F_{(m,n)}$, se cumple que:

Características:

$$E(W)=\frac{n}{n-2}\ ,\quad \text{para } n>2$$

$$V(W)=\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}\text{, para } n>4$$

Propiedad: Si se tienen las variables aleatorias independientes $X \sim N(\mu_X,\ \sigma_X^2)$ y $Y \sim N(\mu_Y,\ \sigma_Y^2)$, y se seleccionan muestras aleatorias con reemplazo de tamaños *m* y *n*, respectivamente. Entonces:

$$\frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} \sim F_{(m-1,n-1)}$$

Ejemplo 10:

La variabilidad en la cantidad de artículos producidos por dos grupos de trabajadores A y B es una característica necesaria de controlar para que la producción marche de manera óptima. Lo deseable es que la producción en ambos grupos tenga una variabilidad baja. Para ello se selecciona una muestra de 15 trabajadores del grupo A; mientras que en el grupo B se selecciona otra muestra de 21 trabajadores.

¿Cuál es la probabilidad de que la varianza muestral para el grupo A sea menor que 2.2 veces la varianza muestral del grupo B, suponiendo que la varianza poblacional en ambos grupos es la misma?

Solución:

Según el enunciado del problema, se supone que σ_A^2 y σ_B^2 son homogéneas y se conoce que:

$n_A = 15 \qquad n_B = 21$

$$P(s_A^2 < 2.2s_B^2) = P\left(\frac{s_A^2}{s_B^2} < 2.2\right) = P\left(\frac{s_A^2\sigma_B^2}{s_B^2\sigma_A^2} < 2.2\,\frac{\sigma_B^2}{\sigma_A^2}\right)$$

$$P(s_A^2 < 2.2s_B^2) = P(F_{(14,20)} < 2.2)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
F distribution with 14 DF in numerator and 20 DF in
denominator
 X           P( X <= x )
 2.2         0.947611
```

$$P(s_A^2 < 2.2s_B^2) = 0.947611$$

8. Distribuciones muestrales de un estimador

Se denomina distribución muestral de un estimador a la distribución de probabilidad que se genera por la extracción de un número muy grande de muestras, calculándose en cada una el estimador respectivo. Las principales distribuciones muestrales se presentan a continuación:

8.1 Distribución de una media muestral con varianza poblacional conocida

Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria con reemplazo de tamaño n obtenida de una distribución Normal $N(\mu, \sigma^2)$. Si $\bar{x}$ es la media muestral, entonces:

$$E(\bar{x}) = \mu \quad \text{y} \quad Var(\bar{x}) = \frac{\sigma^2}{n}$$

Por consiguiente, la variable aleatoria:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Nota: Téngase en cuenta la diferencia entre $\frac{X - \mu}{\sigma}$ (valor individual) y $\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ (promedio)

Ejemplo 11:

Una máquina empaqueta un determinado producto, cuyo peso en gramos se distribuye normalmente con una desviación estándar de 20 gramos y con una media μ de 500, que debe ser regulada constantemente.

a. La media μ se encuentra bien regulada solo si el 1% de los pesos de todos los paquetes que produce la máquina tienen pesos mayores a 546.5 gramos. ¿Cuánto debe valer?
b. Con la media bien regulada, se programa el siguiente control del peso del producto: cada hora se escogen al azar 4 paquetes, si el promedio de los pesos no está entre 480 y 520 gramos, se para la máquina para mantenimiento. En caso contrario, se continúa el proceso. ¿Cuál es la probabilidad de parar la máquina cuando realmente está bien regulada?
c. Si la máquina está bien regulada, ¿con qué tamaño de muestra se consigue que la media muestral sea a lo más 409.2 gramos con probabilidad igual a 0.025?

Solución:

a. Sea X: Peso del producto empaquetado por la máquina, $X \sim N(\mu, 20^2)$.
   Se debe calcular μ tal que: $P(X > 546.5) = 0.01$. Entonces:

$$P(X > 546.5) = 0.01 \Rightarrow \left[1 - P(X \le 546.5)\right] = 0.01$$

$$\Rightarrow P(X \le 546.5) = 1 - 0.01 = 0.99$$

$$P\left(\frac{X-\mu}{\sigma} \le \frac{546.5-\mu}{20}\right) = 0.99 \qquad \Rightarrow P\left(Z \le \frac{546.5-\mu}{20}\right) = 0.99$$

Haciendo uso del Minitab

Inverse Cumulative Distribution Function

```
Normal with mean = 0 and standard deviation = 1
P( X <= x )          x
     0.99        2.32635
```

Luego: $\frac{546.5-\mu}{20} = 2.32635 \qquad \Rightarrow \mu = 546.5 - ((20)2.32635) \cong 500\,\text{gramos}$.

Nota: En una distribución Normal, cuando no se conozca μ, σ o n se debe aplicar la estandarización de la variable.

b. Sea $\overline{x}$ la media de la muestra de tamaño 4

$\overline{x} \to N(500, 10^2)$, donde: $\sigma_{\overline{x}} = \sigma/\sqrt{n} = 20/\sqrt{4} = 10$

La probabilidad de no parar la máquina cuando realmente está bien regulada es:

$$P(480 \le \overline{x} \le 520) = P(\overline{x} \le 520) - P(\overline{x} < 480)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 500 and standard deviation = 10
x          P( X <= x )
480        0.022750
520        0.977250
```

Entonces: $P(480 \le \overline{x} \le 520) = 0.977250 - 0.022750 = 0.9545$

Luego: La probabilidad de parar la máquina es: 1 - 0.9545 = 0.0455

c. $P(\overline{x} \le 490.2) = 0.025$

Estandarizando:

$$P(\overline{x} \le 490.2) = 0.025 \qquad \Rightarrow P\left(\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{490.2-500}{\frac{20}{\sqrt{n}}}\right) = 0.025$$

$$P\left(Z \le \frac{490.2-500}{\frac{20}{\sqrt{n}}}\right) = 0.025 \qquad \Rightarrow P\left(Z \le \frac{\sqrt{n}*(-9.8)}{20}\right) = 0.025$$

Haciendo uso del Minitab

Inverse Cumulative Distribution Function

```
Normal with mean = 0 and standard deviation = 1
P( X <= x )          x
  0.025         -1.95996
```

Luego:

$$\Rightarrow \frac{-9.8\sqrt{n}}{20} = -1.95996 \quad \Rightarrow \sqrt{n} = \frac{(20)1.95996}{9.8} = 3.999918 \quad \Rightarrow n = 15.999347 \approx 16$$

8.2 Distribución de una media muestral con varianza poblacional desconocida

Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria con reemplazo de tamaño n escogida de una distribución Normal $N(\mu, \sigma^2)$, donde la varianza poblacional σ^2 es desconocida, entonces la variable aleatoria:

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

tiene una distribución t con $(n-1)$ grados de libertad.

Nota: Tener presente que la distribución de la variable X debe ser Normal; de otro modo, este resultado es inaplicable.

Ejemplo 12:

Si $\overline{x}$ es la media y s^2 es la varianza de una muestra aleatoria de tamaño $n = 9$ seleccionada de una población Normal con media $\mu = 90$.

Calcule: $P\left(0.2353 \le \frac{\overline{x} - \mu}{s} \le 1.1183\right)$

Solución:

De los datos se tiene que: $\frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t_{(8)}$

Entonces:

$$P\left(0.2353 \le \frac{\bar{x}-\mu}{s} \le 1.1183\right) = P\left(\frac{0.2353}{1/\sqrt{n}} \le \frac{\bar{x}-\mu}{s/\sqrt{n}} \le \frac{1.1183}{1/\sqrt{n}}\right) =$$

$$P((3)0.2353 \le t_{(8)} \le (3)1.1183)$$

$$P\left(0.2353 \le \frac{\bar{x}-\mu}{s} \le 1.1183\right) = P(0.7059 \le t_{(8)} \le 3.3549) =$$

$$P(t_{(8)} \le 3.3549) - P(t_{(8)} < 0.705\ 9)$$

Haciendo uso del Minitab

Cumulative Distribution Function

Student's t distribution with 8 DF

x	P(X <= x)
3.3549	0.994996
0.7059	0.749857

$$P\left(0.2353 \le \frac{\bar{x}-\mu}{s} \le 1.1183\right) = 0.994996 - 0.749857 = 0.245139$$

8.3 Distribución de una proporción muestral

Sea $x_1, x_2, \ldots, x_n$ una muestra aleatoria con reemplazo de tamaño n extraída de una población de Bernoulli $B(1, \pi)$, donde π es la proporción de éxitos en la muestra. La proporción de éxitos en la muestra se encuentra definida como:

$$p = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{Y}{n}$$

$Y =$ Número de éxitos en la muestra.

Nota: Recordar que una variable aleatoria Bernoulli asume únicamente los valores 0 y 1.

Su distribución muestral es:

$$\mu_p = E(p) = E\left(\frac{Y}{n}\right) = \frac{1}{n}E(Y) = \frac{1}{n}(n\pi) = \pi$$

$$\sigma_p^2 = V(p) = V\left(\frac{Y}{n}\right) = \frac{1}{n^2}V(Y) = \frac{1}{n^2}\left[n\pi(1-\pi)\right] = \frac{\pi(1-\pi)}{n}$$

Si n es suficientemente grande, entonces la variable aleatoria:

$$Z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$$

Nota: La desviación estándar de una distribución muestral se denomina también **error estándar**. Para la distribución de la proporción muestral, el error estándar es:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Ejemplo 13:

Según la Asociación de Empresas de Seguros, 4 de cada 10 choferes de vehículos de transporte público ha sufrido algún tipo de accidente.

a. ¿Cuál es la probabilidad de que en una muestra de tamaño 250, la proporción de choferes de vehículos de transporte público que ha sufrido algún tipo de accidente sea mayor que 0.45?
b. ¿Cuántos choferes deben ser seleccionados para que la proporción de choferes de vehículos de transporte público que ha sufrido algún tipo de accidente difiera de la proporción poblacional correspondiente con un margen de error de estimación menor que 2% y con 0.96 de probabilidad?

Solución:

Sea X variable dicotómica $X: \begin{cases} 1, & \text{si el chofer ha sufrido accidente} \\ 0, & \text{caso contrario} \end{cases}$

La distribución de p es:

$p \to N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$, entonces: $p \to N\left(0.4, 0.030984^2\right)$ con

$$\sigma_p = \sqrt{\frac{0.4(0.6)}{250}} = 0.030984$$

a. $P(p > 0.45) = 1 - P(p \le 0.45) = 1 - 0.94671 = 0.05329$

Haciendo uso del Minitab

```
Cumulative Distribution Function
Normal with mean = 0.4 and standard deviation = 0.030984
   x       P( X <= x )
0.45       0.946708
```

b. Se desea determinar el número de choferes de tal manera que:

$$P\left(|p-\pi| < 0.02\right) = 0.96 \quad \Rightarrow P\left(\frac{-0.02}{\sqrt{\frac{0.4(0.6)}{n}}} < \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} < \frac{0.02}{\sqrt{\frac{0.4(0.6)}{n}}}\right) = 0.96$$

$$P\left(|p-\pi|<0.02\right)=2P\left(Z<\frac{0.02}{\sqrt{\frac{0.24}{n}}}\right)-1=0.96 \Rightarrow P\left(Z<\frac{0.02}{\sqrt{\frac{0.24}{n}}}\right)=\frac{1.96}{2}=0.98$$

Haciendo uso del Minitab

Inverse Cumulative Distribution Function

```
Normal with mean = 0 and standard deviation = 1
P( X <= x )       x
      0.98        2.05375
```

Luego: $\frac{0.02}{\sqrt{\frac{0.24}{n}}}=2.05375 \quad \Rightarrow \quad n=2530.733438\approx 2531$ choferes.

Se deben seleccionar 2.531 choferes.

8.4 Distribución de la varianza muestral

Sea $x_1, x_2,\ldots, x_n$ una muestra aleatoria con reemplazo de tamaño *n* escogida de una distribución Normal $N(\mu,\sigma^2)$, y sea: $s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

Entonces: esta variable $\frac{(n-1)s^2}{\sigma^2}$ tiene una distribución Ji Cuadrado con (n –1) grados de libertad.

Propiedades:

$$E\left(s^2\right)=\sigma^2 \qquad\qquad V\left(s^2\right)=\frac{2\sigma^4}{n-1}$$

Ejemplo 14:

En una embotelladora de gaseosas se dispone de una máquina reguladora, de tal manera que llena en promedio μ onzas por botella. El jefe de la planta afirma que la máquina embotelladora llena las botellas de acuerdo con una distribución Normal.

a. Se selecciona una muestra aleatoria de 15 botellas y se mide el contenido de cada botella, encuentre los valores "*a*" y "*b*" tales que $P\left(a\le s^2\le b\right)=0.95$. Considere el 95% de la parte central de la distribución con $\sigma^2 = 1.4$ onzas^2.

b. Considere ahora una muestra de 20 botellas y que $\sigma^2 = 1.12$. Encuentre los valores "*a*" y "*b*" tal que $P\left(2a \le s^2 \le \left(\frac{b-0.075}{3}\right)\right) = 0.90$. Considere el 90% de la parte central de la distribución.

Solución:

a. Hallando a y b tal que: $P(a \le s^2 \le b) = 0.95$

Entonces:

$$P(a \le s^2 \le b) = P\left(\frac{(n-1)a}{\sigma^2} \le \frac{(n-1)s^2}{\sigma^2} \le \frac{(n-1)b}{\sigma^2}\right) = P\left(\frac{14a}{1.4} \le \chi^2_{(15-1)} \le \frac{14b}{1.4}\right)$$

$$P(a \le s^2 \le b) = P(10a \le \chi^2_{14} \le 10b) = P(\chi^2_{14} \le 10b) - P(\chi^2_{14} < 10a) = 0.95$$

Ya que se tiene en la parte central el 95% de las observaciones, tal como se aprecia en la figura 17.

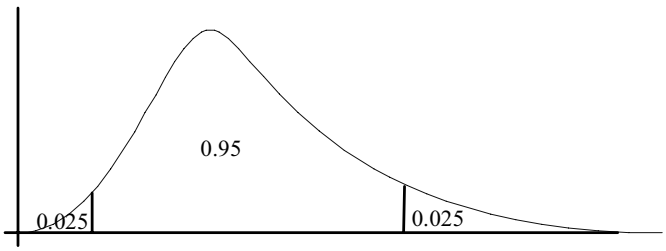


Figura 17.

Haciendo uso del Minitab

Inverse Cumulative Distribution Function

```
Chi-Square with 14 DF
P( X <= x )          x
   0.975          26.1189
   0.025           5.6287
```

$P(\chi^2_{14} \le 10b) = 0.975 \quad \Rightarrow \quad 10b = 26.1189 \quad \Rightarrow \quad b = 2.61189 \approx 2.61$

$P(\chi^2_{14} \le 10a) = 0.025 \quad \Rightarrow \quad 10a = 5.6287 \quad \Rightarrow \quad a = 0.56287 \approx 0.56$

b. Hallando a y b:

$$P\left(2a \le s^2 \le \left(\frac{b-0.075}{3}\right)\right) = P\left(\frac{(19)2a}{1.12} \le \chi^2_{(19)} \le \frac{(19)}{1.12}\left(\frac{b-0.075}{3}\right)\right) = 0.90$$

$$P\left(2a \le s^2 \le \left(\frac{b-0.075}{3}\right)\right) = P\left(33.928571a \le \chi^2_{(19)} \le 5.654762(b-0.075)\right) = 0.90$$

$$P\left(\chi^2_{19} \le (5.654762b - 0.424107)\right) = 0.95 \;\Rightarrow\; (5.6548762b - 0.424107) = 30.1435$$

$$\Rightarrow\; b = 5.405555 \approx 5.41$$

$$P\left(\chi^2_{19} \le 33.928571a\right) = 0.05 \;\Rightarrow 33.928571a = 10.1170 \;\Rightarrow\; a = 0.298185 \approx 0.30$$

9. Distribuciones muestrales de dos muestras

Cuando se trata de comparar dos poblaciones de acuerdo con una característica de interés, se comparan los valores medios poblacionales mediante muestras aleatorias tomadas de ambas poblaciones.

9.1 Diferencia de medias muestrales con varianzas poblacionales conocidas

Sean: $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$ dos variables aleatorias independientes.

Si se toman muestras con reemplazo de tamaño n_x y n_y y se obtienen las distribuciones de sus medias muestrales, entonces:

$$\bar{x} \to N\left(\mu_X, \frac{\sigma_X^2}{n_X}\right) \qquad , \qquad \bar{y} \to N\left(\mu_Y, \frac{\sigma_Y^2}{n_Y}\right)$$

La distribución de la diferencia de las medias muestrales está dada por:

$$(\bar{x} - \bar{y}) \to N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$$

donde la esperanza y varianza se obtienen de la siguiente manera:

$$\mu_{(\bar{x}-\bar{y})} = E(\bar{x} - \bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_X - \mu_Y$$

$$\sigma^2_{(\bar{x}-\bar{y})} = V(\bar{x} - \bar{y}) = V(\bar{x}) + V(\bar{y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$$

Nota: La expresión $(\bar{x} - \bar{y})$ representa una variable aleatoria.

Ejemplo 15:

Los celulares producidos por la empresa Audix tienen una duración media de 80 meses y una desviación estándar de 5 meses, mientras que los celulares fabricados por la empresa Fonorola tienen una duración media de 75 meses y una desviación estándar de 3 meses. ¿Cuál es la probabilidad de que una muestra aleatoria de 36 celulares de la empresa Audix tenga una duración media de al menos tres meses más que la duración media de 49 celulares de la empresa Fonorola? Suponga que las duraciones medias en ambos casos se distribuyen normalmente y son independientes.

Solución:

Sea X: Duración de los celulares Audiovox con la siguiente información:

$\mu_X = 80 \quad \sigma_X = 5 \quad n_X = 36 \quad \bar{x} \sim \left(80, 0.8333^2\right)$, donde: $\sigma_{\bar{x}} = \frac{5}{\sqrt{36}} = 0.8333$

Sea Y: Duración de los celulares Motorola con:

$\mu_Y = 75 \ \sigma_Y = 3 \ \ n_Y = 49 \ \ \bar{y} \sim \left(75, 0.4286^2\right)$ donde: $\sigma_{\bar{y}} = \frac{3}{\sqrt{49}} = 0.4286$

La distribución de la diferencia de medias muestrales es:

$$\left(\bar{x} - \bar{y}\right) \rightarrow N\left(5, 0.937079^2\right)$$

donde: $\sigma^2_{(\bar{x}-\bar{y})} = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y} = \frac{5^2}{36} + \frac{3^2}{49} = 0.878118$

$\sigma_{(\bar{x}-\bar{y})} = \sqrt{\sigma^2_{(\bar{x}-\bar{y})}} = \sqrt{0.878118} = 0.937079$

Entonces: $P\left[\left(\bar{x} - \bar{y}\right) \geq 3\right] = 1 - P\left[\left(\bar{x} - \bar{y}\right) < 3\right]$

Haciendo uso del Minitab

```
Cumulative Distribution Function
Normal with mean = 5 and standard deviation = 0.937079
x   P(X <= x)
3   0.0164095
```

$P\left[\left(\bar{x} - \bar{y}\right) \geq 3\right] = 1 - 0.0164095 = 0.9835905$

9.2 Diferencia de medias muestrales con varianzas poblacionales desconocidas

En el caso de que las varianzas σ_1^2 y σ_2^2 sean desconocidas, estas varianzas pueden ser homogéneas o heterogéneas.

9.2.1 Varianzas poblacionales homogéneas $\left(\sigma_1^2 = \sigma_2^2 = \sigma^2\right)$

Para ese caso, la variable aleatoria $\left(\bar{x}_1 - \bar{x}_2\right)$ tiene la siguiente distribución:

$$\frac{\left(\bar{x}_1 - \bar{x}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)} \qquad \text{Con: } s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

Ejemplo 16:

Inspectel se encuentra realizando un estudio de investigación de mercados con relación al consumo de los clientes y a los costos de los servicios que brindan las empresas de telecomunicaciones A y B. Se determinó que el consumo en nuevos soles por hora de los clientes y los costos por minutos (S/.) de cada una de las empresas presenta el siguiente comportamiento:

Consumo S/. (hora)		
Empresa	Distribución	Media
A	Nomal	22
B	Normal	27

Consumo S/. (minuto)		
Empresa	Distribución	Media
A	Nomal	1.69
B	Normal	1.85

a. Se registró una muestra aleatoria del consumo de 18 clientes de la empresa A y de 20 clientes de la empresa B, los cuales presentan una desviación estándar muestral de 6.18 y 4.39, respectivamente. ¿Cuál es la probabilidad de que la diferencia del consumo promedio muestral de la empresa A con respecto al consumo promedio muestral de la empresa B sea por lo menos S/.1? Suponga varianzas poblacionales desconocidas pero homogéneas.

b. Si se seleccionaron aleatoriamente muestras de 20 y 16 clientes de las empresas A y B respectivamente, obteniéndose $s_A = 0.82$ y $s_B = 0.57$, ¿cuál es la probabilidad de que estos promedios muestrales difieran como máximo en S/.0.5? Suponga varianzas poblacionales desconocidas pero homogéneas.

Solución:

a. La probabilidad pedida es:

$$P(\bar{x}_A - \bar{x}_B \geq 1) = 1 - P\left(\frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{S_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \leq \frac{1 - (\mu_A - \mu_B)}{\sqrt{s_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \right) =$$

$$1 - P\left(t_{(n_A + n_B - 2)} \leq \frac{1 - (22 - 27)}{\sqrt{s_p^2\left(\frac{1}{18} + \frac{1}{20}\right)}} \right)$$

Donde:

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = \frac{(18-1)6.18^2 + (20-1)4.39^2}{18 + 20 - 2} = 28.206686$$

Reemplazando:

$$P(\bar{x}_A - \bar{x}_B \geq 1) = 1 - P\left(t_{(18+20-2)} < \frac{1 - (-5)}{\sqrt{28.206686(0.105556)}} \right) =$$

$$1 - P\left(t_{(36)} < \frac{6}{1.725506} \right)$$

$$P(\bar{x}_A - \bar{x}_B \geq 1) = 1 - P\left(t_{(36)} < 3.477240 \right)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Student's t distribution with 36 DF
     x          P( X <= x )
3.47724       0.999329
```

$$P(\overline{x}_A - \overline{x}_B \geq 1) = 1 - 0.999329 = 0.000671$$

b. Sean las variables aleatorias:

$\overline{x}_A$: El costo promedio muestral que pagan los clientes de la empresa A.

$\overline{x}_B$: El costo promedio muestral que pagan los clientes de la empresa B.

La probabilidad pedida es:

$$P(|\overline{x}_A - \overline{x}_B| \leq 0.5) = P(-0.5 \leq \overline{x}_A - \overline{x}_B \leq 0.5)$$

$$P(|\overline{x}_A - \overline{x}_B| \leq 0.5) = P\left(\frac{-0.5-(\mu_A-\mu_B)}{\sqrt{S_p^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}} \leq \frac{(\overline{x}_A-\overline{x}_B)-(\mu_A-\mu_B)}{\sqrt{S_p^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}} \leq \frac{0.5-(\mu_A-\mu_B)}{\sqrt{S_p^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}}\right)$$

$$P(|\overline{x}_A - \overline{x}_B| \leq 0.5) = P\left(\frac{-0.5-(1.69-1.85)}{\sqrt{S_p^2\left(\frac{1}{20}+\frac{1}{16}\right)}} \leq t_{(20+16-2)} \leq \frac{0.5-(1.69-1.85)}{\sqrt{S_p^2\left(\frac{1}{20}+\frac{1}{16}\right)}}\right)$$

Donde: $s_p^2 = \dfrac{(n_A-1)s_A^2+(n_B-1)s_B^2}{n_A+n_B-2} = \dfrac{(20-1)0.82^2+(16-1)0.57^2}{20+16-2} = 0.519091$

Reemplazando:

$$P(|\overline{x}_A - \overline{x}_B| \leq 0.5) = P\left(\frac{-0.5-(0.16)}{\sqrt{0.519091(0.1125)}} \leq t_{(34)} \leq \frac{0.5-(0.16)}{\sqrt{0.519091(0.1125)}}\right)$$

$$P(|\overline{x}_A - \overline{x}_B| \leq 0.5) = P\left(\frac{-0.66}{0.241656} \leq t_{(34)} \leq \frac{0.34}{0.241656}\right) = P(-2.731155 \leq t_{(34)} \leq 1.406959)$$

$$P(|\overline{x}_1 - \overline{x}_2| \leq 0.5) = P(t_{(34)} \leq 1.406959) - P(t_{(34)} < -2.73115)$$

Haciendo uso del Minitab

Cumulative Distribution Function

Student's t distribution with 34 DF

x	P(X <= x)
1.40696	0.915742
-2.73115	0.004966

$$P(|\bar{x}_1 - \bar{x}_2| \leq 0.5) = (0.915742 - 0.004966) = 0.910776$$

9.2.2 *Varianzas poblacionales heterogéneas* ($\sigma_1^2 \neq \sigma_2^2$)

En este caso la distribución de la variable aleatoria $(\bar{x}_1 - \bar{x}_2)$ es:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim t_{(v)}$$

Donde ν corresponde a los grados de libertad.

$$v \cong \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Ejemplo 17

En Securitex se sabe que el número promedio de miembros por familia afiliados en el rubro seguro hogar es de 7 miembros; y el número promedio de miembros por familia afiliados en el rubro seguro vida es de 4 miembros. Si se toman las siguientes muestras aleatorias: 39 afiliados al rubro Hogar y 34 afiliados al rubro Vida. Obteniéndose $s_1 = 1.4$ y $s_2 = 1.8$.
¿Cuál es la probabilidad de que el promedio muestral del número de miembros por familia de los afiliados al rubro Hogar difiera del promedio muestral de los afiliados al rubro Vida en más de 4 miembros? Suponga varianzas poblacionales desconocidas pero heterogéneas.

Solución:

Sea $\bar{x}_1$: Número promedio de miembros por familia afiliados al rubro seguro hogar.

$\bar{x}_1 \rightarrow N(7, 1.4^2)$ $\qquad$ $n_1 = 39$

$\bar{x}_2$: Número promedio de miembros por familia afiliados al rubro de seguro vida.

$\bar{x}_2 \rightarrow N(4, 1.8^2)$ $\qquad$ $n_2 = 34$

Se desea hallar: $P(|\bar{x}_1 - \bar{x}_2| > 4)$

$$P(|\bar{x}_1 - \bar{x}_2| > 4) = 1 - P(|\bar{x}_1 - \bar{x}_2| \leq 4) = 1 - P(-4 \leq \bar{x}_1 - \bar{x}_2 \leq 4)$$

$$P(|\bar{x}_1 - \bar{x}_2| > 4) = 1 - P\left(\frac{-4 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \leq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \leq \frac{4 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \right)$$

Reemplazando valores:

$$P(|\bar{x}_1 - \bar{x}_2| > 4) = 1 - P\left(\frac{-4 - (7 - 4)}{\sqrt{\left(\frac{1.4^2}{39} + \frac{1.8^2}{34}\right)}} \leq t_{(v)} \leq \frac{4 - (7 - 4)}{\sqrt{\left(\frac{1.4^2}{39} + \frac{1.8^2}{34}\right)}} \right)$$

Donde: $$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2 - 1)}} = \frac{\left(\frac{1.4^2}{39} + \frac{1.8^2}{34}\right)^2}{\frac{\left(\frac{1.4^2}{39}\right)^2}{(39 - 1)} + \frac{\left(\frac{1.8^2}{34}\right)^2}{(34 - 1)}} = 62.008348 \cong 62$$

Reemplazando:

$$P(|\bar{x}_1 - \bar{x}_2| > 4) = 1 - P(-18.348102 \leq t_{(62)} \leq 2.621157)$$

$$P(|\bar{x}_1 - \bar{x}_2| > 4) = 1 - \left[P(t_{(62)} \leq 2.621157) - P(t_{(62)} < -18.348102) \right]$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Student's t distribution with 62 DF
         x      P( X <= x )
  -18.3481      0.000000
    2.6212      0.994496
```

$$P(|\bar{x}_1 - \bar{x}_2| > 4) = 1 - (0.994496 - 0) = 0.005504$$

9.3 Cociente de varianzas muestrales $\left(s_1^2 / s_2^2\right)$.

Si s_1^2 y s_1^2 son las varianzas muestrales de dos muestras aleatorias independientes de tamaños n_1 y n_2 seleccionadas con reemplazo de dos poblaciones normales $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$ respectivamente, entonces la variable aleatoria.

$$\left(\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}\right)=\left(\frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2}\right)$$ tiene una distribución F con $(n_1 - 1)$ y $(n_2 - 1)$ grados de libertad.

Ejemplo 18:

Con la finalidad de renovar sus equipos de cómputo, una empresa de servicios informáticos solicitó asesoramiento técnico para la compra de placa base de Intek o DMA. Se tiene información de que la duración de la placa Intel tiene una desviación estándar de 12 meses (σ_1) y que la placa DMA tiene una duración con desviación estándar de 15 meses (σ_2); se seleccionan muestras aleatorias de 32 placas Intel y 26 placas DMA. Si se desea comparar con una probabilidad de 0.96, que la razón de varianzas $\left(\frac{s_1^2}{s_2^2}\right)$ no excede el valor "k". ¿Cuál es el valor de k?

Solución:

$$P\left(\frac{s_1^2}{s_2^2}\le k\right)=0.96$$

$$P\left(\frac{s_1^2}{s_2^2}\le k\right)=P\left(\frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2}\le\frac{15^2}{12^2}k\right)=0.96$$

$$P\left(\frac{s_1^2}{s_2^2}\le k\right)=P\left(F_{(32-1,26-1)}\le\frac{15^2}{12^2}k\right)=0.96\ \Rightarrow\ P\left(F_{(31,25)}\le 1.5625k\right)=0.96$$

Haciendo uso del Minitab

Inverse Cumulative Distribution Function

```
F distribution with 31 DF in numerator and 25 DF in
denominator
P( X <= x )          x
 0.96          1.99658
```

Entonces: $1.5625k = 1.99658\ \Rightarrow\ k=\frac{1.99658}{1.5625}=1.277811\approx 1.28$

9.4 Diferencias de proporciones muestrales ($p_1 - p_2$)

Sean $x_1, x_2,\ldots, x_n$ e $y_1, y_2,\ldots, y_n$ dos muestras aleatorias, con reemplazo, independientes de tamaño n_1 y n_2 seleccionadas de dos poblaciones independientes de Bernoulli $B(1,\pi_1)$ y $B(1,\pi_2)$, donde π_1 y π_2 son las proporciones poblacionales de éxito respectivamente. Si:

$p_1 = \dfrac{\sum_{i=1}^{n_1} X_i}{n_1} = \dfrac{W}{n_1}$, con: $W \sim B(n_1, \pi_1)$ y $p_2 = \dfrac{\sum_{i=1}^{n_2} Y_i}{n_2} = \dfrac{V}{n_2}$, con: $V \sim B(n_2, \pi_2)$

son las proporciones muestrales respectivas; entonces la variable aleatoria ($p_1 - p_2$) tiene distribución: $p_1 - p_2 \underset{\substack{n_1 \to \infty \\ n_2 \to \infty}}{\to} N\left(\pi_1 - \pi_2, \dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}\right)$

Propiedades:

a. $\mu_{(p_1 - p_2)} = E(p_1 - p_2) = E(p_1) - E(p_2) = \pi_1 - \pi_2$

b. $\sigma^2_{(p_1 - p_2)} = V(p_1 - p_2) = V(p_1) + V(p_2) = \dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}$

c. Para n_1 y n_2 suficientemente grandes, la distribución de la variable aleatoria es:

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1)$$

Ejemplo 19:

Según las encuestas de opinión pública, el 65% de los residentes de la ciudad A están en contra de las privatizaciones de las empresas públicas; de igual manera, el 80% de los residentes de la ciudad B están en contra de las privatizaciones. Suponga que usted selecciona al azar 100 y 200 residentes de A y B, respectivamente.

a. ¿Cuál es la probabilidad de que la proporción muestral de los que están en contra en la ciudad A difiera de la proporción muestral de los que están en contra en la ciudad B en menos de 0.01?
b. ¿Cuál es la probabilidad de que la proporción muestral de los residentes de la ciudad A a favor de las privatizaciones de las empresas públicas sea menor que la proporción muestral de los residentes de la ciudad B en favor de las privatizaciones de las empresas públicas?

Solución:

a. Sea π_1: Proporción de residentes de la ciudad A, en contra de la privatización.

$\pi_1 = 0.65 \qquad n_1 = 100$

π_2 : Proporción de residentes de la ciudad B, en contra de la privatización.

$\pi_2 = 0.8 \qquad n_2 = 200$

Se desea obtener: $P(|p_2 - p_1| \leq 0.01)$

Obteniendo la distribución muestral de la diferencia de proporciones:

$$(p_2 - p_1) \to N\left(\pi_2 - \pi_1, \frac{\pi_2(1-\pi_2)}{n_2} + \frac{\pi_1(1-\pi_1)}{n_1}\right)$$

$$(p_2 - p_1) \to N\left(0.8 - 0.65, \frac{0.8(0.2)}{200} + \frac{0.65(0.35)}{100}\right)$$

$$(p_2 - p_1) \to N\left(0.15, 0.055453^2\right),$$

con: $\sigma_{(p_2-p_1)} = \sqrt{\frac{0.8(0.2)}{200} + \frac{0.65(0.35)}{100}} = 0.055453$

Luego:

$$P(|p_2 - p_1| \le 0.01) = P(-0.01 \le p_2 - p_1 \le 0.01) =$$

$$P(p_2 - p_1 \le 0.01) - P(p_2 - p_1 < -0.01)$$

$$P(|p_2 - p_1| \le 0.01) = P(p_2 - p_1 \le 0.01) - P(p_2 - p_1 < -0.01)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 0.15 and standard deviation = 0.055453

    x      P( X <= x )
-0.01     0.0019550
 0.01     0.0057905
```

$$P(|p_2 - p_1| \le 0.01) \cong 0.005791 - 0.001955 = 0.003836$$

b. Sea π_1: Proporción de residentes de la ciudad A a favor de la privatización.

$\pi_1 = 0.35$ $\quad$ $n_1 = 100$

π_2 Proporción de residentes de la ciudad B a favor de la privatización.

$\pi_2 = 0.2$ $\quad$ $n_2 = 200$

Se desea obtener: $P(p_1 < p_2)$

Hallando la distribución de la diferencia de proporciones:

$$(p_1 - p_2) \to N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

$$(p_1 - p_2) \to N\left(0.35 - 0.20, \frac{0.35(0.65)}{100} + \frac{0.2(0.8)}{200}\right)$$

$$(p_1 - p_2) \to N(0.15, 0.055453^2)\ ,$$

con: $\sigma_{(p_2-p_1)}=\sqrt{\frac{0.35(0.65)}{100}+\frac{0.2(0.8)}{200}}=0.055453$

Luego: $P(p_1<p_2)=P(p_1-p_2<0)$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 0.15 and standard deviation = 0.055453
x     P( X <= x )
0     0.0034153
```

$P(p_1<p_2)\cong 0.0034153.$

PROBLEMAS RESUELTOS

1. Cicsa es una empresa consultora en calidad, la cual realiza una investigación en las empresas que tienen implementados sistemas de calidad en el ámbito nacional. El objetivo principal de la investigación es estimar los costos de calidad, para lo cual seleccionará una muestra aleatoria de 30 empresas que cumplan con el siguiente perfil:

- Empresa de producción (de bienes y servicios)
- Antigüedad no menor de 10 años
- Sistema de calidad implementado

El experto selecciona una empresa del directorio económico y se entrevista con el gerente de operaciones; si cumple con el perfil, la empresa es parte de la muestra; en caso contrario, no interviene en el estudio. Este procedimiento continúa hasta completar la muestra.

Responda las siguientes preguntas justificando su respuesta:

a. Indique el tipo de muestreo utilizado.

Solución:

El muestreo utilizado no es de tipo probabilístico sino de conveniencia.

b. Determine la población objetivo.

Solución:

La población objetivo está definida como todas las empresas del Perú que tienen implementados sistemas de calidad.

2. Determinar la unidad de muestreo, la población objetivo, el marco muestral y el tipo de muestreo más apropiado en los siguientes casos:

a. Se desea estimar el número promedio de componentes electrónicos defectuosos, por tablero, fabricados para la instalación en computadoras.

Solución:

Población: Todos los tableros con componentes electrónicos fabricados para la instalación en computadoras.

Unidad de muestreo: Un tablero con componente electrónico fabricado para la instalación en computadoras.

Marco muestral: Inventario de todos los tableros con componentes electrónicos fabricados para la instalación en computadoras.

Tipo de muestreo: Se puede utilizar el muestreo aleatorio simple o el muestreo sistemático.

b. Se desea saber cuál de las profesiones tiene mayor demanda de los alumnos de quinto de secundaria que van a postular a la Universidad.

Solución:

Población: La profesión de preferencia de todos los alumnos del quinto de secundaria que van a postular a la Universidad.

Unidad de muestreo: Un alumno del quinto de secundaria que va a postular a la Universidad.

Marco muestral: Listado de todos los alumnos del quinto de secundaria proporcionado por las unidades de servicios educativos (USE).

Tipo de muestreo: Se puede utilizar el muestreo estratificado por nivel socioeconómico, ubicación geográfica u otro.

c. Se desea estimar el consumo promedio de agua en Lima metropolitana.

Solución:

Población: El consumo de agua de todas las viviendas de Lima con servicio de agua potable.

Unidad de muestreo: Una vivienda con servicio de agua potable en Lima.

Marco muestral: Listado de todas las viviendas con servicio de agua potable en Lima metropolitana.

Tipo de muestreo: Se puede utilizar un muestreo estratificado por distritos o por niveles socioeconómicos.

3. La distribución del número de hijos por familia de una zona de Lima Metropolitana Norte está dada en la siguiente tabla:

N.º de hijos	Porcentaje
0	0.10
1	0.20
2	0.30
3	0.25
4	0.15
Total	1.00

a. Proporcione en una tabla de doble entrada todas las posibles muestras de dos familias elegidas al azar con reemplazo, que pueden ser formadas con sus respectivas probabilidades de ocurrencia.

Solución:

Sea X: Número de hijos por familia

Nº de hijos	0	1	2	3	4	Total
0	0.01	0.02	0.03	0.025	0.015	0.1
1	0.02	0.04	0.06	0.05	0.03	0.2
2	0.03	0.06	0.09	0.075	0.045	0.3
3	0.025	0.05	0.075	0.0625	0.0375	0.25
4	0.015	0.03	0.045	0.0375	0.0225	0.15
Total	0.1	0.2	0.3	0.25	0.15	1.00

b. Si se hubiera seleccionado una muestra aleatoria de tamaño 4, ¿cuál es la probabilidad de observar la cuaterna (2, 3, 3, 1)?

Solución:

Suponiendo independencia de eventos esto equivale a hallar:

$$P(2,3,3,1) = P(2)P(3)P(3)P(1) = (0.30)(0.25)(0.25)(0.20) = 0.00375$$

4. La Secretaría de Transporte Urbano afirma que la distribución del número de accidentes automovilísticos por hora está dada por:

Número de accidentes automovilísticos por hora	0	1	2	3
Probabilidad	0.35	0.35	0.20	0.10

Suponga que usted registra el número de accidentes ocurridos durante las últimas 25 horas.

a. ¿Cuál es la media de la distribución del número de accidentes?

Solución:

$$\mu = E(X) = \sum_{i=1}^{4} X_i P(x_i) = (0)(0.35) + (1)(0.35) + (2)(0.20) + (3)(0.10) = 1.05 \text{ accidentes.}$$

b. ¿Cuál es la varianza de la distribución del número de accidentes?

Solución:

$$\sigma^2 = V(X) = E(X^2) - [E(X)]^2$$

$$\sigma^2 = V(X) = \left[(0^2)(0.35) + (1^2)(0.35) + (2^2)(0.20) + (3^2)(0.10) \right] - (1.05)^2$$

$$\sigma^2 = V(X) = (2.05) - (1.1025) = 0.9475$$

c. Obtenga la distribución de la media muestral.

Solución:

$$E(\bar{x}) = \mu = 1.05$$

$$V(\bar{x}) = \frac{V(X)}{n} = \frac{\sigma^2}{n} = \frac{0.9475}{25} = 0.0379$$

La distribución de la media muestral es:

$$\bar{x} \to N\left(1.05, (0.194679)^2\right)$$

Con: $\sigma_{(\bar{x})} = \sqrt{\frac{0.9475}{25}} = \sqrt{0.0379} = 0.194679$

d. ¿Cuál es la probabilidad de que el promedio muestral de accidentes de las últimas 25 horas sea por lo menos igual a 1?

Solución:

Por el ítem anterior se tiene que:

$$\bar{x} \to N\left(1.05, 0.194679^2\right)$$

Haciendo uso del Minitab

```
Cumulative Distribution Function
Normal with mean = 1.05 and standard deviation = 0.194679
x     P( X <= x )
1        0.398654
```

Luego:

$$P(\bar{x} \geq 1) = 1 - P(\bar{x} \leq 1) \cong 1 - 0.398654 = 0.601346$$

Terorema Central del Límite-distribución de la media muestral $(\bar{x})$

5. Indusold es una compañía que se dedica a labores de soldadura industrial. El inspector de calidad inspecciona las obras de acuerdo con la siguiente unidad de análisis: UA = 5 metros lineales de soldadura. El número de puntos de soldadura defectuosos por cada unidad de análisis presenta una distribución de Poisson con una media de 4 puntos de soldadura defectuosos. Si se selecciona una muestra de 32 unidades de análisis para la realización de la inspección diaria:

a. ¿Cuál es la probabilidad de que el promedio de puntos de soldadura defectuosos encontrados por cada unidad de análisis sea por lo menos de 5 puntos de soldadura?

Solución:

Sea X: Número de puntos de soldadura.

$X \sim P(\lambda = 4)$

Aplicando Teorema Central del Límite: $\bar{x} \to N(4, 0.353553^2)$

donde: $\sigma_{(\bar{x})} = \sqrt{\frac{4}{32}} = \sqrt{0.125} = 0.353553$

Por consiguiente:

$$P(\bar{x} \geq 5) = 1 - P(\bar{x} < 5) \cong 1 - 0.997661 = 0.002339$$

b. ¿Cuántas unidades de análisis serán necesarias inspeccionar para que con una probabilidad de 0.85, el promedio de puntos de soldadura defectuosos encontrados en la muestra sea mayor al verdadero valor promedio en por lo menos 0.5 puntos de soldadura?

Solución:

$P(\bar{x} - \mu > 0.5) = 1 - P(\bar{x} - \mu \leq 0.5) = 0.85$

$P(\bar{x} - \mu \leq 0.5) = P(Z \leq 0.5(\sqrt{n}/2)) = 0.15$

Haciendo uso del Minitab

```
Inverse Cumulative Distribution Function
Normal with mean = 0 and standard deviation = 1
P( X <= x )              x
        0.15      -1.03643
```

$0.5(\sqrt{n}/4) = -1.03643 \quad \Rightarrow n = 68.748$

Se deben seleccionar 69 unidades de análisis aproximadamente.

6. Un operario del grifo Petrosol indica que el tiempo (en minutos) que demora la atención en el llenado del tanque de un auto, ya sea con gasolina o petróleo, sigue una distribución uniforme en el intervalo de [4,10]. Se tiene información de 42 autos que han llenado el tanque de combustible:

a. Obtenga la distribución de probabilidades del tiempo medio de minutos que demora la atención.

Solución:

Sea X: Tiempo que demora atención de llenado de tanque.

Se tiene: $X \sim U(4,10)$ $\quad n = 42.$

Entonces:

$\bar{x} \to N(7, 0.267261^2)$, donde: $\sigma_{(\bar{x})} = \sqrt{\frac{(10-4)^2}{12(42)}} = \sqrt{0.071429} = 0.267261$

b. Obtenga la probabilidad de que el tiempo promedio muestral de demora en el llenado del tanque de combustible sea mayor a 7.2 minutos.

Solución:

La probabilidad es: $P(\bar{x} > 7.2) = 1 - P(\bar{x} \leq 7.2)$

Haciendo uso del Minitab

```
Cumulative Distribution Function
Normal with mean = 7 and standard deviation = 0.267261
 x       P( X <= x )
7.2      0.772870
```

Obteniendo:

$$P(\bar{x} > 7.2) \cong 1 - 0.772870 = 0.22713$$

7. El contenido de nicotina de un cigarrillo de una marca en particular es una variable aleatoria con media 0.8 mg y desviación estándar 0.1 mg. Si una persona fuma 5 cajetillas de estos cigarrillos por semana, ¿cuál es la probabilidad de que la cantidad total de nicotina consumida en una semana sea por lo menos de 82 mg?

Solución:

Sea X: Contenido de nicotina de un cigarrillo (mg).

Se sabe que $X \sim (0.8, 0.1^2)$

Suponiendo que son cajetillas de 20 cigarrillos, entonces 5 cajetillas contienen 100 cigarrillos.

El contenido total de nicotina de 100 cigarrillos es $\sum_{i=1}^{100} x_i$

$$P\left(\sum_{i=1}^{100} x_i \geq 82\right) = P\left(\frac{\sum_{i=1}^{100} x_i}{n} \geq \frac{82}{100}\right) = P(\bar{x} \geq 0.82)$$

siendo: $\bar{x} \to N(0.8, 0.01^2)$ donde: $\sigma_{(\bar{x})} = \sqrt{\frac{0.01}{100}} = 0.01$

$$P(\bar{x} \geq 0.82) = 1 - P(\bar{x} < 0.82)$$

Haciendo uso del Minitab

```
Cumulative Distribution Function
Normal with mean = 0.8 and standard deviation = 0.01
   x   P( X <= x )
0.82      0.977250
```

$$P(\bar{x} \geq 0.82) \cong 1 - 0.977250 = 0.02275$$

Es decir, el 2,27% de las semanas la persona consume por lo menos 82 mg de nicotina o, dicho de otra forma, en una semana elegida al azar la probabilidad de que la persona consuma por lo menos 82 mg de nicotina es 0.02275.

8. El administrador de una tienda de menudeo afirma que la cajera de la tienda puede atender, sin ningún inconveniente, a 100 clientes en menos de 2 horas. Para comprobar tal afirmación se registraron los tiempos de espera de los 100 clientes que pasaron por la caja registradora donde se encuentra la cajera y se obtuvo una media de 1.5 minutos con desviación estándar de 1 minuto. ¿Cuál es la probabilidad de que la cajera pueda atender a 100 clientes en menos de 2 horas?

Solución:

Sea X: Tiempo de atención en minutos de clientes que pasan por la caja registradora.

$$X \sim (1.5, 1^2)$$

Entonces:

$\bar{x} \to N(1.5, 0.1^2)$, donde: $\sigma_{(\bar{x})} = \sqrt{\frac{1}{100}} = 0.1$

Tiempo de atención para 100 clientes: $\sum_{i=1}^{100} x_i$

$$P\left(\sum_{i=1}^{100} x_i \le 120\right) = P\left(\frac{\sum_{i=1}^{100} x_i}{n} \le \frac{120}{100}\right) = P(\bar{x} \le 1.2)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 1.5 and standard deviation = 0.1
  x   P( X <= x )
1.2     0.0013499
```

$P(\bar{x} \le 1.2) \cong 0.0013499$

9. Luego de estudiar los manuales de manejo y mantenimiento de las máquinas de la empresa Wilco S.A., el ingeniero responsable determinó que:

- El tiempo en minutos que la máquina A demora en realizar una operación de corte tiene distribución Normal con media μ_1 y desviación estándar σ_1.

- El tiempo en minutos que la máquina B demora en realizar una operación de soldadura tiene distribución Normal con media μ_2 y desviación estándar σ_2.
- El número de fallas que tiene la máquina C en el proceso de sellado en un día de trabajo tiene distribución:

X	0	1	2	3
$P(X)$	0.8	0.1	0.05	0.05

Máquina A

a. Si $\mu_1 = 1.2$ y $\sigma_1 = 0.1$, ¿cuál es la probabilidad de que el tiempo de corte de una operación elegida al azar sea mayor de 1.4 minutos?

Solución:

Sea X_1: Tiempo que la máquina A demora en realizar una operación de corte.

Como:

$\mu_1 = 1.2$ y $\sigma_1 = 0.1$ entonces $X_1 \sim N(1.2, 0.1^2)$

$P(X_1 > 1.4) = 1 - P(X_1 \leq 1.4)$

Haciendo uso del Minitab

```
Cumulative Distribution Function
Normal with mean = 1.2 and standard deviation = 0.1
  x      P( X <= x )
1.4      0.977250
```

$P(X_1 > 1.4) = 1 - 0.977250 = 0.02275$

b. Si $\mu_1 = 2$ y $\sigma_1 = 0.2$, ¿cuál es la probabilidad de que el promedio de una muestra de 30 observaciones sea mayor que 2.05?

Solución:

Si $\mu_1 = 2$ y $\sigma_1 = 0.2$, entonces $X_1 \sim N(2, 0{,}2^2)$

Entonces: $\bar{x}_1 \rightarrow N(2, 0.0365^2)$ donde $\sigma_{(\bar{x}_1)} = \sigma_1 / \sqrt{30} = 0.0365$

$P(\bar{x}_1 > 2.05) = 1 - P(\bar{x}_1 \leq 2.05)$

Haciendo uso del Minitab

```
Cumulative Distribution Function
Normal with mean = 2 and standard deviation = 0.0365
   x       P( X <= x )
2.05       0.914635
```

$P(\bar{x}_1 \geq 2.05) = 1 - 0.914635 = 0.085365$

c. Si $\sigma_1 = 0.1$, ¿cuál debe ser el tiempo promedio poblacional tal que la probabilidad de que el promedio muestral de una muestra de 25 observaciones se encuentre entre 2.0208 y 2.0992 con probabilidad igual a 0.95? Asuma puntos simétricos.

Solución:

Como: $\sigma_1 = 0.1 \quad \Rightarrow \; X_1 \sim N(\mu, 0.1^2)$

Entonces: $P\left(2.0208 < \overline{x} < 2.0992\right) = 0.95$

Estandarizando:

$$P\left(\frac{2.0208 - \mu}{0.1/5} < Z < \frac{2.0992 - \mu}{0.1/5}\right) = 0.95$$

$$P\left(Z < \frac{2.0992 - \mu}{0.02}\right) - P\left(Z \leq \frac{2.0208 - \mu}{0.02}\right) = 0.95$$

Asumiendo puntos simétricos:

$$\frac{2.0992 - \mu}{0.02} = -\left(\frac{2.0208 - \mu}{0.02}\right)$$

Entonces: $\frac{2.0992 - \mu}{0.02} = 1.95996$ y $\frac{2.0208 - \mu}{0.02} = -1.95996$

Ambos resultados conducen a que $\mu = 2.06$ minutos.

Máquina B

a. Si $\mu_2 = 20$ y $\sigma_2 = 4$, ¿cuál es la probabilidad de que el tiempo total de soldadura de 10 productos sea mayor que 205 minutos?

Solución:

Sea X_2 : Tiempo que la máquina B demora en realizar una operación de soldadura

Como $\mu_2 = 20$ y $\sigma_2 = 4 \;\Rightarrow\; X_2 \sim N(20, 4^2)$

La distribución de $\overline{x}_2$ es:

$\overline{x}_2 \rightarrow N(20, 1.26^2)$

$\sum_{i=1}^{10} x_{2_i}$: Tiempo total que la máquina B demora en realizar 10 operaciones de soldadura.

$$P\left(\sum_{i=1}^{10} x_{2_i} > 205\right) = P\left(\frac{\sum_{i=1}^{10} x_{2_i}}{n} > \frac{205}{10}\right) = P\left(\overline{x}_2 > 20.5\right)$$

$$P(\bar{x}_2 > 20.5) = 1 - P(\bar{x}_2 \leq 20.5)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 20 and standard deviation = 1.26
   x   P( X <= x )
20.5      0.654252
```

$$P(\bar{x}_2 > 20.5) = 1 - 0.654252 = 0.345748$$

b. Si $\sigma_2 = 10$, ¿cuál es la probabilidad de que la varianza de una muestra de 16 observaciones sea mayor que 50?

Solución:

Si $\sigma_2 = 10$ entonces $X_2 \sim N(\mu, 10^2)$

$$P(s_2^2 > 50) = P\left(\frac{(n-1)s_2^2}{\sigma_2^2} > \frac{(15)50}{100}\right) = P(\chi^2_{(15)} > 7.5)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Chi-Square with 15 DF
  x    P( X <= x )
7.5     0.0577369
```

$$P(s_2^2 > 50) = 1 - 0.0577369 = 0.9422631$$

c. Si $\mu_2 = 20$ y la varianza de una muestra de 40 observaciones es igual a 50, ¿cuál es la probabilidad de que el promedio muestral sea mayor que 22?

Solución:

Como: $\mu_2 = 20 \Rightarrow X_2 \sim N(20, \sigma^2)$

Por otro lado: $s_2^2 = 50$; $n = 40$

$$P(\bar{x}_2 > 22) = P\left(\frac{\bar{x}_2 - \mu_2}{s_2/\sqrt{n}} > \frac{22-20}{\sqrt{50}/\sqrt{40}}\right) = P(t_{(39)} > 1.788854)$$

$$P(\bar{x}_2 > 22) = 1 - P(t_{(39)} \leq 1.788854) = 1 - 0.959296 = 0.040704$$

Máquina C

a. Si se seleccionan al azar una muestra de 100 días y se registra el número de fallas de la máquina C de cada día, ¿cuál es la probabilidad de que el promedio de la muestra sea por lo menos 0.4?

Solución:

Sea X_3: Número de fallas de la máquina C

$$E(X_3)=\sum x_{3_i}p(x_{3_i})=0(0.8)+1(0.1)+2(0.05)+3(0.05)=0.35$$

$$V(X_3)=E(X_3^2)-\left[E(X_3)\right]^2=\sum x_{3_i}^2 p(x_{3_i})-\left[E(X_3)\right]^2$$

$$V(X_3)=\left[0^2(0.8)+1^2(0.1)+2^2(0.05)+3^2(0.05)\right]-(0.35)^2=0.75-0.1225=0.6275$$

$\bar{x}_3 \to N(0.35, 0.0792149^2)$ Con: $\sigma_{(\bar{x}_3)}=\dfrac{\sigma_{X_3}}{\sqrt{n}}=\dfrac{0.792149}{\sqrt{100}}=0.079215$

$$P(\bar{x}\geq 0.4)=1-P(\bar{x}<0.4)\cong 1-0.736056=0.263944$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 0.35 and standard deviation = 0.07921
 x        P( X <= x )
0.4       0.736056
```

$$P(\bar{x}\geq 0.4)=1-P(\bar{x}<0.4)\cong 1-0.736056=0.263944$$

b. Si se selecciona al azar una muestra de 50 días, ¿cuál es la probabilidad de que la máquina C falle en menos de 15 días?

Solución:

Sea Y_3: Número de días que la Máquina C falla en 50 días de trabajo.

P (falle un día cualquiera) = 0.2
P (no falle un día cualquiera) = 0.8

Entonces:

$$P(Y_3<15)=P\left(\frac{Y_3}{50}<\frac{15}{50}\right)=P(p<0.3)$$

La distribución de p es:

$$p\to N\left(0.2,\frac{0.2(0.8)}{50}\right) \quad \Rightarrow p\to N(0.2, 0.056569^2)$$

Luego:

$$P(p<0.3)\cong 0.961449$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 0.2 and standard deviation = 0.056569
  x   P( X <= x )
0.3       0.961449
```

10. El supervisor de una planta de servicio automotriz ha determinado que el tiempo que demora un automóvil en ser atendido se distribuye normalmente con una media de 60 horas y desviación estándar σ horas.

a. Si se supone que el 65% de los tiempos de atención son mayores que 56 horas, ¿cuál es la probabilidad de que el promedio de una muestra de 64 automóviles demore más de 58 horas?

Solución:

Sea: $X \sim N(60,\sigma^2)$

$$P(x > 56) = 0.65$$

$$P\left(Z > \frac{56-60}{\sigma}\right) = 0.65 \quad \Rightarrow \quad P\left(Z > \frac{56-60}{\sigma}\right) = 0.35$$

$$\frac{56-60}{\sigma} = -0.385320 \quad \Rightarrow \quad \sigma = 10.380982$$

Entonces: $X \sim N(60,10.380982^2)$

Luego: $\bar{x} \to N(60,1.297623^2)$ Con: $\sigma_{(\bar{x})} = \frac{\sigma_{\bar{x}}}{\sqrt{n}} = \frac{10.380982}{\sqrt{64}} = 1.297623$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 60 and standard deviation = 1.29762
 x   P( X <= x )
58     0.0616244
```

$$P(\bar{x} > 58) = 1 - P(x \le 58) = 1 - 0.061624 = 0.938376$$

b. Suponga que 15% de los tiempos de atención son mayores que 65 horas y que se extrae una muestra aleatoria de 9 tiempos de demora de igual número de automóviles, ¿cuál es la probabilidad de que el promedio muestral difiera de la media poblacional en menos de 1.25 horas?

Solución:

Obteniendo el valor de σ tal que:

$$P(x > 65) = 0.15$$

$$P\left(Z > \frac{65-60}{\sigma}\right) = 0.15 \quad \Rightarrow P\left(Z \le \frac{65-60}{\sigma}\right) = 0.85$$

$$\frac{65-60}{\sigma} = 1.03643 \qquad \sigma = 4.824252$$

Entonces la distribución de la media es:

$$\bar{x} \to N(60,1.608067^2) \text{ donde: } \sigma_{(\bar{x})} = \sqrt{\frac{4.8242^2}{9}} = 1.608067$$

Luego:

$$P\left(\left|\bar{x}-\mu\right|<1.25\right)=P\left(-1.25<\bar{x}-\mu<1.25\right)=P\left(58.75<\bar{x}<61.25\right)$$

$$P\left(\left|\bar{x}-\mu\right|<1.25\right)=P\left(\bar{x}<61.25\right)-P\left(\bar{x}\leq 58.75\right)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 60 and standard deviation = 1.60807
     x   P( X <= x )
58.75      0.218482
61.25      0.781518
```

$$P\left(\left|\bar{x}-\mu\right|<1.25\right)=0.781518-0.218482=0.563036$$

c. Suponga que $\sigma = 3$ horas. ¿Cuántos tiempos deben seleccionarse para tener un error de estimación del promedio menor que 0.5 horas con probabilidad 0.935?

Solución:

$$P\left(\left|\bar{x}-\mu\right|<0.5\right)=0.935$$

$$P\left(-0.5<\bar{x}-\mu<0.5\right)=P\left(-\frac{0.5}{\frac{3}{\sqrt{n}}}<\frac{\bar{x}-\mu}{\frac{3}{\sqrt{n}}}<\frac{0.5}{\frac{3}{\sqrt{n}}}\right)=0.935$$

$$\frac{0.5}{\frac{3}{\sqrt{n}}}=1.84526 \qquad n=122.579441\approx 123 \text{ tiempos.}$$

11. El tiempo total necesario para procesar una solicitud de préstamo hipotecario en un banco local sigue una distribución Normal con un promedio de 7 días y desviación estándar 3 días.

a. ¿Cuál es la probabilidad de que el tiempo promedio de procesamiento de una muestra de 20 solicitudes, elegidas al azar, sea superior a 9 días?

Solución:

Como $X \sim N(7,3^2)$

La distribución del tiempo promedio $\bar{x} \to N\left(7,(3/\sqrt{20})^2\right)$, entonces: $\bar{x} \to N(7,0.67^2)$

$$P\left(\bar{x}>9\right)=1-P\left(\bar{x}\leq 9\right)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 7 and standard deviation = 0.67
x   P( X <= x )
9      0.998582
```

$$P(\bar{x} > 9) = 1 - P(\bar{x} \leq 9) = 1 - 0.998582 = 0.001418$$

b. ¿Cuántas solicitudes de préstamo se deben seleccionar para encontrar un tiempo promedio de procesamiento inferior a 8 días, con un 97,5% de probabilidad?

Solución:

Se tiene:

$$P(\bar{x} \leq 8) = 0.975 \Rightarrow P\left(\frac{(\bar{x} - \mu_{\bar{x}})}{\sigma_{\bar{x}}} \leq \frac{(8-7)}{3/\sqrt{n}}\right) = 0.975 \Rightarrow P\left(Z \leq \sqrt{n}/3\right) = 0.975$$

Haciendo uso del Minitab

```
Inverse Cumulative Distribution Function
Normal with mean = 0 and standard deviation = 1
P( X <= x )          x
       0.975   1.95996
```

Entonces:

$$\sqrt{n}/3 = 1.95996 \Rightarrow \sqrt{n} = 5.87988 \Rightarrow n = 34.57298881 \approx 35$$

Se deben seleccionar 35 solicitudes.

12. En un estudio realizado acerca de la problemática del transporte urbano, se determinó que el ingreso de los choferes de servicio público tiene distribución Normal con media $\mu = 1200$ nuevos soles con desviación estándar $\sigma = 200$ nuevos soles.

a. Si se selecciona una muestra aleatoria de 20 choferes, ¿cuál es la probabilidad de que el promedio muestral difiera de la media poblacional en menos de 20 nuevos soles?

Solución:

Sea X: Ingreso de los choferes de servicio público, cuya distribución es

$$X \sim N(1200, 200^2)$$

Entonces la distribución de la media es:

$\bar{x} \to N(1200, 44.721359^2)$ donde:

$$P(|\bar{x} - \mu| < 20) = P(|\bar{x} - 1200| < 20) = P(-20 < \bar{x} - 1200 < 20) = P(1180 < \bar{x} < 1220)$$

$$P(|\bar{x} - \mu| < 20) = P(\bar{x} < 1220) - P(\bar{x} \leq 1180) = 0.672640 - 0.327360 = 0.34528$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 1200 and standard deviation =
44.7214
   x  P( X <= x )
1180      0.327360
1220      0.672640
```

b. Determine el número de choferes que deben ser seleccionados para asegurar que el error de estimación del promedio sea menor que 20 con probabilidad 0.98.

Solución:

Choferes seleccionados

$$P\left(|\bar{x}-\mu|<20\right)=0.98 \quad \Rightarrow P\left(-20<\bar{x}-\mu<20\right)=0.98$$

$$P\left(\frac{-20}{44.721359/\sqrt{n}}<\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}<\frac{20}{44.721359/\sqrt{n}}\right)=0.98$$

$$\Rightarrow P\left(-\frac{20\sqrt{n}}{44.721359}<Z<\frac{20\sqrt{n}}{44.721359}\right)=0.98$$

Como la distribución Normal es simétrica:

$$P\left(Z<\frac{20\sqrt{n}}{44.721359}\right)=0.99; \quad P\left(Z\leq-\frac{20\sqrt{n}}{44.721359}\right)=0.01$$

Entonces:

$$\frac{20\sqrt{n}}{44.721359}=2.32635; \quad -\frac{20\sqrt{n}}{44.721359}=-2.32635$$

Luego:

$$n=\left\lceil\frac{2.32635\left(44.721359\right)}{20}\right\rceil^2=\left(5.201877\right)^2=27.059521\approx 28 \text{ choferes.}$$

13. En la librería "El Buen Saber" se venden libros, útiles escolares y accesorios de oficina en general. En la librería se hace entrega de pedidos de libros a domicilio. El tiempo de demora en la entrega de los libros tiene distribución Normal con promedio 25 minutos y desviación estándar 7 minutos.

a. ¿Cuántos pedidos de libros se deben atender, para tener un tiempo promedio de entrega de los libros inferior o igual a 30 minutos, con una probabilidad de 0.98?

Solución:

Sea X: Tiempo empleado en la entrega de libros a domicilio.

$X \sim N(25, 7^2)$

Entonces:

$$P(\bar{x} \leq 30) = 0.98 \quad \Rightarrow P\left(Z \leq \frac{30-25}{7/\sqrt{n}}\right) = 0.98$$

$$\Rightarrow P\left(Z \leq \sqrt{n}(0.714285)\right) = 0.98$$

Haciendo uso del Minitab

Inverse Cumulative Distribution Function

```
Normal with mean = 0 and standard deviation = 1
P( X <= x )          x
   0.98          2.05375
```

$$0.714285(\sqrt{n}) = 2.05375$$

$$\sqrt{n} = 2.87525$$

Luego $n = 8.267063 \approx 8$ pedidos de libros.

b. ¿Cuál es la probabilidad de que el tiempo promedio de entrega de una muestra de 15 pedidos a domicilio, elegidos al azar, sea superior a 25.5 minutos?

Solución:

La distribución del tiempo promedio es:

$\bar{x} \to N(25, 1.807392^2)$ donde: $\sigma_{(\bar{x})} = \sqrt{\frac{7^2}{15}} = 1.807392$

Luego: $P(\bar{x} > 25.5) = 1 - P(\bar{x} \leq 25.5)$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 25 and standard deviation = 1.80739
   x   P( X <= x )
25.5      0.608972
```

$$P(\bar{x} > 25.5) = 1 - 0.608972 = 0.391028$$

Es decir, la probabilidad de que el tiempo promedio de entrega de los pedidos supere a 25.5 minutos es de 0.391028 cuando se atienden 15 pedidos a domicilio.

14. (En relación con el problema anterior). El gerente de ventas de la librería "El Buen Saber" desea mejorar las ventas de libros universitarios, para lo cual cuenta con información de estudios anteriores de que el valor de las ventas diarias de este tipo de libros tiene distribución aproximadamente Normal con varianza 12.93. Si se sabe que en el estudio se tomó una muestra de 25 días, ¿cuál es la probabilidad de que el valor promedio de ventas diarias en la muestra difiera de la media poblacional en menos de 1.5?

Sugerencia: Use el proceso de estandarización.

Solución:

Sea X : venta de libros universitarios en 25 días.

Se sabe que: $X \sim N(\mu, 3.595831^2)$ donde: $\sigma_{(X)} = \sqrt{12.93} = 3.595831$

Entonces: $\bar{x} \to N(\mu, 0.719166^2)$ donde: $\sigma_{(\bar{x})} = \sqrt{\frac{12.93}{25}} = 0.719166$

Hallando: $P(|\bar{x} - \mu| < 1.5)$

$$P(|\bar{x} - \mu| < 1.5) = P(-1.5 < \bar{x} - \mu < 1.5) = P\left(\frac{-1.5}{\frac{3.595831}{\sqrt{25}}} < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{1.5}{\frac{3.595831}{\sqrt{25}}}\right)$$

$$P(|\bar{x} - \mu| < 1.5) = P\left(\frac{-1.5}{0.7191} < Z < \frac{1.5}{0.7191}\right) = P(-2.085748 < Z < 2.085748)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 0 and standard deviation = 1
   x          P( X <= x )
-2.08575      0.018501
 2.08575      0.981499
```

$$P(|\bar{x} - \mu| < 1.5) = 0.981499 - 0.018501 = 0.962998$$

Es decir, la probabilidad de que el valor promedio de ventas en la muestra difiera de la media poblacional en menos de 1.5 es de 0.962998.
Nota: Tener presente que no es necesario conocer μ ya que la expresión de la que se desea calcular la probabilidad es $|\bar{x} - \mu| < 1.5$.

15. El contador de una empresa informa a la Gerencia que el promedio de ventas diarias de libros escolares es S/.6.500. ¿Cuál es la probabilidad de que en 25 días el promedio muestral de las ventas sea mayor que S/. 6.480 si se sabe que la varianza muestral es de 12.930? Suponga que las ventas diarias tienen distribución Normal.

Solución:

Se sabe que:

$\mu = 6500 \qquad s = \sqrt{12930} = 113.710158 \qquad n = 25$

Como no se conoce la desviación estándar poblacional se aplica la distribución *t*.

$$P(\bar{x} > 6480) = 1 - P(\bar{x} \leq 6480) = 1 - P\left(\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \leq \frac{6480 - 6500}{\frac{113.710158}{\sqrt{25}}} \right)$$

$$P(\bar{x} > 6480) = 1 - P\left(t_{(24)} \leq \frac{-20}{22.742032} \right) = 1 - P(t_{(24)} \leq -0.879429)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Student's t distribution with 24 DF
         x   P( X <= x )
-0.879429      0.193945
```

$P(\bar{x} > 6480) = (1 - 0.193945) = 0.806055$

Es decir, la probabilidad de que en 25 días el promedio muestral sea mayor que S/.6.480 es 0.806055.

16. En un estudio de la empresa consultora Consultat S.A. respecto de las tarjetas de crédito realizado en el último verano se analizaron las siguientes variables:

C1 = Tipo de tarjeta (Bizza, Vaster, Credix, Triplex, RMC, etcétera).
C2 = Fecha de corte de pago (10 de cada mes o 25 de cada mes).
C3 = Monto del pago mensual (en nuevos soles).
C4 = Banco.
C5 = Tiempo de uso de la tarjeta (en años).
C6 = Número de veces que perdió la tarjeta.

Si uno de los ejecutivos de Consultat S.A. sostiene que la distribución del monto del pago mensual es Normal, entonces:

a. Calcule μ y σ, tal que el 80% de las observaciones son mayores que 1.500 y que el tercer cuartil es 2.600.

Solución:

Se sabe que:

$P(X > 1500) = 0.8 \quad \Rightarrow \quad P(X \leq 1500) = 0.2$

$\frac{1500 - \mu}{\sigma} = -0.841621 \qquad \Rightarrow 1500 = \mu - (0.841621\ \sigma)$

Por otro lado, por definición del tercer cuartil:

$P(X < 2600) = 0.75 \Rightarrow \frac{2600 - \mu}{\sigma} = 0.67449 \Rightarrow 2600 = \mu - (0.67449\ \sigma)$

Despejando los valores de μ y σ :

$\mu = 2110.63016$ y $\sigma = 725.540544$

b. Suponga que $\mu = 1980$ y $\sigma = 250$; calcular el tamaño de muestra necesario tal que la probabilidad de que el promedio muestral difiera de la media poblacional en menos de 10 nuevos soles sea igual a 0.98.

Solución:

$$P\left(|\bar{x} - \mu| < 10\right) = 0.98 \quad \Rightarrow \quad P\left(-10 < \bar{x} - \mu < 10\right) = 0.98$$

Estandarizando: $P\left(\frac{-10\sqrt{n}}{250} < Z < \frac{10\sqrt{n}}{250}\right) = 0.98$

Haciendo uso del Minitab

Inverse Cumulative Distribution Function

```
Normal with mean = 0 and standard deviation = 1
P( X <= x )          x
        0.99   2.32635
```

Entonces:

$$\frac{10\sqrt{n}}{250} = 2.32635 \quad \Rightarrow n = 3382.4402 \approx 3383 \text{ tarjetas.}$$

c. Si se extrae una muestra de 16 observaciones y se encuentra que la desviación estándar muestral es igual a 100, ¿cuál es la probabilidad de que la media muestral sea mayor que 2.050? Suponga que $\mu = 2.000$.

Solución:

$$P\left(\bar{x} > 2050\right) = P\left(\frac{\bar{x} - 2000}{\frac{100}{4}} > \frac{2050 - 2000}{\frac{100}{4}}\right) = P(t_{(15)} > 2) = 1 - P(t_{(15)} \leq 2)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Student's t distribution with 15 DF
x   P( X <= x )
2       0.968027
```

$$P\left(\bar{x} > 2050\right) = 1 - 0.968027 = 0.031973$$

17. Suponga que los aportes de los fonavistas en el país tienen distribución Normal con media 1.250 nuevos soles y desviación estándar 200 nuevos soles.

a. Si se eligen al azar a diez fonavistas, ¿cuál es la probabilidad de que sus aportes sumen más de 12.000 nuevos soles?

Solución:

Sea X: Aporte de los fonavistas en nuevos soles. Con: $X \sim N(1250, 200^2)$
La distribución muestral de los aportes promedios de los fonavistas es:

$\bar{x} \to N(1250, 63.245553^2)$ donde: $\sigma_{\bar{x}} = \sqrt{\frac{200^2}{10}} = 63.245553$

Entonces:

$$P\left(\sum_{i=0}^{10} x_i > 12000\right) = P(\bar{x} > 1200) = 1 - P(\bar{x} \le 1200)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 1250 and standard deviation = 63.2456
  X       P( X <= x )
1200      0.214598
```

$$P\left(\sum_{i=0}^{10} x_i > 12000\right) = 1 - 0.214598 = 0.785402$$

b. Si se eligen al azar a 100 fonavistas, ¿cuál es la probabilidad de que el promedio muestral sea mayor que 1.200 nuevos soles pero menor que 1.300?

Solución:

$\bar{x} \to N(1250, 20^2)$ donde: $\sigma_{(\bar{x})} = \sqrt{\frac{200^2}{100}} = \sqrt{400} = 20$

Entonces:

$$P(1200 < \bar{x} < 1300) = P(\bar{x} < 1300) - P(\bar{x} \le 1200)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 1250 and standard deviation = 20
   x      P( X <= x )
1200      0.006210
1300      0.993790
```

$$P(1200 < \bar{x} < 1300) = 0.993790 - 0.006210 = 0.98758$$

Distribución de la proporción muestral (p)

18. Juan Pérez, vendedor de insumos químicos en lotes, afirma que solo uno de 20 lotes vendidos tiene al menos un defecto. Un comprador de dichos insumos ha registrado cuidadosamente 200 lotes seleccionados al azar y basado en ello ha elaborado la siguiente distribución:

N.º de defectuosos por lote	0	1	2	3	4 o más
N.º de lotes	180	13	4	2	1

Si lo afirmado por el vendedor se cumple, ¿de cuántos lotes debe estar compuesta una muestra para tener una diferencia de a lo más 4 puntos porcentuales entre la proporción de la muestra y su valor real con probabilidad igual a 0.98?

Solución:

Se desea obtener n tal que se verifica:

$P\left(\left|p-0.05\right|\leq 0.04\right)=0.98$,

Entonces: $P\left(\left|Z\right|\leq \dfrac{0.04}{\sqrt{\dfrac{0.05\left(1-0.05\right)}{n}}}\right)=0.98$.

Haciendo uso del Minitab

```
Inverse Cumulative Distribution Function
Normal with mean = 0 and standard deviation = 1
P( X <= x )          x
    0.99          2.32635
```

Por lo tanto:

$$\frac{0.04}{\sqrt{\dfrac{0.05(1-0.05)}{n}}}=2.32635 ,$$

Entonces $n=\left(\dfrac{2.32635}{0.04}\right)^2 0.05\left(1-0.05\right)=161$, por lo tanto, la muestra debe conformarse de 161 lotes.

19. Un estudio de la consultora MB sobre el nivel de captación y rendimiento de los alumnos de primaria de un distrito de Lima norte informó que un 66% de alumnos captan las clases y rinden en sus evaluaciones. Si se seleccionó una muestra aleatoria de 220 alumnos de primaria en dicho distrito:

a. Obtenga la distribución muestral de la proporción de alumnos de primaria del distrito que captan las clases y rinden en sus evaluaciones en forma adecuada.

Solución:

Sea π : Proporción poblacional de alumnos de primaria de un distrito de Lima norte que captan las clases y rinden en las evaluaciones.

$\pi = 0.66 \quad n = 220$

La proporción muestral tiene la siguiente distribución:

$$p \to N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Reemplazando valores:

$p \to N\left(0.66, 0.031937^2\right)$, donde: $\sigma_p = \sqrt{\frac{0.66(0.34)}{220}} = 0.031937$

b. ¿Cuál es la probabilidad de que la proporción muestral difiera en menos de 0.05 de la proporción poblacional?

Solución:

Se tiene:

$$P\left(|p-\pi| < 0.05\right) \cong P\left(\frac{|p-\pi|}{\sqrt{\frac{\pi(1-\pi)}{n}}} < \frac{0.05}{0.031937}\right) = P\left(|Z| < 1.565561\right)$$

$$P\left(|p-\pi| < 0.05\right) = P\left(-1.565561 < Z < 1.565561\right) =$$

$$P\left(Z < 1.565561\right) - P\left(Z \le -1.565561\right)$$

Haciendo uso del Minitab

```
Cumulative Distribution Function
Normal with mean = 0 and standard deviation = 1
x   P( X <= x )
-1.56556      0.058726
 1.56556      0.941274
```

Reemplazando:

$$P\left(|p-\pi| < 0.05\right) \cong 0.941274 - 0.058726 = 0.882548$$

Es decir:

La probabilidad de que la proporción muestral de los 220 alumnos difiera en menos de 0.05 de la proporción poblacional es de 0.882548.

Distribución de varianza $\frac{(n-1)s^2}{\sigma^2}$

20. El tiempo de atención de un cajero en la ventanilla de un banco es una variable aleatoria Normal con $\sigma = 1.5$ minutos. Este cajero es observado en la atención de 25 clientes seleccionados al azar. ¿Qué valor máximo tomará la desviación estándar de la muestra con probabilidad 0.975?

Solución:

Sea X: Tiempo de atención de cajero en ventanilla

$X \sim N(\mu, 1.5^2)$ Con: $\sigma_X^2 = 1.5^2 = 2.25$

Se desea obtener el valor de k tal que:

$P(s \leq K) = 0.975$

$$P(s^2 \leq K^2) = P\left[\frac{(n-1)s^2}{\sigma^2} \leq \frac{24K^2}{2.25} \right] = P\left(\chi^2_{(24)} \leq \frac{24K^2}{2.25} \right) = 0.975$$

Haciendo uso del Minitab

Inverse Cumulative Distribution Function

```
Chi-Square with 24 DF
P( X <= x )          x
      0.975    39.3641
```

$$\frac{24K^2}{2.25} = 39.3641 \Rightarrow K^2 = 3.690384276 \Rightarrow K = 1.921 \text{ minutos.}$$

21. El consumo diario en galones de petróleo diesel por unidad de transporte público tiene distribución Normal.

a. Si se selecciona una muestra aleatoria de 15 unidades de transporte público y se supone que $\sigma = 5$ galones, ¿cuál es la probabilidad de que la varianza muestral se encuentre entre 10 y 20?

Solución:

Sea X: Consumo diario de petróleo diesel por unidad de transporte público (galones).

Donde: $X \sim N(\mu, 5^2)$

Obteniendo:

$$P(10 < s^2 < 20) = P\left[\frac{(14910}{25} < \frac{(n-1)s^2}{\sigma^2} < \frac{(14)20}{25} \right] = P(5.6 < \chi^2_{(24)} < 11.2)$$

$$P(10 < s^2 < 20) = P(\chi^2_{(24)} < 11.2) - P(\chi^2_{(24)} \leq 5.6)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Chi-Square with 24 DF
   x      P( X <= x )
 5.6     0.0000374
11.2     0.0124873
```

$$P(10 < s^2 < 20) = 0.0124873 - 0.0000374 = 0.01244$$

b. En el siguiente reporte de Minitab:

Descriptive Statistics: Norte, Sur

Variable	*N*	StDev	SE Mean
Norte	10	2.470	0.781

¿Qué significa SE Mean y cómo se interpreta el valor 0.781?

Solución:

SE Mean indica el error estándar de la media muestral, que es calculada de la siguiente manera:

$$\text{S.E.Mean} = \frac{s}{\sqrt{n}} = \frac{2.470}{\sqrt{10}} = \frac{2.470}{3.1622} = 0.781101$$

SE Mean = 0.7811 es un estimador de $\sigma/\sqrt{n}$. Indica la dispersión de los valores de $\bar{x}$ respecto al promedio poblacional.

Distribución de la diferencia de medias $(\bar{x}_1 - \bar{x}_2)$

22. Las bombillas fabricadas por Lumiplus tienen un promedio de vida útil de 6.000 horas con una desviación estándar poblacional de 1.600 horas y las bombillas fabricadas por Fotomax tienen un promedio de vida útil de 8.000 horas con una desviación estándar poblacional de 3.600 horas. Si se seleccionan 200 bombillas fabricadas por Lumiplus y 160 bombillas fabricadas por Fotomax ¿cuál es la probabilidad de que el promedio muestral de vida útil de las bombillas fabricadas por Fotomax no difiera en más de 800 horas del promedio muestral de vida útil de las bombillas fabricadas por Lumiplus? Suponer normalidad en el tiempo de vida de las bombillas.

Solución:

Si:

X_1 : Vida útil bombillas Panasonic $\quad X_1 \sim N(8000, 3600^2) \quad n_1 = 160$

X_2 : Vida útil bombillas Phillips $\quad X_2 \sim N(6000, 1600^2) \quad n_2 = 200$

$$\mu_{(\bar{x}_1 - \bar{x}_2)} = \mu_{(X_1)} - \mu_{(X_2)} = 8000 - 6000 = 2000$$

$$\sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}} = \sqrt{\frac{3600^2}{160} + \frac{1600^2}{200}} = \sqrt{93800} = 306.2679$$

Entonces la distribución de la diferencia de medias sería:

$$\bar{x}_1 - \bar{x}_2 \rightarrow N(2000, 306.2679^2)$$

Luego:

$$P\left(|\bar{x}_1 - \bar{x}_2| \leq 800\right) = P\left(-800 \leq \bar{x}_1 - \bar{x}_2 \leq 800\right) = P\left(\bar{x}_1 - \bar{x}_2 \leq 800\right) - P\left(\bar{x}_1 - \bar{x}_2 < -800\right)$$

Haciendo uso del Minitab

```
Cumulative Distribution Function
Normal with mean = 2000 and standard deviation = 306.268
      x   P( X <= x )
   -800    0.0000000
    800    0.0000446
```

$$P\left(|\bar{x}_1 - \bar{x}_2| \leq 800\right) = 0.0000446 - 0 = 0.0000446$$

23. Los colchones producidos por la empresa Edén tienen una duración media de 80 meses y una desviación estándar de 5 meses, mientras que los colchones fabricados por la empresa Quantum tienen una duración media de 75 meses y una desviación estándar de 3 meses, ¿cuál es la probabilidad de que una muestra aleatoria de 36 colchones de la empresa Edén tenga una duración media de al menos tres meses más que la duración media de 49 colchones de la empresa Quantum? Suponga que las duraciones en ambos casos se distribuyen normalmente y son independientes.

Solución:

Sean:

X_1: Duración de colchones producidos por la empresa Edén (meses).

donde: $X_1 \sim N(80, 5^2)$

X_2: Duración de colchones producidos por la empresa Quantum (meses).

donde: $X_2 \sim N(75, 3^2)$

Como: $\mu_{(\bar{x}_1 - \bar{x}_2)} = 80 - 75 = 5$

Entonces la distribución de la diferencia de medias sería:

$$\bar{x}_1 - \bar{x}_2 \rightarrow N(5, 0.9371^2), \qquad \text{donde: } \sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{5^2}{36} + \frac{3^2}{49}} = 0.9371$$

Luego:

$$P\left[(\bar{x}_1 - \bar{x}_2) \geq 3\right] = 1 - P\left[(\bar{x}_1 - \bar{x}_2) < 3\right]$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 5 and standard deviation = 0.9371
x   P( X <= x )
3     0.0164114
```

$$P\left[(\bar{x}_1 - \bar{x}_2) \geq 3\right] = 1 - 0.0164 = 0.9836$$

24. Una empresa comercializa manzanas de los tipos 1 y 2. El peso de una manzana del tipo 1 tiene distribución Normal con media μ_1 y varianza σ_1^2, lo mismo ocurre para la manzana del tipo 2, es decir, el peso tiene distribución Normal con media μ_2 y varianza σ_2^2. Se seleccionan muestras de 30 manzanas del tipo 1 y 45 manzanas del tipo 2.

a. Los proveedores de ambos tipos de manzanas informan que sus productos tienen las siguientes especificaciones: el tipo 1 tiene $\mu_1 = 80$ gramos y $\sigma_1^2 = 16$ gramos^2, mientras que el tipo 2 tiene $\mu_2 = 85$ y $\sigma_2^2 = 20$ gramos^2, ¿cuál es la probabilidad de que los promedios muestrales de ambos tipos de manzanas difieran en por lo menos 5 gramos?

Solución:

Sean:

X_1: Peso de manzanas Tipo 1, donde $X_1 \sim N(80, 4^2)$

X_2: Peso de manzanas Tipo 2, donde $X_2 \sim N(85, 4.472^2)$

Como:

$\mu_{(\bar{x}_1 - \bar{x}_2)} = 80 - 85 = -5$

Entonces la distribución de la diferencia de medias sería:

$\bar{x}_1 - \bar{x}_2 \rightarrow N(5, 0.9371^2)$, donde: $\sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{16}{30} + \frac{20}{45}} = 0.9888$

Luego:

$$P(|\bar{x}_1 - \bar{x}_2| \geq 5) = 1 - P(|\bar{x}_1 - \bar{x}_2| < 5) = 1 - P(-5 < \bar{x}_1 - \bar{x}_2 < 5)$$

$$P(|\bar{x}_1 - \bar{x}_2| \geq 5) = 1 - [P(\bar{x}_1 - \bar{x}_2 < 5) - P(\bar{x}_1 - \bar{x}_2 \leq -5)]$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = -5 and standard deviation = 0.9888
 x   P( X <= x )
 5           1.0
-5           0.5
```

$$P(|\bar{x}_1 - \bar{x}_2| \geq 5) = 1 - (1 - 0.5) = 1 - 0.5 = 0.5$$

b. Suponga que usted acepta los valores de los promedios poblacionales de los pesos de las manzanas del índice (a) de ambos tipos de manzanas, pero no las varianzas poblacionales; por ello, decide calcular las varianzas muestrales. Los valores calculados fueron $s_1^2 = 36$ gramos^2 y $s_2^2 = 15$ gramos^2. ¿Cuál es la probabilidad de que los promedios muestrales difieran en menos de 5 gramos cuando:

b.1) Las varianzas son desconocidas y homogéneas.

b.2) Las varianzas son desconocidas y heterogéneas.

Solución:

b.1) Las varianzas son desconocidas y homogéneas.

Obteniendo la varianza muestral ponderada:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(29)(36) + (44)(15)}{73} = 23.3425 \quad \Rightarrow \quad s_p = 4.8314$$

Obteniendo:

$$P\left(|\bar{x}_1 - \bar{x}_2| < 5\right) = P\left(-5 < \bar{x}_1 - \bar{x}_2 < 5\right)$$

$$P\left(|\bar{x}_1 - \bar{x}_2| < 5\right) = P\left(\frac{-5+5}{4.8314\sqrt{\frac{1}{30}+\frac{1}{45}}} < t_{(73)} < \frac{5+5}{4.8314\sqrt{\frac{1}{30}+\frac{1}{45}}} \right)$$

$$P(|\bar{x}_1 - \bar{x}_2| < 5) = P(0 < t_{(73)} < 8.7814) = \left[P(t_{(73)} < 8.7814) - P(t_{(73)} \leq 0) \right]$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Student's t distribution with 73 DF
      x   P( X <= x )
0.0000           0.5
8.7814           1.0
```

$$P\left(|\bar{x}_1 - \bar{x}_2| < 5\right) = 1 - 0.5 = 0.5$$

b.2) Las varianzas son desconocidas y heterogéneas.

Obteniendo los grados de libertad ν:

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{36}{30} + \frac{15}{45}\right)^2}{\frac{\left(\frac{36}{30}\right)^2}{29} + \frac{\left(\frac{15}{45}\right)^2}{44}} = \frac{2.351111}{0.052180} = 45.057339 \approx 45$$

$$P\left(|\bar{x}_1 - \bar{x}_2| < 5\right) = P\left(-5 < \bar{x}_1 - \bar{x}_2 < 5\right) = P\left(\frac{-5+5}{\sqrt{\frac{36}{30}+\frac{15}{45}}} < t_{(45)} < \frac{5+5}{\sqrt{\frac{36}{30}+\frac{15}{45}}} \right)$$

$$P(|\bar{x}_1 - \bar{x}_2| < 5) = P(0 < t_{(45)} < 8.075) = P(t_{(45)} < 8.075) - P(t_{(45)} \leq 0)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Student's t distribution with 45 DF
   x         P( X <= x )
0.000       0.50000
8.075       1.00000
```

$$P\left(|\bar{x}_1 - \bar{x}_2| < 5\right) = 1 - 0.5 = 0.5$$

Distribución del cociente de varianzas $\dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$

25. Un experto de control de calidad está investigando el rendimiento de dos marcas de calculadoras: C1 y C2. La variabilidad de los rendimientos de cada una de las marcas es desconocida, pero se suponen homogéneas. Se entregan 20 calculadoras a dos grupos de usuarios de 10 personas cada uno. Cada usuario, luego de un periodo de tiempo de uso, califica su rendimiento en una escala de 0 a 20. Los datos procesados dieron como resultado que $s_1^2 = 5.15$ y $s_2^2 = 6.23$. De acuerdo a la información proporcionada, determine el valor de "k" en la siguiente expresión:

$$P\left(\frac{s_1^2}{s_2^2} \le k\right) = 0.99$$

Solución:

$$P\left(\frac{s_1^2}{s_2^2} \le k\right) = P\left(\frac{s_1^2}{s_2^2}\frac{\sigma_2^2}{\sigma_1^2} \le \frac{\sigma_2^2}{\sigma_1^2}k\right) = P(F_{(9,9)} \le k) = 0.99$$

Haciendo uso del Minitab

```
Inverse Cumulative Distribution Function
F distribution with 9 DF in numerator and 9 DF in denominator
P( X <= x )          x
        0.99   5.35113
Entonces, k = 5.35113.
```

26. El gerente de una empresa dedicada a la investigación de mercados ha recibido el encargo de realizar un estudio relacionado con el comportamiento de los clientes respecto a un grupo de restaurantes de la capital. Para este fin el gerente decidió seleccionar una muestra estratificada aleatoria de 50 clientes, registrando la siguiente información:

- Nombre del restaurante.
- Monto del consumo (en nuevos soles).
- Tipo de carne consumida (carne roja, pollo, pescado).

Luego de procesar los datos se obtuvo el siguiente reporte de Minitab:

Descriptive Statistics: Monto-consumo

Variable	Restaurantes	Count	Mean	StDev	Variance
Monto-consumo	Gaucho del Perú	10	989.7	107.0	1444.7
	Parrilla Argentina	21	1022.3	119.8	14351.2
	Punta Verde	5	841.8	37.3	1392.2
	Meat	9	972.1	111.5	12422.6
	Otros	5	917.40	17.87	319.30

¿Cuál es la probabilidad de que el cociente de la varianza muestral del monto de consumo de los clientes del restaurante "Meat" (1) con respecto a la varianza muestral del monto del consumo de los clientes del restaurante "Punta verde" (2) sea menor o igual que 0.35, asumiendo que $\sigma_1^2 = \sigma_2^2$?

Solución:

Se tiene por el reporte de Minitab que:

$n_1 = 5 \qquad s_1^2 = 37.3^2 \qquad s_1 = 37.3$

$n_2 = 9 \qquad s_2^2 = 111.5^2 \qquad s_2 = 111.5$

Se desea obtener: $P\left(\frac{s_2^2}{s_1^2} \le 0.35\right)$

$$P\left(\frac{s_2^2}{s_1^2} \le 0.35\right) = P\left(\frac{s_2^2\sigma_1^2}{s_1^2\sigma_2^2} \le 0.35\frac{\sigma_1^2}{\sigma_2^2}\right) = P\left(F_{(n_2-1;n_1-1)} \le 0.35\frac{\sigma_1^2}{\sigma_2^2}\right)$$

$$P\left(\frac{s_2^2}{s_1^2} \le 0.35\right) = P(F_{(8;4)} \le 0.35) = P(F_{(8;4)} \le 0.35)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
F distribution with 8 DF in numerator and 4 DF in denominator
    x   P( X <= x )
0.35     0.0963879
```

Entonces: $P\left(\frac{s_2^2}{s_1^2} \le 0.35\right) = 0.0963879$

Distribución de la diferencia de proporciones ($p_1 - p_2$)

27. Según las autoridades del gobierno, el 45% de los choferes de servicio público urbano son propietarios de sus vehículos, mientras que en el caso de los vehículos de servicio público interprovincial lo son el 25% de los choferes. Suponga que se seleccionan 200 choferes de servicio público urbano y 300 choferes de servicio público interprovincial.

a. ¿Cuál es la probabilidad de que la proporción muestral de choferes de servicio público interprovincial que son propietarios de sus vehículos difiera de la correspondiente proporción poblacional en menos de 0.05?

Solución:

Sea:

π_2 : Proporción de choferes de servicio público interprovincial. $\pi_2 = 0.25$

La distribución de la proporción muestral de choferes de servicio público interprovincial está dada por:

$$p_2 \sim N(0.25, 0.025^2) \text{ donde: } \sigma_{(p_2)} = \sqrt{\frac{\pi_2(1-\pi_2)}{n_2}} = \sqrt{\frac{0.25(0.75)}{300}} = 0.025$$

Entonces:

$$P(|\, p_2 - \pi_2 \,| < 0.05) = P(-0.05 < p_2 - \pi_2 < 0.05)$$

$$P(|\, p_2 - \pi_2 \,| < 0.05) = P(-0.05 + 0.25 < p_2 < 0.05 + 0.25) = P(0.2 < p_2 < 0.3)$$

$$P(|\, p_2 - \pi_2 \,| < 0.05) = P(p_2 < 0.3) - P(p_2 \leq 0.2)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 0.25 and standard deviation = 0.025
  x   P( X <= x )
0.2      0.022750
0.3      0.977250
```

Obteniéndose:

$$P(|p_2 - \pi_2| < 0.05) \cong 0.977250 - 0.022750 = 0.9545$$

b. ¿Cuál es la probabilidad de que la proporción muestral de choferes de servicio público interprovincial, que son propietarios de sus vehículos, difiera de la proporción muestral de choferes de servicio público urbano, que son propietarios de sus vehículos en menos de 0.1?

Solución:

Sean:

π_1 : Proporción de choferes de servicio público urbano $\pi_1 = 0.45$.

π_2 : Proporción de choferes de servicio público interprovincial $\pi_2 = 0.25$.

La distribución muestral de la diferencia de proporciones está dada por:

$$p_2 - p_1 \to N\left(-0.20, 0.0432^2\right) \text{ donde:}$$

$$\sigma_{(p_2 - p_1)} = \sqrt{\frac{0.25(0.75)}{300} + \frac{0.45(0.55)}{200}} = 0.0432$$

Se debe obtener: $P(|\, p_2 - p_1 \,| < 0.1)$

$$P(|p_2 - p_1| < 0.1) = P(-0.1 < p_2 - p_1 < 0.1) = P(p_2 - p_1 < 0.1) - P(p_2 - p_1 \leq -0.1)$$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = -0.2 and standard deviation = 0.0432
   x   P( X <= x )
 0.1       1.00000
-0.1       0.98969
```

Obteniendo:

$$P(|\ p_2 - p_1\ | < 0.1) \cong 1 - 0.9897 = 0.0103$$

28. Según el INEI, el 45% de amas de casa del distrito de Surco afirman que el Sistema de Seguridad (SS) de su distrito es satisfactorio, y en el distrito de San Isidro el 25% de amas de casa afirman lo mismo. Se seleccionó una muestra aleatoria de 180 amas de casa del distrito de Surco y 120 amas de casa del distrito de San Isidro.

a. ¿Cuál es la probabilidad de que la proporción muestral de amas de casa del distrito de Surco que afirmen que el sistema de seguridad de su distrito es satisfactoria supere la proporción muestral del distrito de San Isidro en a lo más 0.35?

Solución:

Sea: π_1 la proporción de amas de casa que afirma que SS es satisfactoria en el distrito de Surco

$\pi_1 = 0.45 \qquad n_1 = 180$

π_2: Proporción de amas de casa que afirma que SS es satisfactoria en el distrito de San Isidro

$\pi_2 = 0.25 \qquad n_2 = 120$

La distribución muestral de la diferencia de proporciones está dada por:

$(p_1 - p_2) \to N(0.20, 0.054199^2)$, donde:

$$\sigma_{(p_2 - p_1)} = \sqrt{\frac{0.45(0.55)}{180} + \frac{0.25(0.75)}{120}} = 0.054199$$

Se debe obtener: $P(0 < p_1 - p_2 \ \leq 0.35)$

Haciendo uso del Minitab

Cumulative Distribution Function

```
Normal with mean = 0.2 and standard deviation =
0.054199
  x       P( X <= x )
0.35      0.997176
0.0       0.000112
```

Obteniendo:

$$P(0 < p_1 - p_2 \leq 0.35) = P(p_1 - p_2 \leq 0.35) - P(p_1 - p_2 \leq 0) =$$

$$0.997176 - 0.000112 = 0.997064$$

b. ¿Cuál es la probabilidad de que la proporción muestral de amas de casa del distrito de Surco que afirman que el sistema de seguridad es satisfactoria difiera de la proporción muestral del distrito de San Isidro en a lo más 0.3?

Solución:

La distribución muestral de la diferencia de proporciones está dada por:

$(p_1 - p_2) \rightarrow N(0.20, 0.054199^2)$, donde:

$$\sigma_{(p_2 - p_1)} = \sqrt{\frac{0.45(0.55)}{180} + \frac{0.25(0.75)}{120}} = 0.054199$$

Se debe obtener: $P(|p_1 - p_2| \leq 0.3)$

$$P(|p_1 - p_2| \leq 0.3) = P(-0.3 \leq (p_1 - p_2) \leq 0.3) =$$

$$P((p_1 - p_2) \leq 0.3) - P((p_1 - p_2) < -0.3)$$

Haciendo uso del Minitab

```
Cumulative Distribution Function
Normal with mean = 0.2 and standard deviation = 0.054199
   x  P( X <= x )
-0.3     0.000000
 0.3     0.967485
```

Obteniendo:

$$P(|p_1 - p_2| \leq 0.3) \cong 0.967485 - 0 = 0.967485$$

Miscelánea de problemas

29. Se presentan los siguientes casos:

a. Estudios de retiro de documentos indican que el tiempo de demora de los trámites en la Aduana del Callao para el retiro de documentos tiene distribución Normal con una media de 20 horas y desviación estándar 2 horas, y en la Aduana del Aeropuerto tiene distribución Normal con una media de 30 horas y desviación estándar 5 horas. Si se eligen muestras aleatorias de los tiempos de demora de los trámites de ambas aduanas, 20 observaciones de la Aduana del Callao y 15 del Aeropuerto, ¿cuál es la probabilidad de que la varianza muestral de los tiempos de demora de

los trámites en la Aduana del Callao sea menor que la varianza muestral de los tiempos de demora de los trámites en la Aduana del Aeropuerto?

Solución

Sea X_1: Tiempo de demora de trámites en la Aduana del Callao

Donde: $X_1 \sim N(20, 2^2)$

Sea: X_2: Tiempo de demora de trámites en la aduana del Aeropuerto Jorge Chávez

Donde: $X_2 \sim N(30, 5^2)$

Se desea obtener:

$$P\left(s_1^2 < s_2^2\right) = P\left(\frac{s_1^2}{s_2^2} < 1\right) = P\left(\frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2} < \frac{\sigma_2^2}{\sigma_1^2}\right) = P\left(F_{(19,14)} < \frac{25}{4}\right) = P\left(F_{(19,14)} < 6.25\right)$$

Por lo tanto:

$P(s_1^2 < s_2^2) = 0.9994$

b. El gerente de la Superintendencia Nacional de Aduanas (Sunad) conoce que el tiempo de demora de los trámites en la Aduana del Callao para el retiro de documentos tiene distribución Normal con una media de 28 horas y en la Aduana del Aeropuerto tiene distribución Normal con una media de 30 horas. Si el gerente elige muestras aleatorias de los tiempos de demora de los trámites de ambas aduanas, 24 observaciones de la Aduana del Callao (1) y 18 del Aeropuerto Jorge Chávez (2), encontrando que $s_1^2 = 8$ y $s_2^2 = 30$, ¿cuál es la probabilidad de que las medias muestrales difieran en al menos 1 hora? Suponga varianzas poblacionales desconocidas y homogéneas.

Solución

Se sabe que:

$$\mu_1 = 28 \quad \text{y} \quad \mu_1 = 30 \qquad s_p^2 = \frac{23(8) + 17(30)}{40} = 17.35 \qquad s_p = 4.1653$$

Se desea obtener:

$$P(|\bar{x}_1 - \bar{x}_2| \geq 1) = 1 - P(|\bar{x}_1 - \bar{x}_2| < 1) = 1 - P(-1 < \bar{x}_1 - \bar{x}_2 < 1)$$

$$= 1 - P\left(\frac{-1-(-2)}{4.1653\sqrt{\frac{1}{24}+\frac{1}{18}}} < \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < \frac{1-(-2)}{4.1653\sqrt{\frac{1}{24}+\frac{1}{18}}}\right)$$

$$= 1 - \mathrm{P}(0.7699 < \mathrm{t}_{(40)} < 2.3099)$$

Haciendo uso del Minitab:

Cumulative Distribution Function

```
Student's t distribution with 40 DF
   x       P( X <= x )
0.7699      0.777059
2.3099      0.986932
```

Obteniendo:

$$P\left(|\overline{x}_1 - \overline{x}_2| \geq 1\right) = 1 - \left(0.9869 - 0.7771\right) = 1 - 0.2098 = 0.7902$$

30. Un funcionario del gobierno central que participó en un estudio sobre guarderías infantiles ha registrado los ingresos de dos grupos de mujeres trabajadoras; el primer grupo conformado por 10 mujeres elegidas al azar que tienen a sus niños en guarderías infantiles, y el segundo grupo conformado por 12 mujeres elegidas al azar que no tienen a sus niños en guarderías infantiles. Si el funcionario supone que los ingresos tienen distribución Normal para ambos grupos de mujeres trabajadoras:

a. ¿Cuál será el valor máximo de la varianza muestral de los ingresos de las mujeres con niños en guarderías infantiles con 0.975 de probabilidad? Suponga $\sigma^2 = 20$

Solución

Sea X: Ingresos de las mujeres con niños en guarderías infantiles.

Con: $n = 10$ y $\sigma^2 = 20$

Además: $\dfrac{(n-1)s^2}{\sigma} \sim \chi^2_{(n-1)}$

Se desea obtener el valor de la constante "a" tal que: $P\left(s^2 < a\right) = 0.975$

$$P(s^2 < a) = 0.975 \Rightarrow P\left(\frac{(n-1)s^2}{\sigma^2} < \frac{(n-1)a}{\sigma^2}\right) = 0.975$$

$$\Rightarrow P\left(\chi^2_{(9)} < \frac{(10-1)a}{20}\right) = 0.975$$

Haciendo uso del Minitab:

Inverse Cumulative Distribution Function

```
Chi-Square with 9 DF
P( X <= x )          x
  0.975          19.0228
```

Luego:

$$\frac{(10-1)a}{20} = 19.0228 \quad \Rightarrow \quad a = \frac{20(19.0228)}{9} = 42.2728$$

b. ¿Cuál es la probabilidad de que la varianza muestral de los ingresos del grupo de mujeres con niños en guarderías infantiles (1) sea mayor que la mitad de la varianza muestral de los ingresos del grupo de mujeres sin niños en guarderías infantiles (2)? Considere $\sigma_1^2 = 20$ y $\sigma_2^2 = 55$

Solución:

Se desea obtener: $P\left(s_1^2 > \frac{s_2^2}{2}\right)$

$$P\left(s_1^2 > \frac{s_2^2}{2}\right) = P\left(\frac{s_1^2}{s_2^2} > \frac{1}{2}\right) = 1 - P\left(\frac{s_1^2}{s_2^2} \le \frac{1}{2}\right) = 1 - P\left(\frac{s_1^2 \big/ \sigma_1^2}{s_1^2 \big/ \sigma_2^2} \le \frac{1 \big/ \sigma_1^2}{2 \big/ \sigma_2^2}\right)$$

$$P\left(s_1^2 > \frac{s_2^2}{2}\right) = 1 - P\left(F_{(n_1-1,n_2-1)} \le \frac{\sigma_2^2}{2\sigma_1^2}\right) = 1 - P\left(F_{(9,11)} \le \frac{55}{40}\right) =$$

$$1 - P\left(F_{(9,11)} \le \frac{55}{40}\right) = \ 1 - P(F_{(9,11)} \le 1.375)$$

Haciendo uso del Minitab:

Cumulative Distribution Function

```
F distribution with 9 DF in numerator and 11 DF in
denominator
     x      P( X <= x )
1.375      0.695544
```

$$P\left(s_1^2 > \frac{s_2^2}{2}\right) = 1 - 0.695544 = 0.304456$$

c. ¿Cuál es la probabilidad de que el ingreso medio muestral de las mujeres que tienen a sus hijos en guarderías infantiles (1) sea mayor que el ingreso medio muestral de las mujeres que no tienen a sus hijos en guarderías infantiles (2) en más de 30 nuevos soles? Suponga varianzas desconocidas pero homogéneas con $\mu_1 = 700$ y $\mu_2 = 680$. Además considere que $s_1^2 = 19.1 \quad y \quad s_2^2 = 21.26$

Solución

Se desea obtener: $P(\bar{x}_1 - \bar{x}_2 > 30)$

$$P(\bar{x}_1 - \bar{x}_2 > 30) = 1 - P(\bar{x}_1 - \bar{x}_2 \le 30) = 1 - P\left(\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \le \frac{30 - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\right)$$

$$P(\bar{x}_1 - \bar{x}_2 > 30) = 1 - P\left(t_{(10+12-2)} \leq \frac{30 - (700 - 680)}{\sqrt{s_p^2\left(\frac{1}{10} + \frac{1}{12}\right)}} \right) =$$

$$= 1 - P\left(t_{(20)} \leq \frac{10}{\sqrt{20.288\left(\frac{1}{10} + \frac{1}{12}\right)}} \right)$$

$$P(\bar{x}_1 - \bar{x}_2 > 30) = 1 - P\left(t_{(20)} \leq \frac{10}{1.928592} \right) = 1 - P(t_{(20)} \leq 5.18513)$$

Donde:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(10 - 1)19.1^2 + (12 - 1)21.26^2}{10 + 12 - 2} = 20.288$$

Haciendo uso del Minitab:

```
Cumulative Distribution Function
Student's t distribution with 20 DF
       x  P( X <= x )
5.18513      0.999978
```

Obteniendo:

$$P(\bar{x}_1 - \bar{x}_2 > 30) = 1 - 0.999978 = 0.000022$$

PROBLEMAS PROPUESTOS

1. Para cada uno de los casos indicar la unidad de muestreo, la población objetivo, el marco muestral y el tipo de muestreo más apropiado:

a. Se desea estimar el porcentaje de personas que piensan votar blanco o viciado en las próximas elecciones municipales.

b. Se está interesado en conocer la opinión de los turistas extranjeros sobre la situación del país.

c. Se desea estimar el promedio ponderado obtenido por los alumnos del curso de Estadística y Probabilidad I.

2. El director ejecutivo de una empresa que realiza ventas de teléfono averiguó que a su *call center* entraron el día anterior 6 llamadas. A causa de la insuficiencia del personal las demoras de cada cliente en contactar con la oficina de ventas fueron: 18, 16, 20, 14, 12 y 22 segundos.

a. Si el director ejecutivo tuviera que elegir una muestra aleatoria de 2 llamadas (sin reposición), ¿cuántas muestras puede escoger?

b. En base a las muestras obtenidas de tamaño 2, obtener $E(\bar{x})$ y $V(\bar{x})$.

c. ¿Cuál es la probabilidad de que en (a) se elijan en la muestra las 2 demoras más altas?

d. ¿Cuál es la probabilidad de que en (a) se elija en la muestra la demora de 20 segundos?

Nota: En la parte b) use el teorema de la distribución de la media muestral.

3. Sea $F \sim F_{(15;8)}$, la probabilidad de que la variable aleatoria F sea mayor que 8.94 y menor que 15 es:

a. 0.0002 b. 0.0017 c. 0.9980 d. 0.9997 e. 1.000.

4. En las siguientes expresiones, encuentre los valores que correspondan.

a. Si $X \sim \chi^2_{(13)}$, encontrar el valor de k en la siguiente expresión $P(k < X) = 0.58$.

b. Si $X \sim t_{(11)}$, encontrar el valor de $P(|t| > 3.2)$.

c. Si $X \sim F_{(12;31)}$, encontrar el valor de $P(0.8 < F < 1.1) + P(2.3 < F < 2.7)$.

5. Complete los espacios en blanco de las siguientes expresiones:

a. Si $X \sim \chi^2_{(10)}$, entonces la probabilidad de que X sea mayor que 8 o menor que 3 es: _____

b. Si $X \sim \chi^2_{(14)}$, entonces la probabilidad de que X sea mayor que su promedio es: _____

c. Si $X \sim t_{(13)}$, entonces la probabilidad de que X sea mayor que -2.55 es: _____

d. Si $X \sim t_{(15)}$, entonces el valor del primer cuartil es: _____

e. Si $X \sim F_{(6,13)}$, entonces la probabilidad de que X sea mayor que 3.15 es: _____

f. Si $X \sim F_{(4,9)}$, entonces los valores de a, b y c tales que:

$P(a < X < b) = \frac{19}{45}$, $P(c < X < b) = \frac{47}{90}$ y $P(X > b) = \frac{25}{90}$

son: _____

6. En la fabricación de una alfombra se utiliza una fibra sintética con una resistencia a la tensión que tiene distribución Normal con media 75.50 y desviación estándar 3.5 psi.

a. Encuentre la probabilidad de que en una muestra aleatoria de 6 unidades de fibra, la media de la resistencia a la tensión en la muestra sea mayor que 75.75 psi.

b. ¿Cómo cambia la desviación estándar de la media muestral cuando el tamaño de la muestra aumenta de 6 a 49?

7. El promedio diario de ventas de la empresa Dabarret S.A. es de S/.8.000, con una desviación estándar de S/.1.000. Si la distribución de las ventas es Normal, hallar:

a. La probabilidad de que una venta, en un día cualquiera, se encuentre entre los S/.7.000 y S/.8.600.

b. La probabilidad de que el promedio de ventas en 100 días sea menor que S/.7.950.

8. ¿Cuántas observaciones deben ser seleccionadas de una población Normal con media μ y varianza 9 para que la media muestral difiera de la media poblacional en menos de 2 con probabilidad 0,95?

9. El tiempo de fabricación de una plancha de vidrio de un metro cuadrado es una variable aleatoria con distribución uniforme en el intervalo de 15 a 19 minutos. Si se registran los tiempos de fabricación de 200 planchas de vidrio, ¿cuál es la probabilidad de que el tiempo promedio muestral sea menor que 16.85 minutos?

10. Una máquina embotelladora puede regularse de tal manera que llene en promedio μ onzas por botella. Se ha observado que las onzas de contenido que vacía la máquina embotelladora tienen distribución Normal con $\sigma = 1$ onza. Suponga que se selecciona una muestra aleatoria de 10 botellas y se mide el contenido de cada botella. Utilizando esas 10 observaciones encuentre los números b_1 y b_2 tales que: $P(b_1 \leq s^2 \leq b_2) = 0.90$.

Nota: Suponga el 90% de la parte central de la distribución.

11. Un instituto de apoyo a la mujer informó que el 37% de las mujeres que viven en Lima norte sufren de maltratos físicos y psicológicos por parte de su esposo o cónyuge. Si se selecciona una muestra aleatoria simple de 1.000 mujeres con este tipo de problema:

a. Indique la distribución muestral de la proporción de mujeres que sufren maltratos físicos y psicológicos por parte de su esposo o cónyuge.

b. ¿Cuál es la probabilidad de que la proporción muestral esté a ± 0.03 de la proporción poblacional?

c. Conteste el inciso "b", pero ahora con una muestra aleatoria simple de 500.

12. Los datos de una empresa revelan que el 75% de los clientes contactados por correo compran su producto. Si la empresa envía 200 cartas en las que ofrece vender su producto y si tiene que hacer 160 ventas como mínimo para financiar un negocio que tiene planificado para el próximo mes, ¿cuál es la probabilidad de que pueda financiar el negocio?

13. Un industrial compró la producción total de un año de un componente electrónico para televisores tipo plasma. Los datos técnicos proporcionados por la compañía son los siguientes: la duración media del componente es 2.800 horas y la desviación estándar es 500 horas. Si se consideran dos muestras de componentes electrónicos, una de tamaño 120 y la otra de tamaño 200, calcular:

a. La probabilidad de que la duración media de vida de la primera muestra no sea superior en más de 100 horas a la duración media de la segunda.

b. La probabilidad de que $\bar{x}_1 - \bar{x}_2$ se encuentre entre 100 y 200 horas.

14. Una empresa de investigación de mercado concluyó en su última investigación que el 30% de mujeres y el 20% de hombres consumen cierto producto de aseo personal. Si se hace una encuesta a 200 hombres y 200 mujeres elegidas aleatoriamente, determine la probabilidad de que las mujeres acepten el producto más que los hombres.

15. El director del hospital sostiene, en base a su experiencia, que la edad de los pacientes internados por enfermedades cardiovasculares tiene distribución Normal con una media de 70 años y varianza σ_1^2. También sostiene que la edad de los pacientes internados por enfermedades gastrointestina les tiene distribución Normal con una media de 65 años y varianza σ_2^2. Si se extraen muestras aleatorias de pacientes internados por ambas enfermedades y se obtienen los siguientes resultados:

Descriptive Statistics:

Variable	Count	Variance
Cardiovascular	26	64.64
Gastrointestinal	35	27.841

a. ¿Cuál es la probabilidad de que la edad promedio de los pacientes internados por enfermedades cardiovasculares (1) sea menor que la edad promedio de la muestra de los pacientes internados por enfermedades gastrointestinales (2)? Suponga varianzas poblacionales desconocidas y heterogéneas.

b. Se extrae otra muestra $n_1 = 26$ y $n_2 = 35$. ¿Cuál es la probabilidad de que $P\left(\frac{s_1^2}{s_2^2} \leq 0.8\right)$? Suponga varianzas poblacionales desconocidas y homogéneas.

16. Según un representante del Ministerio de Salud, el 40% de los pacientes internados por enfermedades gastrointestinales (1) gasta al menos 500 nuevos soles en el tratamiento de su enfermedad mientras que el 50% de los pacientes internados por enfermedades respiratorias (2) gasta al menos 400 nuevos soles en el tratamiento de su enfermedad. Si se extraen muestras aleatorias de pacientes internados por ambas enfermedades con $n_1 = 250$ y $n_2 = 300$, ¿cuál es la probabilidad de que las proporciones muestrales difieran en menos de 0.04?

17. Delifrezze es un helado de crema producido por Friola S.A. El gerente de producción de la empresa tuvo interés en determinar la demanda de Fríorrico por el consumidor limeño. Para este fin, el gerente decide realizar un estudio sobre la base de una muestra aleatoria de hogares. Los principales objetivos del estudio fueron:

- Estimar el gasto mensual por hogar en productos Friorrico.
- Estimar la proporción de hogares que conocen la existencia de Fríorrico.

a. Si la estrategia de muestreo consiste en dividir la zona de estudio en 4 zonas de características socioeconómicas similares, seleccionando una muestra aleatoria de viviendas de cada zona, defina los siguientes términos:
   a.1 Unidad de estudio.
   a.2 Variables de estudio.
   a.3 Tipo de muestreo a usar por Fríorrico.

b. El gerente de Friola S.A. supone que en la zona 1 el gasto mensual en productos Fríorrico se distribuye normalmente con un promedio por hogar de S/.250 y con una desviación estándar de S/.40. Si en esta zona se selecciona una muestra de 200 hogares,
   b.1 ¿Cuál es la probabilidad de que un hogar gaste menos de S/.215?
   b.2 ¿Cuál es la probabilidad de que el gasto promedio, obtenido de la muestra, supere los S/.255?

c. Si el gerente de Friola S.A. supone que en la zona 3 el gasto mensual en productos Fríorrico se distribuye como un Ji-Cuadrado con un promedio de S/.300 por hogar, determine lo siguiente:
   c.1 El porcentaje de hogares de la zona 3 que gastan más de S/.320.
   c.2 El valor de la mediana del gasto mensual, en la zona 3.

   Nota: Si $X \sim \chi^2_{(v)}$, entonces $E(X) = v$ y $V(X) = 2v$.

d. ¿Cuántos hogares se deberían seleccionar para que la proporción muestral de hogares que conocen la existencia de Fríorrico difiera de la verdadera proporción en no más de 0.05 con una probabilidad de 0.95? Suponga que $\pi = 0.5$.

18. La compra de medicamentos genéricos permite al usuario ahorrar hasta 60% en promedio. Un medicamento genérico es más barato que uno de marca por la siguiente razón: que pueden ser producidos por cualquier laboratorio, mientras que los de marca tienen derecho de fabricación exclusiva.

a. Un experto internacional de la Organización Mundial de la Salud (OMS) afirma que los precios de los medicamentos genéricos se distribuyen según una distribución exponencial con parámetro 3.5.

Nota: La distribución exponencial es: $X \sim \exp(\beta)$ con $E(X) = \beta$ y $V(X) = \beta^2$.

1. Indique la unidad de análisis y la variable de estudio.
2. Si se seleccionan 80 medicamentos genéricos, ¿cuál es la probabilidad de que el promedio muestral de los precios exceda a su media poblacional en por lo menos 0.01 nuevo sol?

b. De igual forma, el experto internacional de la OMS sostiene que los precios de los medicamentos de marca tienen distribución Normal con media μ y $\sigma^2 = 100$. ¿Cuántos medicamentos deben ser seleccionados para que la probabilidad de que el promedio muestral exceda a la media poblacional en por lo menos 2 nuevos soles sea igual a 0.15?

19. El precio de un barril de petróleo Brent tiene distribución Normal y se registran las cotizaciones de 20 barriles.

a. ¿Entre qué valores se encontrará la varianza muestral con probabilidad 0.8? Suponga que la probabilidad de 0.8 se encuentra en el centro de la distribución y asuma $\sigma^2 = 0.64$.

b. Si se registran las cotizaciones de 18 barriles de petróleo Brent y se determina que $s = 0.4$, ¿cuál es la probabilidad de que el promedio muestral difiera de la media poblacional en menos de 0.15 dólares?

20. La empresa Phoenix (1) sostiene que su gasto anual (en miles de millones de dólares) en investigación científica tiene distribución Normal con media 30. Por otro lado, la empresa TTA (2) sostiene que su gasto anual (en miles de millones de dólares) en investigación tiene distribución Normal con media 40. Luego de seleccionarse muestras aleatorias de los montos dedicados a la investigación por ambas empresas se encontró que:

```
Phoenix (1) n1=18     s=2.25
TTA (2)     n2=14     s=8
```

a. ¿Cuál es la probabilidad de que el promedio de la muestra de Phoenix sea inferior al promedio de la muestra de TTA? Suponga varianzas poblacionales desconocidas y heterogéneas.

b. ¿Cuál es la probabilidad de que el promedio de la muestra de Phoenix difiera del promedio de la muestra de TTA en más de 5? Suponga varianzas poblacionales desconocidas y homogéneas.

21. Según un experto en economía, el 48% de los comerciantes de Tacna han sido beneficiados con la caída del tipo de cambio del dólar; mientras que solo el 38% de los comerciantes de Iquitos han sido beneficiados con la caída del tipo de cambio de la moneda norteamericana.

a. Si se seleccionan al azar a 300 comerciantes de Tacna, ¿cuál es la probabilidad de que al menos 160 respondan que fueron beneficiados con la caída del tipo de cambio del dólar?

b. Si se seleccionan muestras aleatorias de Tacna e Iquitos de tamaño 250 y 400, respectivamente, ¿cuál es la probabilidad de que la proporción de comerciantes de Tacna que fueron beneficiados con la caída del tipo de cambio del dólar supere la proporción de comerciantes de Iquitos que fueron beneficiados con la caída del tipo de cambio de la moneda norteamericana en al menos 0.15?

22. El director de un hospital del sistema de salud pública ha acumulado experiencia, durante sus años de trabajo, acerca del comportamiento de las siguientes variables correspondientes a los pacientes que son internados en su institución:

X_1 = Edad del paciente.

X_2 = Tipo de dolencia del paciente.

X_3 = Número de intervenciones quirúrgicas del paciente.

X_4 = Costo aproximado de la medicina utilizada por el paciente.

a. El director sostiene que según su experiencia $X_1 \sim N(60, 8^2)$. Entonces:
   i) Si se selecciona un paciente al azar, ¿cuál es la probabilidad de que su edad sea menor que 65 años?
   ii) Si se seleccionan al azar a 16 pacientes, ¿cuál es la probabilidad de que el promedio sea mayor que 64 años?
   iii) ¿Cuántos pacientes deben ser seleccionados tal que la probabilidad de que el promedio muestral supere a la media poblacional en al menos 1.5 años sea igual a 0.92?

b. La distribución de la variable X_3 es:

X_3	0	1	2	4	5
$f(X_3)$	0.25	0.4	0.2	0.1	0.05

Si se eligen al azar a 64 pacientes, ¿cuál es la probabilidad de que el promedio muestral sea mayor que 1.2?

c. Si se seleccionan al azar 49 observaciones de la variable X_4 cuya distribución es Normal con media $\mu = 200$ y varianza σ^2, ¿cuál es la probabilidad de que el promedio muestral se encuentre entre 185 y 210 nuevos soles? Suponga que la varianza de la muestra es 1225.

d. ¿Cuál es la probabilidad de que la varianza de la muestra sea mayor que 1.000? Suponga que $\sigma^2 = 1500$ y $n = 49$.

e. El director del hospital también sostiene que según su experiencia con respecto a la variable X_2, el 40% de los pacientes es internado en el hospital por enfermedades gastrointestinales. Si se seleccionan al azar a 300 pacientes, ¿cuál es la probabilidad de que como máximo 100 pacientes sufran de enfermedades gastrointestinales?

Capítulo

2 Estimación de parámetros: Puntual y por intervalos

En este capítulo trataremos los siguientes temas:

- Conceptos relacionados con el problema de estimación
- Estimación puntual
  - Propiedades de los estimadores puntuales
  - Métodos de obtención de estimadores puntuales
- Estimación por intervalos

Los procedimientos para la estimación de parámetros poblacionales a partir de una muestra aleatoria es uno de los temas que trata este capítulo, donde también se detallan las propiedades que debe poseer un buen estimador puntual, así como sus métodos de obtención. Asimismo, se explica la utilidad de la estimación por intervalos, en la cual se calculan dos valores, entre los que se encuentra el verdadero valor del parámetro poblacional, con un nivel de confianza fijado de antemano. A partir de este punto, se hace uso intensivo del software estadístico Minitab, cuya utilización es explicada en detalle con la aplicación de diferentes casos de estudio.

Los datos empleados en este capítulo y en los siguientes se pueden encontrar en la dirección electrónica: <http://downloads.ulima.edu.pe/fondoeditorial/libros/estadprobab002/datos>.

1. Introducción

En el capítulo anterior se mencionó que los objetivos de la inferencia estadística son estimar los valores de los parámetros de una población, como la media o la proporción de la población, entre otros, y realizar la prueba de la hipótesis referente a valores de los parámetros. En este capítulo se estudiará la estimación de los parámetros, puntual y por intervalos.

La estimación de los parámetros es una técnica muy útil en las investigaciones científicas, especialmente en el proceso de la toma de decisiones en los diferentes campos del quehacer de un profesional. Por ejemplo, los ingenieros y los jefes de plantas pueden desear estimar el porcentaje de defectuosos de una línea de ensamblaje para controlar la calidad, el número de horas inoperativas de las máquinas y muchos otros problemas asociados a la industria.

Asimismo, el Gobierno central, para planificar la política económica de un país a corto, mediano y largo plazos, necesita estimaciones, cada cierto tiempo, de los niveles de desempleo, inflación, producción agrícola e industrial, importación y exportación. En las empresas, los directivos piden estimaciones de cierta información para las operaciones cotidianas o para la planificación; un gerente de ventas desea estimar las demandas de sus productos con el fin de tener un *stock* adecuado, etcétera. Estos problemas pueden ser analizados con las técnicas de estimación de parámetros.

2. Definición

La estimación consiste en encontrar un valor que represente una buena aproximación de un parámetro desconocido θ . En la práctica, se supone que la variable aleatoria X (población) tiene una determinada distribución de probabilidad, luego se toma una muestra (o muestras) de n observaciones y a partir de esta se busca estimar los parámetros de dicha población. Esta estimación puede estar

dada por un único valor experimental obtenido a partir de la muestra y se denomina *estimación puntual*, o puede estar dada por un conjunto de valores que constituyen un intervalo experimental, cuyos extremos son obtenidos a partir de la muestra, se espera que dicho intervalo contenga el verdadero valor del parámetro con cierto grado de seguridad o confianza medido en términos de probabilidad; a esta estimación se le denomina como *estimación por intervalos de confianza*.

3. Características de un buen estimador puntual

Un buen estimador debe poseer las siguientes propiedades: insesgabilidad, consistencia, suficiencia y eficiencia.

3.1 Insesgabilidad

Se dice que un estimador $\hat{\theta}$ de un parámetro desconocido θ es insesgado si: $E(\hat{\theta}) = \theta$

Ejemplo 1:

La media muestral $\overline{x}$ es un estimador insesgado de μ, ya que: $E(\overline{x}) = \mu$

Ejemplo 2:

La varianza muestral $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ es un estimador insesgado de la varianza poblacional σ^2.

En efecto:
Se sabe que si $Y \sim \chi^2_{(n)}$; $\Rightarrow E(Y) = n$; por lo tanto la variable:

$U = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$, como: $E(U) = (n-1)$; $\Rightarrow E\left[\frac{(n-1)s^2}{\sigma^2}\right] = (n-1)$ y aplicando propiedades de valor esperado se tiene:

$\frac{(n-1)}{\sigma^2} E(s^2) = (n-1) \Rightarrow E(s^2) = \sigma^2$.

Nota: Si un estimador $\hat{\theta}$ de un parámetro desconocido θ es sesgado $(E(\hat{\theta}) \neq \theta)$, se dice que es asintóticamente insesgado si satisface el siguiente límite: $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$

Ejemplo 3:

La varianza muestral $s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ es un estimador sesgado de σ^2, pues $E(s_1^2) = \frac{n-1}{n}\sigma^2$. Pero: $\underset{n\to\infty}{Lim}\ E(s_1^2) = \underset{n\to\infty}{Lim}\frac{n-1}{n}\sigma^2 = \sigma^2$; por lo tanto, s_1^2 es un estimador asintóticamente insesgado de σ^2.

3.2 Consistencia

Se dice que un estimador $\hat{\theta}$ de un parámetro desconocido θ es consistente si satisface el siguiente límite: $\underset{n\to\infty}{Lim} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$. Es decir, un estimador $\hat{\theta}$ es consistente si a medida que el tamaño n de la muestra aumenta, la probabilidad de que el estimador $\hat{\theta}$ sea igual al parámetro θ tiende a uno. Una manera de ver la consistencia es probar que:

$$\underset{n\to\infty}{Lim} E(\hat{\theta}_n) = \theta \quad y \quad \underset{n\to\infty}{Lim} V(\hat{\theta}_n) = 0$$

Ejemplo 4:

Si X se distribuye con función de densidad no Normal $f(X)$ y $E(X) = \mu$, $V(X) = \sigma^2$. Entonces $\bar{x}$ es un estimador consistente de μ.

En efecto, como $E(\bar{x}) = \mu$ y $V(\bar{x}) = \frac{\sigma^2}{n}$; por el teorema central del límite, entonces:

$$\underset{n\to\infty}{Lim} E(\bar{x}_n) = \underset{n\to\infty}{Lim}\mu = \mu \quad y \quad \underset{n\to\infty}{Lim} V(\bar{x}_n) = \underset{n\to\infty}{Lim}\frac{\sigma^2}{n} = 0$$

Por consiguiente, $\bar{x}$ es un estimador consistente de μ.

3.3 Suficiencia

Un estimador $\hat{\theta}$ de un parámetro desconocido θ es suficiente si a medida que n aumenta proporciona una mayor información de la población. Por ejemplo: $\bar{x}$ (media muestral) al tomar en su cálculo a todos los valores de la muestra, es un estimador suficiente de μ, en tanto que la mediana y la moda no lo son.

3.4 Eficiencia

Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores insesgados diferentes, de un parámetro desconocido θ con varianzas $V(\hat{\theta}_1)$ y $V(\hat{\theta}_2)$, respectivamente; si $V(\hat{\theta}_1) < V(\hat{\theta}_2)$,

entonces $\hat{\theta}_1$ es un estimador más eficiente que $\hat{\theta}_2$. Es decir, un estimador es eficiente si posee la menor varianza. Por ejemplo: $\bar{x}$ es un estimador eficiente de μ.

Ejemplo 5:

La lectura de un voltímetro conectado a un circuito de prueba tiene una distribución Uniforme en el intervalo $(\theta, \theta+1)$, en donde θ es el verdadero voltaje del circuito. Si, $x_1, \cdots, x_n$ es una muestra aleatoria de tales lecturas.

a. Demostrar que $\bar{x}$ es un estimador sesgado de θ y calcular su sesgo.

b. Encontrar una función de $\bar{x}$ que sea un estimador insesgado de θ.

Solución:

a. Obteniendo el valor esperado de $\bar{x}$, en efecto:

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n}\left(\theta + \frac{1}{2}\right) = \theta + \frac{1}{2} \neq \theta$$

Por consiguiente, $\bar{x}$ es un estimador sesgado de θ con sesgo de. 0.5.

b. Como el sesgo es 0.5, el estimador insesgado de θ (usando $\bar{x}$) es $\bar{x} - 0.5$

Ejemplo 6:

Sea $x_1, \cdots, x_n$ una muestra aleatoria extraída de una población $N(\mu, 1)$. Probar que la variable $T = \bar{x}^2 - \frac{1}{n}$ es un estimador insesgado de μ^2.

Solución:

Hay que probar que $E(T) = \mu^2$.

En efecto: $E(T) = E\left(\bar{x}^2 - \frac{1}{n}\right) = E(\bar{x}^2) - \frac{1}{n}$

Por definición se sabe que: $E(\bar{x}^2) - \frac{1}{n} = V(\bar{x}) + \left[E(\bar{x})\right]^2 - \frac{1}{n} = \frac{\sigma^2}{n} + \mu^2 - \frac{1}{n}$

Como $\sigma^2 = 1$, entonces: $E(T) = \frac{1}{n} + \mu^2 - \frac{1}{n} = \mu^2$

Por lo tanto, T es un estimador insesgado de μ^2.

Nota: $V(\bar{x}) = E\left(\bar{x}^2\right) - \left[E(\bar{x})\right]^2$

4. MÉTODOS DE OBTENCIÓN DE ESTIMADORES PUNTUALES

Existen diversos métodos que permiten obtener estimadores puntuales; entre ellos se tienen el método de momentos, de mínimos cuadrados, de máxima verosimilitud, etcétera. El método de máxima verosimilitud es el de mayor aplicación en estadística aplicada.

4.1 Método de máxima verosimilitud

El procedimiento se resume en dos pasos:

a. Sea $x_1, x_2, \cdots, x_n$ una muestra aleatoria de $f(X;\theta)$, determinar la función de verosimilitud, esto es:

$$L(\theta) = g(x_1, x_2, \cdots, x_n, \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

Nota: El $L(\theta)$ depende de θ y no de los valores $x_1, x_2, \cdots, x_n$ porque θ es desconocido.

b. Si $\hat{\theta}$ es el estimador de θ que hace máxima $L(\theta)$, se dice que $\hat{\theta}$ es el estimador máximo verosímil de θ, y se obtiene de la solución de: $\frac{\partial L(\theta)}{\partial \theta} = 0$, siempre que $L(\theta)$ sea derivable. En la mayoría de las situaciones resulta más fácil hallar el máximo del logaritmo neperiano de la función de verosimilitud, esto es, el estimador máximo verosímil es la solución de:

$$\frac{\partial \left[LnL(\theta) \right]}{\partial \theta} = 0$$

Ejemplo 7:

Sea X una variable aleatoria cuya función de probabilidad es:

$f(x,\theta) = \theta(1-\theta)^{x-1}$; para $x = 1, 2, 3, \cdots$ y $0 < \theta < 1$.

Se desea obtener el estimador máximo verosímil de θ.

Solución:

Si toma una muestra aleatoria $x_1, x_2, \cdots, x_n$ de $f(x;\theta)$. Se sabe que

$f(x_i,\theta) = \theta(1-\theta)^{x_i - 1}$; para $x_i = 1, 2, \ldots, \wedge \ 0 < \theta < 1$.

Entonces:

1. $L(\theta) = \prod_{i=1}^{n} f(x_i,\theta) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i - 1} = \theta^n (1-\theta)^{\left(\sum_{i=1}^{n} x_i \right) - n}$

Aplicando logaritmos se tiene: $Ln\left[L(\theta) \right] = nLn(\theta) + \left(\sum_{i=1}^{n} x_i - n \right) Ln(1-\theta)$

2. $\dfrac{\partial Ln\left[L(\theta)\right]}{\partial\theta}=\dfrac{n}{\theta}-\dfrac{\sum_{i=1}^{n}x_i-n}{1-\theta}=0 \quad \Rightarrow n(1-\theta)-\theta\left(\sum_{i=1}^{n}x_i-n\right)=0 \quad \Rightarrow\hat{\theta}=\dfrac{n}{\sum_{i=1}^{n}x_i}=\dfrac{1}{\bar{x}}$

Por consiguiente: $\hat{\theta}=\dfrac{1}{\bar{x}}$

Ejemplo 8:

Una variable aleatoria X tiene la siguiente función de densidad de probabilidad: $f(x)=(\alpha+1)x^{\alpha}$; $0<x<1,\alpha>0$. Obtenga el estimador máximo verosímil de α sobre la base de una muestra aleatoria de tamaño n y luego evalúe la estimación para los siguientes valores muestrales:

0.1, 0.8, 0.27, 0.35, 0.62, 0.55, 0.45, 0.25, 0.32 y 0.60.

Solución:

Como $f(x_i,\alpha)=(\alpha+1)x_i^{\alpha}$; $0<x_i<1,\ \alpha>0$ Entonces:

1. $L(\alpha)=\prod_{i=1}^{n}f(x_i,\alpha)=\prod_{i=1}^{n}(\alpha+1)x_i^{\alpha}=(\alpha+1)^n\prod_{i=1}^{n}x_i^{\alpha}$

Aplicando logaritmo se tiene: $Ln\left[L(\alpha)\right]=nLn(\alpha+1)+\alpha\sum_{i=1}^{n}Ln(x_i)$

2. $\dfrac{\partial Ln\left[L(\alpha)\right]}{\partial\alpha}=\dfrac{n}{\alpha+1}+\sum_{i=1}^{n}Ln(x_i)=0 \quad \Rightarrow\dfrac{n}{\alpha+1}=-\sum_{i=1}^{n}Ln(x_i) \quad \Rightarrow\hat{\alpha}=-\left(1+\dfrac{n}{\sum_{i=1}^{n}Ln(x_i)}\right)$

Para el caso particular de $n=10$ se tiene:

$$\hat{\alpha}=-\left(1+\frac{10}{\sum_{i=1}^{10}Ln(x_i)}\right)=-\left(1+\frac{10}{-8.6972}\right)=-(-0.14979)=0.14979$$

Por consiguiente $\hat{\alpha}=0.14979$ es el estimador máximo verosímil de α, esto también significa que $\hat{\alpha}=0.14979$ es el valor más probable de α que ha generado la muestra aleatoria.

5. Estimación por intervalos

La estimación puntual se realiza utilizando un único valor para estimar el parámetro correspondiente; pero, ¿qué tan precisa es la estimación? Indudablemente, no es posible contestar este tipo de preguntas con los conocimientos adquiridos hasta este punto; por ello, es necesario introducir el concepto de estimación por intervalos, en el cual la estimación está dada por un conjunto de valores.

Por ejemplo, suponga que un supermercado tiene información de las últimas 40 semanas acerca del número de unidades de un producto que vende semanalmente y desea conocer la demanda promedio para establecer un stock adecuado. Con esta información se calcula la media muestral en $\bar{x} = 50$; es decir, esta es una estimación puntual del valor de la demanda promedio verdadera o poblacional. Sobre la base de esta estimación no es posible afirmar que la demanda promedio verdadera no será menor de 30 o mayor de 80. Imposible saberlo, porque no se ha cuantificado el error en la estimación puntual; este error se mide por la variabilidad del correspondiente estimador puntual.

Ahora, si se supone que la desviación estándar muestral de la media muestral $\bar{x}$ es 20, por el teorema central del límite se puede afirmar que $\bar{x} \to N(\mu, 20^2)$. Además, si $\bar{x}$ se distribuye aproximadamente como Normal, entonces, la probabilidad de que $\bar{x}$ se encuentre dentro de 2 desviaciones estándar es aproximadamente 0.96. Por lo tanto:

$$P\left(|\bar{x} - \mu| \leq 40\right) = 0.96 \Rightarrow P\left(-40 \leq \bar{x} - \mu \leq 40\right) = 0.96 \ .$$

De lo cual se puede concluir que $\mu \in \bar{x} - 40, \bar{x} + 40$ con un 96% de confianza; lo que muestra que es posible que la demanda promedio poblacional sea tan pequeña como 10 o tan grande como 90, si $\sigma_{\bar{x}} = 20$ y $\bar{x} = 50$.

Pero si la desviación estándar de la media muestral es de 5 unidades, entonces se tendría: $\bar{x} - 40, \bar{x} + 40$ con un 96% de confianza; y se aprecia que ahora es casi imposible que la media sea tan pequeña como 10 o tan grande como 90. Esto indica que la variabilidad del estimador juega un papel muy importante en la estimación.

Para entender la estimación por intervalos, se sabe que $\bar{x}$ es una variable aleatoria, por lo tanto, el intervalo $\langle \bar{x} - 40, \bar{x} + 40 \rangle$ es un intervalo aleatorio y la confianza de que este intervalo contenga al verdadero valor de μ es de 96%. Es decir, si se extraen muestras de tamaño 40 de esta población y se reemplaza el valor de $\bar{x}$ en el intervalo $\langle \bar{x} - 40, \bar{x} + 40 \rangle$ entonces se espera que el 96% de estos intervalos contengan el valor de la media poblacional μ , que es desconocida.

La estimación por intervalos tiene las siguientes ventajas sobre la estimación puntual: precisión, dada por la amplitud del intervalo, y confiabilidad, expresada en términos de probabilidad.

- **Definición**. La estimación por intervalos consiste en encontrar, mediante una muestra aleatoria, dos valores *a* y *b* tales que: $\theta \in \langle a, b \rangle$ con una confianza del 96%.

  Donde θ es el parámetro por estimar y $(1-\alpha)*100\%$ se denomina nivel de confianza; *a* y *b* son los límites del intervalo de confianza que varían de una muestra a otra.

  Si *a* y *b* son funciones de las observaciones para muestras de tamaño *n*, para una determinada muestra asumen valores específicos.

- **Intervalo aleatorio**. Es un intervalo finito o infinito, donde por lo menos uno de sus extremos es una variable aleatoria.
  Ejemplos:
  $\langle -\infty, X \rangle$, donde X es una variable aleatoria.
  $\langle X, Y \rangle$, donde X e Y son variables aleatorias.
  $\langle Z, \infty \rangle$, donde Z es una variable aleatoria.

- **Intervalo de confianza**. Sea $x_1, x_2, \cdots, x_n$ una muestra aleatoria extraída de una población con función de densidad $f(X;\theta)$ y sean $L_1 = l_1(x_1, x_2, \cdots, x_n)$ y $L_2 = l_2(x_1, x_2, \cdots, x_n)$ dos estadísticas tales que: $L_1 < L_2$.
  Se dice que $I = \langle L_1, L_2 \rangle$ es un intervalo de confianza para el parámetro θ con coeficiente de confianza $(1-\alpha)*100\%$, si $\theta \in \langle L_1, L_2 \rangle$ con una confianza del $(1-\alpha)*100\%$.
- **Interpretación**: Si se obtuviesen 100 muestras de tamaño n de la misma población y si se calculase el intervalo para cada muestra, el $(1-\alpha)*100\%$ de estos intervalos contendrían el verdadero valor del parámetro θ.
- **Cantidad pivotal.** Sea $x_1, x_2, \cdots, x_n$ una muestra aleatoria extraída de una población con función de densidad $f(X,\theta)$. Sea $Q = q(x_1, x_2, \cdots, x_n)$ una función de los valores muestrales y del parámetro. Si la distribución de probabilidad de Q no depende de θ, entonces se dice que Q es una cantidad pivotal.
- **Método de la cantidad pivotal para construir intervalos de confianza**
  Sea $Q = q(x_1, x_2, \cdots, x_n)$ una cantidad pivotal con función de densidad $f(q)$.

  Si se supone que $(1-\alpha)$ es fijo y que existen números reales q_1 y q_2 tal que: $Q \in (q_1, q_2)$ con una confianza del $(1-\alpha)*100\%$, entonces se desean obtener estos valores.

  Si para los valores muestrales $x_1, x_2, \cdots, x_n$ se puede expresar o transformar $q_1 < Q < q_2$ en la forma: $L_1 = l_1(x_1, x_2, \cdots, x_n) < \theta < L_2 = l_2(x_1, x_2, \cdots, x_n)$ donde L_1 y L_2 son estadísticas, entonces: $I = \langle L_1, L_2 \rangle$ es un intervalo de confianza para el parámetro θ con una confianza del $(1-\alpha)*100\%$.

5.1 Intervalo de confianza para la media poblacional (μ)

Se desea estimar la media poblacional con una confianza del $(1-\alpha)*100\%$ para muestras de tamaño n, esto equivale a encontrar dos valores a y b tales que: $\mu \in \langle a, b \rangle$ con una confianza del $(1-\alpha)*100\%$. Como las distribuciones muestrales asociadas a μ son dos, se tiene que especificar ambos casos.

5.1.1 Cuando la varianza poblacional (σ^2) es conocida

Para determinar las expresiones de a y b se utilizará el procedimiento de la cantidad pivotal.

i) El estimador puntual de la media poblacional μ es $\bar{x}$.

ii) La distribución muestral del estimador es: $\bar{x} \to N\left(\mu, \frac{\sigma^2}{n}\right)$. Por lo tanto, se utiliza la cantidad pivotal: $Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$ para construir el intervalo. Se tiene:

$P(-Z_0 \le Z \le Z_0) = 1-\alpha$; reemplazando Z se obtiene $P\left(-Z_0 \le \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \le Z_0\right) = 1-\alpha$,

entonces: $\mu \in \langle \overline{x} - Z_0\sigma_{\overline{x}}, \overline{x} + Z_0\sigma_{\overline{x}} \rangle$ con una confianza del $(1-\alpha)*100\%$; como $Z_0 = Z_{(1-(\alpha/2))}$ (existe una área de $(1-\alpha/2)$ a la izquierda del punto) concluyéndose que $\langle \overline{x} - Z_0\sigma_{\overline{x}}, \overline{x} + Z_0\sigma_{\overline{x}} \rangle$; donde: $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$.

- Tamaño de muestra y error de estimación

  Se tiene que $\mu \in \langle \overline{x} - Z_0\sigma_{\overline{x}}, \overline{x} + Z_0\sigma_{\overline{x}} \rangle$ con una confianza del $(1-\alpha)*100\%$, conclusión que proviene de: $P\left(\left| \overline{x} - \mu \right| \leq Z_{(1-\alpha/2)} \dfrac{\sigma}{\sqrt{n}} \right) = 1-\alpha$. Trabajando con la expresión dentro del paréntesis se tiene: $\left| \overline{x} - \mu \right| \leq Z_{(1-\alpha/2)} \dfrac{\sigma}{\sqrt{n}}$. Si se define $E = \left| \overline{x} - \mu \right|$ error de estimación, entonces se tiene:

  $E \leq Z_{(1-\alpha/2)} \dfrac{\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} \leq Z_{(1-\alpha/2)} \dfrac{\sigma}{E} \Rightarrow n \leq \left(Z_{(1-\alpha/2)} \dfrac{\sigma}{E} \right)^2$ de donde:

  $n = \left(Z_{(1-\alpha/2)} \dfrac{\sigma}{E} \right)^2$

  Nota: Cuando el muestreo es sin reposición y la población es finita de tamaño N, el tamaño de muestra queda definido por: $n = \dfrac{Z^2_{(1-\alpha/2)}\sigma^2 N}{(N-1)E^2 + Z^2_{(1-\alpha/2)}\sigma^2}$

5.1.2 Cuando la varianza poblacional (σ^2) es desconocida

En este caso, si la variable X tiene una distribución Normal, se utiliza la distribución *t* de Student con (n – 1) grados de libertad, y se concluye que:

$\mu \in \left\langle \overline{x} - t_{(1-\alpha/2,n-1)} \dfrac{s}{\sqrt{n}}, \overline{x} + t_{(1-\alpha/2,n-1)} \dfrac{s}{\sqrt{n}} \right\rangle$ con una confianza del $(1-\alpha)*100\%$.

Ejemplo 9:

Un comerciante mayorista compra latas de conserva de atún de la marca A. Según la indicación de la etiqueta el peso aproximado promedio por lata es μ onzas. Se supone que la población de los pesos es Normal con $\sigma = 2$ onzas. Si de un envío reciente el comerciante escoge al azar 20 latas y encuentra que el peso promedio es de 18,5 onzas:

a. Determine el intervalo de confianza al 90% para el peso promedio de todas las latas de conserva (μ).
b. Si el comerciante no conoce la desviación estándar poblacional y encuentra que $s = 2.0$ onzas, construya el intervalo de confianza de μ al 90%.
c. ¿Cuánto debió ser el tamaño de muestra si al estimar a μ se quiere un error no superior a 0.98 con confianza del 95%? Use $\sigma = 2$ onzas.

Solución:

a. Se desea hallar los valores de a y b tal que: $\mu \in \langle a,b \rangle$ con una confianza del 90%. Aplicando la fórmula que utiliza la distribución Z (la varianza poblacional es conocida) se tiene:

$$I = \left\langle \overline{x} \pm Z_{0.95}\sigma_{\overline{x}} \right\rangle = \left\langle 18.5 \pm 1.645(2/\sqrt{20}) \right\rangle = \langle 17.76, 19.24 \rangle$$

Se puede decir que con 90% de confianza el peso promedio (μ) de las latas se encuentra entre 17.76 y 19.24 onzas.

b. Ahora, como α es desconocida, se usa la distribución t para construir el intervalo que está dado por:

$$I = \left\langle \overline{x} \pm t_{(0.95,19)}s_{\overline{x}} \right\rangle = \left\langle 18.5 \pm 1.72913(2/\sqrt{20}) \right\rangle = \langle 17.7267, 19.2733 \rangle$$

Con 90% de confianza el peso promedio (μ) de las latas se encuentra entre 17.73 y 19.27 onzas. Note que este intervalo tiene mayor amplitud que el del inciso (a), esto se debe a que se usó la distribución t que presenta una mayor dispersión que la Normal.

c. $$n = \left(\frac{Z_{0.975}(\sigma)}{E} \right)^2 = \left(\frac{(1.96)2}{0.98} \right)^2 = 16$$

Ejemplo 10:

Un fabricante de fibras sintéticas desea estimar la tensión de ruptura media de una fibra, para lo cual diseña un experimento en el que se observan las tensiones de ruptura, en libras, de 16 hilos del proceso seleccionado al azar. Las medidas de las tensiones son: 20.8, 20.6, 21, 20.9, 19.9, 20.2, 19.8, 19.6, 20.9, 21.1, 20.4, 20.6, 19.7, 19.6, 20.3 y 20.7. Si se supone que la tensión de ruptura de una fibra se encuentra modelada por una Normal con $\sigma = 0.45$ libras, construya un intervalo de confianza del 95% para el valor real de la tensión de ruptura promedio de la fibra.

Solución:

Para lograr este objetivo haciendo uso del Minitab se realiza el siguiente procedimiento:

1. Primero se ingresan los datos en una columna de la ventana <Worksheet> del Minitab. El nombre de la columna permitirá hacer referencia a esta variable (vector de datos).

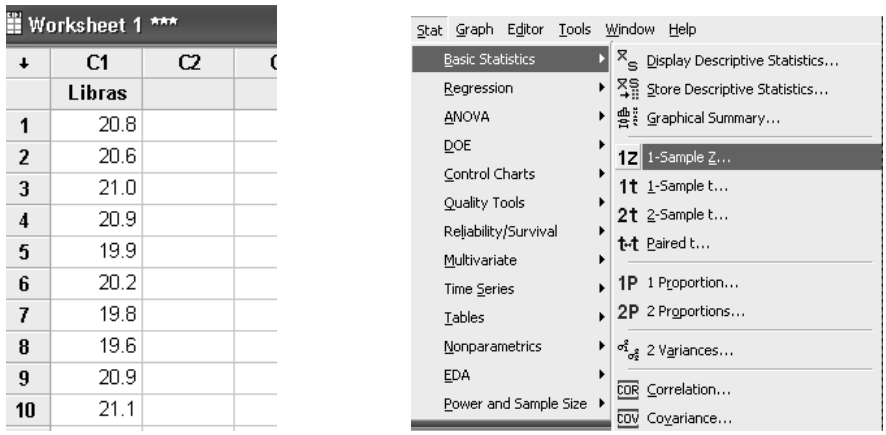


Figura 1. Ingreso de datos y opción "1 Sample Z".

2. Se seleccionan las opciones Stat/Basic Statistics/1 Sample Z … del menú de la ventana principal del Minitab. Debe recordarse que se usa una distribución Z para hacer la estimación de la media, ya que σ es conocida.
3. En la ventana se selecciona la variable (columna) que contiene la información requerida en el campo <Samples in columns>, y se ingresa el valor de σ en el campo de <Standard Deviation>.

Nota: El Minitab ofrece dos opciones para el ingreso de información cuando se desea construir el intervalo de confianza de la media. Si se cuenta con datos originales se emplea el campo <Samples in columns> y si se cuenta con información resumida, como el tamaño de muestra y media muestral, se emplea el campo <Summarized data>.

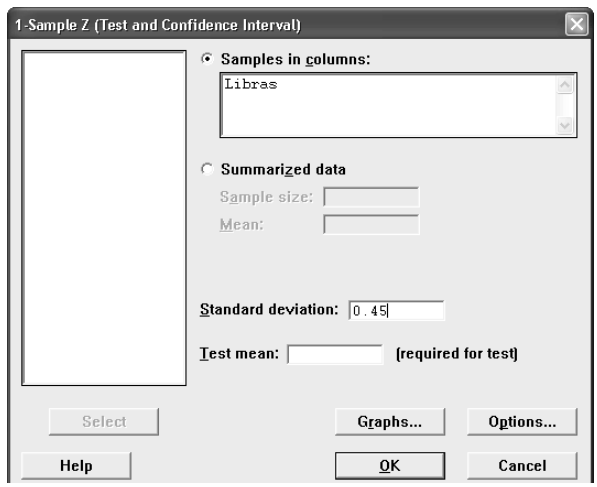


Figura 2. Ventana para el ingreso de datos.

4. Luego de seleccionar el botón de <Options> aparece la siguiente ventana.

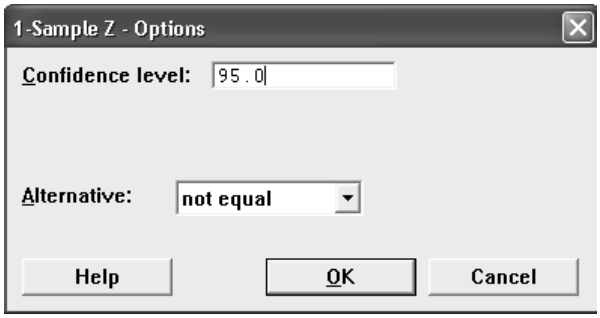


Figura 3. Ventana para la elección del nivel de confianza.

En el campo de <Confidence Level> se indica el nivel de confianza, que es del 95% en este caso.

5. Los resultados se muestran en la ventana <Session>, y son los siguientes:

```
One-Sample Z: Libras
The assumed standard deviation = 0.45
Variable  N   Mean    StDev  SE Mean      95% CI
Libras   16 20.3812  0.5231  0.1125 (20.1608, 20.6017)
```

Los límites confidenciales para la tensión de ruptura son 20.1608, 20.6017.

Esto significa que se tiene 95% de confianza de que el parámetro se encuentra en el intervalo 20.1608, 20.6017

5.2 Intervalo de confianza para proporción (π)

Si se desea estimar la proporción poblacional π con una confianza del $(1-\alpha)*100\%$ para muestras de tamaño $n \geq 30$; esto equivale a encontrar dos valores *a* y *b* tales que: $\pi \in \langle a,b \rangle$ con una confianza del $(1-\alpha)*100\%$.

Si n es un valor grande, se puede usar la distribución *Z*, concluyéndose que:

$$\pi \in \left\langle p - Z_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}}, p + Z_{(1-\alpha/2)}\sqrt{\frac{p(1-p)}{n}} \right\rangle$$ con una confianza del $(1-\alpha)*100\%$.

- Tamaño de muestra

a. Si la población es infinita o el muestreo es con reposición:

$$n = p(1-p)\frac{Z^2_{(1-\alpha/2)}}{E^2} .$$

Nota: Si no se conoce el valor de p, usar $p = 0.5$, en la fórmula anterior.

b. Si la población es finita de tamaño N y el muestreo es sin reposición:

$$n = \frac{Z^2_{(1-\alpha/2)}p(1-p)N}{(N-1)E^2 + Z^2_{(1-\alpha/2)}p(1-p)}$$

Ejemplo 11:

Se desea realizar un estudio de mercado para estimar la proporción poblacional de amas de casa que prefieren un producto de la competencia. Asimismo, se requiere que el error al estimar la proporción no sea mayor de 0.04 con un nivel de confianza del 95%. El departamento de ventas estima que cerca del 20% de las amas de casa podrían preferir el producto. Si cuesta S/.500 poner en marcha el estudio de mercado y S/.10 por entrevista, ¿cuál será el costo total de la encuesta?

Solución:

Sea C_T : Costo total del proyecto, entonces: $C_T = C_F + C_V = 500 + 10n$
Donde:

C_F : Costo fijo

C_V : Costo variable

n : Número de entrevistas por realizar

$$n = p(1-p)\frac{Z^2_{(0.975)}}{E^2} = 0.2(0.8)\left(\frac{1.96}{0.04}\right)^2 = 384.16 \approx 385$$

Por lo tanto, el costo total es $C_T = 500 + 10(385) = \text{S/. } 43.50$.

Ejemplo 12:

El gerente de producción de una planta ensambladora de artefactos eléctricos garantiza que el 95% de los artefactos que se producen están de acuerdo con las especificaciones estándares exigidas. Al examinar una muestra de 200 unidades de dichos artefactos se encontró que 25 son defectuosos. Si se pone en duda la afirmación del gerente de producción, ¿cuál será el intervalo de confianza del 98% para la proporción de artefactos defectuosos?

Solución:

Empleando el Minitab se tiene:

1. Elegir las opciones
2. Stat / Basic Statistics / 1 proportion de la ventana principal del Minitab, como se presenta en la figura 4.

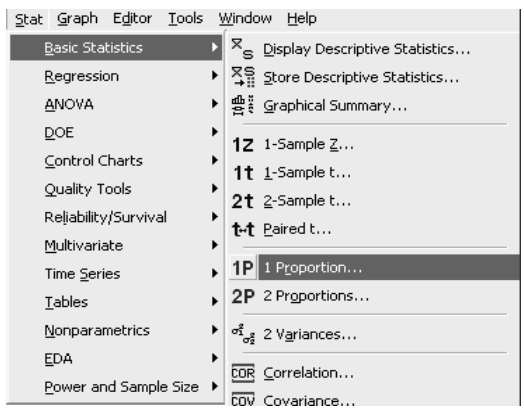


Figura 4. Selección de la Opción: 1 proportion.

3. En la ventana se activa <Sumarized data> se ingresa n (Number of

trials) y x (Number of events), luego de lo cual se debe pulsar el botón <Options> (véase la figura 5).

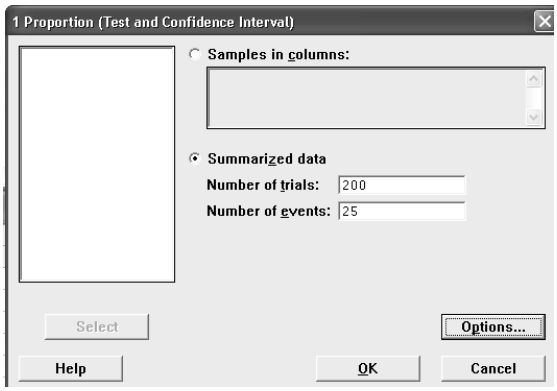


Figura 5. Ingreso de datos.

4. Aparece la siguiente ventana, que se presenta en la figura 6, en el primer campo se ingresa el nivel de confianza, en este caso 98%, y también se activa la última opción.

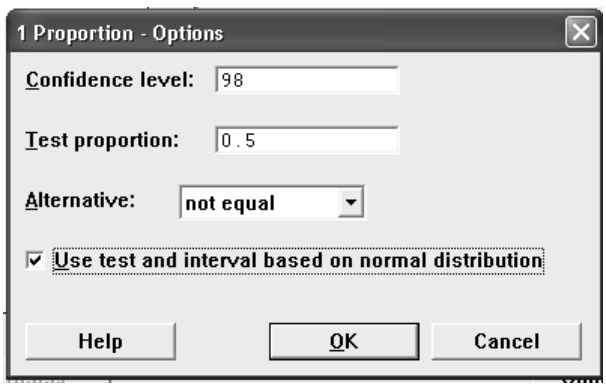


Figura 6. Elección del nivel de confianza.

5. Luego de seleccionar OK aparecerán los resultados siguientes:

Test and CI for One Proportion

```
Test of p = 0.5 vs p not = 0.5
Sample X   N  Sample p        98% CI        Z-Value  P-Value
1      25  200 0.125000 (0.070598, 0.179402) -10.61    0.000
```

El intervalo contiene valores del porcentaje de defectuosos que supera el 5% por lo tanto lo que afirma el gerente no es cierto.

Nota: Esta opción permite dos formas de ingresar la información. La pri-

mera fue usada en el ejemplo anterior; la segunda, para información original, se debe ingresar en una columna, luego usar la opción <Samples in columns>, como se muestra en la figura 5.

5.3 Intervalo de confianza para la varianza poblacional (σ^2)

Se desea estimar la varianza poblacional σ^2 con una confianza del $(1-\alpha)*100\%$ para muestras de tamaño n; esto equivale a encontrar dos valores a y b tales que: $\sigma^2 \in \langle a, b \rangle$ con una confianza del $(1-\alpha)*100\%$. En este caso se emplea la distribución Ji-cuadrado con (n–1) grados de libertad; concluyéndose que:

$$\sigma^2 \in \left\langle \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2,n-1)}}, \frac{(n-1)s^2}{\chi^2_{(\alpha/2,n-1)}} \right\rangle \text{ con una confianza del } (1-\alpha)100\%.$$

Ejemplo 13:

Según registros del departamento de calidad de una compañía, el peso de ciertos paquetes tiene una distribución Normal con $\mu = 40$ gramos y $\sigma = 0.25$ gramos. Una muestra aleatoria de 20 paquetes arrojó $s = 0.32$ gramos. ¿Con 95% de confianza se podría concluir que la variabilidad de los paquetes se ha incrementado? Justifique.

Solución:

Se debe construir un intervalo del 95% para σ^2.

En este caso: $\sigma^2 \in \left\langle \frac{(n-1)s^2}{\chi^2_{(0.975,19)}}, \frac{(n-1)s^2}{\chi^2_{(0.025,19)}} \right\rangle$ y reemplazando valores se obtiene

$$\left\langle \frac{19(0.32)^2}{32.8523}, \frac{19(0.32)^2}{8.90652} \right\rangle = \langle 0.05922, 0.21845 \rangle$$

Como $\sigma^2 = 0.25^2 = 0.0625$ se encuentra contenido en el intervalo hallado, entonces, con una confianza del 95% se puede afirmar que la variabilidad no se ha incrementado.

Empleando el Minitab:

1. Elija las opciones Stat/Basic Statistics/1 variance de la ventana principal del Minitab, como se presenta en la figura 7.
2. En la ventana que se aprecia en la figura 7 se activa <Sumarized data>, se ingresa el valor de n en <Sample size> y la desviación estándar en <Sample standard deviation>, luego de lo cual se debe pulsar el botón <Options>.
3. La figura 8 presenta la ventana de una varianza. En el primer campo se

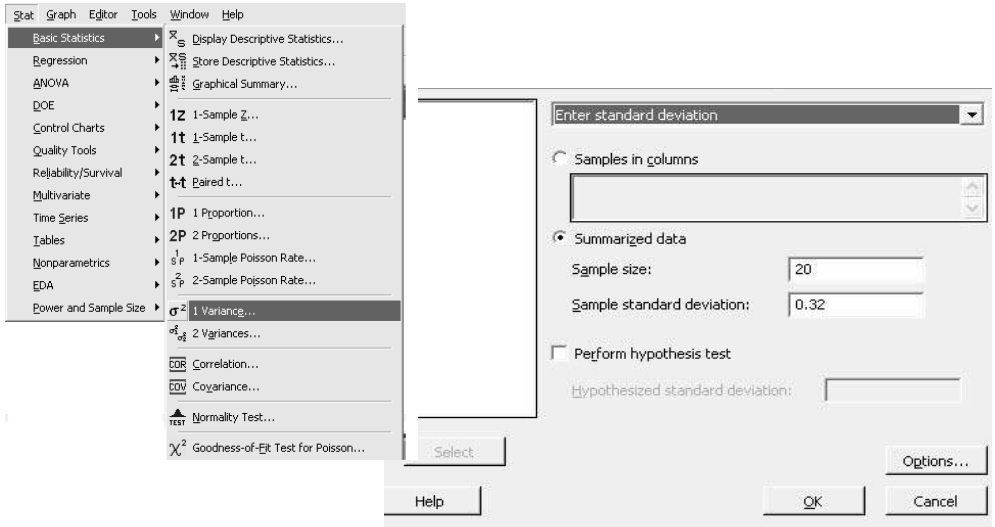


Figura 7. Selección de la opción: 1 variance, e ingreso de datos.

ingresa el nivel de confianza, en este caso 95%, tal como se muestra a continuación.

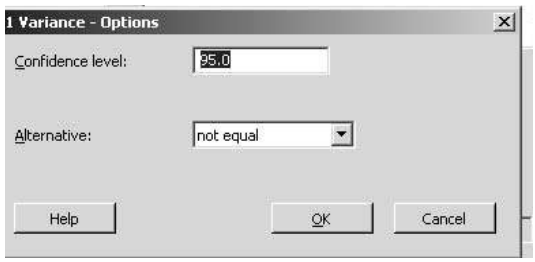


Figura 8. Selección del nivel de confianza.

4. Luego de seleccionar OK, aparecerán los siguientes resultados:

```
95% Confidence Intervals
Method          CI for StDev          Variance
Standard      (0.243, 0.467)      (0.059, 0.218)
```

Esto quiere decir que con un 95% de confianza, la varianza poblacional se encuentra contenida en el intervalo 0.059, 0.218.

5.4 Intervalo de confianza para la diferencia de proporciones ($\pi_1 - \pi_2$)

Si se desea estimar la diferencia de proporciones poblacionales $\pi_1 - \pi_2$, con una confianza del $(1-\alpha)100\%$ para muestras de tamaño $n_1 \geq 30$ y $n_2 \geq 30$; esto equivale a encontrar dos valores *a* y *b* tales que: $(\pi_1 - \pi_2) \in \langle a, b \rangle$ con una confianza del $(1-\alpha)100\%$.

Si $n_1 \to \infty$ y $n_2 \to \infty$, se puede usar la distribución *Z*; concluyéndose que:

$$(\pi_1-\pi_2)\in \left\langle (p_1-p_2)-Z_{(1-\alpha/2)}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}},(p_1-p_2)+Z_{(1-\alpha/2)}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}\right\rangle$$

con una confianza del $(1-\alpha)100\%$.

Nota: Si el intervalo contiene al valor cero, se concluye que las proporciones poblacionales podrían ser iguales.

Ejemplo 14

Una firma distribuye dos marcas de cerveza. En una reciente encuesta se encontró que 60 de 120 prefieren la marca A y 50 de 80 prefieren la marca B. Obtenga un intervalo para la diferencia de las proporciones poblacionales, con una confianza del 99%, para determinar si ambas marcas tienen la misma preferencia.

Solución:

Los datos son: $n_A=120;\ n_B=80;\ x_A=60;\ x_B=50;\ p_A=\frac{x_A}{n_A}=\frac{60}{120}=0.5$; $p_B=\frac{x_B}{n_B}=\frac{50}{80}=0.625$

El intervalo de confianza queda definido por:

$$\left\langle (p_A-p_B)\pm Z_{(0.995)}\sqrt{\frac{p_A(1-p_A)}{n_A}+\frac{p_B(1-p_B)}{n_B}}\right\rangle$$

Al reemplazar los valores se obtiene:

$$\left\langle (0.5-0.625)\pm Z_{(0.995)}\sqrt{\frac{0.5(1-0.5)}{120}+\frac{0.625(1-0.625)}{80}}\right\rangle = \langle -0.3073757,\ 0.0573757\rangle$$

Por lo tanto, como el intervalo contiene al cero, se concluye que ambas marcas podrían tener la misma preferencia.

El procedimiento usando el Minitab es:

1. Ingresar a la opción /Stat/ Basic Statistics / 2 proportions, como aparece en la figura 9.

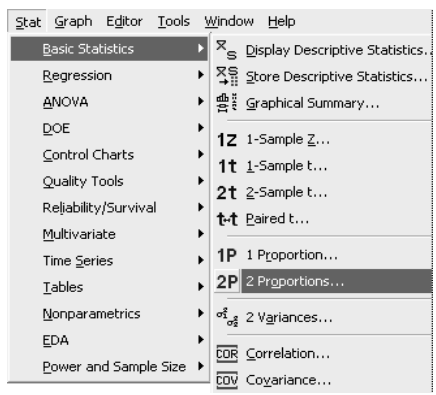


Figura 9. Procedimiento para dos proporciones.

2. En la figura 10 se observa la ventana para el ingreso de los datos. Luego

de activar el campo <Sumarized data> ingresar en First los datos para la marca A, en <Trials> se ingresa el tamaño de muestra (120), y en Events los casos a favor (60); en <Second> ingresar los datos de la marca B, similarmente en <Trials> ingresar el tamaño de muestra (80) y en Events los casos favorables (50); luego activar <Options>.

Nota: Esta opción permite tres formas de ingresar la información. La primera fue usada en este ejemplo, las otras dos dependen de que la información original se encuentre en una sola columna, para la cual se usa la opción <Samples in one column>, o si está en dos columnas, para lo cual se emplea <Samples in different columns>.

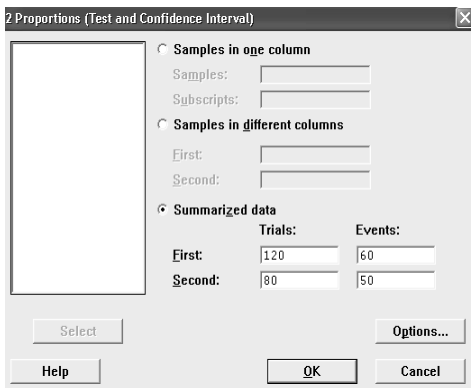


Figura 10. Ingreso de datos.

3. En la figura 11 aparece la ventana de dos proporciones, en ella se escoge el nivel de confianza, en este caso 99%, tal como muestra a continuación.

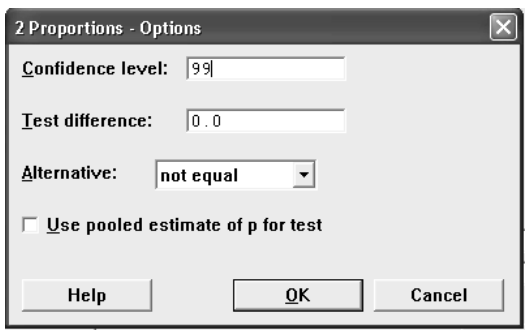


Figura 11. Elección del nivel de confianza.

4. Luego de dar OK aparecen los resultados siguientes:

```
Test and CI for Two Proportions
```

```
Sample    X     N   Sample p
1        60   120   0.500000
2        50    80   0.625000
Difference = p (1) - p (2)
Estimate for difference:  -0.125
99% CI for difference:  (-0.307376, 0.0573756)
Test for difference = 0 (vs not = 0):  Z = -1.77
P-Value = 0.077
```

En este caso, la respuesta es similar a la obtenida con la aplicación de las fórmulas.

5.5 Intervalo de confianza para una razón de varianzas poblacionales $\left(\sigma_1^2/\sigma_2^2\right)$

Se desea estimar la razón de varianzas σ_1^2/σ_2^2, con una confianza del $(1-\alpha)100\%$ para muestras de tamaño n_1 y n_2; es decir, encontrar dos valores a y b tales que: $(\sigma_1^2/\sigma_2^2) \in \langle a, b \rangle$ con una confianza del $(1-\alpha)100\%$. En este caso se usará la distribución *F*, concluyéndose que:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left\langle \frac{s_1^2}{s_2^2} F_{(\alpha/2, n_2-1, n_1-1)}, \frac{s_1^2}{s_2^2} F_{(1-\alpha/2, n_2-1, n_1-1)} \right\rangle \text{ con una confianza del } (1-\alpha)100\%.$$

Nota: Si el intervalo contiene al valor uno, se concluye que las varianzas podrían ser iguales, en caso contrario se dice que son diferentes.

Ejemplo 15:

En una empresa de fabricación de textiles se emplean dos máquinas: M1 y M2, que producen independientemente en forma Normal. Se han tomado dos muestras aleatorias, una de la Máquina 1 y otra de la Máquina 2, obteniéndose los siguientes tiempos de producción en minutos:

Máquinas	Tiempo de producción en minutos									
Máquina N.º 1	18	20	16	17	12	16	21	15	22	14
Máquina N.º 2	16	22	14	13	20	17	15	19		

Construir el intervalo de confianza del 94% para σ_1^2/σ_2^2.

Solución:

Sean X e Y las variables aleatorias que representan los tiempos en fabricar los productos de la Máquina 1 y la Máquina 2, respectivamente. Se supone que las distribuciones de X e Y son normales.
Se sabe que:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left\langle \frac{s_1^2}{s_2^2} F_{(\alpha/2, n_2-1, n_1-1)}, \frac{s_1^2}{s_2^2} F_{(1-\alpha/2, n_2-1, n_1-1)} \right\rangle$$

Se hallan las desviaciones estándar muestrales, haciendo uso del Minitab:

Descriptive Statistics: Maquina N°1, Maquina N°2

Variable	N	Mean	StDev	Variance
Máquina N°1	10	17.10	3.18	10.10
Máquina N°2	8	17.00	3.12	9.71

Reemplazando:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left\langle \frac{3.18^2}{3.12^2} F_{(0.03,8-1,10-1)}, \frac{3.18^2}{3.12^2} F_{(0.97,8-1,10-1)} \right\rangle$$

$$\frac{\sigma_1^2}{\sigma_2^2} \in \langle 1.038831 * 0.222209,\ 1.038831 * 3.94654 \rangle$$

$$\frac{\sigma_1^2}{\sigma_2^2} \in \langle 0.230838,\ 4.099790 \rangle$$

Es decir, el intervalo del cociente de varianzas poblacionales de los tiempos en producir los productos de la Máquina 1 y la Máquina 2 va entre 0.23 hasta 4.09, con un nivel de confianza del 94%.

5.6 Intervalo de confianza para la diferencia de medias $(\mu_1 - \mu_2)$

Si se desea estimar la diferencia de dos medias poblacionales $(\mu_1 - \mu_2)$ con una confianza del $(1-\alpha)100\%$ para muestras de tamaño n_1 y n_2, esto equivale a encontrar dos valores *a* y *b* tales que: $(\mu_1 - \mu_2) \in \langle a, b \rangle$ con una confianza del $(1-\alpha)100\%$.
Como son dos las distribuciones de muestreo asociadas a $(\mu_1 - \mu_2)$, a continuación se presentan las condiciones de su aplicación:

5.6.1 Varianzas poblacionales conocidas

Las poblaciones de donde provienen las variables pueden presentar o no una distribución Normal, pero sus varianzas poblacionales son conocidas. En este caso se emplea la distribución *Z*, concluyéndose que:

$$(\mu_1 - \mu_2) \in \left\langle (\bar{x}_1 - \bar{x}_2) - Z_{(1-\alpha/2)} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + Z_{(1-\alpha/2)} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\rangle$$

con una confianza del $(1-\alpha)100\%$

5.6.2 *Varianzas poblacionales desconocidas*

Las poblaciones de donde provienen las variables deben presentar una distribución Normal y sus varianzas poblacionales ser desconocidas. Se utiliza la distribución *t*; pero sus grados

de libertad y su fórmula varían, por lo que se presentan los siguientes casos:

a. Muestras independientes y varianzas poblacionales iguales.
Se deben encontrar los valores *a* y *b* tales que: $(\mu_1-\mu_2)\in\langle a,b\rangle$ con una confianza del $(1-\alpha)100\%$. Usando la distribución t con (n_1+n_2-2) grados de libertad, se concluye que:

$$(\mu_1-\mu_2)\in\left\langle(\bar{x}_1-\bar{x}_2)-t_{(1-\alpha/2,n_1+n_2-2)}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},(\bar{x}_1-\bar{x}_2)+t_{(1-\alpha/2,n_1+n_2-2)}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right\rangle$$

con una confianza del $(1-\alpha)100\%$.

donde : $s_p^2=\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$.

b. Muestras independientes y varianzas poblacionales diferentes.
Se deben encontrar los valores a y b tales que: $(\mu_1-\mu_2)\in\langle a,b\rangle$ con una confianza del $(1-\alpha)100\%$. Usando la distribución *t* con (ν) grados de libertad, se concluye que:

$$(\mu_1-\mu_2)\in\left\langle(\bar{x}_1-\bar{x}_2)-t_{(1-\alpha/2,v)}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},(\bar{x}_1-\bar{x}_2)+t_{(1-\alpha/2,v)}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right\rangle$$

con una confianza del $(1-\alpha)100\%$

donde $v=\dfrac{\left(\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1-1}+\dfrac{\left(s_2^2/n_2\right)^2}{n_2-1}}$

c. Muestras de datos pareados.
Se define la variable aleatoria $\mu_d=(\mu_1-\mu_2)$. Se deben buscar dos valores a y b tal que $\mu_d\in\langle a,b\rangle$ con una confianza del $(1-\alpha)100\%$. Utilizando la distribución t de Student con (n–1) grados de libertad y el procedimiento adecuado se concluye que:

$\mu_d\in\left\langle\bar{d}-t_{(1-\alpha/2,n-1)}\dfrac{s_d}{\sqrt{n}},\bar{d}+t_{(1-\alpha/2,n-1)}\dfrac{s_d}{\sqrt{n}}\right\rangle$ con una confianza del $(1-\alpha)100\%$.

Nota: El Minitab construye intervalos de confianza para $(\mu_1 - \mu_2)$ solamente empleando la distribución t.

Ejemplo 16:

Las siguientes son 16 determinaciones independientes de puntos de fusión en °C de un compuesto, los cuales siguen una distribución Normal; 8 fueron efectuadas por un inspector y 8 por otro.

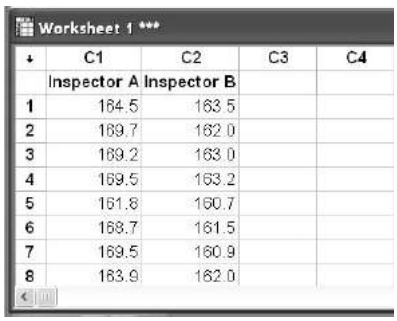
Inspector A	164.5	169.7	169.2	169.5	161.8	168.7	169.5	163.9
Inspector B	163.5	162.0	163.0	163.2	160.7	161.5	160.9	162.0

¿Se puede concluir, a partir de estos datos, que hay cierta tendencia de un inspector a obtener resultados más altos que el otro?
Nota: Suponga que las varianzas poblacionales son desconocidas pero iguales; emplee 95% de confianza.

Solución:

Empleando el Minitab:

1. Se ingresan los datos en dos columnas en la ventana Worksheet (véase la figura 12). El nombre de las columnas permite hacer referencia a las variables (vector de datos) tal como aparece en la figura.


Worksheet 1 ***

↓	C1	C2	C3	C4
	Inspector A	Inspector B		
1	164.5	163.5		
2	169.7	162.0		
3	169.2	163.0		
4	169.5	163.2		
5	161.8	160.7		
6	168.7	161.5		
7	169.5	160.9		
8	163.9	162.0		

Figura 12. Ingreso de datos.

Nota: El Minitab permite tres formas de ingresar la información. Si se tienen los datos originales de ambas muestras en una sola columna se usa la opción <Samples in one column>. Si se tienen los datos en dos columnas diferentes se usa la opción <Samples in different columns>. Si la información está resumida (tamaño de ambas muestras, medias y varianzas muestrales) se usa el campo de <Summarized data>.

2. Se eligen las opciones Stat / Basic Statistics / 2 Sample t de la ventana principal del Minitab. Esta es la opción para muestras independientes y varianzas desconocidas. (véase la figura 13).

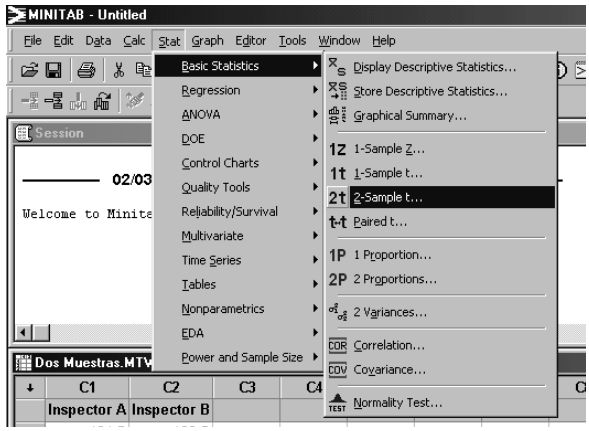


Figura 13. Selección de la opción: 2 Sample t.

3. En la ventana que aparece, ingrese las dos columnas que contiene la información requerida en el campo <Samples in different columns>; como por hipótesis las varianzas son iguales, marque el campo de <Assume equal variantes> (véase la figura 14).

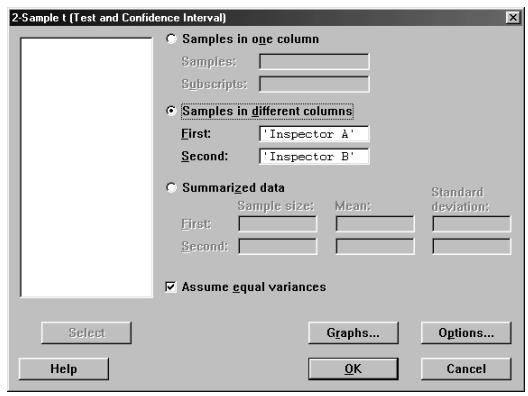


Figura 14. Ventana de la opción: 2 Sample t.

4. Se pulsa el botón <Options> y aparece la siguiente ventana (véase la figura 15), donde en el campo <Confidence level> se especifica el nivel de confianza deseado; por defecto indica un nivel de confianza del 95%.

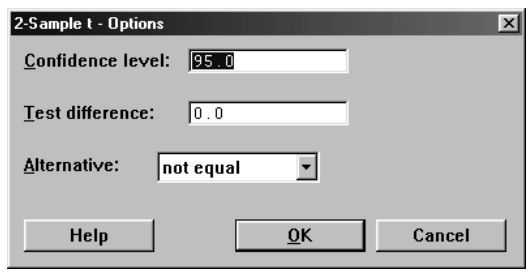


Figura 15. Elección del nivel de confianza.

5. Los resultados se muestran en la ventana <Session> y son los siguientes:

```
Two-Sample T-Test and CI: Inspector A, Inspector B
Two-sample T for Inspector A vs Inspector B
                 N        Mean      StDev     SE Mean
Inspector A      8      167.10       3.17         1.1
Inspector B      8      162.10       1.05        0.37
Difference = mu (Inspector A) - mu (Inspector B)
Estimate for difference:  5.00000
95% CI for difference:  ( 2.46711,  7.53289 )
T-Test of difference = 0 (vs not =):  T-Value = 4.23
P-Value = 0.001     DF = 14
Both use Pooled StDev = 2.3619
```

Como los límites del intervalo son ambos positivos, se puede concluir que el inspector A tiene cierta tendencia a obtener resultados mayores que el inspector B.

Ejemplo 17:

La compañía ABC envasa su producto en frascos de 400 gramos. En la etiqueta de cada frasco se afirma que su contenido es, en promedio, 50% de maní y 50% de nueces. Un investigador examina el contenido de 5 frascos, obteniendo los siguientes porcentajes de maní: 51%, 47%, 45%, 43% y 47%. Usando los datos recabados, construya un intervalo de 96% para la diferencia de promedios reales de los contenidos de maní y nuez; de acuerdo con el intervalo construido señale si es cierto o no lo que afirma la compañía.

Solución:

Como el mismo frasco proporciona los porcentajes de maní y nueces, se tiene un caso de datos pareados; suponiendo normalidad en los contenidos de nuez y maní se usa la estadística t para datos pareados. A continuación se determinan, primero, los contenidos de maní y nuez y se obtienen los siguientes valores:

Contenido de maní: 204, 188, 180, 172, 188.
Contenido de nuez: 196, 212, 220, 228, 212.

Sea:

μ_M : Contenido medio de maní de toda la producción.

μ_N : Contenido medio de nuez de toda la producción. Se desea que:

$\mu_d \in \langle a,b \rangle$ con una confianza del 96%, donde: $\mu_d = \mu_M - \mu_N$.

El procedimiento con el Minitab es el siguiente:

1. Ingrese los datos en la ventana <Data>. Tal como se muestra en la figura 16.

↓	C1	C2
	Mani	Nuez
1	204	196
2	188	212
3	180	220
4	172	228
5	188	212

Figura 16. Ingreso de datos.

2. Ingrese a las opciones Stat / Basic Statistics / Paired t de la ventana principal del Minitab como se presenta en la figura 17.

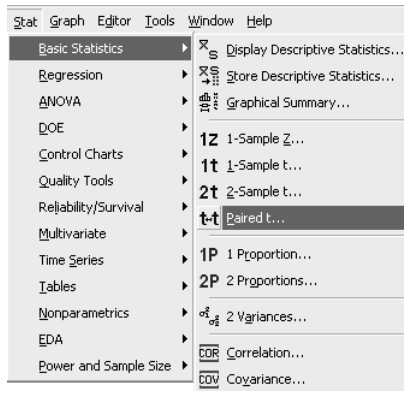


Figura 17. Opción Paired t.

3. Aparece la ventana <Paired t>; en el campo <Samples in columns> ingrese las dos columnas que contienen los datos (véase la figura 18).

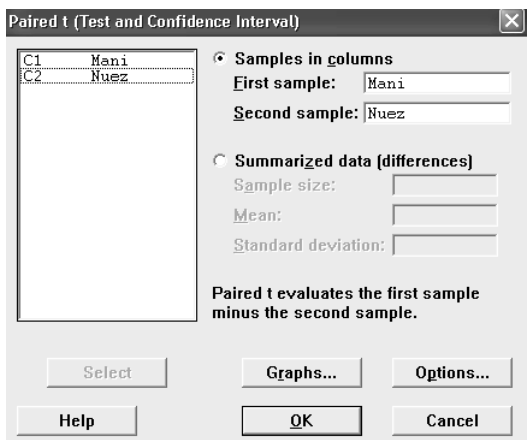


Figura 18. Ingreso de valores.

4. Pulsar el botón <Ok> para obtener los resultados siguientes:

```
Paired T-Test and CI: Maní, Nuez
Paired T for Maní - Nuez
              N         Mean       StDev      SE Mean
Maní          5      186.400      11.866        5.307
Nuez          5      213.600      11.866        5.307
Difference  5  -27.2000  23.7318  10.6132
95% CI for mean difference: ( - 56.6670,  2.2670 )
T-Test of mean difference = 0 (vs not = 0):
T-Value = -2.56       P-Value = 0.062
```

Como el intervalo contiene al cero, se concluye que μ_d podría ser 0 y, por lo tanto, lo que afirma la compañía es verdad y los porcentajes de maní y nuez son iguales en los frascos.

Ejemplo 18:

Con la finalidad de probar dos métodos de enseñar a deletrear, 40 alumnos fueron asignados al azar en 2 secciones, usándose un método en cada sección. Al finalizar la enseñanza de ambas metodologías se procedió a una evaluación. Los siguientes son los registros de las dos evaluaciones:

Método A	Método B
10, 20, 25, 30, 33, 37, 41, 43, 46, 46	20, 27, 35, 40, 41, 50, 50, 54, 56, 57
48, 50, 51, 52, 54, 56, 57, 65, 73, 86	57, 60, 63, 64, 65, 67, 67, 73, 83. 95

a. Calcule los límites de confianza del 90%, por separado para ambos métodos; ingresando a Stat/Basic Statistics/1-sample t. Indique sus conclusiones.
b. Construya un intervalo de confianza del 95% para la diferencia de los rendimientos promedio de ambos métodos y diga si los métodos de enseñar a deletrear son igualmente efectivos.

Solución:

a. Empleando el Minitab se tiene:

```
Variable   N   Mean   StDev  SE Mean     90,0 % CI
Método A  20  46,15  17,78    3,98    (39.27 ; 53.03)
Variable   N   Mean   StDev  SE Mean     90.0 % CI
Método B  20  56,20  17,96    4,02   (49.26 ; 63.14).
```

Al comparar por separado ambos intervalos se concluye que los rendimientos de los estudiantes de la sección del método B fueron mejores.

b. Como no se sabe nada con respecto a las varianzas poblacionales, primero se construye un intervalo para la razón de varianzas con una confianza del 95%:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left\langle \frac{316.13}{2.51(322.56)}, \frac{316.13(2.51)}{322.56} \right\rangle \Rightarrow \frac{\sigma_1^2}{\sigma_2^2} \in \langle 0.39,\ 2.46 \rangle$$

Como el intervalo contiene el valor 1, se concluye que las varianzas poblacionales podrían ser iguales y se debe utilizar la siguiente expresión:

$$\left\langle (\bar{x}_1 - \bar{x}_2) \mp t_{\left(1-\frac{\alpha}{2},\, n_1+n_2-2\right)} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right\rangle$$

Empleando nuevamente el Minitab se obtiene el siguiente intervalo:

```
Two-Sample T-Test and CI: Metodo A, Metodo B
Two-sample T for Metodo A vs Metodo B
             N   Mean    StDev   SE Mean
Metodo A    20   46.2     17.8        4.0
Metodo B    20   56.2     18.0        4.0
Difference = mu (Metodo A) - mu (Metodo B)
Estimate for difference:  - 10.0500
95% CI for difference:  ( - 21.4903,  1.3903 )
```

Como el intervalo contiene el valor 0 se concluye que ambos métodos podrían ser igualmente efectivos con 95% de confianza.

PROBLEMAS RESUELTOS

1. Sean $x_1, x_2, \cdots, x_n$ una muestra aleatoria de tamaño n, obtenida de una población, donde: $E(X_i) = \mu$ y $V(X_i) = \sigma^2$. Se definen como estimadores de μ las siguientes expresiones:

$\hat{\theta}_1 = x_1; \quad \hat{\theta}_2 = \frac{1}{2}(x_1 + x_2); \quad \hat{\theta}_3 = \frac{1}{2}(x_1 + 2x_5); \quad \hat{\theta}_4 = \bar{x}$, ¿cuál estimador utilizaría y por qué?

Solución:

Para saber cuál es mejor se determinará el sesgo y la eficiencia de los estimadores en forma conjunta, para lo cual se analizará primero el sesgo, determinando el valor esperado de cada uno de ellos, en efecto:

$$E(\hat{\theta}_1) = E(x_1) = \mu$$

$$E(\hat{\theta}_2) = E\left[\frac{1}{2}(x_1 + x_2) \right] = \frac{1}{2}(\mu + \mu) = \mu$$

$$E(\hat{\theta}_3) = E\left[\frac{1}{2}(X_1 + 2X_5) \right] = \frac{1}{2}(\mu + 2\mu) = \frac{3}{2}\mu$$

$$E(\hat{\theta}_4) = E(\bar{x}) = \mu$$

Como la esperanza del estimador 3 es diferente de la media poblacional, se concluye que $\hat{\theta}_3$ es sesgado y todos los demás son insesgados.
Para analizar la eficiencia, se calcula la varianza de cada estimador; en efecto:

$$V(\hat{\theta}_1) = V(x_1) = \sigma^2$$

$$V(\hat{\theta}_2) = V\left[\frac{1}{2}(x_1 + x_2)\right] = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{\sigma^2}{2}$$

$$V(\hat{\theta}_3) = V\left[\frac{1}{2}(X_1 + 2X_5)\right] = \frac{1}{4}(\sigma^2 + 4\sigma^2) = \frac{5}{4}\sigma^2$$

$$V(\hat{\theta}_4) = V(\bar{x}) = \frac{1}{n}\sigma^2$$

Analizando los resultados se tiene que $\hat{\theta}_4$ es el mejor (insesgado y eficiente) estimador, pues posee mínima varianza si n es mayor que 2.

2. Sea x_1, x_2, x_3, x_4 una muestra aleatoria proveniente de una distribución uniforme en el intervalo $[\theta - 2, \theta + 2]$. Si se considera el siguiente estimador para el parámetro θ :

$\hat{\theta} = \frac{1}{4}\sum_{i=1}^{4} iX_i$; ¿es $\hat{\theta}$ un estimador insesgado de θ ?

Solución:

Se debe calcular $E(\hat{\theta})$. En efecto:

$$E(\hat{\theta}) = E\left(\frac{1}{4}\sum_{i=1}^{4} iX_i\right) = \frac{1}{4}\sum_{i=1}^{4} iE(X_i) = \frac{1}{4}\sum_{i=1}^{4} i\theta = \frac{\theta}{4}\sum_{i=1}^{4} i = \frac{\theta}{4}10 = \frac{5}{2}\theta \neq \theta \quad .$$

Por lo tanto, $\hat{\theta}$ es un estimador sesgado.

3. Complete los siguientes espacios en blanco:

a. Las propiedades de un estimador máximo verosímil son: _____

b. La función de verosimilitud de un parámetro θ se define como: _____

c. Las propiedades que debe tener un buen estimador puntual son: _____

Solución:

a. Son aproximadamente insesgados.
   - Poseen mínima varianza
   - Consistentes
   - Suficientes

b. La función de probabilidad conjunta de la muestra, es decir:

$$L(\underset{\sim}{\theta}) = f(X_1,\underset{\sim}{\theta})f(X_2,\underset{\sim}{\theta})\cdots f(X_n,\underset{\sim}{\theta}) = \prod_{i=1}^{n} f(X_i,\underset{\sim}{\theta})$$

c. Insesgamiento, consistencia, suficiencia y eficiencia.

4. La duración de cierto tipo de componente electrónico X, en horas, tiene distribución Gamma con parámetros $\alpha = 2$ y β. Al probarse tres de tales componentes de manera independiente, se registraron las siguientes duraciones de 120, 130 y 128 horas. Basándose en esta muestra, obtenga el estimador máximo verosímil de β.
Nota: Primero se obtiene el estimador máximo verosímil de β, usando una muestra aleatoria de tamaño n y luego se utiliza el resultado para la muestra dada.

Solución:

El estimador máximo verosímil de β es:

Si $x_i \sim G(2,\beta)$, entonces $f(x_i) = \frac{1}{\beta^2} x_i e^{\frac{-x_i}{\beta}}$; $x_i > 0, \quad \beta > 0$

1. $L(\beta) = \prod_{i=1}^{n} f(x_i,\beta) = \prod_{i=1}^{n} \frac{1}{\beta^2} x_i e^{\frac{-x_i}{\beta}} = \frac{1}{\beta^{2n}} e^{\frac{-\sum_{i=1}^{n} x_i}{\beta}} \prod_{i=1}^{n} x_i$

$$Ln\left[L(\beta)\right] = -2nLn(\beta) - \frac{\sum_{i=1}^{n} x_i}{\beta} + Ln\prod_{i=1}^{n} x_i$$

2. $\frac{\partial Ln\left[L(\beta)\right]}{\partial \beta} = -\frac{2n}{\beta} + \frac{\sum_{i=1}^{n} x_i}{\beta^2} = 0 \quad \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} x_i}{2n}$

Luego para $n = 3$, $\hat{\beta} = \frac{\sum_{i=1}^{3} x_i}{2(3)} = \frac{120+130+128}{6} = 63$ horas

5. Una fábrica trabaja con dos tipos de máquinas: A y B, de manera independiente. El costo semanal Y de reparación de las máquinas A tiene distribución Normal con media μ_1 y varianza σ^2. El costo semanal X de reparación de las máquinas del tipo B también tiene distribución Normal con media μ_2 y varianza $3\sigma^2$. También se conoce que el costo medio semanal para la fábrica es: $2\mu_1 + \mu_2$.

a. Si se extrae una muestra aleatoria $y_1, y_2, \cdots, y_n$ de costos semanales para las máquinas del tipo A y una muestra aleatoria $x_1, x_2, \cdots, x_m$ de costos semanales para las máquinas del tipo B. Obtenga una expresión para

el intervalo de confianza del 95% para $2\mu_1 + \mu_2$ (costo medio semanal de la fábrica).

b. Si $n = 20;\quad \bar{x}_1 = \$\ 250;\quad m = 25;\quad \bar{x}_2 = \$\ 280$; y además $\sigma^2 = 120$, hallar los valores de los límites en (a).

Solución:

a. Para obtener una expresión para el intervalo de confianza, se tiene:

i) Si el costo medio semanal de la fabrica esta dado por: $C = 2\mu_1 + \mu_2$; su estimador puntual es: $\hat{C} = 2\bar{x}_1 + \bar{x}_2$.

ii) La distribución muestral de este estimador es:

$\hat{C} \sim N\left(2\mu_1 + \mu_2, \sigma^2\left(\frac{4}{n} + \frac{3}{m}\right)\right)$, pues las poblaciones son normales e independientes, y las varianzas poblacionales son conocidas.

En efecto: $E(\hat{C}) = E\left(2\bar{x}_1 + \bar{x}_2\right) = 2\mu_1 + \mu_2$

$$V(\hat{C}) = V(2\bar{x}_1 + \bar{x}_2) = 4V\left(\bar{x}_1\right) + V\left(\bar{x}_2\right) = \frac{4\sigma^2}{n} + \frac{3\sigma^2}{m} = \sigma^2\left(\frac{4}{n} + \frac{3}{m}\right)$$

iii) Usando como cantidad pivotal $Z = \dfrac{\hat{C} - C}{\sigma_{\hat{C}}}$ para construir el intervalo.

Si se sabe que: $P\left(-Z_0 \le Z \le Z_0\right) = 1 - \alpha$; reemplazando Z se obtiene:

$$P\left(-Z_0 \le \frac{\hat{C} - C}{\sigma_{\hat{C}}} \le Z_0\right) = 1 - \alpha \ \Rightarrow \qquad C \in \left\langle \hat{C} - Z_0\sigma_{\hat{C}}, \hat{C} + Z_0\sigma_{\hat{C}} \right\rangle$$

con una confianza del $(1-\alpha)100\%$.

Entonces, para un nivel de confianza del 95%, se tiene:

$$2\mu_1 + \mu_2 \in \left\langle 2\bar{x}_1 + \bar{x}_2 - 1.96\sigma\sqrt{\frac{4}{n} + \frac{3}{m}}, 2\bar{x}_1 + \bar{x}_2 + 1.96\sigma\sqrt{\frac{4}{n} + \frac{3}{m}} \right\rangle$$

b. Si $n = 20;\quad \bar{x}_1 = \$\ 250;\quad m = 25;\quad \bar{x}_2 = \$\ 280;\quad \sigma^2 = 120.$ Entonces se concluye que:

$2\mu_1 + \mu_2 \in \langle 767.854,\ 792.1496 \rangle$ con una confianza del 95%.

6. Uno de los aspectos relevantes en toda organización empresarial es la de otorgar incentivos a sus colaboradores más destacados. En este sentido, Beauty S.A. premia a sus distribuidores a escala nacional con dinero, según el volumen de ventas alcanzado en un período dado. Se obtuvo una muestra compuesta por 50 distribuidores de la empresa, de los cuales 30 recibieron bonificaciones monetarias, en dólares, por el nivel de ventas alcanzado. La mencionada información se presenta a continuación:

60	54	76	80	77	93	77	81	84	64
97	94	88	79	70	76	73	80	88	69
71	68	65	81	69	68	70	75	93	95

a. Obtenga e interprete, en términos del enunciado, el intervalo de confianza del 95% para el monto promedio pagado a todos los distribuidores que recibieron esta bonificación extra.

b. Si Beauty S.A. tiene 1.200 distribuidores a escala nacional, con 95% de confianza, ¿entre qué valores se hallará el número de distribuidores de la empresa que reciben incentivos monetarios?

Solución:

a. Con ayuda del Minitab se obtienen los siguientes valores:

n	$\overline{x}$	s	$t_{(29,0.975)}$
30	77.1667	10.9011	2.04523

Empleando la fórmula:

$$\mu \in \left\langle \overline{x} \pm t_{\left(1-\frac{\alpha}{2},n-1\right)} \frac{s}{\sqrt{n}} \right\rangle = \left\langle 77.1667 \pm t_{(0.975,29)} \frac{10.9011}{\sqrt{30}} \right\rangle$$

Se concluye que: $\mu \in \langle 73.0961, 81.2372 \rangle$ con una confianza del 95%.

Entonces, basado en una muestra de 30 distribuidores se puede afirmar, con un 95% de confianza, que el monto promedio de las bonificaciones recibidas se encuentra entre 73.1 y 81.2 dólares.

b. Nuevamente, empleando Minitab se obtiene:

N	n	x	p	$Z_{(0.975)}$
1200	50	30	0.6	1.96

Primero se hallan los límites para la proporción de distribuidores que reciben incentivos, para ello se usa la fórmula: $\left\langle p \pm Z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\frac{p(1-p)}{n}} \right\rangle$

y al reemplazar valores se concluye que: $\pi \in \langle 0.46421, 0,73579 \rangle$ con una probabilidad del 95%.

Para obtener un intervalo con el 95% de confianza para el total (X) de distribuidores que recibieron incentivos se multiplican los límites del intervalo anterior por $N = 1200$, luego se concluye que:

$X \in \langle 557,883 \rangle$ con una confianza del 95%.

7. Gayoso S.A. comercializa bicicletas de cierta marca en dos colores: azul y rojo. El administrador sostiene que la preferencia de los clientes por alguno de estos colores se encuentra muy pareja.

a. En una muestra de 50 bicicletas vendidas, ¿qué valor máximo tomará la diferencia entre la proporción de bicicletas azules de la muestra y su valor real con una probabilidad de 0.96?

b. Si en una muestra de 180 bicicletas vendidas este año, se encuentra que 72 fueron para el color rojo. Con 95% de confianza, ¿estaría usted en condiciones de afirmar que actualmente hay mayor preferencia por el color azul que por el rojo? Justifique su respuesta.

c. Si para comprobar la afirmación del administrador se tomó una muestra de 400 bicicletas vendidas y para la estimación se empleó un error de 0.06, ¿qué nivel de confianza se empleó?

Solución:

a. Para obtener el valor de E se debe cumplir:

$$P\left(|p-\pi| \le E\right) = 0.96 \quad \Rightarrow P\left(\frac{-E}{\sqrt{\frac{0.5(0.5)}{50}}} \le Z \le \frac{E}{\sqrt{\frac{0.5(0.5)}{50}}} \right) = 0.96$$

$$\Rightarrow Z = \frac{E}{\sqrt{\frac{0.5(0.5)}{50}}} = 2.05375 \qquad \Rightarrow E = 0.14522$$

b. Con ayuda del Minitab se obtiene:

n	x	p	$Z_{(0.975)}$
180	108	0.6	1.96

Para hallar el intervalo de confianza para proporción de ventas del color azul se usa la expresión: $\left\langle p \pm Z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\frac{p(1-p)}{n}} \right\rangle$. Al reemplazar valores se concluye que:

$\pi \in \langle 0.5284,\ 0.6716 \rangle$ con una confianza del 95%.

Se observa que el intervalo incluye valores de π mayores de 0.5, por lo que con 95% de confianza, y sobre la base de los registros de 180 bicicletas vendidas, se concluye que el color azul tiene más preferencia que el color rojo.

c. Usando la fórmula del intervalo de confianza para la proporción se tiene:

$$E = \pm 0.06 = \pm Z_{(1-\alpha/2)}\sqrt{\frac{0.5(0.5)}{400}} \quad \Rightarrow Z_{(1-\alpha/2)} = 2.40$$

nivel de confianza $(1-\alpha)100\%$ se encuentra dado por:

$$1-\alpha = P(Z < 2.40) - P(Z < -2.40) = 0.991802 - 0.0081975 = 0.9836$$

Luego, el nivel de confianza es 98.36%.

8. SGC S.A., empresa consultora, se encarga de elaborar y administrar pruebas de aptitud laboral a trabajadores del sector empresarial como paso previo para obtener certificaciones de calidad ISO. Dada la experiencia de la empresa en estos tipos de test, la gerencia considera que una desviación estándar de 6 en las puntuaciones es un buen indicador de la homogeneidad del grupo. Al evaluar a 200 trabajadores de La Exclusiva elegidos al azar se encontró una puntuación media de 77 con una desviación estándar de 8.2.

a. Como encargado del análisis de la información proporcionada y empleando 96% de confianza, ¿qué podría usted afirmar respecto a la varianza poblacional?

b. Utilizando su conclusión del intervalo anterior, estime —con 97% de confianza— la puntuación media de todos los trabajadores evaluados en La Exclusiva. Interprete su resultado.

Solución:

a. Usando la siguiente expresión para el intervalo de confianza de la varianza poblacional: $\left\langle \frac{(n-1)s^2}{\chi^2_{\left(1-\frac{\alpha}{2},n-1\right)}}, \frac{(n-1)s^2}{\chi^2_{\left(\frac{\alpha}{2},n-1\right)}} \right\rangle$; entonces, con una confianza del 96%, se concluye que:

$$\sigma \in \left\langle \sqrt{\frac{199(8.2^2)}{242.084}}, \sqrt{\frac{199(8.2^2)}{160.203}} \right\rangle = \langle 7.4346,\ 9.1391 \rangle$$

De acuerdo con la indicación de la consultora, y con 96% de confianza, se puede afirmar que los trabajadores de La Exclusiva no tienen puntuaciones homogéneas debido a que el intervalo contiene valores mayores de 6.

b. Como no se puede comprobar que $\sigma = 6$, entonces se construirá el intervalo de confianza para la puntuación media, haciendo uso de la desviación estándar muestral y la distribución *t* de Student. Con ayuda del Minitab se obtiene:

n	$\overline{x}$	s	$t_{(199,0.985)}$
200	77	8.2	2.18576

Luego:

$$\left\langle \overline{x} \pm t_{\left(1-\frac{\alpha}{2},n-1\right)} \frac{s}{\sqrt{n}} \right\rangle = \left\langle 77 \pm t_{(0.975,199))} \frac{8.2}{\sqrt{200}} \right\rangle$$

$$\mu \in \langle 75.7326,\ 78.2674 \rangle$$

Con 97% de confianza, la puntuación media de todos los trabajadores evaluados en La Exclusiva se encuentra entre 75.74 y 78.27 puntos.

9. La empresa Hobbits S.A. ha firmado un contrato de consultoría para analizar el comportamiento en el mercado nacional de cinco diferentes tipos de detergentes, comercializados en bolsas de aproximadamente un kilogramo. Luego de fijar los objetivos del estudio, el gerente de la empresa decidió seleccionar muestras aleatorias de los detergentes en diferentes lugares de comercialización. Las variables estudiadas fueron:

X1 = Marca del detergente.
X2 = Peso del detergente (en kilogramos).
X3 = Precio del detergente (en nuevos soles).
X4 = Lugar de comercialización.

Utilice los datos del archivo Detergente.MTW y $\alpha = 0.03$, para responder las siguientes preguntas:

a. ¿Cuál es el estimador del error estándar de la media de los precios del detergente Aze?

b. ¿Cuál es el estimador puntual de la varianza poblacional de los pesos del detergente White?

c. ¿Puede usted afirmar que el detergente Haryel es más caro que el detergente White?

d. ¿Puede usted afirmar que el detergente Clean se vende más en mercados que en *minimarkets*?

e. ¿Puede usted afirmar que el detergente Tyde es más uniforme en el peso que el detergente Haryel?

Solución:

a. Empleando el Minitab se tiene:

```
Variable      Marca      StDev       SE Mean
Precio        Aze        0.2594      0.0499
```

El estimador es: $s_{\overline{x}} = \dfrac{0.2594}{\sqrt{27}} = 0.0499$

b. De acuerdo con el reporte del Minitab:

```
Variable  Marca   N   Mean    Median  TrMean   StDev
Peso      White  11   0.9955  1.0000  0.9956   0.0463
```

Entonces: $s^2 = 0.00214369$

c. El reporte de Minitab muestra:

Two-sample T for Precio_Haryel vs Precio_White

```
               N      Mean       StDev       SE Mean
Precio_H       32     5.053      0.239        0.042
Precio_W       11     5.035      0.213        0.064
Estimate for difference:  0.0189
97% CI for difference: (-0.1644, 0.2022)
T-Test of difference = 0 (vs not =): T-Value = 0.23
P-Value = 0.818  DF = 41
Both use Pooled StDev = 0.233
```

Conclusión: No se puede afirmar, con 97% de confianza, que el detergente Haryel es más caro que White.

d. El Minitab proporciona los siguientes resultados

Test and CI for Two Proportions

```
 Sample        X       N       Sample p
    1          8       38       0.210526
    2          5       25      0.200000
Estimate for p(1) - p(2):  0.0105263
97% CI for p(1) - p(2):  (-0.214723, 0.235775)
Test for p(1) - p(2) = 0 (vs not = 0):  Z = 0.10  P-Value = 0.919
```

Conclusión: No se puede afirmar, con 97% de confianza, que el detergente Clean se vende más en mercados que en *minimarkets*.

e. El Minitab proporciona:

```
             Tyde=1        Haryel =2
Variable  Marca   N    Mean    Median   TrMean   StDev
Peso      Haryel  32  1.0147   1.0200   1.0132   0.0408
           Tyde   11  1.0027   0.9800   1.0022   0.0559
```

Usando el Minitab para calcular las probabilidades, se concluye que:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \langle 1.87716864(0.3636), 1.87716864(3.8042) \rangle = \langle 0.68253852,\ 7.14112493 \rangle$$

donde: $\frac{s_1^2}{s_2^2} = 1.87716864$

Conclusión: No se puede afirmar, con 97% de confianza, que el detergente Tyde tenga un peso más uniforme que el detergente Haryel, ya que el cociente podría ser uno.

10. Suponga que usted seleccionó muestras aleatorias de deudores en el mes de setiembre del 2008 en Piura y Arequipa, y construyó la siguiente distribución de frecuencias:

Tipo de deuda	Piura	Arequipa
	N.º deudores	N.º deudores
Comercial y microempresarial	120	150
Agropecuario	25	65
Hipotecario	50	35
Consumo	205	100
Total	400	350

a. ¿Qué puede usted concluir con respecto al tipo de deuda agropecuaria en Piura versus el tipo de deuda agropecuaria en Arequipa con 98% de confianza?

b. ¿Qué puede usted concluir con respecto a los deudores de Piura con deuda de consumo versus los deudores de Piura con deuda agropecuaria, con 96% de confianza?

Solución:

a. Se tiene que:
p_1: Porcentaje de deudores agropecuarios en Piura.
p_2: Porcentaje de deudores agropecuarios en Arequipa.

$$p_1 = \frac{25}{400} = 0.0625;\ p_2 = \frac{65}{350} = 0.18571$$

De acuerdo con la expresión para el intervalo de la diferencia de proporciones, con una confianza del 98% se tiene:

$$(\pi_1 - \pi_2) \in \left\langle \left(\frac{25}{400} - \frac{65}{350} \right) \pm 2.326348 \sqrt{\frac{(0.0625)(1-0.0625)}{400} + \frac{(0.18571)(1-0.18571)}{350}} \right\rangle$$

Empleando el Minitab se tiene:

Test and CI for Two Proportions

```
Sample       X         N       Sample p
1           25       400       0.062500
2           65       350       0.185714
Difference = p (1) - p (2)
Estimate for difference:  -0.123214
98% CI for difference:(-0.179170, -0.0672583)
```

Conclusión: La proporción de deudores del tipo agropecuario en Arequipa es mayor que la proporción de deudores del tipo agropecuario en Piura, con 98% de confianza.

b. Se tiene que:

p_1: Porcentaje de deudores de Piura con deuda de consumo.

p_2: Porcentaje de deudores de Piura con deuda agropecuaria.

$$p_1 = \frac{205}{305} = 0.67213;\ p_2 = \frac{25}{90} = 0.27778$$

De acuerdo con la expresión para el intervalo de la diferencia de proporciones, con una confianza del 96%, se tiene:

$$(\pi_1 - \pi_2) \in \left\langle \left(\frac{205}{305} - \frac{25}{90}\right) \pm 2.053749 \sqrt{\frac{(0.67213)(1-0.67213)}{305} + \frac{(0.27778)(1-0.27778)}{90}} \right\rangle$$

Utilizando el Minitab:

Test and CI for Two Proportions

```
Sample     X       N       Sample p
1        205     305       0.672131
2         25      90       0.277778
Difference = p (1) - p (2)
Estimate for difference:  0.394353
96% CI for difference: (0.282776, 0.505931)
```

Conclusión: La proporción de deudores en consumo es mayor que la proporción de deudores en agropecuario en la ciudad de Piura, con 96% de confianza.

11. En un estudio realizado durante el último fin de semana para estudiar el comportamiento de los clientes de Mega Plaza y Plaza San Miguel se seleccionaron, al azar, muestras aleatorias de clientes de ambos centros comerciales, registrándose los montos (en nuevos soles) de consumo en alimentos. Los datos obtenidos fueron los siguientes:

Centro comercial	Consumo de alimentos (S/.)							
Mega plaza	40	45	35	45	60	80	45	65
Plaza San Miguel	35	68	62	78	65	60	20	

¿Qué puede concluir usted respecto a los promedios de los gastos en alimentos de los clientes de Mega Plaza y Plaza San Miguel, con 96% de confianza? Indique claramente las suposiciones necesarias para realizar su análisis.

Solución:

Para contestar a la pregunta se debe hallar a y b tal que: $\mu_1 - \mu_2 \in \langle a, b \rangle$ con una confianza del 96%. Como la fórmula por emplear es la distribución t para muestras independientes, no se conoce si las varianzas poblacionales son iguales o diferentes, entonces se debe determinar primero el intervalo de confianza para la razón de varianzas. El reporte del Minitab muestra:

$s_1^2 = 228.125$; $s_2^2 = 415.952381$ y usando la fórmula del intervalo para la razón de varianzas se tiene: $\left\langle \frac{228.125}{415.952} F_{(0.02;\,6,7)}, \frac{228.125}{415.952} F_{(0.98;\,6,7)} \right\rangle$

De donde se tiene que:

$\frac{\sigma_1^2}{\sigma_2^2} \in \langle 0.08773875,\ 3.05788332 \rangle$ con un 96% de confianza.

Conclusión. Las varianzas poblacionales de los gastos de consumo en alimentos podrían ser iguales en ambos centros comerciales, al 96% de confianza.

Por lo tanto, la fórmula por utilizar sería: $\left\langle (\bar{x}_1 - \bar{x}_2) - t_{(0.98,\,13)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\rangle$

Nuevamente, empleando el Minitab se obtiene:

```
Two-sample T for Mega Plaza vs San Miguel
             N      Mean     StDev     SE Mean
Mega Plaza   8      51.9      15.1         5.3
San Miguel   7      55.4      20.4         7.7
Estimate for difference:  -3.55357
96% CI for difference:  (- 24.50527, 17.39813)
Both use Pooled StDev = 17.7430
```

Conclusión: En ambos centros comerciales se gasta en promedio la misma cantidad, con 96% de confianza.

12. Ante la interrogante: ¿Cuántos hogares tienen computadoras en Lima metropolitana?, se procedió a seleccionar muestras aleatorias de hogares de los niveles alto y medio, obteniéndose la siguiente información:

Nivel	Tamaño de la muestra	Cantidad de hogares que tienen computadoras
Alto	400	300
Medio	750	350

Use $\alpha = 0.02$ en sus cálculos.

a. ¿Cuál es el valor del estimador puntual de la diferencia entre la proporción poblacional de hogares de niveles alto y medio que tienen computadoras?

b. Si el total de hogares de nivel medio de Lima metropolitana son 250.000, ¿cuántos hogares de nivel medio estima usted que tienen computadoras?

Solución:

a. Como: $p_1 = \frac{300}{400}$ y $p_2 = \frac{350}{750}$; entonces: $p_1 - p_2 = \frac{300}{400} - \frac{350}{750} = 0.2833333$

b. Primero se construye el intervalo de confianza para la proporción con ayuda del Minitab:

Test and CI for One Proportion

```
Test of p = 0.5 vs p not = 0.5
Sample  X    N   Sample p         98% CI         P-Value
1       350  750  0.466667  (0.423894, 0.509797)  0.074
```

Multiplicando este intervalo por 250 000: 0.423894*250 000, 0.509797* 250 000 = 105 973, 127 449

Por lo tanto, el número de hogares de nivel medio en Lima metropolitana que tienen computadoras se encuentra entre 105.973 y 127.449, con 98% de confianza.

13. La siguiente noticia apareció en un diario de Lima:

> Envíos *courier* tardan más de dos días
>
> Mientras que en otros países de la región los trámites en Aduanas para el retiro de documentos demoran apenas seis horas, en nuestro país esta operación puede durar entre un día y medio y dos días, afirmó el gerente de Operaciones de DHL, Orlando Cevallos. Agregó que esta situación ha originado que la industria *courier* en nuestro país genere sobrecostos en las empresas, como la contratación de más personal e inversión en tecnología para agilizar el desaduanaje. "Existe un vacío legal que incluye el envío de documentos como si fueran mercancías de importación, por lo que Aduanas pide facturas comerciales de compra y venta como si se tratara de una importación", comentó.

En el archivo Aduanas.MTW se presentan los datos de un estudio realizado en octubre del 2008 en las diferentes aduanas del país. De estas se seleccionaron muestras aleatorias y las variables que se registraron fueron:

C1: Ubicación de la aduana.
C2: Tiempo de demora (en horas) de los trámites para el retiro de documentos.
C3: ¿Son documentos de importación?
C4: Empresa courier (DHL, UPS, Federal Express, Serpost, otros). Utilice $\alpha = 0.03$.

En cada caso presente la fórmula estadística con sus valores respectivos.

a. ¿Es el tiempo promedio de demora de los trámites para el retiro de documentos mayor que 35 horas en la aduana de Tacna?
b. ¿Puede afirmarse que DHL tiene una proporción de documentos de importación mayor que UPS?
c. Las autoridades de la Superintendencia de Aduanas del Perú han decidido repetir el estudio en octubre del 2009. En la Aduana de Iquitos, ¿qué tamaño de muestra recomendaría usted para estimar la proporción poblacional de documentos de Federal Express, con un margen de error de 3% y nivel de confianza del 98%? Utilice el archivo Aduanas.MTW para determinar el valor de p correspondiente.
d. ¿Es el tiempo de demora promedio de los trámites para el retiro de documentos de la aduana de Tumbes menor que el tiempo de demora

promedio de los trámites para el retiro de documentos de la aduana de Pucallpa?

Solución:

a. Se debe construir un intervalo de confianza para μ y la fórmula por utilizar es:

$\left\langle \bar{x} \pm t_{(1-\alpha/2;n-1)} \frac{s}{\sqrt{n}} \right\rangle$; reemplazando valores con ayuda del Minitab se tiene:

$$\left\langle 23.7538 \pm 2.19436 \frac{5.7628}{\sqrt{130}} \right\rangle = \langle 22.6447,\ 24.863 \rangle$$

Interpretación: No es mayor que 35 horas. El tiempo promedio se encuentra en el intervalo 22.6447, 24. 863, con 97% de confianza.

b. Se debe determinar el intervalo de confianza del 97% para la diferencia de proporciones. Definiendo:

DHL = 1 UPS = 2; entonces: $p_1 = \frac{62}{180} = 0.34444;\ p_2 = \frac{67}{179} = 0.37430$

La fórmula por utilizar es:

$\left\langle p_1 - p_2 \pm Z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right\rangle$ y reemplazando valores se obtiene:

$$\left\langle \left(\frac{62}{180} - \frac{67}{179} \right) \pm 2.17009 \sqrt{\frac{0.34444(1-0.34444)}{180} + \frac{0.37430(1-0.37430)}{179}} \right\rangle$$

$$\Rightarrow \pi_1 - \pi_2 \in \langle -0.139717,\ 0.080002 \rangle$$

Interpretación: Las proporciones de documentos de importación de DHL y UPS son iguales, con 97% de confianza.

c. La fórmula por usar es: $n = \left(\frac{Z}{E} \right)^2 p(1-p)$

$$n = \left(\frac{2.32635}{0.03} \right)^2 \frac{6}{69} \left(1 - \frac{6}{69} \right) = 477.420671 \approx 478$$

d. Para contestar la pregunta se deben hallar dos valores a y b tal que: $\mu_1 - \mu_2 \in \langle a, b \rangle$, con una confianza del 97%. Como se emplea la distribución t para muestras independientes y no se sabe si las varianzas poblacionales son iguales o diferentes, primero se debe determinar el intervalo de confianza para la razón de varianzas. Se sabe que:

$\frac{\sigma_1^2}{\sigma_2^2} \in \left\langle \left(\frac{s_1^2}{s_2^2} \right) F_{\left(\frac{\alpha}{2}, n_2-1, n_1-1\right)}, \left(\frac{s_1^2}{s_2^2} \right) F_{\left(1-\frac{\alpha}{2}, n_2-1, n_1-1\right)} \right\rangle$ y con ayuda del Minitab se tiene lo siguiente:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left\langle \left(\frac{25.147}{29.326} \right) 0.516501, \left(\frac{25.147}{29.326} \right) 1.87383 \right\rangle$$

$\frac{\sigma_1^2}{\sigma_1^2} \in \langle 0.44289881,\ 1.60680635 \rangle$, entonces, las varianzas podrían ser iguales.

Por lo tanto, las fórmulas por usar son: $\left\langle (\bar{x}_1 - \bar{x}_2) \pm t_{\left(1-\frac{\alpha}{2}, n_1+n_2-2\right)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\rangle$.

Reemplazando valores se tiene: $(22.82 - 23.22) \pm 2.20282(5.1854)\sqrt{\frac{1}{57} + \frac{1}{41}}$

El intervalo de confianza es: –2.734029, 1.944127.
Por lo tanto, los tiempos promedio de demora de las Aduanas de Tumbes y Pucallpa son iguales, con 97% de confianza.

14. El señor Ángeles es un empresario que está estudiando la posibilidad de invertir en la comercialización de café instantáneo para la próxima temporada de invierno. Con este fin, decide utilizar técnicas estadísticas que le permitan conocer el perfil de los consumidores de café. Una muestra aleatoria de los consumidores de café de diferentes centros comerciales, mercados y bodegas proporcionó la siguiente información:

C1 = Edad.
C2 = Género.
C3 = Marca de café instantáneo que consume regularmente.
C4 = Gasto mensual (en nuevos soles) en café instantáneo.

Los datos fueron procesados en el Minitab, obteniéndose los resultados que aparecen en el anexo siguiente.

Suponga que el señor Ángeles le plantea a usted las siguientes preguntas:

1. ¿Entre qué valores se encontrará el gasto promedio poblacional en café instantáneo de la marca Kores? Utilice 97% de confianza.
   a) Fórmula estadística con sus valores respectivos.
   b) Interpretación.

2.- ¿Cuál es el intervalo de confianza para la proporción poblacional del mercado perteneciente al café instantáneo Ragel? Utilice 96% de confianza.
   a) Fórmula estadística con sus valores respectivos.
   b) Interpretación.

3.- ¿Es la desviación estándar poblacional de los gastos de la marca Buin mayor que 3 nuevos soles? Utilice 98% de confianza.
   a) Fórmula estadística con sus valores respectivos.
   b) Interpretación.

4.- ¿Será el promedio poblacional de la edad de los consumidores de Mug (1) mayor que el promedio poblacional de la edad de los consumidores de Rangel (2)? Utilice 97% de confianza.
   a) Fórmula estadística con sus valores respectivos.
   b) Interpretación.

5.- ¿Puede afirmarse que la proporción poblacional de hombres que consumen el café de la marca Buin (1) es mayor que la proporción poblacional de hombres que consumen el café de la marca Mug(2)? Utilice 96.4% de confianza.

a) Fórmula estadística con sus valores respectivos.
b) Interpretación.

Solución:

1. a) La fórmula estadística con sus valores respectivos son:

$$\left\langle \bar{x} \pm t_{(0.985,193)} s_{\bar{x}} \right\rangle = \left\langle 30.072 \pm 2.18625(0.177) \right\rangle = \left\langle 30.072 \pm 0.374997 \right\rangle$$
$$= \left\langle 2929.6848,\ 30.4592 \right\rangle$$

b) **Interpretación**. El gasto promedio de café instantáneo de la marca Kores se encuentra entre 29.097 y 30.447 nuevos soles, con una confianza del 97%.

2. a) La fórmula estadística con sus valores respectivos es:

$$\left\langle p \pm Z_{(0.98)} \sqrt{\frac{p(1-p)}{n}} \right\rangle = \left\langle 0.17 \pm 2.05375 \sqrt{\frac{0.17(0.83)}{400}} \right\rangle$$
$$= \left\langle 0.17 \pm 0.0385728 \right\rangle = \left\langle 0.131427, 0.208573 \right\rangle$$

donde: $p = \dfrac{68}{400} = 0.17$

b) **Interpretación**. La proporción del mercado correspondiente al café instantáneo Ragel se encuentra entre 0.13 y 0.208, con una confianza del 96%.

3. a) La fórmula estadística con sus valores respectivos es:

$$\left\langle \sqrt{\frac{(n-1)s^2}{\chi^2_{(0.99,117)}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{(0.01,117)}}} \right\rangle = \left\langle \sqrt{\frac{117(6.212)}{155.496}}, \sqrt{\frac{117(6.212)}{84.3768}} \right\rangle =$$
$$\left\langle \sqrt{4.6741}, \sqrt{8.61379} \right\rangle = \left\langle 2.162,\ 2.935 \right\rangle$$

.

b) **Interpretación**. No, la desviación estándar es menor de 3, con 98% de confianza.

4. a) Se pide a y b tal que: $(\mu_1 - \mu_1) \in \langle a, b \rangle$, con una confianza del 97%. Primero se construye un intervalo de confianza para la razón de varianzas para determinar que fórmula se va a usar:

$$s_1^2 = 528.968; \quad s_2^2 = 278.235; \quad n_1 = 20; \quad n_2 = 68; \quad \bar{x}_1 = 37.7; \quad \bar{x}_2 = 39.0588$$

En efecto:

Por lo tanto, como el intervalo contiene a la unidad, las varianzas son iguales. Se debe utilizar la siguiente fórmula:

$$\left\langle (\bar{x}_1 - \bar{x}_2) \pm t_{(n_1+n_2-2,0.985)} s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \right\rangle$$ y al reemplazar valores se obtiene:

$$\left\langle (37.7 - 39.0588) \pm 2.20669(18.2655)\sqrt{\frac{1}{20}+\frac{1}{68}} \right\rangle = \langle -11.61166,\ 8.89406 \rangle$$

b) Interpretación. No es mayor, los promedios son iguales al 97% de confianza.

5. a) Los datos del problema son:

$$p_1 = \frac{51}{118} = 0.4322; \quad p_2 = \frac{10}{20} = 0.50$$

La fórmula por utilizar es: $$\left\langle (p_1 - p_2) \pm Z_{(0.982)} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right\rangle$$

y al reemplazar valores se obtiene:

$$\left\langle (0.4322 - 0.5) \pm 2.09693 \sqrt{\frac{0.4322(1-0.4322)}{118} + \frac{0.5(1-0.5)}{20}} \right\rangle$$

$$= \langle -0.32099,\ 0.1854 \rangle$$

b) Interpretación. No, porque las proporciones son iguales.

Anexo

Descriptive Statistics: Edad

Variable	Total Count	∑X Sum	∑X² Sum of Squares
Edad	400	15089,0	691687,0

Descriptive Statistics: Edad

Variable	Género	Total Count	∑X Sum	∑X² Sum of Squares
Edad	Hombre	180	6697,00	305153,00
	Mujer	220	8392,00	386534,00

Descriptive Statistics: Edad

Variable	Marca	Total Count	∑X Sum	∑X² Sum of Squares
Edad	Buin	118	4327,00	188713,00
	Kores	194	7352,00	342116,00
	Mug	20	754,00	38476,00
	Ragel	68	2656,00	122382,00

Descriptive Statistics: Gasto

Variable	Total Count	Mean	SE Mean	StDev	Variance
Gasto	400	30.023	0.123	2.455	6.027

Descriptive Statistics: Gasto

Variable	Género	Total Count	Mean	SE Mean	StDev	Variance
Gasto	Hombre	180	29.928	0.186	2.493	6.213
	Mujer	220	30.100	0.164	2.427	5.889

Descriptive Statistics: Gasto

Variable	Marca	Total Count	Mean	SE Mean	StDev	Variance
Gasto	Buin	118	30.042	0.229	2.492	6.212
	Kores	194	30.072	0.177	2.467	6.088
	Mug	20	29.650	0.437	1.954	3.818
	Ragel	68	29.956	0.306	2.524	6.371

Tabulated statistics: Género; Marca

Rows: Género Columns: Marca

	Buin	Kores	Mug	Ragel	All
Hombre	51	90	10	29	180
Mujer	67	104	10	39	220
All	118	194	20	68	400

15. Corporación Fénix S.A. es un consorcio de empresas de diversa índole, que en total tiene 5.000 trabajadores. Manuel Díaz, recientemente promovido a supervisor de personal, desea analizar algunos indicadores para tener una visión más precisa de la empresa y tomar decisiones adecuadas. Con tal fin, el señor Díaz solicitó a Juan Valdés su colaboración con el estudio. Valdés, estudiante de ingeniería, consideró conveniente seleccionar la ficha de 200 trabajadores y registró información de las siguientes variables:

X1: Número de años en la compañía.
X2: Sexo (1: mujer; 2: hombre).
X3: Número de horas extras trabajadas en los últimos 6 meses.
X4: Número de días de inasistencia en los últimos 6 meses.
X5: Nivel de instrucción (secundaria (0), universitaria incompleta (1), titulado (2) y posgrado (3).
X6: Salario anual (nuevos soles).

Utilizando la información que se encuentra en el archivo Corporación Fénix.MTW, colabore usted con Luis Valdés preparando respuestas para las siguientes preguntas:

a. ¿Es el número promedio de años en la corporación mayor de 12 años?
b. ¿Entre qué valores se encuentra el total de pagos anuales efectuados por la corporación a los trabajadores?
c. ¿Es la proporción de mujeres que laboran en la corporación menor que 40%?

d. Suponga que Manuel Díaz ha decidido repetir el estudio. ¿Qué tamaño de muestra recomendaría para estimar la proporción poblacional de trabajadores que son titulados, con un margen de error del 4% y una confianza del 93%? Utilice el archivo Corporación Fénix.MTW para determinar el valor de *p* correspondiente.
e. ¿Es la variabilidad del número de horas extras de los últimos 6 meses inferior a 50 h?
f. En promedio, ¿los varones tienen más tiempo de servicio que las mujeres?
g. ¿La proporción de hombres con 10 años o más de servicio en la corporación es superior a la proporción de mujeres con 10 o más años de servicio?

Donde no se indique usar un nivel de confianza del 98%.

Solución:

a. Hay que determinar *a* y *b* tal que: $\mu \in \langle a,b \rangle$ con una confianza del 98%

$n = 200; \quad \overline{x} = 12.64; \quad s = 7.688$

La fórmula por utilizar es: $\left\langle \overline{x} \pm t_{(0.99,199)} \frac{s}{\sqrt{n}} \right\rangle$, reemplazando: 11.3651, 13.9149

Falso, no es mayor de 12, con 98% de confianza.

b. Para responder a la pregunta se deben hallar de *a* y *b* tales que:

$N_\mu \in \langle a,b \rangle$, con una confianza del 98%.

$n = 200; \quad \overline{x} = 34010; \quad s = 7554; \quad t_{(0.99,199)} = 2.34523$

Primero se determina el intervalo para

$\mu \in \left\langle \overline{x} \pm t_{(0.99,199)} \frac{s}{\sqrt{n}} \right\rangle = \langle 32757.3,\ 35262.7 \rangle$ y luego se multiplica por

$N = 5000$, de donde:

$\langle 32757.3(5000), 35262.7(5000) \rangle = \langle 163786500,\ 176313500 \rangle$

c. Hay que determinar *a* y *b* tal que: $\pi \in \langle a,b \rangle$ con una confianza del 98%.

$n = 200; \quad x = 81; \quad Z_{(0.99)} = 2.32635$

$$\left\langle p \pm Z_{(0.99)} \sqrt{\frac{p(1-p)}{n}} \right\rangle = \langle 0.324249, 0.485751 \rangle$$

Falso, no es menor, con 98% de confianza.

d. $$P = \frac{47}{200} = 0.235 \quad \Rightarrow n = p(1-p)\left(\frac{Z_{(0.965)}}{0.04}\right)^2 = 0.235(0.765)\left(\frac{1.81191}{0.04}\right)^2 = 369$$

f. Hay que determinar a y b tal que: $\sigma \in \langle a,b \rangle$, con una confianza del 98%.

Utilizando la fórmula: $\left\langle \frac{(n-1)s^2}{\chi^2_{\left(1-\frac{\alpha}{2},n-1\right)}}, \frac{(n-1)s^2}{\chi^2_{\left(\frac{\alpha}{2},n-1\right)}} \right\rangle$ y reemplazando valores se obtiene que $\sigma^2 \in \langle 56.64,\ 71.56 \rangle$

No, es mayor, con 98% de confianza.

g. Hay que determinar *a* y *b* tal que: $(\mu_1 - \mu_2) \in \langle a, b \rangle$ con una confianza del 98%.

$n_1 = 81; \quad \bar{x}_1 = 13; \quad s_1 = 7.353; \quad n_2 = 119; \quad \bar{x}_2 = 12.395; \quad s_2 = 9.927$

Primero se determina el intervalo para el cociente de varianzas, con una confianza del 98%.

$$\left\langle \left(\frac{7.353^2}{7.927^2}\right) F_{(0.01,118,80)}, \left(\frac{7.353^2}{7.927^2}\right) F_{(0.99,118,80)} \right\rangle = \langle 0.53837,\ 1.406273 \rangle$$

Entonces: $\sigma_1^2 = \sigma_2^2$. Por lo tanto se usa la siguiente fórmula:

$$\left\langle (\bar{x}_1 - \bar{x}_2) \pm t_{(0.99,198)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\rangle = \langle -1.996918,\ 3.206919 \rangle$$

Como el valor de cero pertenece al intervalo se concluye que $\mu_1 = \mu_2$ y que el tiempo promedio de servicio de hombres y mujeres es el mismo, con 98% de confianza.

g. Hay que determinar *a* y *b* tal que: $(\pi_1 - \pi_2) \in \langle a, b \rangle$, con una confianza del 98%.

$n_1 = 81;\ x_1 = 52;\ p_1 = \frac{52}{81} = 0.641975;$

$n_2 = 119;\ x_2 = 69;\ p_2 = \frac{69}{119} = 0.579832;$

La fórmula por utilizar es: $\left\langle (p_1 - p_2) \pm Z_{(0.99)} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right\rangle$

Al reemplazar valores se tiene: $(\pi_1 - \pi_2) \in \langle -0.100449,\ 0.224736 \rangle$

Entonces, $\pi_1 = \pi_2$ y la proporción de hombres con 10 años o más de servicio en la corporación es igual a la proporción de mujeres con 10 o más años de servicio, con 98% de confianza.

16. El Ministerio de Comercio Exterior y Turismo ha realizado recientemente un estudio dirigido a los vistantes provenientes de diferentes partes del mundo que han visitado las diferentes zonas turísticas del país: centro (costa y sierra) norte, nororiente, sur y suroriente. Para la realización del mencionado estudio se tomaron muestras independientes en seis zonas turísticas. Los encuestadores seleccionaban aleatoriamente un turista y se le solicitaba su cooperación para el llenado de un cuestionario. Las principales variables registradas fueron:

C1: Zona turística visitada
C2: Procedencia.
C3: Gasto estimado de estadía en la zona turística (en dólares)
C4: Opinión sobre la atención brindada al turista.

La información se presenta en el archivo Turismo.MTW. Recupere este archivo para contestar lo siguiente:

a. En la zona turística nororiental, ¿puede afirmarse que el gasto promedio de estadía de los turistas europeos es menor de $395, con 94% de confianza? Presente la fórmula estadística con sus valores respectivos. Interprete.
b. ¿Puede afirmarse que la proporción poblacional de turistas asiáticos que opinan que la atención al turista es buena es mayor del 24%, con 97% de confianza? Presente la fórmula estadística con sus valores respectivos. Interprete.
c. ¿Puede afirmarse que la proporción poblacional de turistas que opinan que la atención es mala en la zona turística centro (sierra) es menor que la proporción poblacional de turistas que opinan que la atención es mala en la zona turística nororiental, con 96% de confianza? Presente la fórmula estadística con sus valores respectivos. Interprete.
d. En relación a los turistas que opinan que la atención al turista es buena, ¿es el gasto promedio de estadía en la zona centro (costa) mayor que el gasto promedio de estadía en la zona sur con 95% de confianza? Presente la fórmula estadística con sus valores respectivos. Interprete.

Solución:

a. Hay que determinar *a* y *b* tal que: $\mu \in \langle a,b \rangle$, con una confianza del 94%.

$$n = 253; \quad \bar{x} = 392.71; \quad s = 83.68$$

$$\left\langle \bar{x} \pm t_{(0.97,253))} \frac{s}{\sqrt{n}} \right\rangle = \langle 382.771,\ 402.649 \rangle$$

Falso, no es menor de 395, con 94% de confianza.

b. Hay que determinar *a* y *b* tal que: $\pi \in \langle a,\ b \rangle$, con una confianza del 97%.

$$n = 252; \quad x = 78; \quad Z_{(0.985)} = 2.17009$$

$$\left\langle p \pm Z_{(0.985)} \sqrt{\frac{p(1-p)}{n}} \right\rangle = \langle 0.246326,\ 0.372721 \rangle$$

Sí, es mayor, con 97% de confianza.

c. Hay que determinar *a* y *b* tal que: $(\pi_1 - \pi_2) \in \langle a,\ b \rangle$, con una confianza del 96%.

$$n_1 = 121;\ x_1 = 27;\ p_1 = \frac{27}{121} = 0.223140;\ n_2 = 253;\ x_2 = 48;\ p_2 = \frac{48}{253} = 0.189723;$$

$$\left\langle (p_1 - p_2) \pm Z_{(0.98)} \sqrt{\left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \right)} \right\rangle = \langle -0.059349,\ 0.126183 \rangle$$

Entonces, $\pi_1 = \pi_2$, con 96% de confianza.

d. Hay que determinar a y b tal que: $(\mu_1 - \mu_2) \in \langle a,\ b \rangle$ con una confianza del 95%.

Results for Atención al turista = Buena

Variable	Zona visitada	Count	Mean	SE Mean	StDev
Gasto estadía	Centro (costa)	34	433.6	24.0	139.7
	Centro (sierra)	38	413.1	17.6	108.4
	Nor oriental	69	395.6	9.9	82.3
	Norte	28	410.3	20.0	105.7
	Sur	52	426.6	11.8	85.0
	Sur oriental	55	398.1	10.4	77.0

Primero determinar a y b, tal que: $\frac{\sigma_1^2}{\sigma_2^2} \in \langle a, b \rangle$ con una confianza del 98%; entonces:

$$\left\langle \left(\frac{139.7^2}{85^2}\right) F_{(0.025,51,33)}, \left(\frac{139.7^2}{85^2}\right) F_{(0.975,51,33)} \right\rangle = \langle 1.47295,\ 5.17657 \rangle ;$$

Entonces: $\sigma_1^2 \neq \sigma_2^2$. Por lo tanto:

$$\left\langle (\bar{x}_1 - \bar{x}_2) \pm t_{(0.975,49)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right\rangle = \langle -46.65773,\ 60.65773 \rangle$$

Por lo tanto $\mu_1 > \mu_2$ es falso, con relación a los turistas que opinan que la atención es buena, los que van a la zona centro (costa) gastan en promedio igual que los que van a la zona sur, con 95% de confianza.

PROBLEMAS PROPUESTOS

1. Sea $\hat{\theta}$ cualquier estimador de un parámetro desconocido θ. Se define el error cuadrático medio de $\hat{\theta}$, como: $ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$

a. Demostrar que $ECM(\hat{\theta}) = V(\hat{\theta}) + (\theta - E(\hat{\theta}))^2$

b. Suponga que se tiene una muestra aleatoria de tamaño 2 extraída de una población exponencial X con parámetro θ. Si se proponen 2 estimadores para el parámetro θ, dados por las siguientes expresiones:

$\hat{\theta}_1 = \frac{x_1 + x_2}{2}$; y $\hat{\theta}_2 = \sqrt{x_1 x_2}$. Usando (a) demostrar que:

$ECM(\hat{\theta}_1) < ECM(\hat{\theta}_2)$.

2. Suponga la siguiente relación entre las variables X e Y, $Y = \beta X + e$, donde X no es una variable aleatoria y $e \sim N(0,\sigma^2)$ Se definen dos estimadores mínimos cuadráticos para el parámetro desconocido β.

$\beta^* = \dfrac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}$ y $\hat{\beta} = \dfrac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2}$. ¿Cuál de ellos es insesgado?

3. Se tienen dos muestras independientes de tamaños n_1 y n_2 obtenidos de una población con media μ y varianza σ^2. Sean $\bar{x}_1$ y $\bar{x}_2$; s_1^2 y s_2^2 los estimadores insesgados basados en ambas muestras. Demostrar que:

$\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$ y $s^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$ son estimadores insesgados de μ y σ^2, respectivamente; halle la varianza de $\bar{x}$.

4. Demuestre que: $\hat{p} = \dfrac{X + \sqrt{n}/2}{n + \sqrt{n}}$ es un estimador consistente de π, donde X es una variable aleatoria binomial.

5. Suponga que se observan n mediciones independientes de la vida útil de componentes que se comportan como una variable aleatoria de Weibull con ley de probabilidad:

$f(x) = \dfrac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^\alpha}$. Hallar el estimador máximo verosímil de β sabiendo que $\alpha = 2$.

6. Sean $x_1, x_1, \cdots, x_n$ y $y_1, y_1, \cdots, y_n$. 2 muestras aleatorias independientes, de poblaciones normales con medias μ_1 y μ_1 y varianzas $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Demuestre que:

$s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (y_i - \bar{y})^2}{2n-2}$, es un estimador consistente de σ^2.

7. Con ayuda de un aparato se obtienen n mediciones independientes con respecto a la duración de un tipo de componente electrónico. Suponiendo que X tiene distribución Normal y la varianza de la i-ésima medición es σ_i^2. Hallar el estimador máximo verosímil del parámetro de la variable aleatoria X. ¿Es insesgado este estimador?

8. Si X es el número de intentos o ensayos en el que se encuentra la primera pieza defectuosa en una serie de pruebas independientes de control de calidad, determine el estimador máximo verosímil de π, la probabilidad verdadera de observar una pieza defectuosa.

9. El tiempo de duración, en meses, de un componente electrónico es una variable aleatoria T que se comporta según una distribución exponencial con parámetro β. Para estimar β se prueban 30 componentes y se encuentra que 18 fallan antes de los 6 meses.

a. Utilizando el método de máxima verosimilitud, estimar el porcentaje de componentes que fallan antes de los 6 meses.
b. Estimar por máxima verosimilitud el parámetro β.

10. Complete los espacios en blanco según corresponda:

a. Una variable de tipo _____ debe ser usada para construir un intervalo de confianza para π con _____ de confianza.
b. En muestreo, un estrato está conformado por unidades que son _____
c. La longitud de un intervalo de confianza es el doble del _____
d. El intervalo de confianza para datos pareados debe usarse cuando las muestras son _____
e. El error estándar del mejor estimador puntual de una proporción poblacional es _____
f. Un estimador es _____

11. Responda las siguientes preguntas:

a. Señale dos razones que justifiquen el uso del muestreo.
b. ¿Qué significa la propiedad de insesgado?
c. ¿Cuál es la importancia del teorema del límite central?
d. ¿Qué es inferencia estadística?

12. Un analista de investigación de mercados quiere estimar el promedio del ingreso familiar mensual de una determinada población. Determine el intervalo de confianza (IC) del 95% si en una muestra aleatoria de tamaño 100 de esa población se encontró que el promedio del ingreso familiar era de S/.1500. Suponga que el ingreso se distribuye normalmente con desviación estándar igual a S/.300.

13. Si en el problema anterior la población consiste en 2000 ingresos familiares, construya un intervalo de confianza del 95% para la estimación del ingreso total.

14. Una casa comercial tiene 500 clientes con cuenta de crédito. Para estimar el total adeudado por estos clientes se selecciona una muestra aleatoria de 49 cuentas, la cual arroja $\bar{x} = S/.\,950$ y $s = \text{S/}.\,300$. Construya un intervalo del 98% para estimar la cantidad adeudada por todos los clientes.

15. Para la estimación del parámetro μ de una distribución Normal que tiene σ conocida se elabora un intervalo de confianza al 90%, determine el tamaño de la muestra para aumentar el nivel de confianza de dicho intervalo al 95% y obtenga la misma variabilidad.

16. La duración de cierta marca de pilas es una variable aleatoria cuya distribución se supone Normal. Inicialmente se estima que la duración media es de 500 horas y que el 95% duran entre 480.4 y 519.6 horas. Si se eligen 9 pilas al azar y se encuentra que la duración media es 480 horas, utilizando un intervalo de confianza del 95% para la media μ, ¿se debería concluir que la duración media es diferente de 500 horas?

17. Un fabricante de automóviles desea estimar el kilometraje medio por litro de gasolina que sus clientes obtendrán con su nuevo modelo. ¿Cuántos viajes de prueba debe efectuar a fin de que su estimación tenga una precisión de 0.2 km/l, con una confianza del 99%? Considere σ = 1.5.

18. Suponga que durante los períodos de gran demanda, el tiempo de espera (en horas) en un banco está distribuido de manera aproximadamente Normal, con una varianza de 2.25 horas.

a. Una muestra de 20 clientes revela un tiempo medio de espera de 1.52 horas, determine el intervalo de confianza del 95% para estimar la media de la población.
b. Suponga que la media de 1.52 horas ha resultado de una muestra de 32 clientes. Obtenga el intervalo de confianza del 95% para estimar la media poblacional.
c. ¿Qué efecto tiene un tamaño de muestra más grande sobre el intervalo citado? Explique.

19. Para discutir la conveniencia de aumentar sus instalaciones una empresa desea estimar la demanda que espera recibir. Para ello selecciona a 10 de sus clientes habituales, observando que el número de unidades adquiridas por ellos en el último semestre se distribuye de la forma siguiente:

Núm. unidades	1000	1002	1004	1006	1008	1010	1012
Núm. clientes	1	2	1	2	1	2	1

a. Determine los límites de confianza del 95% para estimar la demanda promedio.
b. Calcule los límites de confianza del 95% para estimar la desviación

estándar.

20. A una muestra de 457 consumidores potenciales se les pidió que calificasen en una escala de uno (totalmente en desacuerdo) a cinco (totalmente de acuerdo) la siguiente afirmación: "Si se reduce la proporción de productos defectuosos, se incrementarían las ventas del producto". La media y la desviación de las respuestas fueron de 3.59 y 1.045, respectivamente. A partir de estos resultados, se calculó el intervalo de confianza de 3.49 a 3.69 entre el cual estaría fluctuando la media de las respuestas de todos los consumidores del mercado.

a. ¿A qué nivel de confianza se cálculo del intervalo de confianza?
b. ¿De qué tamaño debería ser la muestra si se asume un error de estimación equivalente a las tres cuartas partes del error calculado en el apartado (a), así como el mismo nivel de confianza?

21. Se introduce en el mercado un nuevo tipo de leche evaporada en cajas cuyo contenido promedio se especifica en 0.5 litros. La aceptación del producto se probó durante un mes, en 25 bodegas de un distrito de Lima, verificándose que en promedio se vendieron 145 cajas con una desviación estándar de 9 cajas.

a. Si en un distrito existen 2.000 bodegas, entre qué valores se encuentra el número de cajas necesarias para abastecer dicho mercado. Use $\alpha = 0.02$.
b. Si el contenido de las cajas es observado con detalle por los consumidores, pues algunos consideran que el contenido de las cajas está por debajo de lo especificado en las etiquetas. Para tal efecto, se tomó una muestra de 10 cajas, comprobándose su contenido y los resultados fueron los siguientes: 0.48, 0.51, 0.49, 0.52, 0.45, 0.48, 0.502, 0.498, 0.520, 0.503. Diga si existe suficiente evidencia como para afirmar que el contenido está por debajo de lo especificado en las cajas con 98% de confianza.
c. Determine un intervalo del 95% para la desviación estándar de los contenidos de las cajas.

22. Si X es una variable aleatoria, con distribución Normal. En una muestra de tamaño 22 se obtuvo: $\sum x = 397.3$ y $\sum x^2 = 7374.09$.

a. ¿Cuál es la estimación por intervalo de 0.98 para la varianza poblacional?
b. ¿Cuál es la estimación por intervalo de 0.95 para la media de la población?

23. En 16 recorridos de prueba, el consumo de gasolina de un motor experimental tuvo una desviación estándar de 2.2 galones. Construya un intervalo de confianza del 99% para σ al medir el consumo de gasolina para este motor.

24. Diez objetos de forma cilíndrica elegidos al azar entre los producidos en cierta planta industrial han mostrado los siguientes diámetros en centímetros: 10.1, 9.7, 10.3, 10.4, 9.9, 9.8, 9.9, 10.1, 10.3, 9.9, encuentre un intervalo de confianza del 95% para la varianza de los diámetros de todos los objetos producidos por esta planta. Suponga que los diámetros de tales objetos se distribuyen normalmente.

25. Se quiere estimar el ingreso mensual de un sector de comerciantes informales. Se tomó una muestra aleatoria de 100 de ellos y se encontró entre otros datos los siguientes: un ingreso medio de S/.1.800 con una desviación estándar de S/.150 y solo el 30% tiene ingresos superiores a S/.2.100.

a. Hallar los límites de confianza del 95% para la estimación del ingreso medio; ¿cuál es el error máximo de estimación?
b. Estimar la proporción de todos los comerciantes con ingresos superiores a S/.2.100, con un intervalo del 90%.
c. Si la proporción de comerciantes con ingresos superiores a S/.2.100 se estima en 30%, ¿qué tan grande debe ser la muestra para que el error de estimación no sea superior a 0.04 al 95%?

26. Un distribuidor mayorista de artículos de oficina acaba de recibir un lote de 100.000 lapiceros de cierta marca. Al seleccionar una muestra aleatoria para estimar la proporción de lapiceros con fallas, empleando 90% de confianza, concluye que el número de lapiceros con fallas está entre 786 y 7214.

a. Realice la estimación del número de lapiceros con fallas, empleando 95% de confianza.
b. Si la muestra que tomó para hacer las estimaciones anteriores fue de 100 lapiceros, ¿qué alternativas propondría usted para mejorar la estimación?

27. Se observa una máquina en puntos aleatorios del tiempo para estimar la proporción de tiempo en que no se encuentra en trabajo productivo. Suponga que la proporción de tiempo fuera de producción para la máquina es realmente $\pi = 0.01$.

a. ¿Cuál debe ser el tamaño de la muestra, si la desviación estándar de la proporción muestral p es igual a 0.01?
b. ¿Cuál será el tamaño de la muestra si se desea una precisión de 0.04 con una probabilidad de 0.95?

28. Para estimar el número de trabajadores desempleados en el Perú, un economista seleccionó al azar 400 personas de la clase trabajadora de las cuales 65 no tenían trabajo.

a. Estime la proporción de trabajadores sin empleo, usando 95% de confianza.
b. Si se quiere estimar, ahora, la proporción de trabajadores desempleados, con 95% de confianza, ¿a cuántas personas hay que seleccionar si el error de estimación no debe ser mayor del 2%?

29. "Gym Club" es una empresa que brinda el servicio de gimnasia. La empresa está promocionando un riguroso programa de entrenamiento bajo el título de "En un mes, usted será capaz de hacer 8 planchas más, en promedio, de las que podía hacer al inicio del programa". Para verificar esta afirmación el Indecopi selecciona una muestra aleatoria de 10 participantes y registra las planchas que estos hicieron antes y después del programa de entrenamiento. Los datos son los siguientes:

Participante	1	2	3	4	5	6	7	8	9	10
Antes	38	11	34	25	17	38	12	27	32	29
Después	45	24	41	39	30	44	30	39	40	41

¿Es válida la afirmación de "Gym Club"? ¿Por qué? Use $\alpha = 0.025$.
Nota: Señale las fórmulas y los cálculos correspondientes que sustenten su respuesta.

30. Fortunita S.A. es una empresa que se dedica a la compra y venta de gas doméstico en balones de 25 y 50 libras. La empresa tiene cuatro sucursales, distribuidas estratégicamente en: San Borja, San Isidro, La Molina y Surco. Al inicio de la presente semana, el gerente de la empresa presentó, en una reunión de Directorio, las siguientes conclusiones, obtenidas de un estudio realizado con una muestra aleatoria de clientes.

Nota: Señale las fórmulas y los cálculos correspondientes que sustenten su respuesta.

a. Conclusión 1: El 65% de los clientes opina que el servicio que brinda la empresa es "aceptable". El reporte de Minitab que sustenta la afirmación del gerente es:

```
Test and CI for One Proportion
Simple       X      Sample           p          96% CI
     1     150    0.428571
```

Complete los espacios en blanco. ¿Acepta usted la afirmación del gerente? ¿Por qué?

b. Conclusión 2: La dispersión de los gastos mensuales de los clientes en los balones de 50 libras es mayor que la dispersión de los gastos mensuales de los clientes en los balones de 25 libras

El reporte de Minitab que sustenta la afirmación del gerente es:

Descriptive Statistics: Gastos 25 libras, Gastos 50 libras

Variable	Count	Total	Sum	Sum of Squares
Gastos	25 libras	10	654.52	43710.28
Gastos	50 libras	15	1266.27	114678.08

¿Es correcta la afirmación del gerente, al 93.8% de confianza? ¿Por qué? Suponga que los gastos mensuales tienen distribución Normal.

c. Conclusión 3: Los clientes que consumen balones de 25 libras gastan en promedio más que los clientes que consumen balones de 50 libras. ¿Es correcta la afirmación del gerente al 95.6% de confianza? ¿Por qué?
Nota. Utilice los valores de la conclusión 2 y su respuesta formulada en (b).

d. **Conclusión 4**: Las sucursales de San Borja y San Isidro venden la misma cantidad de balones de 25 libras. El reporte de Minitab que sustenta la afirmación del gerente es:

Test and CI for Two Proportions

Sample	N	Sample p
San Borja	150	0.533333
San Isidro	280	0.428571

Estimate for difference: ☐

98.4% CI for difference: (☐, ☐)

Complete los espacios en blanco. ¿Acepta usted la conclusión del gerente? ¿Por qué?

31. Frazadas S.A. es una empresa que se dedica a la producción y comercialización de frazadas de los siguientes tamaños: una plaza, plaza y media y dos plazas. Para la temporada de invierno del 2008, el gerente ha decidido analizar la información registrada en las facturas emitidas por la empresa, utilizando técnicas estadísticas. Para este fin, el gerente seleccionó una muestra aleatoria de las facturas y definió las siguientes variables:

C1 = Monto total de la factura (en nuevos soles). Suponga normalidad de los datos.
C2 = Lugar de venta (Lima, Arequipa, Puno, Cusco).
C3 = Tamaño de las frazadas (una plaza, plaza y media y dos plazas).
C4 = Vendedor (José, Víctor, Esperanza, Andrés, Luz).
Los datos obtenidos se encuentran en el archivo Frazadas.MTW.
Use $\alpha = 0.02$.
En cada caso muestre la fórmula estadística con sus valores correspondientes.

a. ¿Qué puede concluir usted respecto del monto promedio poblacional de las facturas?
b. El gerente afirma que la proporción de frazadas de una plaza vendidas en Lima es mayor que la proporción de frazadas de una plaza vendidas en Puno. ¿Es correcta esta afirmación?
c. ¿Es correcto sostener que el monto promedio de las facturas del vendedor Víctor es menor que el monto promedio de las facturas del vendedor José?

32. El gerente de una empresa productora de planchas de madera para casas prefabricadas desea comprobar si la resistencia (lb/pulg^2) de su producto es la misma en toda la plancha. Para este fin, el gerente selecciona al azar 12 planchas de madera y a continuación cada plancha es dividida en dos partes. Cada parte es sometida a pruebas de resistencia, obteniéndose los siguientes resultados:

Plancha	1	2	3	4	5	6	7	8	9	10	11	12
Parte 1	32	35	25	18	19	20	45	56	27	9	14	21
Parte 2	28	32	20	25	23	15	38	45	26	12	16	17

¿Puede concluirse que la resistencia es la misma en toda la plancha? Use $\alpha = 0.02$.

33. Complete los espacios en blanco del siguiente reporte de Minitab:

Two-sample T for La Molina vs San Isidro

Variable	Count	SE Mean	Sum
La Molina	172	12.1	342746.9
San Isidro	58	21.6	115919.6

Difference = mu (La Molina) - mu (San Isidro)
Estimate for difference: [ ]
97% CI for difference: [ ]
Both use Pooled StDev [ ]

34. En un proceso de envasado de frascos de champú se utilizan dos máquinas envasadoras. De acuerdo con las especificaciones técnicas, ambas máquinas deben llenar los frascos en un contenido promedio de 400 ml. El gerente de producción afirma que no existe diferencia significativa entre ambas máquinas en el proceso de envasado. En tal sentido, selecciona al azar 10 frascos de champú producidos por una máquina, observando que el contenido promedio fue de 403.34 ml, con una desviación estándar de 2.4 ml. Del mismo modo, escogió aleatoriamente 9 frascos de la otra máquina, comprobando que el contenido promedio de estos fue de 398.75 ml, con una desviación de 6.8 ml. ¿Cree usted que tiene sustento la afirmación del gerente de producción, si este asume un nivel de confianza del 99%?

35. Una empresa fabrica el mismo producto en dos máquinas. Una muestra aleatoria de 9 productos de la máquina 1 ha dado los siguientes tiempos de fabricación en segundos: 12, 28, 10, 25, 24, 19, 22, 33, 17; mientras que una muestra aleatoria de 8 productos de la máquina 2 ha dado los siguientes tiempos de fabricación del producto en segundos: 16, 20, 16, 20, 16, 17, 15, 21. Mediante un intervalo de confianza del 95% para la diferencia de los tiempos promedio de fabricación, ¿se puede concluir que la máquina 1 tiene diferente promedio de tiempo de fabricación que la máquina 2? Explique.

36. Una comparación de los tiempos de reacción a dos estímulos diferentes en un experimento psicológico de asociación de palabras aplicado a una muestra aleatoria de 16 personas produjo los resultados (en segundos) que se muestran en la siguiente tabla. Usando un intervalo del 95%, ¿se puede decir que un estímulo es diferente que el otro?

Estímulo 1	Estímulo 2
1, 2, 3, 1, 2, 3, 1, 2.	4, 1, 2, 2, 3, 3, 3, 3.

37. Un ingeniero cree que un programa de entrenamiento puede acortar el tiempo de ensamble de los obreros para cierto mecanismo. Una muestra aleatoria de 5 obreros han dado los siguientes tiempos de empleo (minu-

tos) antes (X) y después (Y) de que hayan terminado el programa de entrenamiento:

Obrero	1	2	3	4	5
X_i	10	10	12	12	14
Y_i	6	10	15	9	10

Con un grado de confianza del 95%, ¿se puede afirmar que el programa de entrenamiento reduce el tiempo medio de ensamble? Suponga distribución Normal de los tiempos de ensamble.

38. Dos ayudantes de laboratorio han determinado la amplitud de oscilaciones. El primer ayudante, al realizar 10 observaciones, ha obtenido un valor medio de la amplitud de 81 mm. El segundo ayudante, al realizar 15 observaciones, ha obtenido una media de 84 mm. Suponiendo que las desviaciones estándar se conocen y son iguales a 8 mm y 9 mm, respectivamente, hallar el intervalo de confianza del 99% para la diferencia de medias. ¿Se puede considerar que los resultados de los ayudantes de laboratorio son realmente diferentes? Explique.

39. En una fábrica de cartón que cuenta con dos máquinas corrugadoras se obtuvo la siguiente información sobre el tiempo de corrugado por metro cuadrado al emplear cartulina Liner.

Máquina	n (días)	$\overline{x}$ (seg.)	s^2 (seg.^{2})
I	9	24.3	30
II	11	25.6	36

a. Calcule e interprete un intervalo de confianza del 90% para la verdadera media del tiempo de corrugado de la máquina I.
b. Calcule e interprete un intervalo de confianza del 95% para la desviación estándar del tiempo de corrugado de la máquina II.
c. Calcule e interprete un intervalo de confianza del 95% para la diferencia en el tiempo medio de corrugado de ambas máquinas.

40. Una máquina produce barras metálicas empleadas en el sistema de suspensión de automóviles. Se selecciona una muestra aleatoria de 15 barras y se mide el diámetro (en metros). Los datos son los siguientes: 8.24 8.23 8.20 8.20 8.21 8.28 8.23 8.26 8.24 8.25 8.19 8.25 8.26 8.23 8.24.

a. Construya un intervalo de confianza del 98% para el diámetro medio de las barras. Interprete.
b. Obtenga un intervalo de confianza del 99% para σ Interprete.

41. Si la desviación estándar de una medición específica de un componente metálico es suficientemente pequeña, entonces el componente metálico es empleado para ensamblajes; sin embargo, se está considerando la adopción de un nuevo componente siempre que su desviación estándar sea menor que la actual. Se registran al azar 100 observaciones de nuevo componente, encontrando $s_2^1 = 0.00041$. De igual manera, se registran al azar 100 observaciones del componente actual y se encuentra que $s_1^1 = 0.00057$. Calcule un intervalo de confianza del 90% para σ_2/σ_1 .

42. Se ha realizado un estudio para comparar el contenido de nicotina (en gramos) de dos marcas de cigarrillos A y B, sobre la base base de muestras de $n_A = 10$ y $n_B = 8$ cigarrillos, respectivamente. Luego de procesar los datos se obtuvieron los siguientes resultados:

$\bar{x}_A = 3.1; \quad \bar{x}_B = 2.7; \quad s_A = 0.5; \quad s_B = 0.7$.

a. Obtenga un intervalo de confianza del 98% para la razón de varianzas poblacionales.
b. Construya un intervalo de confianza del 95% para la diferencia real en el contenido promedio de nicotina de las dos marcas de cigarrillos.

43. Dos marcas de refrigeradoras A y B tienen, ambas, una garantía de un año. En una muestra aleatoria de 50 refrigeradoras de la marca A, 12 se malograron antes de terminar el periodo de garantía. Por otro lado, en una muestra aleatoria de 60 refrigeradoras de la marca B 12 se malograron durante el periodo de garantía. Estime la diferencia real entre las proporciones de fallas durante el periodo de garantía con 98% de confianza.

44. En muestras tomadas de 200 tractores de una línea de ensamblaje y 400 tractores de otra, había, respectivamente, 16 y 20 tractores que requerían de ajustes antes de su comercialización, ¿se puede afirmar que existe una diferencia significativa en la calidad del trabajo de las dos líneas de ensamblaje con un 95% de confianza?

45. En un estudio para analizar los montos diarios de retiros en cajeros automáticos($) y la proporción de retiros por montos mayores de $300 efectuados por los clientes de un banco en 2 zonas distantes, se encontraron los siguientes resultados:

Zona	n_i	Σx	Σx^2	más de $ 300
1	21	4 200	938 000	10
2	16	2 880	596 160	7

Halle e interprete un intervalo de confianza:

a. Del 98% de confianza para la razón de varianzas poblacionales σ_1^2/σ_2^2.
b. Del 90% para la diferencia de los promedios de los montos diarios de retiros de las zonas 1 y 2.
c. Del 95% de confianza para la diferencia en la proporción de retiros diarios para montos mayores de $300, en ambas zonas.
d. Estime, mediante un intervalo del 95% de confianza, el monto total de dinero, para satisfacer las solicitudes diarias de retiros de la zona 1, de un total de 100 clientes. Interprete sus resultados.

Capítulo

3 *Prueba de hipótesis*

En este capítulo trataremos los siguientes temas:

- *Definición.*
- *Hipótesis estadística. Tipos.*
- *Tipos de prueba de hipótesis.*
- *Tipos de errores estadísticos.*
- *Prueba de hipótesis para diferentes parámetros.*
- *Funciones potencia y característica de operación.*
- *Prueba de bondad de ajuste.*
- *Prueba de independencia.*

A partir del análisis de una muestra se probará si una población es caracterizada por algún parámetro poblacional. Además, se abordan los conceptos relacionados con los tipos de hipótesis estadística, las pruebas de hipótesis y los tipos de errores estadísticos; se detallan las diferentes pruebas de hipótesis para los parámetros poblacionales, mediante una gran variedad de ejercicios resueltos en forma pormenorizada y apoyados en el uso del software Minitab, mostrando las interpretaciones de los reportes obtenidos. Finalmente, se desarrollan las funciones potencia y característica de operación, así como las pruebas de bondad de ajuste y la prueba de independencia, en las que se emplea la distribución Chi Cuadrado para determinar si los datos se ajustan a un tipo de distribución o si dos variables cualitativas (o cuantitativas pero categorizadas) son independientes, respectivamente.

1. Introducción

La prueba de hipótesis es la segunda rama de la estadística inferencial (inductiva clásica), cuyo objetivo está orientado a la elección de una de las acciones: a_0 o a_1; ambas asociadas al parámetro de una población. La distribución de una población comúnmente no es conocida, por lo tanto, en la práctica generalmente se obtiene una muestra aleatoria de la población y se elegirá entre a_0 o a_1, basándose en la información que proporcione la muestra. Por ejemplo:

1. El gerente de una empresa ha establecido la norma de aceptar los lotes de cierto producto si este no tiene más del 2% de defectuosos. Si en un lote de 400 encuentra que 7 son defectuosos, ¿debe ser aceptado o rechazado el lote?
2. Un político se presentará a las elecciones municipales de cierto distrito si considera que más del 35% de los electores votarían por él. Una encuesta de 200 electores indica que 80 están a favor de su candidatura. ¿Debe considerar el político, como prueba, de que más del 35% de los electores de su distrito votarían por él?

2. Definición

Una hipótesis estadística es una aseveración acerca de la distribución de una variable aleatoria o de los parámetros de una población, que puede ser verdadera o no. Se puede especificar una hipótesis estadística dando el tipo de distribución y el valor o valores del parámetro o parámetros que la definen, respectivamente. Por ejemplo:

a. X tiene una distribución binomial con $\pi = 0.2$.
b. X tiene una distribución normal con $\mu = 10$ y $\sigma = 4$.

En los casos en que la distribución es supuestamente conocida, una hipótesis estadística se especifica con el valor o los valores del parámetro. Por ejemplo:

a. El promedio de ingresos por semana de los obreros de la industria textil es $1.500, es decir: H: $\mu = 1500$.
b. El porcentaje de personas atacadas por cierta epidemia en una universidad es del 5%, es decir: H: $\pi = 0.05$.

3. Clases de hipótesis

Una forma sencilla de especificar con precisión un procedimiento de prueba de hipótesis es determinando las hipótesis estadísticas; a estas se les llama hipótesis nula (H_0) e hipótesis alternativa (H_1). Estas hipótesis deben ser formuladas antes de seleccionar los datos.

La prueba de hipótesis parte del principio de que la H_0 es verdadera, el objetivo de la prueba de hipótesis es rechazar la H_0 .

3.1 Hipótesis nula (H_0)

Es la afirmación de que se va a someter a prueba para ser aceptada o rechazada, es decir que se desea verificar. Representa lo conocido e indica que todo sigue igual.

3.2 Hipótesis alternativa (H_1)

Es aquella que es aceptada si H_0 es rechazada. H_1 representa el cambio, lo sospechoso, lo novedoso, etcétera.

Nota: La hipótesis nula o alternativa puede ser simple si solamente asume un valor, o puede ser compuesta si asume más de un valor.

3.3 Prueba estadística de una hipótesis

Definición. La prueba estadística de una hipótesis es una norma o regla que luego de aplicarse a valores experimentales conduce a una decisión: aceptar (no rechazar) o no la hipótesis nula. En una prueba de hipótesis, a pesar de probar una hipótesis nula contra una alternativa bajo el supuesto de que la hipótesis nula es cierta, lo que realmente ocurre es que se está tomando una decisión entre dos acciones o entre H_0 y H_1 .

4. Tipos de prueba

Existen tres tipos de prueba de hipótesis, cada uno identificado por la forma como se formulan H_0 y H_1 .

4.1 Prueba de cola izquierda o inferior

Las hipótesis se formulan de la forma:

$H_o : \theta \geq \theta_0$

$H_1 : \theta < \theta_0$

Existe un punto crítico c y se rechaza H_0

si $\hat{\theta} < c$

RC: Región crítica.
RA. Región de aceptación.

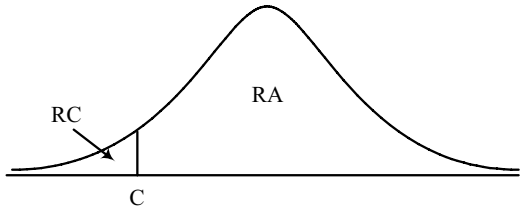


Figura 1. Prueba de cola izquierda.

4.2 Prueba de cola derecha o superior

Las hipótesis se definen por:

$H_o : \theta \leq \theta_0$

$H_1 : \theta > \theta_0$

Existe un punto crítico c y se rechaza H_0

si $\hat{\theta} > c$.

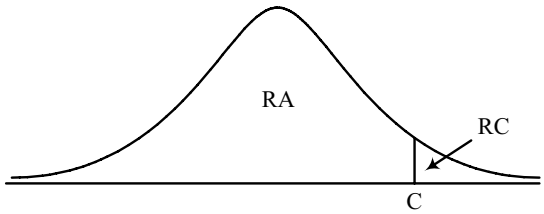


Figura 2. Prueba de cola derecha.

4.3 Prueba de dos colas o bilateral

Las hipótesis se formulan de la forma:

$H_o : \theta = \theta_0$

$H_1 : \theta \neq \theta_0$

Existen dos puntos críticos y se rechaza H_0

si: $\hat{\theta} < c_1$ o $\hat{\theta} > c_2$

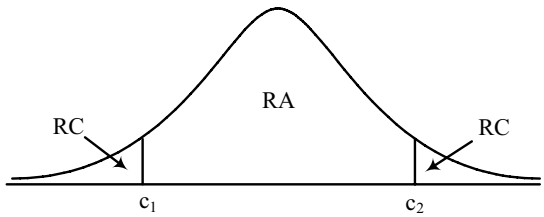


Figura 3. Prueba de dos colas.

5. Tipos de errores

En una prueba de hipótesis se pueden presentar las siguientes situaciones:

1. Rechazar una H_0 que es verdadera (decisión incorrecta).
2. Aceptar una H_0 que es verdadera (decisión correcta).
3. Rechazar una H_0 que es falsa (decisión correcta).
4. Aceptar una H_0 que es falsa (decisión incorrecta).

El rechazo de una H_0 que es verdadera se llama error de tipo I, y la aceptación de una H_0 que es falsa se llama error de tipo II. Estas dos posibilidades incorrectas, junto con las dos decisiones correctas, se pueden resumir en el siguiente cuadro:

Muestra	H_0 verdadera	H_0 falsa (H_1 verdadera)
Aceptar H_0	Decisión correcta	Error tipo II
Rechazar H_0 (Aceptar H_1)	Error tipo I	Decisión correcta

El responsable de la toma de decisiones deberá reducir al máximo las probabilidades de cometer estos dos tipos de errores, que en la práctica no es fácil, porque las probabilidades de cometer estos tipos de errores son inversamente proporcionales para cualquier prueba dada y tamaño de muestra fijo n . De ahí que cuanto menor es el riesgo de cometer un error de tipo I tanto mayor es la probabilidad de cometer un error de tipo II, y viceversa.

5.1 Nivel de significación

Definición. El nivel de significación se denota por α y se define como la máxima probabilidad de cometer un error de tipo I, es decir la probabilidad de rechazar H_0 siendo esta verdadera.

Se denota como:

α = P(error tipo I) = P(Rechazar H_0/H_0 es verdadero) = P(Aceptar H_1/H_1 es falso).

La probabilidad de cometer error de tipo II se denota por β , es decir la probabilidad de aceptar H_0 siendo esta falsa.

β = P(Aceptar H_0/H_0 es falso) = P(Rechazar H_1/H_1 es verdadero)

Nota: Se debe tener presente que $\alpha + \beta \neq 1$

5.2 Región crítica

Definición. Región crítica (RC) es la región de la distribución muestral, que de acuerdo con una prueba definida conduce al rechazo de la hipótesis nula bajo consideración (H_0).

5.3 Región de aceptación

Definición. Región de aceptación (RA) es la región de la distribución muestral, que de acuerdo con una prueba definida conduce a no rechazar la hipótesis nula (H_0).

Procedimiento para realizar una prueba de hipótesis referente a un parámetro

Los pasos para hacer una prueba de hipótesis relativa al parámetro θ de una población (con un tamaño de muestra fijo n) son:

i. Formular las hipótesis de acuerdo con el problema que se tiene; es decir la hipótesis nula debe ser: $H_o : \theta \geq \theta_0$ o $H_o : \theta \leq \theta_0$ y H_1 puede ser $(\theta < \theta_0, \ \theta > \theta_0, \ \theta \neq \theta_0)$.
ii. Escoger el nivel de significación o riesgo α. Plantear las suposiciones de la prueba.
iii. Escoger la estadística de prueba adecuada, cuya distribución por muestreo sea conocida, bajo el supuesto de que H_0 sea cierta.
iv. Establecer la región crítica, es decir determinar el valor(es) crítico(s) que depende de la hipótesis alternativa, el nivel de significación y la regla de decisión.
v. Calcular el valor de la estadística de prueba, sobre la base de una muestra tomada al azar.
vi. Decisión y conclusión; rechazar H_0, si la estadística de prueba tiene un valor en la región crítica y no rechazarla en otro caso.

6. Prueba de hipótesis para los parámetros

6.1 Prueba de hipótesis para la media poblacional (μ)

6.1.1 Cuando la varianza poblacional es conocida

Cuando en ciertas situaciones se desea probar alguna hipótesis relacionada con la media poblacional de una variable X (μ_x). Por ejemplo: tiempo de vida de un transistor, tiempos de espera para pagar en un centro comercial, etcétera; en que la variable X se distribuye según una normal con media desconocida, pero de la cual se conoce su varianza poblacional debido a estudios similares. Entonces, se puede analizar una muestra aleatoria de la población: $x_1, x_2, ..., x_n$; y se puede afirmar que la siguiente expresión:

$$\left(\frac{\bar{x} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \right),$$

se distribuye según una normal con media cero (0) y una desviación estándar igual a la unidad (1): N $(0,1)$. Esta es la distribución estadística que se debe emplear para la prueba de hipótesis correspondiente.

Suposición básica:

X: puede tener cualquier distribución y σ conocida (si no es normal se aplica el Teorema Límite Central)

1. Planteamiento de hipótesis. Pueden ser:

Pruebas unilaterales		Prueba bilateral
1.1 $\mathrm{H}_o : \mu \leq \mu_o$	1.2 $\mathrm{H}_o : \mu \geq \mu_o$	1.3 $\mathrm{H}_o : \mu = \mu_o$
$\mathrm{H}_1 : \mu > \mu_o$	$\mathrm{H}_1 : \mu < \mu_o$	$\mathrm{H}_1 : \mu \neq \mu_o$

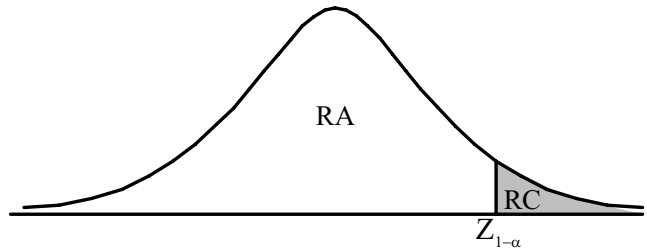
2. Elegir nivel de significación α.

3. Estadística de prueba: $Z_0 = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

4. Región crítica (RC) y regla de decisión.

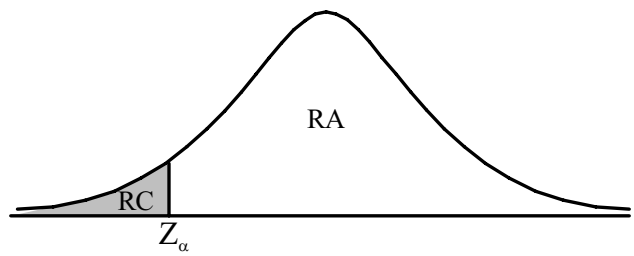
   4.1 $\text{RC} = \langle Z_{(1-\alpha)}, \infty \rangle$

   Rechazar H_0 si $Z_o \in \text{RC}$, es decir, si: $Z_o > Z_{(1-\alpha)}$



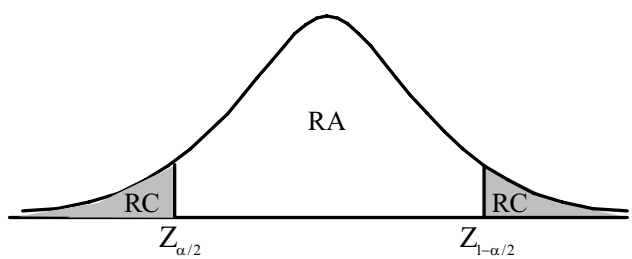
   4.2 $\text{RC} = \langle -\infty, Z_{(\alpha)} \rangle$

   Rechazar H_0 si $Z_o \in \text{RC}$, es decir, si: $Z_o < Z_{(\alpha)}$



   4.3 $\text{RC} = \langle -\infty, Z_{(\alpha/2)} \rangle \cup \langle Z_{(1-\alpha/2)}, \infty \rangle$

   Rechazar H_0 si $Z_o \in \text{RC}$, es decir, si: $Z_o < -Z_{(1-\alpha/2)}$ ó $Z_o > -Z_{(1-\alpha/2)}$



5. Se determina el valor de la estadística de prueba usando la información muestral.

6. Decisión y conclusión.

Ejemplo 1:

El jefe de seguridad de Galicia afirma que el estacionamiento es usado por personas que no son clientes, en promedio por más de 80 minutos con varianza de 20 minutos^2. Si se toma una muestra de 25 vehículos que se encontraban en el estacionamiento y que pertenecen a personas que no son clientes, y se encuentra un promedio de 78 minutos, esta muestra sustenta lo que el jefe de seguridad afirma con $\alpha = 0.05$.

Solución:

1. Formulación de las hipótesis:

   $\mathrm{H_o} : \mu \leq 80$ (el tiempo promedio de los no clientes no es mayor a 80 minutos)

   $\mathrm{H_1} : \mu > 80$ (el tiempo promedio de los no clientes es mayor a 80 minutos)

2. $\alpha = 0.05$.

3. La estadística de prueba es: $Z_0 = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$, y se supone que los datos tienen una distribución Normal.

4. Región crítica: $\mathrm{RC} = \langle Z_{(1-0.05)}, \infty \rangle = \langle 1.645, \infty \rangle$, rechazar $\mathrm{H_o}$ si $\mathrm{Z_o} \in \mathrm{RC}$

5. Valor de la estadística de prueba: $Z_0 = \dfrac{78-80}{4.4721/5} = -2.236 \qquad \sigma = \sqrt{20} = 4.4721$

6. Decisión y conclusión: Como $\mathrm{Z_o} = -2.236 \notin \mathrm{RC}$; por lo tanto no se rechaza la hipótesis nula; entonces, las personas que utilizan el estacionamiento y no son clientes en promedio presentan un tiempo de uso no mayor de 80 minutos.

Tamaño de muestra (basado en los riesgos α y β fijos)

Suponga que se tienen que probar las hipótesis:

$\mathrm{H_o} : \mu = \mu_0$ (prueba unilateral)

$\mathrm{H_1} : \mu < \mu_0$ (puede escribirse con $\mu = \mu_1$ donde $\mu_1 < \mu_0$)

Si $\mu_1 < \mu_0$, la región crítica es: $\mathrm{RC} = \langle -\infty, \mathrm{C} \rangle$. Si se supone que la varianza poblacional es conocida, entonces:

$$\alpha = P(RH_0/\mu = \mu_0) = P(\bar{x} < C/\mu = \mu_0) = P\left(Z < \frac{C - \mu_0}{\sigma/\sqrt{n}}\right) \Rightarrow Z_{(\alpha)} = \frac{C - \mu_0}{\sigma/\sqrt{n}}$$, de donde:

$$C = \mu_0 + Z_{(\alpha)} \frac{\sigma}{\sqrt{n}} \qquad (1)$$

$$\beta = P(AH_0/\mu = \mu_1) = P(\bar{x} \geq C/\mu = \mu_1) = P\left(Z \geq \frac{C - \mu_1}{\sigma/\sqrt{n}}\right) \Rightarrow Z_{(1-\beta)} = \frac{C - \mu_1}{\sigma/\sqrt{n}}$$, de donde

$$C = \mu_1 + Z_{(1-\beta)}\frac{\sigma}{\sqrt{n}} \qquad (2)$$

Al igualar (1) con (2) se tiene: $\mu_0 + Z_{(\alpha)}\frac{\sigma}{\sqrt{n}} = \mu_1 + Z_{(1-\beta)}\frac{\sigma}{\sqrt{n}}$, operando:

$\mu_0 - \mu_1 = -Z_{(\alpha)}\frac{\sigma}{\sqrt{n}} + Z_{(1-\beta)}\frac{\sigma}{\sqrt{n}} = \frac{\sigma\left(-Z_{(\alpha)} + Z_{(1-\beta)}\right)}{\sqrt{n}}$ y despejando n se tiene:

$$\sqrt{n} = \frac{\sigma\left(-Z_{(\alpha)} + Z_{(1-\beta)}\right)}{\mu_0 - \mu_1};\ \text{entonces:}\ n = \frac{\left(Z_\alpha + Z_\beta\right)^2\sigma^2}{\left(\mu_0 - \mu_1\right)^2} \qquad (3)$$

Donde μ_0 es el valor hipotético fijado en la hipótesis nula y μ_1 es el valor que se supone es el verdadero y que se fija para el cálculo de β ; la fórmula es válida para una prueba de cola derecha o para una prueba de cola izquierda.

Para una prueba de dos colas la fórmula es:

$$n = \frac{\left(Z_{\alpha/2} + Z_\beta\right)^2\sigma^2}{\left(\mu_0 - \mu_1\right)^2} \qquad (4)$$

Ejemplo 2:

Se tiene que probar las hipótesis: $\mathrm{H}_0 : \mu \geq 12$
$\mathrm{H}_1 : \mu < 12$

Suponiendo que la varianza de la población es de 3, el error de tipo I del 5% y un error de tipo II del 5%, cuando realmente el promedio es de 13. Determine el tamaño de la muestra.

Solución:

Usando la fórmula definida en (3) se tiene:

$$n = \frac{\left(Z_{0.05} + Z_{0.05}\right)^2\sigma^2}{\left(\mu_0 - \mu_1\right)^2} = \frac{\left(1.64485 + 1.64485\right)^2 3}{\left(12 - 13\right)^2} = 32$$

Por lo tanto $n = 32$.

6.1.2 Cuando la varianza poblacional es desconocida

Cuando se desea probar alguna hipótesis relacionada a la media poblacional de una variable $X\ (\mu_x)$ que se distribuye según una distribución normal con media desconocida, y donde además se desconoce su varianza poblacional debido a que no existen estudios similares. Entonces también se analiza una muestra aleatoria de una población: $x_1, x_2, ..., x_n$; de la que se puede calcular su varianza muestral s_X^2 , y se sabe que la siguiente expresión:

$$\left(\frac{\bar{x} - \mu_X}{\frac{s_X}{\sqrt{n}}} \right),$$

se distribuye según una t de Student con $n - 1$ grados de libertad $[t_{(n-1)}]$. Siendo esta la distribución estadística que se debe emplear para la prueba de hipótesis correspondiente.

Suposición básica: $X \sim N(\mu, \sigma^2)$

1. Planteamiento de hipótesis. Estas pueden ser:

Pruebas unilaterales		Prueba bilateral
1.1 $H_o : \mu \leq \mu_o$	1.2 $H_o : \mu \geq \mu_o$	1.3 $H_o : \mu = \mu_o$
$H_1 : \mu > \mu_o$	$H_1 : \mu < \mu_o$	$H_1 : \mu \neq \mu_o$

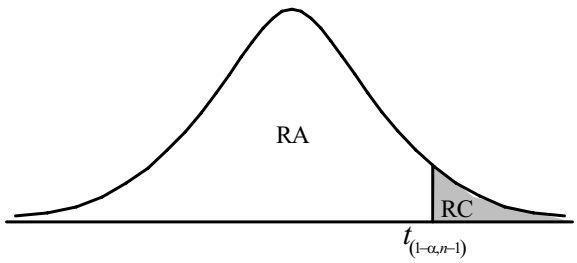
2. Elegir nivel de significación α.

3. Estadística de prueba: $t_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{(n-1)}$ y se supone que los datos tienen distribución Normal.

4. Región crítica (RC) y regla de decisión.

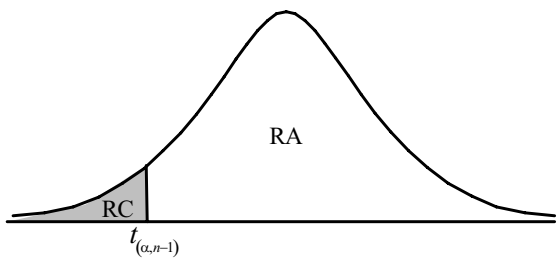
   4.1 $RC = \langle t_{(1-\alpha, n-1)}, \infty \rangle$

   Rechazar H_0 si $t_0 \in RC$, es decir, si: $t_0 > t_{(1-\alpha, n-1)}$



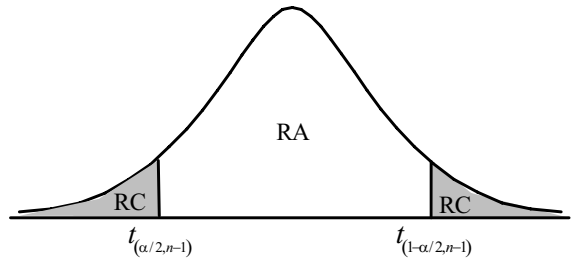
   4.2 $RC = \langle -\infty, t_{(\alpha, n-1)} \rangle$

   Rechazar H_0 si $t_0 \in RC$, es decir, si: $t_0 < t_{(\alpha, n-1)}$



4.3 $\mathrm{RC} = \langle -\infty, \mathrm{t}_{(\alpha/2,\, n-1)} \rangle \cup \langle \mathrm{t}_{(1-\alpha/2,\, n-1)}, \infty \rangle$

Rechazar H_0 si $t_0 \in \mathrm{RC}$, es decir, si: $t_0 < t_{(\alpha/2,n-1)}$ ó $t_0 > t_{(1-\alpha/2,n-1)}$



5. Se determina el valor de la estadística de prueba usando la información muestral.
6. Decisión y conclusión.

Ejemplo 3:

Un fabricante de pilas indica que el tiempo de duración de las pilas que fabrica sigue una distribución normal y que el promedio de duración es de al menos 55 horas. Un cliente mayorista de la calle Capón ha hecho un pedido de pilas, pero antes de aceptar el pedido analiza una muestra de seis pilas cuyos resultados son los siguientes: 55, 48, 46, 47, 50, 49. ¿Qué decisión tomará el mayorista al 5%?

Solución:

1. Se formulan las hipótesis:
   $\mathrm{H}_o : \mu \geq 55$ (acepta el pedido)
   $\mathrm{H}_1 : \mu < 55$ (no acepta el pedido)
2. $\alpha = 0.05.$
3. Estadística de prueba: $t_0 = \dfrac{x - \mu_0}{s/\sqrt{n}} \sim t_{(5)}$ y se asume que los datos tienen distribución Normal.
4. Región crítica: $\mathrm{RC} = \langle -\infty, t_{(0.05,6-1)} \rangle = \langle -\infty, -2.01505 \rangle$, rechazar H_0 si $t_0 \in \mathrm{RC}$
5. Valor de la estadística de prueba: $t_0 = \dfrac{49.17 - 55}{3.19/\sqrt{6}} = \dfrac{5.83}{1.30} = -4.48$
6. Decisión y conclusión: Como $t_0 = -4.48 \in \mathrm{RC}$; por lo tanto se rechaza la hipótesis nula y el cliente mayorista no debe aceptar el pedido.

Concepto del P-value

El P-value se define como la probabilidad de tener un valor más extremo (más grande o más pequeño) que el observado.

Así, para una prueba de cola derecha el P-value se define como: $P(\hat{\theta} > \hat{\theta}_0)$. Para una prueba de cola izquierda el P-value se define como:

$$P(\hat{\theta} < \hat{\theta}_0) \ .$$

Por ejemplo, para probar un valor del parámetro, cuando se conoce la varianza poblacional, se usa la distribución Z, y para una prueba de cola derecha si ocurre que:

$$Z_{CALCULADA} < Z_{CRITICA}$$

entonces no se rechaza la hipótesis nula H_0 . Este concepto se ilustra en la figura 4.

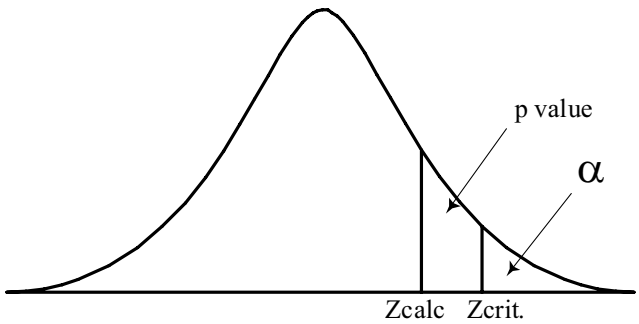


Figura 4. Gráfico de α y del P-value.

Como esta es una prueba de cola derecha su región de rechazo se encuentra a la derecha y por lo tanto la hipótesis nula H_0 no puede ser rechazada porque $Z_{Calculada} < Z_{Crítica}$. En forma equivalente, el P-value daría la misma decisión si se usa el criterio:

$$\text{P- Value} \geq \alpha$$

En términos generales, los valores altos del P-value están relacionados con la decisión de no rechazar la hipótesis nula, mientras que los valores bajos del P-value están asociados con el rechazo de la hipótesis nula. Frecuentemente, el P-value es comparado con el nivel α establecido para decidir sobre el rechazo o no de la hipótesis nula.

La mayoría de los programas estadísticos proporcionan el P-value, para que el usuario tome la decisión de rechazar o no la hipótesis nula al compararlo con el valor α.

El criterio de decisión al comparar el P-value con α es:

Sí P- Value < α, entonces se rechaza la hipótesis nula (H_0)

Sí P- Value ≥ α, entonces no se rechaza la hipótesis nula (H_0)

Nota: A continuación se presenta el procedimiento de una prueba de hipótesis para una media poblacional usando Minitab. Se debe tener presente que a diferencia del procedimiento tradicional (manual), donde la prueba se hace en seis pasos, una prueba de hipótesis usando un programa estadístico (en particular el Minitab en este caso) se hace en dos pasos.

- Se formulan las hipótesis al igual que el procedimiento tradicional.
- Se compara el P-value que proporciona el Minitab con el valor de α, para tomar una decisión, basándose en el criterio de decisión de la definición del P-value.

Ejemplo 4:

El procedimiento de solución del problema 3 usando el Minitab es:

1. Las hipótesis por probar son:

$$H_o : \mu \geq 55 \text{ (acepta el pedido)}$$
$$H_1 : \mu < 55 \text{ (no acepta el pedido)}$$

   a. Se ingresan los datos a la hoja de cálculo, luego se sigue el mismo procedimiento que se usó para construir un intervalo de confianza para μ. En este caso, como la varianza poblacional es desconocida, se emplea la distribución t y se procede como se muestra en las siguientes figuras 5 y 6.

   (Nótese que ahora en el campo de <test mean> se ingresó el valor de 55).

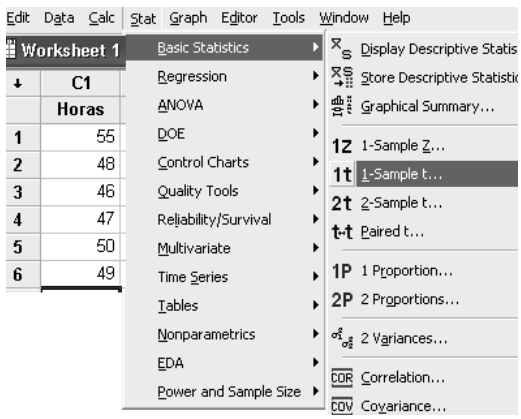


Figura 5. Prueba t para una muestra.

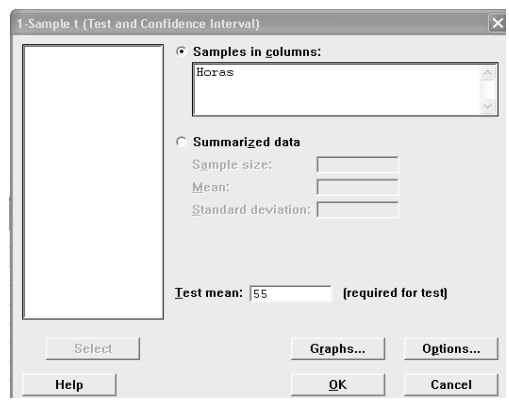


Figura 6. Valor para la prueba.

   b. Luego activar <Options>, en la ventana siguiente se escoge el tipo de prueba, como en este caso se trata de cola inferior, en <Alternative> se escoge <less than>, finalmente elegir <Ok>. Véase la figura 7.

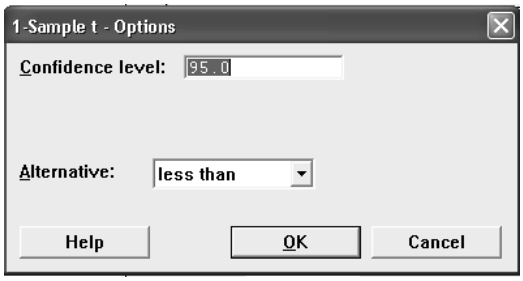


Figura 7. Elección del tipo de prueba.

2. Se obtienen los siguientes resultados:

One-Sample T: Horas

```
Test of mu = 55 vs < 55

                                                95%Upper
Variable  N    Mean    StDev   SE Mean   Bound     T       P
Horas     6  49.1667  3.1885   1.3017   51.7897 -4.48   0.003
```

Como el P-value = 0.003 es menor que $\alpha = 0.05$, entonces se rechaza la hipótesis nula y se concluye igual que en el problema 3.

$$\mathrm{P}(t < -4.48) = \text{P-value} = 0.003259.$$

6.2 Prueba de hipótesis para una proporción poblacional (π)

Suponga que la variable cualitativa X representa el estado de un artículo: defectuoso o no defectuoso, funciona o no funciona, vendido o no vendido, etcétera; y de la cual se desea contrastar una hipótesis relacionada con la proporción de determinada categoría de la variable; entonces se analiza una muestra aleatoria de la población $x_1, x_2,\ldots, x_n$, que se distribuye según una binomial con parámetro π (proporción desconocida). Si se desea probar que la proporción muestral p es igual a un valor π_0; entonces, se sabe que la siguiente expresión:

$$\left(\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \right),$$

se distribuye aproximadamente según una normal con media cero (0) y varianza igual a la unidad (1): $N(0,1)$, que es la distribución estadística que se debe emplear para la prueba de hipótesis correspondiente.

Supuesto: la variable debe ser cualitativa (o cuantitativa pero previamente categorizada).

1. Formulación de las hipótesis:

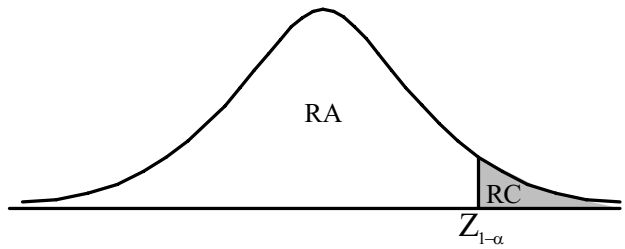
Pruebas unilaterales		Prueba bilateral
1. $\mathrm{H_o}: \pi \le \pi_o$	2. $\mathrm{H_o}: \pi \ge \pi_o$	3. $\mathrm{H_o}: \pi = \pi_o$
$\mathrm{H_1}: \pi > \pi_o$	$\mathrm{H_1}: \pi < \pi_o$	$\mathrm{H_1}: \pi \ne \pi_o$

2. Elegir α.

3. La estadística de prueba es: $Z_0 = \dfrac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} \to N(0,1)$ suponiendo que los datos tienen distribución binomial.
4. Región crítica (RC) y regla de decisión.

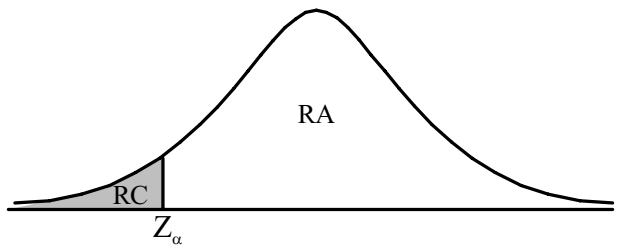
4.1 $\mathrm{RC} = \langle Z_{(1-\alpha)}, \infty \rangle$

Rechazar H_0 si $Z_0 \in \mathrm{RC}$, es decir, si: $Z_0 > Z_{(1-\alpha)}$



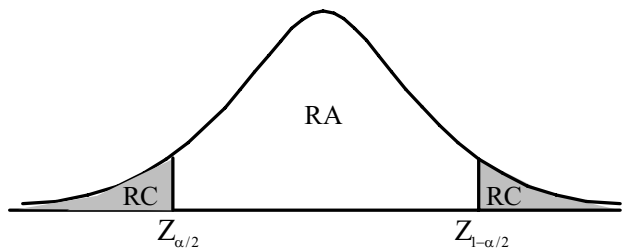
4.2 $\mathrm{RC} = \langle -\infty, Z_{(\alpha)} \rangle$

Rechazar H_0 si $Z_0 \in \mathrm{RC}$, es decir, si: $Z_0 < Z_{(\alpha)}$



4.3 $\mathrm{RC} = \langle -\infty, Z_{(\alpha/2)} \rangle \cup \langle Z_{(1-\alpha/2)}, \infty \rangle$

Rechazar H_0 si $Z_0 \in \mathrm{RC}$, es decir, si: $Z_0 < -Z_{(1-\alpha/2)}$ ó $Z_0 > Z_{(1-\alpha/2)}$



5. Se determina el valor de la estadística de prueba usando la información muestral.
6. Decisión y conclusión.

Ejemplo 5:

De 100 compras efectuadas en una tienda 10 fueron hechas por internet, según datos de ventas del año pasado proporcionados por el gerente. Este año se seleccionó una muestra de 200 compras para determinar qué proporción de compras se efectuaron por internet. Esta muestra indicó que el 20% de las compras fueron por internet. A nivel de significancia del 1%, ¿puede concluirse que la proporción de compras por internet en la tienda ha cambiado significativamente?

Solución:

1. Formulación de la hipótesis:

   $H_0 : \pi = 0.10$ (la proporción de compras por internet no ha cambiado)

   $H_1 : \pi \neq 0.10$ (la proporción de compras por internet ha cambiado)
2. $\alpha = 0.01$.
3. Estadística de prueba:

$$Z_0 = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \rightarrow N(0,1)$$

   y los datos tienen una distribución binomial.
4. Región crítica:

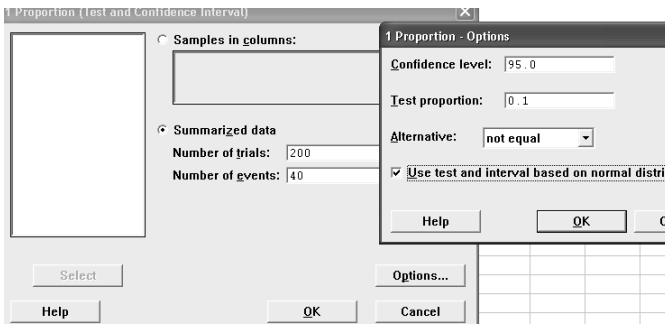


Figura 8. Ingreso de nivel de confianza.

$$RC = \langle -\infty, Z_{(0.01/2)} \rangle \cup \langle Z_{(1-0.01/2)}, \infty \rangle = \langle -\infty, 2.57583 \rangle \cup \langle 2.57583, \infty \rangle ,$$

   rechazar H_o si $Z_o \in RC$
5. Valor de la estadística de prueba: $Z_0 = \frac{0.2 - 0.1}{\sqrt{\frac{0.1(1-0.1)}{200}}} = 4.71$
6. Decisión y conclusión: Como $Z_o = 4.71 \in RC$, se rechaza la hipótesis nula, entonces, la proporción de compras por internet ha cambiado, con 1% de nivel de significación.

El procedimiento usando Minitab es:

1. Formulación de las hipótesis:

   $H_o : \pi = 0.10$ (la proporción de compras por internet no ha cambiado)

   $H_1 : \pi \neq 0.10$ (la proporción de compras por internet ha cambiado)

   a. Se sigue el procedimiento usado en la construcción del intervalo de confianza para π, se ingresa en el campo de <Sumarized data> n y x, se activa <Options> y aparece la ventana de la izquierda, en ella se ingresa el valor de $\pi_0 = 0.10$ en <Test proportion>, se escoge el tipo de prueba <not equal>, se activa la última opción y finalmente <Ok>.

b. Los resultados son los siguientes:

```
Test and CI for One Proportion
Test of p = 0.1 vs p not = 0.1
Sample X  N  Sample p     95% CI         Z-Value P-Value
1      40 200 0.200000 (0.144564, 0.255436) 4.71  0.000
```

Como el P-value = 0.00 es menor que $\alpha = 0.05$, entonces se rechaza la hipótesis nula y se concluye que la proporción de compras por internet ha cambiado, es decir no es verdad que $\pi = 0.10$.

6.3 Prueba de hipótesis para la varianza poblacional (σ^2)

Suponga que la variable X representa la longitud de una pieza mecánica o el peso de cierto componente presente por unidad de masa, etcétera; que se distribuye según una normal con media conocida y varianza desconocida. Entonces, es de interés realizar algunas hipótesis con respecto a su varianza poblacional. Para estas situaciones se analiza una muestra aleatoria de la población: $x_1, x_2, \ldots, x_n$; de la que se puede calcular su varianza muestral s_X^2, y se sabe que la siguiente expresión:

$$\left(\frac{(n-1)s_X^2}{\sigma_X^2}\right),$$

se distribuye según un Ji Cuadrado con n–1 grados de libertad:

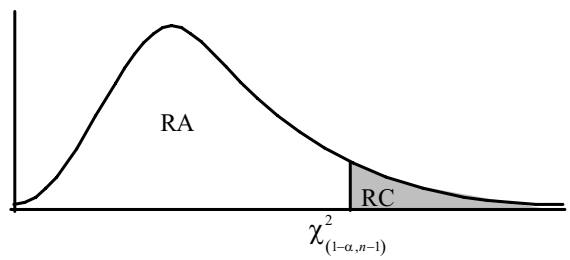
$\left[\chi^2_{(n-1)}\right]$. Esta es la distribución estadística que se debe emplear para la prueba de hipótesis correspondiente.

1. Formulación de hipótesis: Estas pueden ser:

Pruebas unilaterales		Prueba bilateral
1.1 $H_o : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 > \sigma_0^2$	1.2 $H_o : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 < \sigma_0^2$	1.3 $H_o : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$

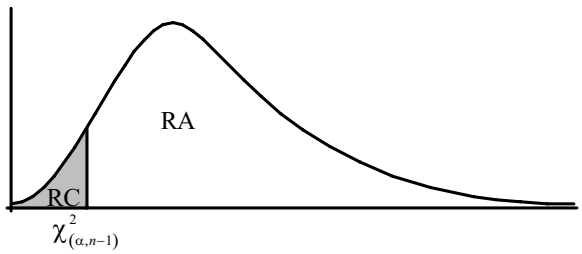
2. Elegir α.
3. La estadística de prueba: $\chi_0^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$ y los datos tienen una distribución Normal.
4. Región crítica (RC) y regla de decisión.
   4.1 $RC = \langle \chi^2_{(1-\alpha, n\text{-}1)}, \infty \rangle$

   Rechazar H_o si $\chi_0^2 \in RC$, es decir, si: $\chi_0^2 > \chi^2_{(1-\alpha, n-1)}$



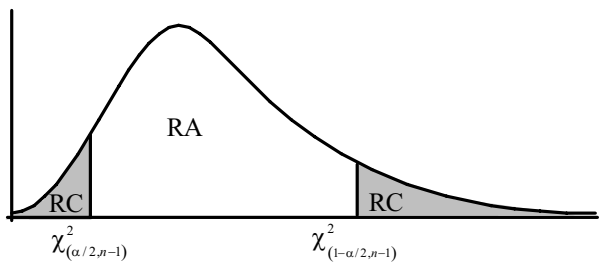
4.2 $\text{RC} = \langle -\infty, \chi^2_{(\alpha, n-1)} \rangle$

Rechazar H_o si $\chi_0^2 \in \text{RC}$, es decir, si: $\chi_0^2 < \chi^2_{(\alpha, n-1)}$



4.3 $\text{RC} = \langle -\infty, \chi^2_{(\alpha/2,n-1)} \rangle \cup \langle \chi^2_{(1-\alpha/2,n-1)}, \infty \rangle$.

Rechazar H_o si $\chi_0^2 \in \text{RC}$, es decir, si: $\chi_0^2 < \chi^2_{(\alpha/2,n-1)}$ *ó* $\chi_0^2 > \chi^2_{(1-\alpha/2,n-1)}$



5. Se determina el valor de la estadística de prueba usando la información muestral.
6. Decisión y conclusión.

Ejemplo 6:

El director de un colegio está interesado en que sus alumnos que terminen la secundaria ingresen a la universidad, e indica que una varianza superior en 50 en las notas de una prueba de conocimientos se considera negativa. Propone un programa de preparación a la universidad y al final del curso se elige una muestra de 10 alumnos y se toma un test de conocimientos cuyos resultados son los siguientes:

84, 70, 90, 92, 85, 75, 93, 80, 93, 76

Pruebe la hipótesis de que la varianza poblacional no es superior a 50 con un nivel de significación del 5%.

Solución:

1. Las hipótesis son:

   $H_o : \sigma^2 \leq 50$ (la varianza poblacional no es superior a 50)

   $H_1 : \sigma^2 > 50$ (la varianza poblacional es superior a 5 0)

2. $\alpha = 0.05.$

3. La estadística de prueba es: $\chi_0^2 = \dfrac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{(9)}$ y las notas de la prueba de conocimiento tienen distribución Normal.

4. Región crítica: $RC = \langle \chi^2_{(1-0.05,\,10-1)}, \infty \rangle = \langle 16.919, \infty \rangle$, rechazar H_o si $\chi_0^2 \in RC$

5. Valor de la estadística de prueba: $\chi_0^2 = \dfrac{9(68.844)}{50} = 12.39$

6. Decisión y conclusión: el valor observado, estadística de prueba, no pertenece a la región crítica, es decir, pertenece a la región de aceptación, no se rechaza la hipótesis nula. Se concluye que la varianza no es superior a 50.

6.4 Prueba de hipótesis para una razón de varianzas (σ_1^2/σ_2^2)

Cuando se desea determinar si la variabilidad, desviación estándar o varianzas, es similar para dos variables X e Y, sabiendo que estas se distribuyen normalmente, entonces se pueden analizar muestras aleatorias independientes de ambas poblaciones en cuestión: $x_1, x_2, \ldots, x_n$ y $y_1, y_2, \ldots, y_m$; de las que se pueden calcular sus varianzas muestrales s_X^2 y s_Y^2 . Además, se sabe que la siguiente expresión:

$$\left(\frac{s_X^2 / \sigma_X^2}{s_Y^2 / \sigma_Y^2} \right),$$

se distribuye según una distribución F de Snedecor con $(n-1)$ y $(m-1)$ grados de libertad: $F_{(n-1,m-1)}$. Esta es la distribución estadística que se debe emplear para la prueba de hipótesis correspondiente.

1. Formulación de hipótesis:

	Pruebas unilaterales		Prueba bilateral
1.1	$H_o : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 > \sigma_2^2$	1.2 $H_o : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 < \sigma_2^2$	1.3 $H_o : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$

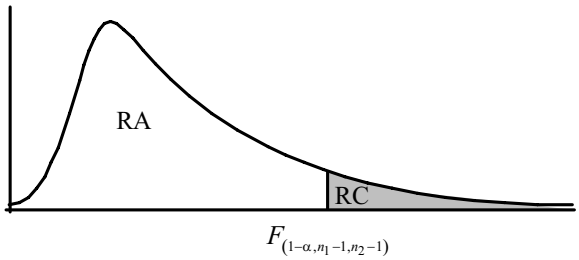
2. Elegir α.

3. Estadística de prueba: $F_0 = \dfrac{s_1^2}{s_2^2} \sim F_{(n_1-1;\, n_2-1)}$ y los datos tienen una distribución Normal.

4. Región crítica (RC) y regla de decisión.

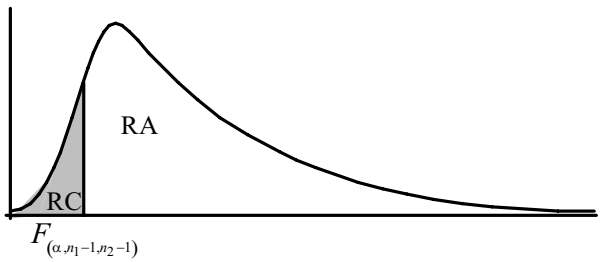
4.1 $\mathrm{RC} = \langle F_{(1-\alpha, n_1-1, n_2-1)}, \infty \rangle$

Rechazar $\mathrm{H_o}$ si $F_0 \in \mathrm{RC}$, es decir, si: $F_0 > F_{(1-\alpha, n_1-1, n_2-1)}$



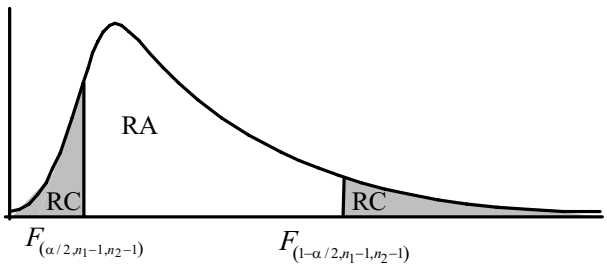
4.2 $\mathrm{RC} = \langle -\infty, F_{(\alpha, n_1-1, n_2-1)} \rangle$

Rechazar $\mathrm{H_o}$ si $F_0 \in \mathrm{RC}$, es decir, si $F_0 < F_{(\alpha, n_1-1, n_2-1)}$



4.3 $\mathrm{RC} = \langle -\infty, F_{(\alpha/2, n_1-1, n_2-1)} \rangle \cup \langle F_{(1-\alpha/2, n_1-1, n_2-1)}, \infty \rangle$

Rechazar $\mathrm{H_o}$ si $F_0 \in \mathrm{RC}$, es decir, si $F_0 < F_{(\alpha/2, n_1-1, n_2-1)}$ ó $F_0 < F_{(1-\alpha/2, n_1-1, n_2-1)}$



5. Se determina el valor de la estadística de prueba usando la información muestral.
6. Decisión y conclusión.

Ejemplo 7:

A los alumnos ingresantes a una universidad se les aplicó un examen escrito acerca del uso de un determinado programa informático que utilizaron en quinto de secundaria. Los alumnos se juntaron en dos grupos: los que habían usado y los que no habían usado el mencionado programa. Luego se aplicó una prueba a cada grupo y se obtuvieron los siguientes resultados:

Con conocimiento del programa	18, 15, 13, 12, 18, 16, 20, 18, 16
Sin conocimiento previo del programa	16, 17, 14, 11, 20, 18, 15, 14, 13

Pruebe la hipótesis de que las variabilidades en ambos grupos son iguales con un nivel del 5%.

Solución:

1. Formulación de la hipótesis:

   $\mathrm{H_o}: \sigma_1^2 = \sigma_2^2$ (la homogeneidad de ambos grupos es la misma)

   $\mathrm{H_1}: \sigma_1^2 \neq \sigma_2^2$ (la homogeneidad de ambos grupos no es la misma)
2. $\alpha = 0.05$ y las notas son independientes con distribución Normal.
3. La estadística de prueba es: $F_0 = \dfrac{s_1^2}{s_2^2} \sim F_{(9-1,9-1)}$
4. Región crítica:

   $\mathrm{RC} = \langle -\infty, F_{(0.05/2,9-1,9-1)} \rangle \cup \langle F_{(1-0.05/2,9-1,9-1)}, \infty \rangle = \langle -\infty, 0.226 \rangle \cup \langle 4.433, \infty \rangle$,

   rechazar $\mathrm{H_o}$ si $F_0 \in \mathrm{RC}$
5. Valor de la estadística de prueba: $F_0 = \dfrac{6.694}{7.5} = 0.8925$
6. Decisión y conclusión: Como $F_0 \notin \mathrm{RC}$, por lo tanto se concluye que no hay evidencia estadística para rechazar la $\mathrm{H_o}$, es decir, la variabilidad de ambos grupos es la misma.

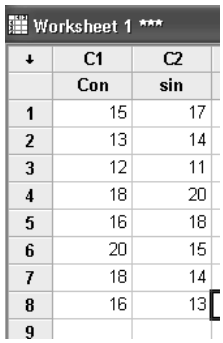
El procedimiento con el Minitab es el siguiente:

1. Formulación de la hipótesis:

   $\mathrm{H_o}: \sigma_1^2 = \sigma_2^2$ (la homogeneidad de ambos grupos es la misma)

   $\mathrm{H_1}: \sigma_1^2 \neq \sigma_2^2$ (la homogeneidad de ambos grupos no es la misma)

a. Se ingresan los datos en la hoja de cálculo tal como se muestra en la figura 9.



Worksheet 1 ***

	C1	C2
	Con	sin
1	15	17
2	13	14
3	12	11
4	18	20
5	16	18
6	20	15
7	18	14
8	16	13
9		

Figura 9. Ingreso de datos.

b. Ingresar a las opciones Stat / Basic statistics / 2 variances… tal como se presenta en la figura 10:

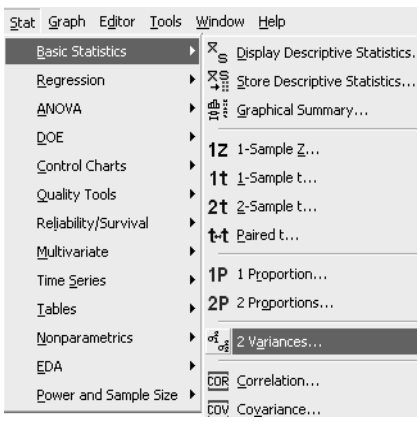


Figura 10. Opción 2 variances.

c. Al ingresar a esta opción aparece la siguiente ventana de la figura 11, en ella se seleccionan las dos columnas que contienen la información requerida.

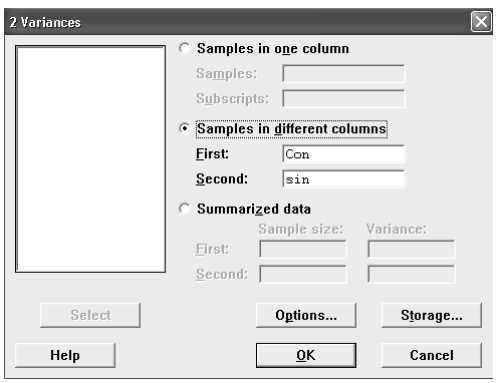


Figura 11. Ingreso de la información requerida.

d. Luego de dar <Ok> aparece la ventana de la figura 12.

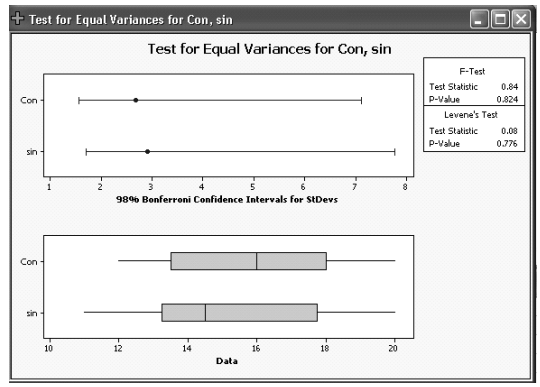


Figura 12. Resultados de la prueba.

En la parte superior de la derecha se encuentra el valor del P-value = 0.824 (F-test); por lo tanto, no se rechaza H_0 y se concluye que los datos muestrales son consistentes con la H_0 y no hay evidencia para rechazarla.

Nota: La opción de prueba de varianzas en el Minitab solamente permite hacer la prueba de dos colas.

6.5 Prueba de hipótesis para la diferencia de dos medias

Se considerará el caso de una prueba de hipótesis para comparar las medias poblacionales de dos variables X_1 y X_2, las cuales pueden presentar o no una distribución Normal. En el caso de que no lo sean se aplicará el Teorema del Límite Central en la resolución de los problemas.

6.5.1 Varianzas conocidas y muestras independientes

Cuando se dispone de dos variables X e Y, que se distribuyen normalmente y sus varianzas σ_X^2 y σ_Y^2 son conocidas; entonces, se puede desear contrastar hipótesis para las diferencias de las medias poblacionales de ambas variables. Analizando muestras aleatorias independientes de ambas poblaciones: $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_m$; se pueden calcular sus medias muestrales: $\overline{x}$ e $\overline{y}$. Además, se sabe que la siguiente expresión:

$$\left(\frac{(\overline{x}-\overline{y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \right),$$

se distribuye según una normal con media cero (0) y varianza igual a la unidad (1): $N(0,1)$. Esta es la distribución estadística que se debe emplear para la prueba de hipótesis correspondiente.

Suposición básica: X_1 y X_2 pueden ser Normales o no. En caso de que no sean distribuciones normales se aplica el teorema de límite central.

1. Formulación de las hipótesis:

	Pruebas unilaterales		Prueba bilateral
1.1	$H_0 : \mu_1 \leq \mu_2$	1.2 $H_0 : \mu_1 \geq \mu_2$	1.3 $H_0 : \mu_1 = \mu_2$
	$H_1 : \mu_1 > \mu_2$	$H_1 : \mu_1 < \mu_2$	$H_1 : \mu_1 \neq \mu_2$

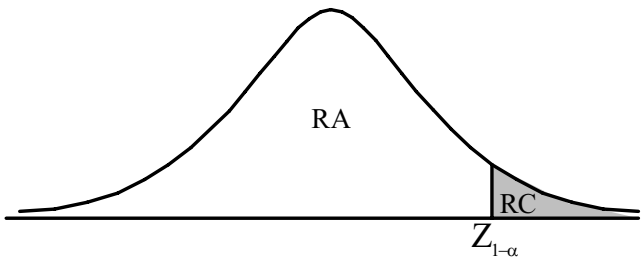
2. Elegir α.

3. Estadística de prueba: $Z_0 = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$ e indicar la distribución de los datos.

4. Región crítica (RC) y regla de decisión.

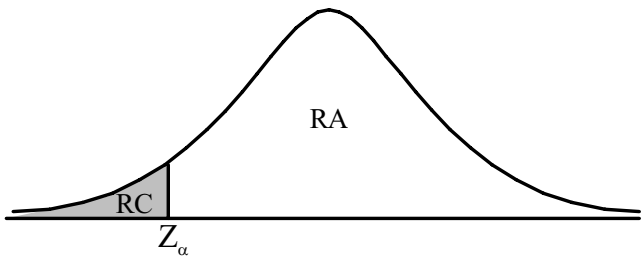
   4.1 $RC = \langle Z_{(1-\alpha)}, \infty \rangle$

   Rechazar H_o si $Z_o \in RC_o$, es decir, si: $Z_o > Z_{(1-\alpha)}$



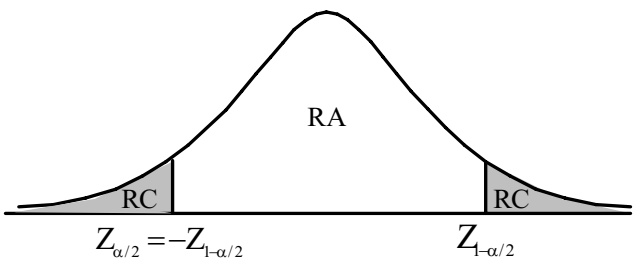
   4.2 $RC = \langle -\infty, Z_{(\alpha)} \rangle$

   Rechazar H_o si $Z_o \in RC_o$, es decir, si: $Z_o < Z_{(\alpha)}$



   4.3 $RC = \langle -\infty, Z_{(\alpha/2)} \rangle \cup \langle Z_{(1-\alpha/2)}, \infty \rangle$

   Rechazar H_o si $Z_o \in RC_o$, es decir, si: $Z_o < -Z_{(1-\alpha/2)}$ ó $Z_o > Z_{(1-\alpha/2)}$



5. Se determina el valor de la estadística de prueba usando la información muestral.
6. Decisión y conclusión.

6.5.2 Varianzas desconocidas y muestras independientes

Cuando se dispone de dos variables X e Y, que se distribuyen normalmente y sus varianzas σ_X^2 y σ_Y^2 son desconocidas; entonces se puede desear contrastar hipótesis para las diferencias de las medias poblacionales de ambas variables. Para esta situación existen dos posibles situaciones:

Varianzas desconocidas e iguales

Analizando muestras aleatorias independientes de ambas poblaciones: $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_m$; se pueden calcular sus medias y varianzas muestrales: $\bar{x}$, $\bar{y}$, s_X^2 y s_Y^2. Luego se calcula la siguiente expresión:

$$s^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

Además, se sabe que la siguiente expresión:

$$\left(\frac{(\bar{x}-\bar{y}) - (\mu_X - \mu_Y)}{\sqrt{s^2 \frac{n+m}{nm}}} \right),$$

se distribuye según una t de *Student* con $(n+m-2)$ grados de libertad. Esta es la distribución estadística que se debe emplear para la prueba de hipótesis correspondiente.

Varianzas desconocidas y diferentes

Si se analizan muestras aleatorias independientes de ambas poblaciones: $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_m$; se pueden calcular sus medias y varianzas muestrales: $\bar{x}$, $\bar{y}$, s_X^2 y s_Y^2. Luego se calcula la siguiente expresión:

$$v = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m} \right)^2}{\frac{\left(\frac{s_X^2}{n} \right)}{n-1} + \frac{\left(\frac{s_Y^2}{m} \right)}{m-1}}$$

Además, se sabe que la siguiente expresión:

$$\left(\frac{(\bar{x}-\bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \right),$$

se distribuye según una t de Student con (v) grados de libertad. Esta es la distribución estadística que se debe emplear para la prueba de hipótesis correspondiente.

1. Formulación de hipótesis:

	Pruebas unilaterales		Prueba bilateral
1.1	$H_0 : \mu_1 \leq \mu_2$	1.2 $H_0 : \mu_1 \geq \mu_2$	1.3 $H_0 : \mu_1 = \mu_2$
	$H_1 : \mu_1 > \mu_2$	$H_1 : \mu_1 < \mu_2$	$H_1 : \mu_1 \neq \mu_2$

2. Elegir nivel de significación α.

3. Estadística de prueba:

Si las varianzas son desconocidas e iguales

$$(\sigma_1^2 = \sigma_2^2)$$

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

donde: $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$

Si las varianzas son desconocidas y diferentes

$$(\sigma_1^2 \neq \sigma_2^2)$$

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{(v)}$$

donde: $v = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1-1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2-1)}}$

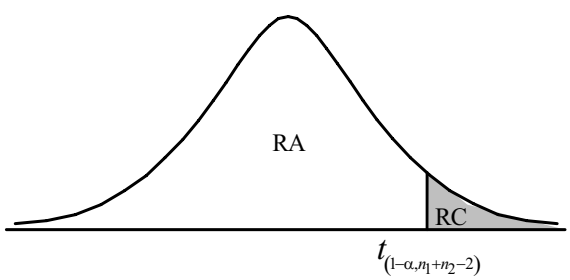
Los datos tienen distribución normal.

4. Región crítica (RC) y regla de decisión.

Si: $(\sigma_1^2 = \sigma_2^2)$, las regiones críticas son:

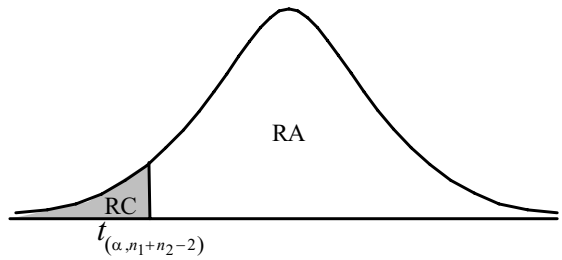
4.1a $RC = \langle t_{(1-\alpha, n_1+n_2-2)}, \infty \rangle$

Rechazar H_o si $t_o \in RC$, es decir, si: $t_o > t_{(1-\alpha, n_1+n_2-2)}$



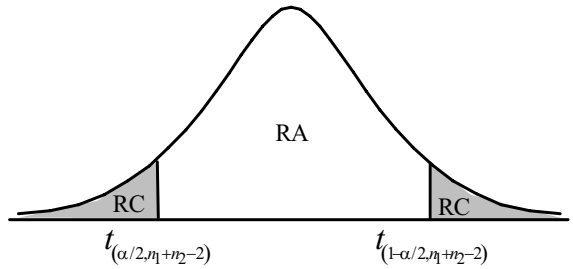
4.2 $\mathrm{RC} = \langle -\infty,\ t_{(\alpha, n_1+n_2-2)} \rangle$

Rechazar $\mathrm{H_o}$ si $t_o \in \mathrm{RC}$, es decir, si: $t_o < t_{(\alpha, n_1+n_2-2)}$



4.3 $\mathrm{RC} = \langle -\infty,\ t_{(\alpha/2, n_1+n_2-2)} \rangle \cup \langle t_{(1-\alpha/2, n_1+n_2-2)},\ \infty \rangle$

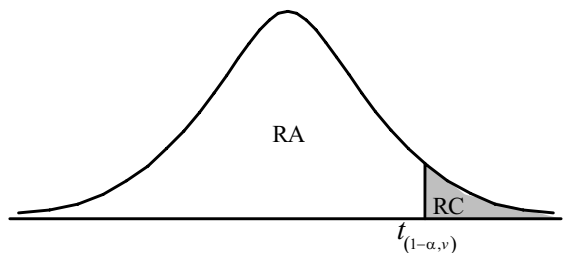
Rechazar $\mathrm{H_o}$ si $t_o \in \mathrm{RC}$, es decir, si: $t_0 < t_{(\alpha/2, n_1+n_2-2)}$ ó $t_0 > t_{(1-\alpha/2, n_1+n_2-2)}$



Si: $(\sigma_1^2 \neq \sigma_2^2)$, las regiones críticas son:

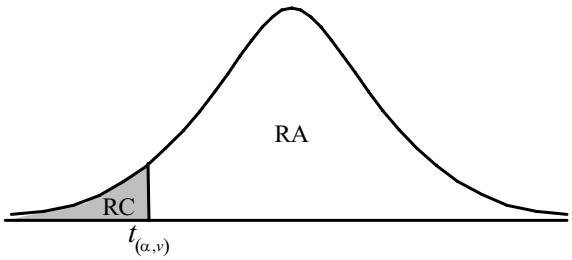
4.1 $\mathrm{RC} = \langle t_{(1-\alpha, v)},\ \infty \rangle$

Rechazar $\mathrm{H_o}$ si $t_o \in \mathrm{RC}$, es decir, si: $t_0 > t_{(1-\alpha, v)}$



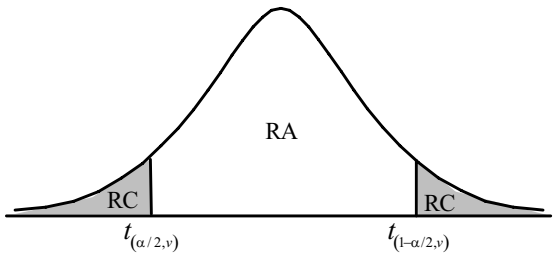
4.2 $\mathrm{RC} = \langle -\infty, t_{(\alpha,v)} \rangle$

Rechazar $\mathrm{H_o}$ si $t_o \in \mathrm{RC}$, es decir, si: $t_0 < t_{(\alpha,v)}$



4.3 $\mathrm{RC} = \langle -\infty,\ t_{(\alpha/2,v)} \rangle \cup \langle\ t_{(1-\alpha/2,v)}, \infty \rangle$

Rechazar $\mathrm{H_o}$ si $t_0 \in \mathrm{RC}$, es decir, si: $t_0 < t_{(\alpha/2,v)}$ ó $t_0 > t_{(1-\alpha/2,v)}$



5. Se determina el valor de la estadística de prueba usando la información muestral.
6. Decisión y conclusión.

Ejemplo 8:

El encargado de la sección de tarjetas de crédito desea saber si existe diferencia significativa en el promedio de saldos de tarjetas de crédito de dos sucursales. Se eligieron muestras aleatorias e independientes de cada sucursal, cuya información se presenta a continuación:

Sucursal	Saldos de las tarjetas de crédito		
	Tamaño muestra	Media	Varianza
1	20	$550	400
2	20	$570	324

Con un nivel de significación del 5%, ¿existe diferencia significativa entre ambos promedios?

Solución:

1. Formulación de la hipótesis:

   $H_0 : \mu_1 = \mu_2$ (no existe diferencia en el promedio de saldos de ambas sucursales)

   $H_1 : \mu_1 \neq \mu_2$ (sí existe diferencia en el promedio de saldos de ambas sucursales)

2. $\alpha = 0.05$ y los datos tienen distribución Normal.
3. Para determinar la estadística de prueba por usar, primero se debe hacer una prueba de igualdad de varianzas. En efecto:

   a. $H_0 : \sigma_1^2 = \sigma_2^2$

      $H_1 : \sigma_1^2 \neq \sigma_2^2$

   b. $\alpha = 0.05$

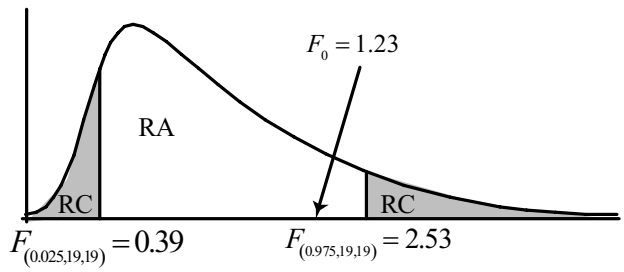
   c. La estadística de prueba es: $F_0 = \frac{s_1^2}{s_2^2} \sim F_{(20-1,20-1)}$

   d. Región crítica:

      $RC = \langle -\infty, F_{(0.05/2,19,19)} \rangle \cup \langle F_{(1-0.05/2,19,19)}, \infty \rangle = \langle -\infty, 0.3958 \rangle \cup \langle 2.5264, \infty \rangle$

      rechazar H_o si $F_0 \in RC$

   e. Valor de la estadística de prueba: $F_0 = \frac{s_1^2}{s_2^2} = \frac{400}{324} = 1.2345$



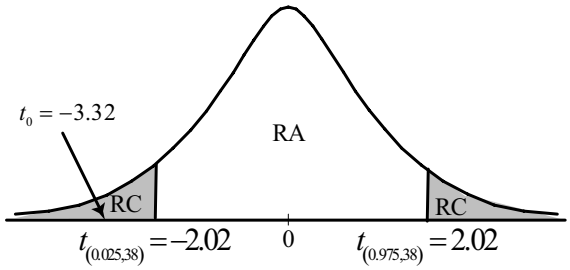
   f. Decisión y conclusión: como $F_0 = 1.2345 \notin RC$. Por lo tanto se concluye que no hay evidencia estadística suficiente para rechazar la hipótesis nula: $H_0 : \sigma_1^2 = \sigma_2^2$

      Entonces, se usará en la prueba: $t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{(n_1+n_2-2)}$

4. Región crítica: $RC = \langle -\infty, t_{(0.025,\,38)} \rangle = \langle -\infty, -2.0244 \rangle$ ó

   $\langle -\infty, t_{(0.975,\,38)} \rangle = \langle -\infty, 2.0244 \rangle$, rechazar H_o si $t_0 \in RC$

5. Valor de la estadística de prueba: $t_0 = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} = \dfrac{550-570}{\sqrt{s_p^2\left(\dfrac{1}{20}+\dfrac{1}{20}\right)}} = -3.32$



6. Decisión y conclusión: como $t_0 = -3.32 \in \text{RC}$, se concluye que los saldos de las tarjetas de crédito podrían ser distintas en ambas sucursales.

El procedimiento con Minitab es:

1. Formulación de la hipótesis:
   $\text{H}_0 : \mu_1 = \mu_2$ (no existe diferencia en el promedio de saldos de ambas sucursales)
   $\text{H}_1 : \mu_1 \neq \mu_2$ (sí existe diferencia en el promedio de saldos de ambas sucursales)

2. Como no se tienen los datos originales se sigue el mismo proceso que se empleó para construir un intervalo de confianza para la diferencia de medias. Aparecen las ventanas que se presentan en las figuras 13 y 14. (Nótese que ahora se activa el campo de <summarized data> para ingresar la información resumida, se activa el último campo para varianzas iguales.)

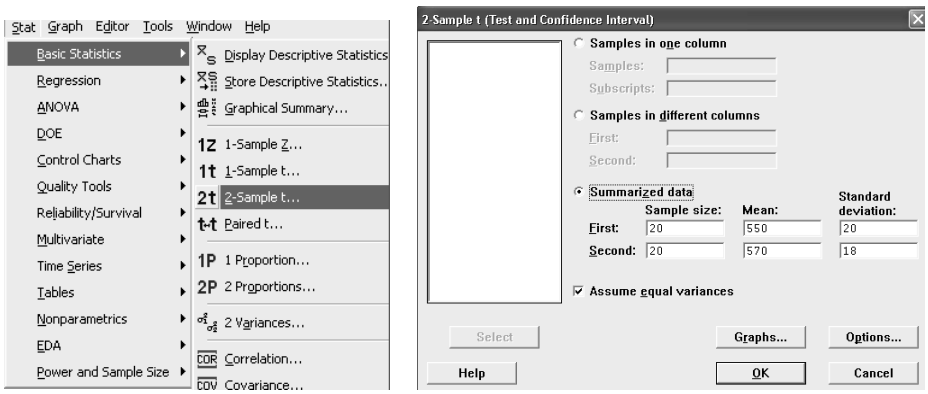


Figura 13. Opción 2-Sample t... e Ingreso de valores para la prueba.

Se activa <Options>, se escoge el tipo de prueba, en este caso de dos colas, en la ventana <Alternative> se escoge <not equal>, finalmente <Ok>.

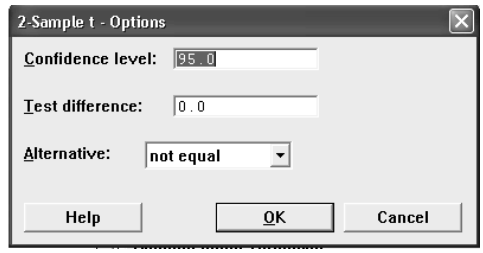


Figura 14. Elección del tipo de prueba.

Los resultados son los siguientes:

```
Two-Sample T-Test and CI
Sample      N      Mean      StDev     SE Mean
    1      20      550.0     20.0       4.5
    2      20      570.0     18.0       4.0
Difference = mu (1) - mu (2)
Estimate for difference:  -20.0000
95% CI for difference:  (-32.1801, -7.8199 )
T-Test of difference = 0 (vs not =):T-Value=-3.32 P-
Value=0.002 DF=38
Both use Pooled StDev = 19.0263
```

Como el P-value = 0.002 es menor que = 0.05, se concluye que el saldo de las tarjetas de crédito podría ser distinto en ambas sucursales.

Ejemplo 9:

El gerente de una entidad financiera sostiene que en la actualidad se realizan muchas operaciones no presenciales vía internet y que los promedios del número de operaciones van creciendo a través del tiempo. Se han elegido aleatoriamente cuatro sucursales y registrado el número de operaciones realizadas durante los meses de enero y abril. A un nivel de significancia del 5% probar las hipótesis correspondientes.
Nota: Suponga varianzas iguales.

Banca virtual	
Enero	Abril
1.250	2.000
1.000	1.500
900	950
1.750	1.550

Solución:

1. Formulación de la hipótesis:

$H_0: \mu_1 = \mu_2$ (las transacciones vía internet no crecieron)

$H_1: \mu_1 < \mu_2$ (las transacciones vía internet crecieron)

2. $\alpha = 0.05$.

3. La salida Minitab proporciona:

Two-Sample T-Test and CI: Enero, Abril

```
Two-sample T for Enero vs Abril
              N      Mean      StDev      SE Mean
Enero         4      1225       380        190
Abril         4      1500       430        215
Difference = mu (Enero) - mu (Abril)
Estimate for difference:  - 275.000
95% CI for difference:  282
T-Test of difference = 0 (vs <): T-Value = -0.96
P-Value = 0.187 DF = 6
Both use Pooled StDev = 405.6887
```

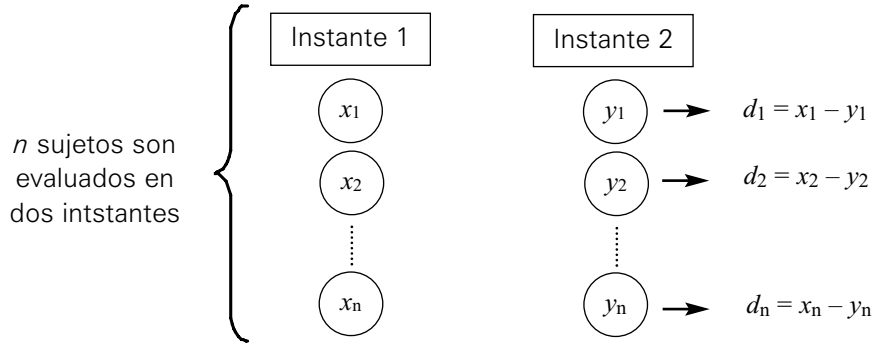
El P-value = 0.187 es mayor que $\alpha = 0.05$, entonces se concluye que no hay evidencia estadística para rechazar la H_0. El gerente de la entidad financiera se equivoca al indicar que el promedio del número de operaciones ha crecido a través del tiempo.

6.5.3 *Muestras pareadas o dependientes*

Uno de los supuestos sobre los que habitualmente se fundamentan las pruebas estadísticas de contraste es que las observaciones pertenecientes a cada una de las muestras son independientes entre sí; siendo precisamente ese uno de los objetivos de la elección aleatoria de los sujetos o unidades de observación. Sin embargo, la falta de independencia entre las observaciones de los grupos puede ser una característica de diseño del estudio para buscar una mayor eficiencia del contraste estadístico al disminuir la variabilidad. Lo que se busca con el diseño pareado es dar una mayor validez a las inferencias obtenidas, controlando o eliminando la influencia de variables extrañas cuyo efecto ya es conocido o sospechado, y se desea que no intervenga en el estudio actual, pudiendo enmascarar el efecto del tratamiento o de la variable de interés.

Si se desea comparar un resultado cuantitativo en dos grupos de datos, a partir de muestras X e Y extraídas en forma aleatoria de una población Normal, siendo n el tamaño de la muestra X, y n el de la muestra Y: $x_1, x_2, ..., x_n$ y $y_1, y_2, ..., y_n$. Suponga que las poblaciones tienen distribución normal con medias (μ_X y μ_Y) y varianzas (σ_X^2 y σ_Y^2) desconocidas entonces se pueden calcular sus medias y varianzas muestrales: $\bar{x}$, $\bar{y}$, s_X^2 y s_Y^2. Las muestras son de la misma población y son dependientes porque el mismo individuo genera un valor x_i y un valor y_i.

La metodología de datos pareados es desarrollada para comparar el comportamiento de una variable o factor en dos instantes y/o contextos, y con el propósito de medir los cambios ocurridos.



Sea la variable D = X – Y, la cual también presenta una distribución Normal, entonces se calcula su media muestral $\overline{d}$ y su varianza muestral s_D^2.

Luego, la siguiente expresión:

$$\left(\frac{\overline{d} - (\mu_X - \mu_Y)}{\frac{s_D}{\sqrt{n}}} \right),$$

se distribuye según una *t* de Student con $(n-1)$ grados de libertad. Esta es la distribución estadística que se debe emplear para la prueba de hipótesis correspondiente.

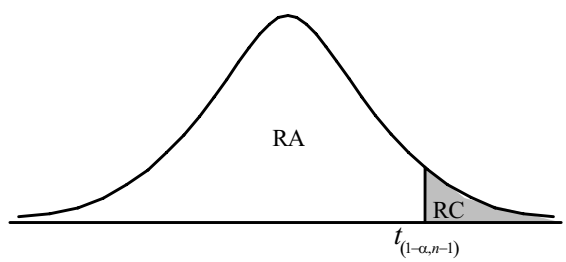
1. Formulación de las hipótesis, donde: $\mu_d = \mu_1 - \mu_2$

	Pruebas unilaterales		Prueba bilateral
1.1	$H_0 : \mu_d \leq 0$ $H_0 : \mu_d > 0$	1.2 $H_0 : \mu_d \geq 0$ $H_0 : \mu_d < 0$	1.3 $H_0 : \mu_d = 0$ $H_0 : \mu_d \neq 0$

Nota: En caso $\mu_d = \mu_2 - \mu_1$, las desigualdades anteriores pueden cambiar de acuerdo con el problema.

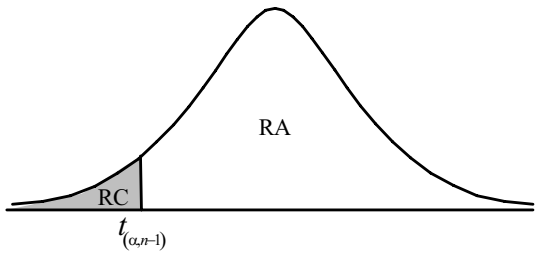
2. Elegir α.
3. La estadística de prueba: $t_0 = \dfrac{\overline{d}}{s_d / \sqrt{n}} \sim t_{(n-1)}$ y los datos deben tener una distribución normal.
4. Región crítica (RC) y regla de decisión.
   4.1 $RC = \langle t_{(1-\alpha,n-1)}, \infty \rangle$

   Rechazar H_o si $t_o \in RC$, es decir, si: $t_0 > t_{(1-\alpha,n-1)}$



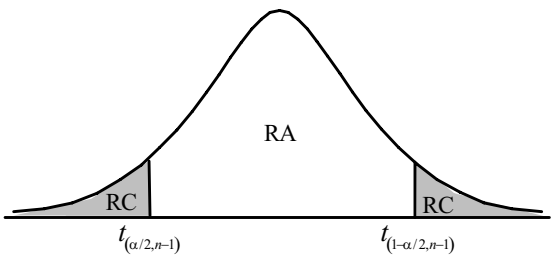
4.2 $\text{RC} = \langle -\infty, t_{(\alpha, n-1)} \rangle$

Rechazar H_0 si $t_o \in \text{RC}$, es decir, si: $t_0 < t_{(\alpha, n-1)}$



4.3 $\text{RC} = \langle -\infty, t_{(\alpha/2, n-1)} \rangle \cup \langle t_{(1-\alpha/2, n-1)}, \infty, \rangle$

Rechazar H_0 si $t_o \in \text{RC}$, es decir, si: $t_0 < t_{(\alpha/2, n-1)}$ ó $t_0 > t_{(1-\alpha/2, n-1)}$



5. Se determina el valor de la estadística de prueba usando la información muestral.
6. Decisión y conclusión.

Ejemplo 10:

Un club de salud ha estado anunciando un riguroso programa de acondicionamiento físico. El club sostiene que al cabo de un mes en el programa, el participante en promedio será capaz de hacer más planchas en dos minutos de las que podía hacer al inicio del programa. ¿Apoya esta afirmación del club la muestra aleatoria de 10 participantes, cuya información se presenta a continuación? Use $\alpha = 0.025$.

Participante	1	2	3	4	5	6	7	8	9	10
Número de planchas antes del programa	38	11	34	25	17	38	12	27	32	29
Número de planchas después del programa	45	24	41	39	30	44	30	39	40	41

Solución:

1. Las hipótesis son: (donde $\mu_D = \mu_d - \mu_a$)

   $H_o : \mu_D = 0$ (lo que afirma el club no es cierto)

   $H_1 : \mu_D > 0$ (lo que afirma el club es cierto)

2. $\alpha = 0.025$ y el número de planchas se distribuyen normalmente. Supuesto: el número de planchas se distribuye normalmente.

3. La estadística de prueba es: $t_0 = \dfrac{\bar{d} - 0}{S_d / \sqrt{n}} \sim t_{(9)}$

4. Región crítica: $RC = \langle t_{(1-0.05,10-1)}, \infty \rangle = \langle 1.83311, \infty \rangle$, rechazar H_o si $t_o \in RC$

5. Valor de la estadística de prueba: $t_0 = \dfrac{11-0}{3.86/\sqrt{10}} = 9.01$

6. Decisión y conclusión: Como $t_o = 9.01 \in RC$; entonces se rechaza H_0 y se concluye que lo que afirma el club es cierto.

Haciendo uso del programa informático Minitab se tiene:

1. El primer paso es ingresar los datos en la ventana <Data>, tal como se muestran en la figura 15.

Worksheet 1 ***

↓	C1	C2
	Antes	Despues
1	38	45
2	11	24
3	34	41
4	25	39
5	17	30
6	38	44
7	12	30
8	27	39
9	32	40
10	29	41

Figura 15. Ingreso de datos para la prueba.

   a. Ingresar a las opciones Stat / Basic Statistics / Paired t… como se indica en la figura 16. Esta es la opción para muestras dependientes o pareadas.

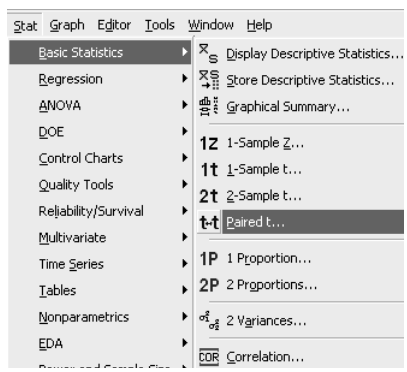


Figura 16. Secuencia de comandos.

b. Aparece la ventana <Paired t>. En el campo <samples in columns> se ingresan las dos columnas que contienen los datos; luego elegir <options> (véase la figura 17).

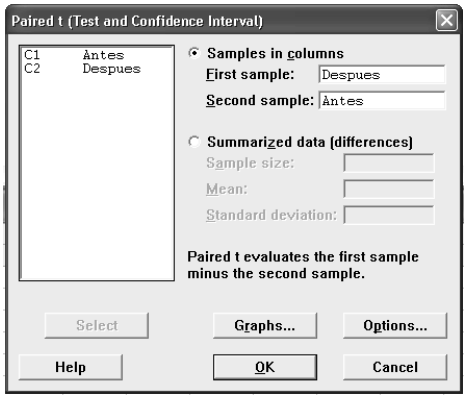


Figura 17. Ingreso de columnas que contienen los datos.

c. En el campo <Alternative> se escoge el tipo de prueba; como en este caso se trata de prueba de cola derecha se escoge *greater than*, finalmente elegir el botón <Ok>.

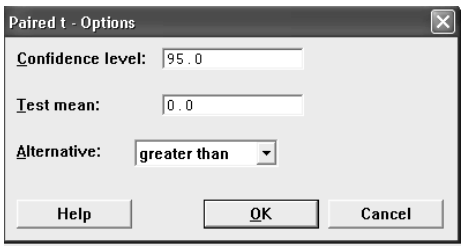


Figura 18. Elección del tipo de prueba.

d. Los resultados se presentan en la ventana de <Session> y son los siguientes:

Paired T-Test and CI: Despues, Antes

Paired T for Despues - Antes

	N	Mean	StDev	SE Mean
Despues	10	37.3000	6.8969	2.1810
Antes	10	26.3	10.0	3.16
Difference	10	11.0	3.86	1.22

95% lower bound for mean difference: 8.76

T-Test of mean difference = 0 (vs > 0):

T-Value = 9.01 P-Value = 0.000

Como el P-value es igual a cero, entonces se rechaza H_0 y se concluye que el programa de acondicionamiento físico es eficaz, es decir, el participante es capaz de hacer más planchas en dos minutos que al inicio del programa.

6.6 Prueba de hipótesis para la diferencia de dos proporciones

Suponga que se desea contrastar si dos proporciones de determinada categoría de una variable cualitativa procedentes de dos poblaciones diferentes X e Y: $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_m$ que se distribuyen según una Binomial: $X \sim B(n, \pi_X)$ e $Y \sim B(m, \pi_y)$, siendo desconocidas las proporciones π_X y π_Y. De las muestras se pueden calcular sus proporciones muestrales p_X y p_Y. Entonces se puede calcular la siguiente expresión:

$$p = \frac{np_X + mp_Y}{n+m}$$

Además, se sabe que la siguiente expresión:

$$\left(\frac{(p_X - p_Y) - (\pi_X - \pi_Y)}{\sqrt{p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)}} \right),$$

se distribuye aproximadamente según una Normal con media cero (0) y varianza igual a la unidad (1): $N(0,1)$. Siendo esta la distribución estadística que se debe emplear para la prueba de hipótesis correspondiente.

Nota: Se aplica a variables cualitativas o cuantitativas (previamente categorizadas).

1. Formulación de las hipótesis:

Pruebas unilaterales		Prueba bilateral
1.1 $H_0 : \pi_1 \leq \pi_2$	1.2 $H_0 : \pi_1 \geq \pi_2$	1.3 $H_0 : \pi_1 = \pi_2$
$H_1 : \pi_1 > \pi_2$	$H_1 : \pi_1 < \pi_2$	$H_1 : \pi_1 \neq \pi_2$

2. Elegir α.

3. La estadística de prueba: $Z_0 = \dfrac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \rightarrow N(0,1)$; donde $p = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \dfrac{x_1 + x_2}{n_1 + n_2}$.

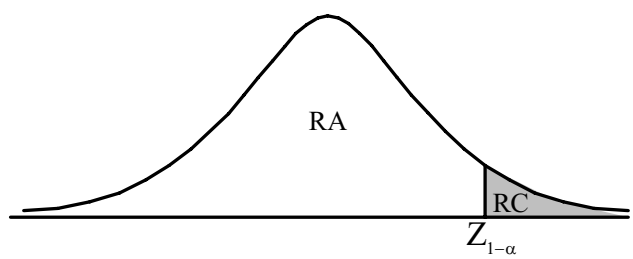
   Supuestos: la distribución de los datos es Binomial.

   Nota: Cuando se usa Minitab se debe marcar la opción <Use pooled estimate of p for test>, para que el programa haga uso del p estimado.

4. Región crítica (RC) y regla de decisión.

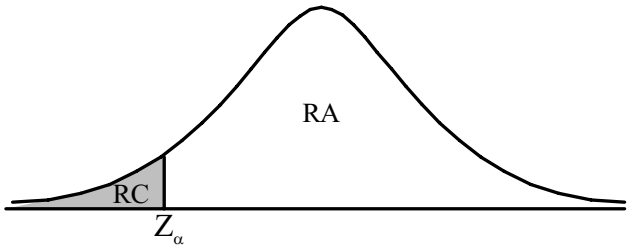
   1. $RC = \langle Z_{(1-\alpha)}, \infty \rangle$

      Rechazar H_0 si $Z_0 \in RC$, es decir, si: $Z_0 > Z_{(1-\alpha)}$



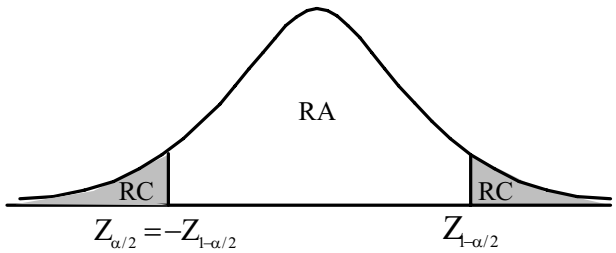
4.2 $\text{RC} = \langle -\infty, Z_{(\alpha)} \rangle$

Rechazar H_0 si $Z_0 \in \text{RC}$, es decir, si: $Z_0 < Z_{(\alpha)}$



4.3 $\text{RC} = \langle -\infty, Z_{(\alpha/2)} \rangle \cup \langle Z_{(1-\alpha/2)}, \infty \rangle$

Rechazar H_0 si $Z_0 \in \text{RC}$, es decir, si: $Z_0 < -Z_{(1-\alpha/2)}$ ó $Z_0 > Z_{(1-\alpha/2)}$



5. Se determina el valor de la estadística de prueba usando la información muestral.
6. Decisión y conclusión.

Ejemplo 11:

Una muestra aleatoria de 400 amas de casa seleccionadas por una organización de investigación de mercados indicó que el 20% prefería la marca de café C a todas las otras marcas. Después de una campaña intensiva a través de la prensa, la radio y la televisión, se seleccionó una segunda muestra, esta vez de 600 amas de casa, la cual dio un 22% de preferencia para la marca C. Si usted está dispuesto a rechazar la hipótesis nula en no más de una vez, en cada 10 veces que repita el experimento, ¿estaría en condiciones de afirmar que la campaña fue eficaz?

Solución:

π_1: Proporción poblacional de amas de casa que prefirió la marca de café C antes de la campaña publicitaria.

π_2: Proporción poblacional de amas de casa que prefirió la marca de café C después de la campaña publicitaria.

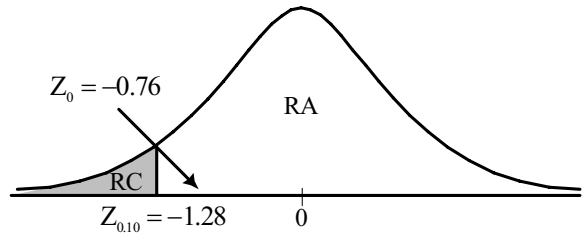
1. Formulación de la hipótesis:

$H_0 : \pi_1 \geq \pi_2$ (la campaña fue ineficaz)
$H_1 : \pi_1 < \pi_2$ (la campaña fue eficaz)

2. $\alpha = 0.10$.
3. La estadística es:

$$Z_0 = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \rightarrow N(0,1) \quad p = \frac{400(0.2) + 600(0.22)}{400 + 600} = 0.212$$ y los datos tienen distribución Binomial.

4. Región crítica: $RC = \langle -\infty, Z_{(0.10)} \rangle = \langle -\infty, -1.28155 \rangle$, rechazar H_0 si $Z_0 \in RC$
5. Valor de la estadística de prueba: $$Z_0 = \frac{0.20 - 0.22}{\sqrt{0.212(0.788)\left(\frac{1}{400} + \frac{1}{600}\right)}} = -0.7581$$



6. Decisión y conclusión: Como $Z_0 = -0.76 \notin RC$; por lo tanto no se rechaza H_0, es decir, la campaña fue ineficaz.

7. Funciones potencia y característica de operación

El análisis de la prueba de hipótesis que se ha presentado ha sido bajo el supuesto de que el tamaño de muestra y el nivel de significación α son fijos. La evaluación del error tipo II o β puede hacerse de dos maneras. La primera es evaluando a β con n y α fijos; este análisis lleva a definir las funciones potencia y característica de operación (OC) de una prueba. La segunda se orienta a determinar un tamaño de muestra óptimo que satisfaga niveles fijos de α y β.

Por lo que se ha tratado en los conceptos anteriores sobre hipótesis, se observa que solo una de las cuatro consecuencias siguientes puede suceder en una prueba dada (debido a que la decisión se toma sobre la base de la muestra):

1. Se puede rechazar H_0 siendo cierta. Probabilidad de tomar esta decisión incorrecta: α.
2. Se puede no rechazar H_0 siendo cierta. Probabilidad de tomar esta decisión correcta: $1 - \alpha$.

3. Se puede no rechazar H_0 siendo falsa. Probabilidad de tomar esta decisión incorrecta: β.
4. Se puede rechazar H_0 siendo falsa. Probabilidad de tomar esta decisión correcta: $1-\beta$.

La potencia de la prueba se define como la probabilidad de rechazar una H_0 falsa (o de aceptar una H_1 cierta), es decir $FP = 1-\beta$, depende del valor de β, así si se prueba:

$$H_0 = \theta = \theta_0$$
$$H_1 = \theta = \theta_1$$

La potencia de la $\text{prueba} = P(\hat{\theta} \in RC / \theta = \theta_1) = 1-\beta$, que es la probabilidad de rechazar H_0 cuando es falsa (H_1 es cierta).

La potencia de la prueba $1-\beta$ representa la sensibilidad de la prueba estadística para detectar cambios que se presentan al medir la probabilidad de rechazar la hipótesis nula cuando debería ser rechazada. La potencia de la prueba estadística depende de que tan diferente en realidad sea el valor verdadero del parámetro del valor supuesto. Además, como α y β tienen una relación inversa, y esta última es el complemento de la función potencia, entonces α y la función potencia varían en una proporción directa; es decir, si α aumenta entonces la potencia de la prueba también aumenta.

Para calcular la probabilidad de error tipo II o β se debe especificar la hipótesis alternativa como simple. Pero, en la mayoría de las situaciones, esta se plantea como compuesta. Al definirse una hipótesis alternativa como compuesta no se puede calcular la probabilidad de error tipo II asociado con esta prueba; para salvar esta dificultad lo que se hace es asignarle varios valores al parámetro de interés en la hipótesis alternativa y construir una curva con los valores de $1-\beta$ y los valores del parámetro, a esta curva se le llama función potencia.

Se define a la función característica de operación (OC) como la probabilidad de aceptar H_0; es decir, es el complemento de la función potencia.

Para entender mejor estos conceptos y ver el uso de las funciones potencia y característica de operación en pruebas de hipótesis, se tiene el siguiente ejemplo.

Ejemplo 12:

Suponga que un vendedor de baterías ha hecho un pedido de 5.000 unidades de este producto a un fabricante; pero el hecho de solicitar el pedido no significa que este debe ser aceptado literalmente, e indudablemente, se debe verificar si dichas baterías son buenas o no, para aceptar el pedido. El vendedor selecciona al azar 25 baterías y sobre la base de la prueba de esta muestra decidirá si acepta o no el pedido. Además, el fabricante sabe por experiencia que una batería es buena si en promedio dura 50 o más horas, Con $\sigma = 4$ y $\alpha = 0.05$, ¿qué decisión debe tomar el vendedor?

Para ayudar a decidir al vendedor se deben probar las siguientes hipótesis:

$$H_0 : \mu \geq 50 \text{ (aceptar el lote)}$$
$$H_1 : \mu < 50 \text{ (no aceptar el lote)}$$

Para esta prueba los riegos α y β son:

Acción	Población	
	H_0: $\mu \geq 50$ (H_0 verdadera)	H_1: $\mu < 50$ (H_1 verdadera)
Aceptar H_0	Decisión correcta	β = P(Aceptar $H_0/\mu < 50$)
No aceptar H_0	α = P(No aceptar $H_0/\mu \geq 50$)	Decisión correcta

Asumiendo que la duración de las baterías es Normal, naturalmente el estimador que se va a utilizar para realizar la prueba es la media de la muestra. Para prueba de cola izquierda, con $\alpha = 0.05$, la decisión será, rechazar H_0 si:

$Z_0 < Z_{0.05} \rightarrow \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < -1.64485 \rightarrow \bar{x} < \mu - 1.64485\left(\frac{\sigma}{\sqrt{n}}\right)$ y reemplazando valores:

$\bar{x} < 50 - 1.64485\left(\frac{4}{\sqrt{25}}\right) \Rightarrow \bar{x} < 48.68412$ horas.

La regla de decisión así establecida en términos de la media muestral es:

Si $\bar{x} \geq 48.68412$, entonces no se rechaza H_0, es decir, se debe aceptar el pedido.
Si $\bar{x} < 48.68412$, entonces no se rechaza H_1, por lo tanto, no se debe aceptar el pedido.

Ahora se analizarán las funciones potencia y característica de operación. Se sabe que para una prueba de cola izquierda, se rechaza H_0 si y solo si $\hat{\theta} < c$. Así, en nuestro ejemplo, si $\mu = 50$, entonces:

$FP(50) = P(\bar{x} < 48.68412 | \mu = 50)) = P(\bar{x} \leq 48.68412) = 0.05 = \alpha$; donde:

$\bar{x} \rightarrow N(50, 0.8^2)$;

Por lo tanto: $OC(50) = 1 - FP(50) = 1 - \alpha = 1 - 0.05 = 0.95$.

El siguiente cuadro contiene las funciones potencia y característica de operación para algunos valores escogidos del parámetro μ.
Valores de las curvas característica de operación y potencia para una prueba de cola superior.

Hipótesis verdadera	μ	FP(μk)	OC(μ)
H_1	48.0	0.80376 = 1–β	0.19624 = β
H_1	48.5	0.59101 = 1–β	0.40899 = β
H_1	49.0	0.34647 = 1–β	0.63533 = β
H_1	49.5	0.15389 = 1–β	0.84611 = β
H_0	50.0	0.05000 = α	0.95000 = 1–α
H_0	50.5	0.01161 = α	0.98839 = 1–α
H_0	51.0	0.00189 = α	0.99811 = 1– α

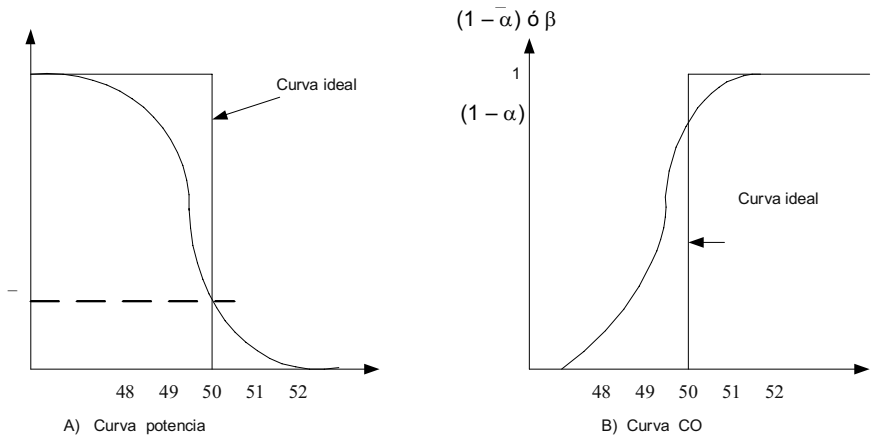


Figura 19. Gráfico de la curva potencia y curva característica de operación.

En la gráfica de la izquierda los valores de la curva ideal de la función potencia son uno en los valores cubiertos por H_1 y cero en los valores cubiertos por H_0, por lo tanto la curva potencia debe ser baja sobre H_0 y alta sobre H_1. Es decir, la potencia de la prueba aumenta cuando el valor del parámetro se desvía más y más de los valores cubiertos por H_0.

Una forma de verificar la idoneidad de la prueba es graficando la curva y ver que su gráfica esté cercana a la curva ideal.

La función característica de operación, como el complemento de la función potencia, debe ser alta sobre H_0 y baja sobre H_1. Idealmente, para una prueba de cola izquierda debe ser uno en la zona cubierta por H_0 y cero en la zona cubierta por H_1.

Para resumir la decisión establecida, basándose en la media muestral, se puede decir que el criterio de decisión es:

1. Tomar una muestra aleatoria de tamaño $n = 25$ y calcular la media de la muestra.
2. Fijar $\alpha = 0.05$
3. Regla de decisión.
   Si $\bar{x} \geq 48.68412$, entonces no se rechaza H_0, es decir, se debe aceptar el pedido.
   Si $\bar{x} < 48.68412$, entonces no se rechaza H_1, es decir, no se debe aceptar el pedido.

El Minitab permite hallar la potencia de la prueba, el procedimiento es el siguiente:

a. Ingresar a las opciones Stat / Power and sample size / 1 Sample Z... tal como se muestra en la figura 20.

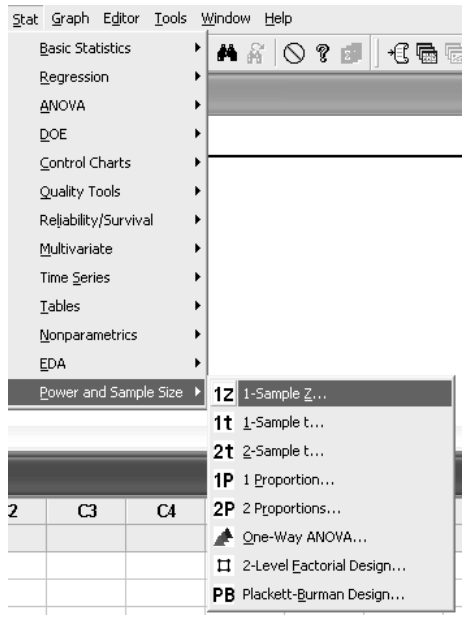


Figura 20. Opción 1 Sample Z...

Luego aparece la ventana de la figura 21, donde se ingresa el valor de n, $d = 0.5$, que se obtiene de restar 50.5 (valor de μ para el cual se quiere hallar la potencia de la prueba) del valor de μ definido en H_0, 50 en este caso.

El siguiente campo queda libre (allí debe ir el valor de la función potencia); luego se ingresa el valor de σ en el último campo, seguidamente dar clic en <Options>.

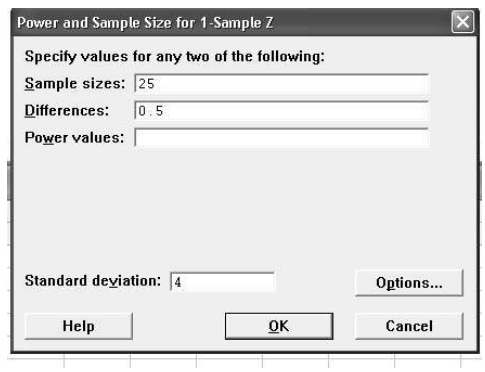


Figura 21. Ingreso de valores.

b. Aparece la ventana de la figura 22, donde se escoge el tipo de prueba, luego se ingresa el valor de α, y finalmente, si se desea que la respuesta la presente en la hoja de cálculo, se ingresan las columnas que deben contener las respuestas.

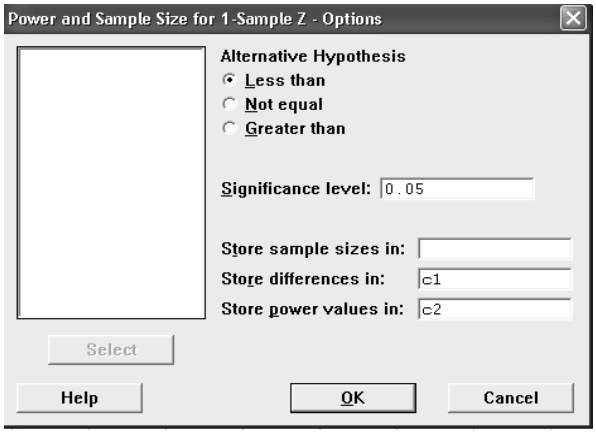


Figura 22. Elección del tipo de prueba.

Finalmente, aparece la respuesta (el valor de función potencia) como se muestra en la figura 23.

	C1	C2
	Mu	FP
1	0.5	0.0116082
2		

Figura 23. Resultados de la operación.

Nota: Esta opción permite determinar simultáneamente más de un valor de la función potencia, para ello se debe ingresar en el campo de <differences> el valor mínimo de d, el valor máximo y la diferencia entre uno y otro valor (Min:Max/diferencia).

c. Si se repite el paso 1 y en el campo de <differences> se ingresa –2:1/0.5 en vez de 0.5 los resultados serán como se muestra en la figura 24. Se usó la calculadora para obtener los valores de la diferencia a sus valores originales y calcular los valores de la curva OC.

	C1	C2	C3
	Mu	FP	OC
1	48.0	0.803765	0.196235
2	48.5	0.591011	0.408989
3	49.0	0.346475	0.653525
4	49.5	0.153899	0.846101
5	50.0	0.050000	0.950000
6	50.5	0.011608	0.988392
7	51.0	0.001897	0.998103

Figura 24. Resultados de la operación.

8. Prueba de bondad de ajuste

Suponga una empresa que tiene cuatro tiendas ubicadas en cuatro zonas de Lima metropolitana, el gerente de la empresa conoce que la distribución del porcentaje de sus clientes es 20%, 30%, 40% y 10% en sus tiendas, respectivamente; pero sospecha que debido a un cambio de imagen se ha producido una modificación en la distribución porcentual de sus clientes, ¿cómo se podrá verificar si se produjo este cambio o no? Este es un típico caso donde es útil usar la prueba de bondad de ajuste.

Cuando se toman decisiones en la vida real generalmente se necesita escoger cierta distribución de probabilidad para aproximar la distribución de los datos que se está analizando.

La prueba de bondad de ajuste sirve para probar si un conjunto de datos puede considerarse como una muestra tomada al azar de una población con una distribución específica (por ejemplo Normal, Poisson, etcétera, o alguna distribución en particular). La información obtenida de una muestra para hacer la prueba de bondad de ajuste se puede organizar en la siguiente tabla:

Tabla para una prueba de bondad de ajuste

Categoría	Frecuencia observada	P_i	Frecuencia esperada
A_1	O_1	p_1	E_1
.	.	.	.
.	.	.	.
.	.	.	.
A_k	O_k	p_k	E_k
Total	*n*		*n*

El procedimiento general para realizar la prueba es:

1. Formulación de las hipótesis.
   H_0: Los datos de la muestra se ajustan a la distribución teórica escogida.
   H_1: Los datos de la muestra no se ajustan a la distribución teórica escogida.
2. Se fija el nivel de significación α.
3. Calcular el valor de la estadística de prueba $\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(k-p-1)}$

   Donde:
   E_i= npi: valores esperados (teóricos)
   O_i: frecuencias observadas (reales)
   p: número de parámetros por estimarse a partir de los datos muestrales.
   k: número de categorías (clases).
   p_i: probabilidad de que la variable x pertenezca a la categoría A_i.
4. Determinar la región crítica y regla de decisión:

$RC = \langle \chi^2_{(1-\alpha,k-p-1)}, \infty \rangle$, se rechaza H_0 si: $\chi^2_0 > \chi^2_{(1-\alpha,k-p-1)}$, en caso contrario no se rechaza.

5. Decisión y conclusión: Rechazar H_0, si la estadística de prueba pertenece a la región crítica.

   Nota: Si alguna frecuencia esperada (E_i) es menor que cinco, se debe eliminar esa clase, y sumar la frecuencia observada a una clase contigua, esta restricción de la estadística de prueba se aplica para mejorar la aproximación de la estadística de prueba.

Ejemplo 13:

La fabricación de tuberías de acero requiere de una soldadura continua. Anteriormente se habían modelado los defectos en la soldadura de los tubos, mediante una distribución de Poisson con una media de tres defectos por tubo. Actualmente se está utilizando un nuevo tipo de máquina soldadora. Si en 100 soldaduras con este nuevo tipo de máquinas se obtienen los siguientes datos:

Número de defectos	0	1	2	3	4	5	6	7	8	9	10 o más
Frecuencia	5	14	16	20	18	17	3	2	4	0	1

¿La nueva máquina soldadora sigue produciendo soldaduras cuyo número de defectos tiene una distribución de Poisson con media 3?

Solución:

1. Formulación de la hipótesis:

   H_0 : El número de defectos se distribuye como Poisson con $\lambda = 3$.
   H_1 : El número de defectos no se distribuye como Poisson con $\lambda = 3$.

2. $\alpha = 0.05$

3. Cálculo de la estadística de prueba: $\chi^2 = \sum_{i=1}^{7} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(6)}$ y $\chi^2_{obs} = 7.288$

4. Región crítica: $RC = \langle \chi^2_{(1-0.05,7-0-1)}, \infty \rangle = \langle 15.592, \infty \rangle$, en este caso p = 0 porque ningún parámetro ha sido estimado; se rechaza H_0 si $\chi^2_0 \in RC$

5. Decisión y conclusión: Como $\chi^2_{obs} = 7.288 \notin RC$; por lo tanto no se rechaza H_0, es decir, que el número de defectos se distribuye como Poisson con $\lambda = 3$.

   Nota: Debido a que los E_i de $X = 6$ a más tuvieron frecuencias esperadas menores que 5, se juntaron en una sola clase (igual se hizo en la hoja de cálculo del Minitab). A continuación se presentan los cálculos, con ayuda del programa informático Minitab.

Nº de defectos (X_i)	Frecuencia (O_i)	P_i	E_i	$(O_i - E_i)^2/E_i$
0	5	0.050	5.00	0.000
1	14	0.149	14.9	0.054
2	16	0.224	22.4	1.829
3	20	0.224	22.4	0.257
4	18	0.168	16.8	0.086
5	17	0.101	10.1	4.714
6 o más	10	0.083	8.30	0.348
Total	100	1.000	100	7.288

Con ayuda del Minitab se procede de la siguiente manera:

a. En la hoja de cálculo del Minitab se ingresan en una columna los números del 0 al 10 (tal como aparece en la ventana de la izquierda), ingresar a la opción Calc / Probability / distributions / Poisson, aparece la ventana de la derecha, en ella se ingresan los datos tal como aparecen y luego se da <Ok> (véase la figura 25).

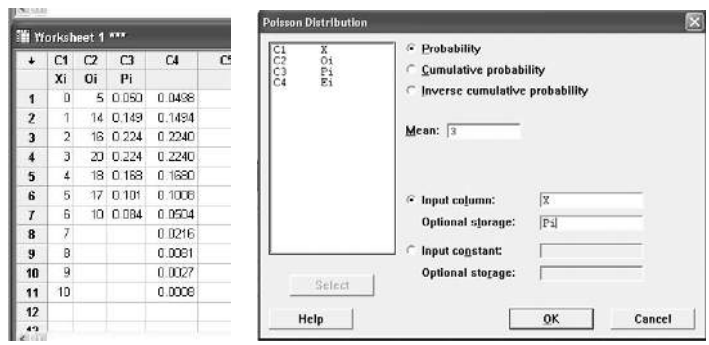


Figura 25. Ingreso de datos y Ventana de la distribución de Poisson.

b. Aparecerán las probabilidades individuales de los diferentes valores de X en la columna de C4, tal como se muestra en la ventana anterior de la izquierda. Luego Ingresar a las opciones Stat / Tables / Chi- Square goodnes-of –Fit (one variable)… (figura inferior de la izquierda), aparece la ventana de la derecha, en ella, en el campo de <observed counts>, se ingresa la columna de Oi y en el campo de <Specific proportions> se ingresa la columna de Pi, tal como se muestra en la figura 26.

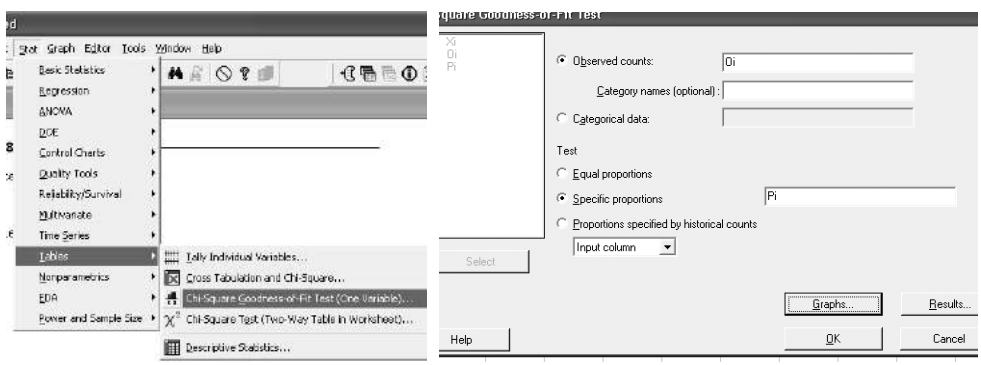


Figura 26. Comandos de Minitab para realizar la prueba de bondad de ajuste.

c. Luego de dar <Ok> aparecen los resultados siguientes:

```
Category  Observed  Test Proportion  Expected  Contribution
                                              to Chi-Sq
0            5        0.050            5.0       0.00000
1           14        0.149           14.9       0.05436
2           16        0.224           22.4       1.82857
3           20        0.224           22.4       0.25714
4           18        0.168           16.8       0.08571
5           17        0.101           10.1       4.71386
6           10        0.084            8.4       0.30476
             N        DF     Chi-Sq    P-Value
           100        6      7.24441   0.299
```

Como la estadística de prueba (7.24441) es menor que el valor observado (15.592) no se rechaza la hipótesis nula, es decir se concluye que el número de defectos se distribuye como Poisson con $\lambda = 3$.

Ejemplo 14:

Los siguientes datos se refieren a los montos en nuevos soles pagados en impuestos por un grupo de personas naturales seleccionadas al azar de entre todos los contribuyentes de quinta categoría de Lima.

Clase	Punto medio	Frecuencia
200-400	300	30
400-600	500	50
600-800	700	90
800-1.000	900	45
1.000-1.200	1.100	35

Pruebe si estos datos provienen de una distribución Normal con $\mu = 704$ y $\sigma = 240$.

Solución:

1. Las hipótesis por probar son:
   H_0: Los montos pagados en impuestos presentan una distribución Normal con parámetros $\mu = 704$ y $\sigma = 240$.
   H_1: Los montos pagados en impuestos no presentan una distribución Normal con parámetros $\mu = 704$ y $\sigma = 240$.
2. $\alpha = 0.01$

   Los datos fueron obtenidos al azar e independientemente.
3. La estadística de prueba: $\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(5-0-1)}$ (p = 0 porque ningún parámetro ha sido estimado).

4. Región crítica: $\mathrm{RC} = \langle \chi^2_{(1-0.01,5-0-1)}, \infty \rangle = \langle 13.27, \infty \rangle$, se rechaza H_0 si $\chi_0^2 \in \mathrm{RC}$
5. Valor de la estadística de prueba: usando el Excel para los cálculos (véase la figura 27).

Figura 27. Calcula de la estadística de prueba con Excel.

Lim inf	Lim sup	X_i	O_i	P(X<Lim sup)	p_i	E_i	$(E_i-O_i)^2/E_i$
200	400	300	30	0.10264	0.10264	25.65933	0.73429
400	600	500	50	0.33239	0.22975	57.43726	0.96301
600	800	700	90	0.65542	0.32303	80.75884	1.05746
800	1000	900	45	0.89127	0.23585	58.96313	3.30663
1000	1200	1100	35	1.00000	0.10873	27.18144	2.24865
Total		-	250		1	250	8.310341255

Se observa que el valor de la estadística de prueba es: $\chi^2 = 8.310341225$.

6. Decisión y conclusión: como $\chi_0^2 \notin \mathrm{RC}$, entonces los montos adeudados presentan una distribución Normal con parámetros $\mu = 704$ y $\sigma = 240$.

Para obtener la tabla anterior en Excel se procede de la siguiente manera:

a. Se ingresan los datos en las cuatro primeras columnas de la hoja de cálculo.
b. Para calcular las probabilidades acumuladas de la columna 5 se usa la distribución Normal, tal como se muestra en la ventana de la izquierda de la figura 28. Luego de elegir <Aceptar> aparece la ventana de la derecha, en ella se llenan los campos en la forma como se presenta a continuación.

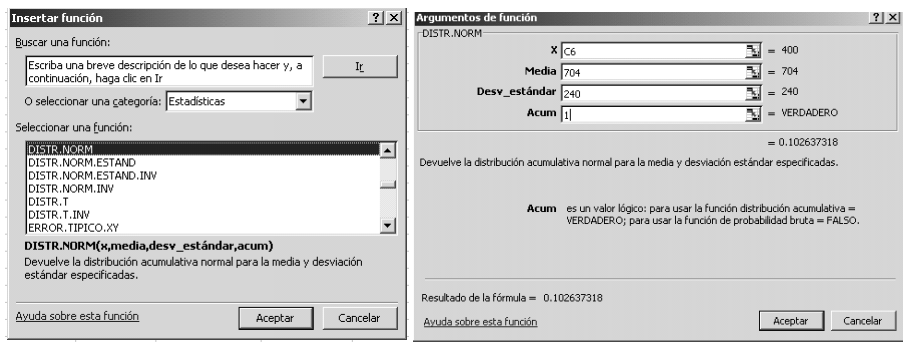


Figura 28. Funciones estadísticas de Excel.

c. En los campos se llenan primero el valor que se va a evaluar (x), después la media de la distribución Normal (Media), la desviación estándar de la distribución Normal (Desv_estandar), y por último, si la probabilidad va a ser exacta o acumulativa (0 o 1).
Una vez llenados los campos se selecciona el botón <Aceptar>.

9. Prueba de independencia

En algunas aplicaciones estadísticas el objetivo se centra en determinar si dos variables están relacionadas (estas pueden ser cuantitativas o cualitativas). Si las variables son cuantitativas continuas se puede utilizar el coeficiente de correlación muestral para realizar la prueba de hipótesis de independencia respectiva. Pero si son cualitativas (o cuantitativas categorizadas) se aplica la prueba de independencia que se presenta a continuación.

Por ejemplo, las siguientes situaciones pueden ser de interés:

¿Están relacionadas las enfermedades del corazón con el uso del tabaco?, ¿están relacionadas las calificaciones de los alumnos con el hábito de estudio?, ¿es independiente el sexo de una persona con la preferencia por el color?

La tabla que se presenta a continuación (tabulación de *n* objetos seleccionados al azar y clasificados de acuerdo con dos criterios diferentes), donde Oij representa el número de objetos pertenecientes a la celda (ij) de la tabla IxJ; puede usarse para probar la hipótesis de que las dos clasificaciones representadas por filas y columnas son estadísticamente independientes. La prueba exacta para la independencia es difícil de aplicar; sin embargo, si *n* es suficientemente grande, un procedimiento razonablemente bien aproximado consiste en calcular la estadística chi-cuadrado.

Tabla para realizar un aprueba de independencia

						Totales
	Y_j / X_i	Y_1	Y_2	. . .	Y_J	$O_{i.}$
	X_1	O_{11}	O_{12}	. . .	O_{1J}	$O_{1.}$
	. .	. .	. .	. . .	. .	. .
	X_I	O_{I1}	O_{I2}	. . .	O_{IJ}	$O_{I.}$
Totales	$O_{.j}$	$O_{.1}$	$O_{.2}$	. . .	$O_{.J}$	$O_{..} = n$

Los pasos para realizar una prueba de independencia son:

1. Formulación de las hipótesis.
   H_0: las variables *X* e *Y* son independientes (no tienen relación)
   H_1: Las variables *X* e *Y* no son independientes (tienen relación), es decir existe un grado de dependencia entre ambas variables.
   Una forma equivalente de formular las hipótesis es:

$$H_0: \pi_{ij} = \pi_i \text{ x } \pi_j \quad \forall\ i = 1,2,...,I; \quad \forall\ j = 1,2,...,J$$

$$H_1: \pi_{ij} \neq \pi_i \text{ x } \pi_j \quad \text{para algún (i, j)}$$

2. Se elige α.
3. Se calcula el valor de la estadística de prueba: $\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(I-1)(J-1)}$

Donde:
$E_{ij} = (O_{i.}O_{.j})/n$: número esperado de la celda (ij).
O_{ij}: Número de observaciones de la celda (ij).
$O_{i.}$: Número de observaciones de la fila i.
$O_{.j}$: Número de observaciones de la columna j.

4. Determinar la región crítica y regla de decisión:

   $\mathrm{RC} = \langle \chi^2_{(1-\alpha,(I-1)(J-1))}, \infty \rangle$, se rechaza H_0 si: $\chi^2_0 > \chi^2_{(1-\alpha,(I-1)(J-1))}$, en caso contrario no se rechaza.

5. Decisión y conclusión. Rechazar H_0, si la estadística de prueba pertenece a la región crítica.

   Nota 1: Si alguna frecuencia esperada (E_{ij}) es menor que cinco, se debe efectuar una corrección de la estadística de prueba de independencia, que consiste en eliminar aquella celda que contiene la frecuencia esperada menor a 5.

   Nota 2: La estadística de prueba definida en el paso 4 también es empleada cuando se prueba la hipótesis de que *k* poblaciones binomiales (prueba de homogeneidad de proporciones) tienen el mismo parámetro π, esto es:

   $$\mathrm{H}_0 : \pi_1 = \pi_2 = \ldots \ \pi_k = \pi$$

   $$\mathrm{H}_1 : \text{al menos dos son diferentes. } (\pi_i \neq \pi_j)$$

Ejemplo 15:

Un gerente preocupado por las ventas de su producto desea estudiar la relación existente entre la región de procedencia de los clientes y su decisión de comprar o no el producto. Para tal fin, selecciona una muestra aleatoria simple de tamaño 400, encontrando los siguientes resultados:

Región	Compra el producto	No compra el producto
Norte	40	56
Centro	58	88
Sur	30	38
Oriente	75	15

De acuerdo con lo presentado, responda a las siguientes preguntas:
a. Formule las hipótesis nula y alternativa adecuadas.
b. Calcule las frecuencias teóricas y el valor muestral de la estadística de prueba.
c. Con $\alpha = 0.05$, ¿se rechazará la hipótesis nula?

Solución:

a. H_0: La compra del producto es independiente de la región.
   H_1: La compra del producto depende de la región.

b. Usando Minitab se obtiene:

   Chi-Sq = 49.693. DF = 3, P-Value = 0.000.

   (El procedimiento se muestra en líneas posteriores)

c. Como P-value = 0.000 es menor que $\alpha = 0.05$, se rechaza H_0, y se concluye que la región y la decisión de compra no son independientes.

El procedimiento de la prueba de independencia con Minitab es:

a. Se ingresan los datos en la ventana <Data>, tal como se muestra en la ventana de la izquierda de la figura 29. Luego se ingresa a la opción Stat / Tables / Chi-square test… tal como se muestra en la ventana de la derecha.

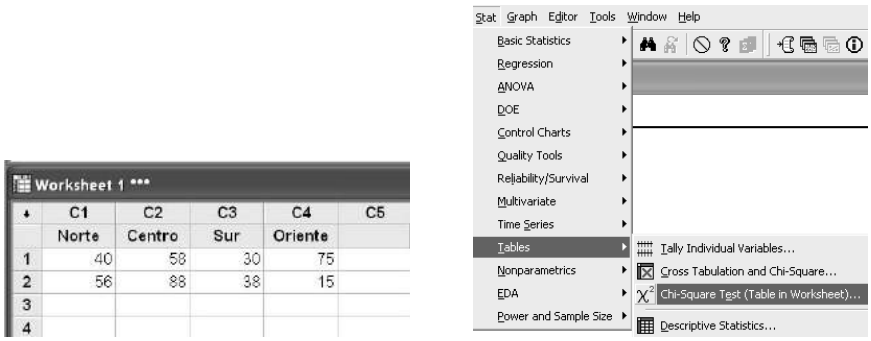


Figura 29. Ingreso de datos y Prueba Chi cuadrado.

b. Aparece la ventana de la figura 30, en el campo <Columns containing the table> se introducen las columnas que contienen los datos, luego dar <Ok>.

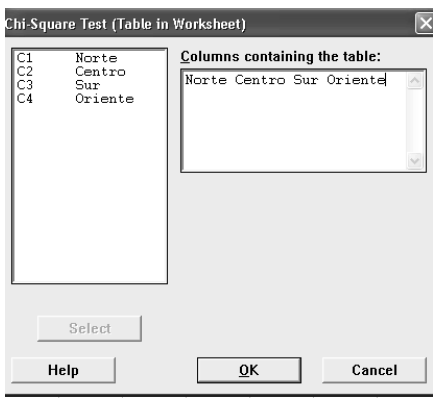


Figura 30. Ingreso de las columnas que contienen los datos.

c. Los resultados de la prueba se presentan a continuación:

Chi-Square Test: Norte, Centro, Sur, Oriente

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected
counts

       Norte    Centro      Sur Oriente     Total
  1       40        58       30      75       203
        48.72     74.09    34.51   45.67
        1.561     3.496    0.589  18.828

  2       56        88       38      15       197
        47.28     71.91    33.49   44.33
        1.608     3.603    0.607   9.401
Total     96       146       68      90       400
Chi-Sq = 49.693, DF = 3, P-Value = 0.000
```

Como P-value = 0.000 es menor que $\alpha = 0.05$, se rechaza H_0, y se concluye que la región y la decisión de compra no son independientes.

Ejemplo 16:

Mediante un nuevo proceso se preparan tres tipos de lubricantes. Cada uno de los lubricantes se prueba con cierto número de automóviles, y el resultado es clasificado como aceptable o inaceptable. Los datos se presentan en el siguiente cuadro:

	Lubricante 1	Lubricante 2	Lubricante 3
Aceptable	144	152	140
Inaceptable	56	48	60
Total	200	200	200

a. Con $\alpha = 0.05$, ¿se puede concluir que la proporción de resultados aceptables es distinta en los tres lubricantes?

b. Según los fabricantes, si la proporción de lubricantes aceptables es superior al 70%, el lubricante es introducido al mercado para su venta. Con $\alpha = 0.025$, ¿hay pruebas suficientes como para introducir al mercado el lubricante 1?

Solución:

a. 1. Las hipótesis son:
H_0: $\pi_1 = \pi_2 = \pi_3$ (la proporción de lubricantes aceptables es la misma)
H_1: al menos un π_i es diferente (la proporción de lubricantes aceptables es diferente)

2. $\alpha = 0.05$
3. La estadística de prueba es: $\chi_0^2 = \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(2)}$
4. Región crítica: $\text{RC} = \langle \chi^2_{(1-0.05,(2-1)(3-1))}, \infty\rangle = \langle 5.991, \infty\rangle$, rechazar H_0 si: $\chi_0^2 \in \text{RC}$
5. El valor de la estadística de prueba es: 1.880.

   Usando Minitab:

Chi-Square Test: LUBRICANTE 1, LUBRICANTE 2, LUBRICANTE 3

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts
      LUBRICANTE 1  LUBRICANTE 2   LUBRICANTE 3     Total
   1           144           152            140       436
            145.33        145.33         145.33
             0.012         0.306          0.196
   2            56            48             60       164
             54.67         54.67          54.67
             0.033         0.813          0.520
Total           200           200            200       600
Chi-Sq = 1.880, DF = 2, P-Value = 0.391
```

6. Decisión y conclusión: Como $\chi_0^2 \notin \text{RC}$, entonces no se rechaza H_0 y se concluye que la proporción de lubricantes aceptables es igual en los tres tipos de lubricantes.

b. 1. Las hipótesis son:

$\text{H}_0: \pi_1 = 0.7$ (no se introduce en el mercado el lubricante 1)

$\text{H}_1: \pi_1 > 0.7$ (se introduce en el mercado el lubricante 1)

2. $\alpha = 0.05$ y los datos tienen distribución Binomial.
3. La estadística de prueba es: $Z_0 = \frac{p_1 - \pi_1}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}}}$
4. Región crítica: $\text{RC} = \langle 1.645, \infty\rangle$, rechazar H_0 si: $\text{Z}_0 \in \text{RC}$
5. Valor de la estadística de prueba: $Z_0 = \frac{0.72 - 0.70}{\sqrt{\frac{0.70(1-0.70)}{200}}} = 0.617$ donde:

$p_1 = \frac{144}{200} = 0.72$

6. Decisión y conclusión: Como $\text{Z}_{0bs} \notin \text{RC}$, por lo tanto no se rechaza H_0 y el lubricante 1 no debe ser introducido en el mercado.

PROBLEMAS RESUELTOS

1. El director de una agencia publicitaria está preocupado por la efectividad de un anuncio en televisión. ¿Qué hipótesis nula está probando si comete:

a. Un error tipo I cuando afirma erróneamente que el comercial es efectivo.
b. Un error de tipo II cuando afirma erróneamente que el comercial es efectivo?

Solución:

a. H_0: El comercial no es efectivo.
b. H_0: El comercial es efectivo.

2. Considere el siguiente caso como una prueba de hipótesis. Se acaba de recibir un paracaídas sobre el cual un inspector postula la siguiente hipótesis nula: "este paracaídas funcionará".

a. Defina en términos del problema el error tipo I y el error tipo II.
b. ¿Cuál error es más grave?
c. Si se pudiesen controlar estadísticamente α y β, ¿qué conjunto de probabilidades preferiría el usuario del paracaídas?

(1) $\alpha = 0.001$ y $\beta = 0.10$ (2) $\alpha = 0.05$ y $\beta = 0.05$ (3) $\alpha = 0.10$ y $\beta = 0.001$

Solución:

a. ETI: Afirmar erróneamente que el paracaídas no funcionará.
ETII: Afirmar erróneamente que el paracaídas funcionará.
b. El error más grave sería el del tipo II.
c. (3).

3. Responda las siguientes preguntas:

a. El Congreso de la República aprobó la adenda al Tratado de Libre Comercio (TLC) con Estados Unidos. Si se supone que el TLC traerá consigo el desarrollo económico del país, ¿cuál es el tipo de error que está cometiendo un líder político al sostener que el TLC no es bueno para el país?

b. En una prueba de hipótesis, una disminución de la probabilidad de un tipo de error siempre resulta en _____ de la probabilidad del otro, siempre que no cambie el tamaño de la muestra.

c. Por sensibilidad se entiende la habilidad que tiene una prueba para

_____

Solución:

a. Error tipo I.
b. Un aumento.
c. Detectar diferencias.

4. La editorial ABC S.A. dedica su producción a la edición de textos universitarios y debe decidir si publica un libro de estadística escrito por cierto profesor. Con base en los costos de publicación, la editorial ha llegado a la siguiente conclusión: sí existe evidencia de que más del 15% de los centros superiores de enseñanza del país deciden usar este libro, entonces se publicará. Si no es así, entonces no se publicará. Se selecciona una muestra aleatoria de 100 centros superiores de todo el país.

a. Explique el significado de los errores tipo I y II de este problema.
b. Qué error sería más importante para la editorial y por qué; y para el profesor y por qué.
c. Si la muestra de 100 centros superiores señala que 25 consideran adoptar este texto, ¿debe publicarlo la editorial? Use $\alpha = 0.05$.

Solución:

a. ETI: Publicar el libro cuando menos del 15% de los centros superiores lo pedirán.
   ETII: No publicarlo cuando el 15% o más lo solicitarán.

b. Para la editorial el error más importante es el ETI, porque al no ser demandado el libro pierde parte de su inversión. Para el profesor el error más importante es el ETII.

c. Sobre la base de una prueba de hipótesis se tiene:

   1. $H_0 : \pi = 0.15$
      $H_1 : \pi > 0.15$

   2. $\alpha = 0.05$ y los datos tienen distribución Binomial.

   3. $RC = \langle 1.645, \infty \rangle$

   4. $Z_0 = 2.8$

   5. Como $Z_0 \in RC$, entonces se rechaza H_0 y se debe publicar el libro.

5. Se está estudiando un nuevo fármaco para utilizarlo en el tratamiento del cáncer de piel. Se espera que sea eficaz en la mayoría de los pacientes sobre los que se aplica, la compañía que produce el fármaco quiere obtener alguna prueba estadística que apoye tal afirmación. Si se selecciona una muestra aleatoria de 35 pacientes y en 18 de ellos el fármaco fue eficaz, ¿existe suficiente evidencia de que el fármaco es eficaz en la mayoría de los pacientes?

Solución:

Sea: proporción de pacientes que se curan con el fármaco.

1. Las hipótesis son:

   $H_0 : \pi = 0.5$ (el fármaco no es eficaz en la mayoría de los pacientes)
   $H_1 : \pi > 0.5$ (el fármaco es eficaz en la mayoría de los pacientes)

2. $\alpha = 0.05$ y los datos tienen distribución Binomial.

3. $\text{RC} = \langle 1.64485, \infty \rangle$

4. $Z_0 = 0.169$.

5. Como Z_0 pertenece a la región de aceptación, entonces se concluye que el fármaco no es eficaz en la mayoría de los pacientes.

6. Sea $X_1, X_2, ..., X_n$ una muestra aleatoria extraída de una población $N(\mu, \sigma^2)$.

a. Suponiendo que se desea probar: $H_0 : \mu = 3.5$ vs $H_1 : \mu = 5$.

Para $n = 4$, $\sigma = 1$, $\alpha = 0.05$; se rechaza H_0 si $\bar{x} > C$ Calcular C.

b. Sean las hipótesis $H_0 : \sigma^2 = 2$ vs. $H_1 : \sigma^2 = 3$. Para $n = 8$, se rechaza H_0, sí:

$\sum_{i=1}^{8} (x_i - \bar{x})^2 > 28,134$. Hallar:

b.1 P(ETI).
b.2 La potencia de la prueba.

Solución:

a. Como la prueba es de cola superior, entonces, $\text{RC} = \langle Z_{0.95}, \infty \rangle = \langle 1.645, \infty \rangle$:

Rechazar $H_0 = \dfrac{\bar{x} - 3.5}{1/2} > 1.645 \Rightarrow$ Rechazar $H_0 : \bar{x} > 4.3225$ (1). Rechazar

H_0 si $\bar{x} > C$ (2); igualando (1) con (2) entonces C = 4.3225.

b. b.1 $\alpha = P\left(\sum_{i=1}^{8} (x_i - \bar{x})^2 > 28{,}134 / \sigma^2 = 2 \right) = P(\chi^2_{(7)} > 14{,}067) = 0.05$

b.2 $1 - \beta = P\left(\sum_{i=1}^{8} (x_i - \bar{x})^2 > 28{,}134 / \sigma^2 = 3 \right) = P(\chi^2_{(7)} > 9{,}378) = 0.226$.

7. Para probar la hipótesis nula $H_0 : \mu = 50$; vs. $H_1 : \mu \neq 50$; se sabe que la desviación estándar es 18 y se extrae una muestra aleatoria de tamaño $n = 36$. Si se decide aceptar H_0 cuando $43 \leq \bar{x} \leq 57$:

a. Hallar la probabilidad de error tipo I.
b. Hallar β si realmente $\mu = 61$.

Solución:

a. P (ETI) = P(Rechazar H_0 / H_0 es Verdadero) =

$P(\bar{x} \leq 43) + P(\bar{x} > 57) = 0.0098153 + 0.0098153 = 0.0196306.$

donde: $\sigma = 18$; $n = 36$; $\bar{x} \rightarrow N(50; 3^2)$

b. P (ETII) = P(Aceptar $H_0 / \mu = 61$) =

$$P(43 \leq \bar{x} \leq 57) = P(\bar{x} \leq 57) - P(\bar{x} \leq 57) = 0.0912112 - 0 = 0.0912112.$$

donde: $\sigma = 18$; $n = 36$; $\bar{x} \rightarrow N(61; 3^2)$

8. Una industria lechera desea adquirir una máquina embotelladora y somete a consideración dos modelos distintos: el A y el B. El gerente de mercadeo consigue información de una muestra de 16 registros de leche embotellada por la máquina A y una muestra de 10 registros de leche embotellada por la máquina B. Los resultados de ambas muestras son:

$$\bar{x}_A = 5.1 \qquad \bar{x}_B = 4.2 \qquad s_A^2 = 0.027 \qquad s_B^2 = 0.065$$

Suponga que las máquinas tienen el mismo costo, por lo tanto se preferirá aquel modelo que cumpla los siguientes requisitos:

a. Tenga menor variabilidad en la cantidad embotellada; y
b. Tenga mayor capacidad de embotellamiento.

¿Hay suficiente evidencia para afirmar que la máquina A es preferible a la máquina B?

Use $\alpha = 0.05$.

Solución:

a. Se deben probar las hipótesis:

1. $H_0 : \sigma_A^2 = \sigma_B^2$ (la variabilidad de ambas máquinas es la misma)
   $H_1 : \sigma_A^2 < \sigma_B^2$ (la variabilidad de máquina A es menor)
2. $\alpha = 0.05$.
3. $F = \dfrac{s_A^2}{s_B^2} \sim F_{(15,9)}$
4. $RC = \langle -\infty, 0.386 \rangle$
5. $F_0 = 0.415$
6. $F_0 \in RA$, entonces no se rechaza H_0; A no cumple el primer requisito, es decir ambas máquinas tienen la misma variabilidad en la cantidad de leche embotellada con nivel del significación 0.05.

b. Se prueban las hipótesis:

1. $H_0 : \mu_A = \mu_B$
   $H_1 : \mu_A > \mu_B$
2. $\alpha = 0.05$ y las cantidades de leche embotellada tengan distribución Normal.
3. Para determinar la estadística de la prueba que se va a usar, primero se debe hacer una prueba de varianzas.
   - $H_0 : \sigma_A^2 = \sigma_B^2$
     $H_1 : \sigma_A^2 \neq \sigma_B^2$

- $\alpha = 0.05$.
- $\text{RC} = \langle -\infty, 0.3205\rangle \cup \langle 3.77, \infty\rangle$
- $F_{obs} = 0.415$ y $F \notin \text{RC}$, por lo tanto no se rechaza H_0 y se concluye que $\sigma_A^2 = \sigma_B^2$

  Por lo tanto se usa la prueba $t_0 = \dfrac{\overline{x}_A - \overline{x}_B}{s_{\overline{x}_A - \overline{x}_B}} \sim t_{(n_A + n_B - 2)}$

4. $\text{RC} = \langle t_{(0.05,24)}, \infty\rangle = \langle 1.711, \infty\rangle$.
5. $t_0 = 10.79$.
6. $t_0 \in \text{RC}$, entonces se rechaza H_0 y el modelo A cumple con el segundo requisito.

Conclusión: A pesar de que A cumple con el segundo requisito, se puede decir que no hay evidencia suficiente para afirmar que se debe preferir la máquina A.

9. Nueve distribuidores de equipos de cómputo fueron elegidos al azar y se les preguntó acerca de los precios de dos impresoras semejantes de inyección de tinta. Los resultados de esta encuesta se presentan a continuación. Con $\alpha = 0.01$ ¿es razonable afirmar que en promedio la impresora A cuesta menos que la impresora B? Suponga variaciones poblacionales desconocidas e iguales.

Distribuidor	1	2	3	4	5	6	7	8	9
Precio A	350	419	385	360	405	395	389	409	375
Precio B	370	425	369	375	389	385	395	425	400

Solución:

Aplicando prueba de hipótesis para decidir se tendría:

$$H_0 : \mu_A = \mu_B$$
$$H_1 : \mu_A < \mu_B$$

Utilizando Minitab se tiene:

	N	Mean	StDev	SE Mean
Precio A	9	387.4	22.8	7.6
Precio B	9	392.6	21.2	7.1

T-Test mu Precio A = mu Precio M (vs <): T = –0.49 P-value = 0.31
DF = 16
Comparando el P-value = 0.31 con $\alpha = 0.01$, no se rechaza H_0, y se concluye que la impresora A no cuesta menos que la impresora B.

10. Para comparar la cantidad de hierro se utilizan dos métodos: A y B. Para ello se toman 12 muestras de mineral y a cada muestra se le aplican ambos métodos. Los resultados del experimento son los siguientes:

Método A

38.2 31.6 26.6 41.2 44.8 46.3 35.4 38.4 42.6 46.7 29.2 30.7

Método B

38.2 31.7 26.6 41.3 44.8 46.4 35.5 38.4 42.7 46.8 29.2 30.8

Si se desea determinar si existe evidencia de que el método B proporciona un promedio de hierro más alto que el método A, y se define: $\mu_D = \mu_A - \mu_B$, se obtiene la siguiente salida con Minitab. Con $\alpha = 0.05$.

T-Test of the Mean

Test of μ_D = 0.0 vs μ_D < 0.0

Variable	N	Mean	StDev	SE Mean	T	P
d	12	- 0.0583	0.0515	0.0149	- 3.92	0.001

Conteste lo siguiente:

a. ¿Son las hipótesis adecuadas? Si no lo son, defínalas correctamente.
b. Usando el P-value, ¿cuál sería su conclusión si $\alpha = 0.05$?, ¿llegaría usted a la misma conclusión que cuando emplea $\alpha = 0.01$?
c. Establezca una regla de decisión en base a la media muestral de las diferencias "d", empleando $\alpha = 0.01$.
d. Enuncie, en términos del problema, en qué consisten los errores tipo I y tipo II.

Solución:

a. Sí, las hipótesis están correctamente definidas.

b. Con $\alpha = 0.05$, el P-value $< \alpha$, entonces, se rechaza H_0; con $\alpha = 0.01$, se acepta H_0.

c. $RC = \langle -\infty, t_{(0.01,11)} \rangle = \langle -\infty, -2.718 \rangle \Rightarrow \dfrac{\bar{d} - 0}{0.00772} < -2.718 \Rightarrow \bar{d} < -0.020983$.

Entonces la regla de decisión será:

1. Extraer una muestra aleatoria de 12 y aplicar los métodos A y B y luego calcular $\bar{d}$.
2. Si $\bar{d} < -0.020983$, se concluye que el método B produce en promedio más hierro que A, en caso contrario producen en promedio cantidades iguales.

d. ETI: Concluir que el método B produce más hierro que A, cuando producen igual.

ETII: Concluir que ambos métodos producen igual, cuando B produce más que A.

11. En el mercado hay dos fabricantes nacionales de ampolletas: A y B. Una muestra aleatoria de 16 ampolletas producidas por A arroja una duración con $\bar{x}_A = 1190$ horas y $s_A = 90$ horas. Por otro lado, una muestra aleatoria de 11 unidades producidas por B arrojan una duración de $\bar{x}_B = 1230$ horas y $s_B = 120$ horas. Luego de conocer los resultados, el gerente general de A afirma: "Tal vez la duración promedio de nuestras ampolletas sea levemente inferior pero en cambio la varianza es muchísimo menor". Si usted fuera dueño de la compañía B, ¿qué respondería y cómo lo justificaría?

Sugerencia. Aplique prueba de hipótesis, usando 5% de nivel de significación.

Solución:

a. Primero se prueba la afirmación de A, respecto a que su varianza es muchísimo menor.

1. $\mathrm{H}_0 : \sigma_A^2 = \sigma_B^2$
   $\mathrm{H}_1 : \sigma_A^2 < \sigma_B^2$
2. $\alpha = 0.05$ y las duraciones de ambas marcas de ampolletas se distribuyen normalmente.
3. $F_0 = \dfrac{s_A^2}{s_B^2} \sim F_{(15,10)}$
4. $\mathrm{RC} = \langle -\infty, 0.3937 \rangle$
5. $F_0 = \dfrac{90^2}{120^2} = 0.5625$
6. $F_0 \notin \mathrm{RC}$; por lo tanto no se rechaza H_0, entonces las varianzas pueden considerarse iguales. Es falso lo que dice el gerente de A.

b. Ahora se prueba que la media es también menor:

1. $\mathrm{H}_0 : \mu_\mathrm{A} = \mu_\mathrm{B}$
   $\mathrm{H}_1 : \mu_\mathrm{A} < \mu_\mathrm{B}$
2. $\alpha = 0.05$.
3. $t_0 = \dfrac{\bar{x}_A - \bar{x}_B}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{(25)}$, por a) $\sigma_A^2 = \sigma_B^2$, y las duraciones de ambas marcas de ampolletas se distribuyen normalmente.
4. $\mathrm{RC} = \langle -\infty, -1.708 \rangle$
5. $t_0 = \dfrac{1190 - 1230}{40.363} = -0.99099$
6. $t_0 \notin \mathrm{RC}$; por lo tanto no se rechaza H_0, se concluye que las medias son iguales con $\alpha = 0.05$, es decir el gerente de A no tiene razón.

12. La AFP "A" cuenta con 100.000 afiliados, que hacen sus aportaciones con regularidad. Se tomó una muestra de los aportes realizados por 1.500 afiliados durante el último trimestre, y se obtuvo los siguientes datos: $\bar{x} = \$850$ y $s = \$225$. Para el mismo período, y basado en una muestra de 1.200 afiliados, AFP "B" indica que el promedio de las aportaciones es de \$970, con $s = \$180$. Con esta información, ¿"B" puede afirmar que en promedio sus afiliados tienen mayores aportaciones que los de "A"? Use $\alpha = 0.05$. Suponga que las varianzas son iguales.

Solución:

Las hipótesis por probar son:

a. $\mathrm{H}_0 : \mu_\mathrm{B} = \mu_\mathrm{A}$

$\mathrm{H}_1 : \mu_\mathrm{B} > \mu_\mathrm{A}$

b. $\alpha = 0.05$

c. $t_0 = \dfrac{\bar{x}_B - \bar{x}_A}{s_p\sqrt{\dfrac{1}{n_B} + \dfrac{1}{n_A}}} = \dfrac{970 - 850}{206.2177\sqrt{\dfrac{1}{1200} + \dfrac{1}{1500}}} = 15.02$

d. $\mathrm{RC} = \langle 1.645, \infty \rangle$

e. $t_0 = 15.02$

f. Como t_0 pertenece a la región crítica, se rechaza H_0 y se acepta la afirmación de la AFP B.

13. Una compañía de ferrocarriles instaló dos conjuntos con 50 traviesas de roble cada una. Los dos conjuntos fueron tratados con creosota, empleando para cada uno un procedimiento diferente. Después de cierto número de años en servicio, se observó que 22 traviesas del primer conjunto y 18 del segundo conjunto estaban aún en buenas condiciones. ¿Está justificado afirmar que no hay diferencia real entre las proporciones de traviesas en buen estado de los dos procesos? Use $\alpha = 0.05$.

Solución:

Sean:

π_1: Proporción de traviesas en buen estado con el primer procedimiento de creosota.

π_2: Proporción de traviesas en buen estado con el segundo procedimiento de creosota.

a. Las hipótesis son:

$\mathrm{H}_0 : \pi_1 = \pi_2$

$\mathrm{H}_1 : \pi_1 \neq \pi_2$

b. $\alpha = 0.01$.

c. La estadística de prueba es: $Z_0 = \dfrac{p_1 - p_2}{\sqrt{p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \rightarrow N(0,1)$ donde:

$p = \dfrac{n_1 p_1 - n_2 p_2}{n_1 + n_2}$; y los datos tienen distribución Binomial Puntual.

d. $\text{RC} = \langle -\infty, -2.57538\rangle \cup \langle 2.57538, \infty\rangle$

e. El valor de la estadística de prueba es: $Z_0 = \dfrac{0.44 - 0.36}{\sqrt{0.4(0.6)\left(\dfrac{1}{50} + \dfrac{1}{50}\right)}} = 0.8164$

f. $Z_0 \notin \text{RC}$; por lo tanto, no existe diferencia real en las propiedades preservativas de ambos procesos.

14. En un estudio de artículos de baño se seleccionaron al azar a 100 consumidores y se les formuló las siguientes preguntas:

X1: ¿Utiliza usted el champú CRN? (Sí / No)
X2: ¿Cuánto es su gasto mensual en champú? (En nuevos soles)
X3: Sexo (H / M)
Use $\alpha = 0.04$.

a. Si el reporte de Minitab de la variable X1 es:

Tally for Discrete Variables: X1

X1	Count
No	30
Si	70
N=	100

¿Puede afirmarse que más del 50% de los consumidores prefiere el champú CRN?

b. Si el reporte de Minitab de la variable X2 es:

Descriptive Statistics: X2

Variable	N	Mean	Median	TrMean	StDev	SE Mean
X2	100	57.20	56.50	57.13	5.71	0.571

¿Puede afirmarse que el promedio de gasto mensual en champú es 60 nuevos soles?

c. Si el reporte de Minitab de las variables X1 y X3 es:

Tabulated Statistics: X1, X3

Rows: X1 Columns: X3

	H	M	All
No	10	20	30
Si	40	30	70
All	50	50	100

¿Son independientes las variables "uso del champú CRN" y "sexo"?

d. Si el reporte de Minitab de las variables X2 y X3 es:

Descriptive Statistics: X2 by X3

Variable	X3	N	Mean	Median	TrMean	StDev
X2	H	50	60.00	60.00	60.00	3.81
	M	50	54.40	52.00	54.40	6.27

¿Puede concluirse que las mujeres gastan en champú más que los hombres? Suponga varianzas homogéneas.

Solución:

a. 1. Las hipótesis por probar son:

$$H_0 : \pi \leq 0.5$$
$$H_1 : \pi > 0.5$$

2. Usando Minitab se obtiene el siguiente reporte:

Test and CI for One Proportion

```
Test of p = 0.5 vs p > 0.5
Sample  X    N   Sample p  96.0%Lower Bound  Z-Value  P-Value
   1    70  100  0.700000       0.619773      4.00      0.000
```

Como el P-value es aproximadamente cero, se rechaza H_0 y se puede afirmar que más del 50% de los consumidores prefiere el champú CRN.

b. 1. Las hipótesis por probar son:

$$H_0 : \mu = 60$$
$$H_1 : \mu \neq 60$$

2. $\alpha = 0.04$ y el gasto mensual en champú tiene distribución Normal.

3. Región crítica: $RC = \langle -\infty, -2.0812 \rangle \cup \langle 2.0812, \infty \rangle$, con $t_{(99)}$

4. La estadística es: $t_0 = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} = \dfrac{57.2 - 60}{5.71/\sqrt{100}} = -4.90367776$

5. $t_0 \in RC$; por lo tanto no se puede afirmar que el gasto mensual en champú sea 60 nuevos soles, con un nivel de significación del 4%.

c. 1. Las hipótesis son:

H_0: Uso del champú CRN y sexo son independientes.

H_1: Uso del champú CRN y sexo no son independientes

2. $\alpha = 0.04$.

3. La región crítica es: $\chi^2 > 4.2179$

4. Usando Minitab se obtiene

```
Expected counts are printed below observed counts
             C6          C7       Total
1            10          20          30
             15.00       15.00
2            40          30          70
             35.00       35.00
Total        50          50         100
Chi-Sq = 4.762, DF = 1, P-Value = 0.029
```

Como el valor crítico 4.2179 es menor que el valor observado 4.762, se rechaza la H_0. Es decir, las variables "uso del champú CRN" y "sexo" no son independientes.

d. 1. Se formulan las hipótesis

$$H_0 : \mu_H \geq \mu_M$$
$$H_1 : \mu_H < \mu_M$$

2. $\alpha = 0.04$ y ambas muestras son aleatorias e independientes.

3. Región crítica: $RC = \langle -\infty, -1.769 \rangle$, usando $t_{(98)}$

4. El valor t_0 es: $t_0 = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \dfrac{60 - 54.4}{5.18791866 \sqrt{\dfrac{1}{50} + \dfrac{1}{50}}} = 5.39715$

donde: $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \dfrac{49(3.81)^2 + 49(6.27)^2}{98} = 26.9145$

5. Como la estadística de prueba (5.3971) es mayor que el valor crítico (–1.769), entonces no se rechaza la H_0. Es decir, las mujeres no gastan más en champú que los hombres.

15. El Gobierno central resuelve identificar a la población de beneficiarios del programa social Pro Perú, y para este fin decide recolectar información al azar acerca de las siguientes variables:

C1: Departamento donde reside la familia beneficiaria.
C2: Jefe de familia (padre, madre, otros).
C3: Ingreso mensual de la familia beneficiaria.
C4: Número de hijos en la familia beneficiaria.
C5: Número de hijos en edad escolar.
C6: Padres alcohólicos (Sí = 1, No = 0)

Para conocer con mayor detalle a las familias beneficiarias usted debe responder las siguientes preguntas utilizando los reportes de Minitab que se adjuntan a cada pregunta.

Pregunta 1. Para probar la hipótesis de que la varianza de los ingresos es 25.000 se desea aplicar la siguiente regla de decisión: si $s^2 \leq 26500$, entonces, no se rechaza la hipótesis nula. Si $n = 1.000$, calcule el error tipo I.

Solución:

Usando la definición de error tipo I se tiene:

$$\alpha = P(\text{Rechazar } H_0 / H_0 \text{ es verdadera}) = P(s^2 > 26500 / \sigma^2 = 25000)$$

$$\alpha = P\left(\chi^2 > \frac{26500(999)}{25000}\right) = P(\chi^2 > 1058.94) = 1 - 0.908223 = 0.091777$$

Pregunta 2. ¿Es el promedio de los ingresos de Ayacucho mayor que el promedio de los ingresos de Tumbes? Use $\alpha = 0.04$.

Descriptive Statistics: Ingreso_Ayacucho, Ingreso_Tumbes

Variable	Count	Mean	Sum of Squares	S2	S
Ingreso_Ayacucho	128	252.3	13005665.9	38250.3054	195.57685
Ingreso_Tumbes	51	211.9	2978855.4	13777.4658	117.37745

Solución:

a. Para responder a la pregunta se deben formular las hipótesis:

$\mathrm{H}_0 : \mu_1 = \mu_2$

$\mathrm{H}_1 : \mu_1 > \mu_2$

b. $\alpha = 0.04$ y las muestras son aleatorias e independientes.

c. La estadística que se usa es la t para muestras independientes, pero como no se sabe si las varianzas son iguales o diferentes se debe hacer la prueba de varianzas. En efecto:

1. $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$

   $\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2$

2. Región crítica:

   $F_{\left(\frac{\alpha}{2}, n_1-1, n_2-1\right)} = F_{(0.02,127,50)} = 0.62873$ y $F_{\left(1-\frac{\alpha}{2}, n_1-1, n_2-1\right)} = F_{(0.98,127,50)} = 1.6737$

   $\mathrm{RC} = \langle -\infty, 0.62873\rangle \cup \langle 1.6737, \infty\rangle$

3. $F_0 = \dfrac{s_1^2}{s_2^2} \sim F_{(127,50)}$ y el valor de la estadística es: $F_0 = 2.776295$.

4. Se rechaza la H_0. Las varianzas no son homogéneas, se usa la *t* de varianzas diferentes, es decir: $t_0 = \dfrac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \sim t_{(v)}$

d. Región crítica: $\mathrm{RC} = \langle t_{(1-\alpha, v)}, \infty\rangle$

donde: $v = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{(n_1-1)} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{(n_2-1)}} = \dfrac{\left(\dfrac{38250.3054}{128} + \dfrac{13777.4658}{51}\right)^2}{\dfrac{\left(\dfrac{38250.3054}{128}\right)^2}{(127)} + \dfrac{\left(\dfrac{13777.4658}{51}\right)^2}{(50)}} = 146.4528$

Entonces:

$t_{(1-\alpha, v)} = t_{(0.96,146.4528)} = 1.76292$

Por lo tanto, $\mathrm{RC} = \langle 1.76292, \infty\rangle$

e. El valor de la estadística de prueba es: $t_0 = 1.69$.

f. Como la estadística de prueba (1.69) es menor que el valor crítico 1.76292, no se rechaza la hipótesis nula. Es decir, los promedios pueden ser iguales.

Pregunta 3. ¿Es posible afirmar que el número de hijos tiene distribución de Poisson con parámetro 3? Use $\alpha = 0.05$.

Número de hijos	1	2	3	4	5	6	7	8
Familias beneficiarias	60	107	127	356	256	48	34	12

Solución:

a. Las hipótesis a probar son:

H_0 : El número de hijos tiene distribución de Poisson con parámetro 3.

H_1 : El número de hijos no tiene distribución de Poisson con parámetro 3.

b. $\alpha = 0.05$.

c. La estadística de prueba es: $\chi_0^2 = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(7)}$

d. Región crítica: $RC = \langle \chi^2_{(1-\alpha,k-p-1)}, \infty \rangle$

$\chi^2_{(1-\alpha,k-p-1)} = \chi^2_{(1-0.05,8-0-1)} = \chi^2_{(0.95,7)} = 14.0671$

Por lo tanto, $RC = \langle 14.0671, \infty \rangle$

e. El valor de la estadística de prueba es: $\chi_0^2 = 656.758216$

Los cálculos se muestran en la siguiente tabla:

Nº hijos	O_i	p_i	E_i	$\frac{(O_i - E_i)^2}{E_i^2}$
1	60	0.19914827	199.148273	97.2252567
2	107	0.22404181	224.041808	61.143877
3	127	0.22404181	224.041808	42.0328354
4	356	0.16803136	168.031356	210.271536
5	256	0.10081881	100.818813	238.85622
6	48	0.05040941	50.4094067	0.11516185
7	34	0.02160403	21.6040315	7.11256307
8	12	0.01190450	11.9045039	0.00076606
	1.000		1.000	656.758216

f. Se rechaza la H_0. Los datos no pueden modelarse como Poisson con $\lambda = 3$.

Pregunta 4. ¿Es la proporción de familias beneficiarias de Lambayeque con dos hijos en edad escolar menor que la proporción de familias beneficiarias de Tacna con dos hijos en edad escolar? Use $\alpha = 0.03$ y utilice la siguiente información para hacer la prueba.

Hijos Edad escolar_Lambayeque	Count	Hijos Edad escolar_Tacna	Count
0	58	0	20
1	23	1	13
2	10	2	9
N =	91	N =	42

Solución:

a. Las hipótesis por probar son:

$H_0 : \pi_1 = \pi_2$

$H_1 : \pi_1 < \pi_2$

b. $\alpha = 0.03$.

c. La estadística a usarse es: $Z_0 = \dfrac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim N(0,1)$

d. Región crítica: $RC = \langle -\infty, Z_{(\alpha)} \rangle$

$Z_{(\alpha)} = Z_{(0.03)} = -1.88079$

$RC = \langle -\infty, -1.88079 \rangle$

e. Valor de la estadística de prueba: $Z_0 = \dfrac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = -1.60$

f. No se rechaza la hipótesis nula.

Por lo tanto, las proporciones de familias con dos hijos de ambas ciudades son las mismas con $\alpha = 0.03$.

La salida del Minitab es:

```
Test and CI for Two Proportions
Sample     X      N  Sample p
1         10     91  0.109890
2          9     42  0.214286
Difference = p (1) - p (2)
Estimate for difference:  -0.104396
97% upper bound for difference:  0.0297042
Test for difference = 0 (vs < 0):  Z = - 1.60 P-Value = 0.055.
```

16. Una empresa consultora realizó una encuesta a consumidores de bebidas rehidratantes para conocer las características relevantes de este mercado. Las variables registradas fueron:

C1: Género del consumidor (0 = Masculino, 1 = Femenino)
C2: Nivel socioeconómico (1 = Nivel A, 2 = Nivel B)
C3: Edad del consumidor (1 = 13-24, 2 = 25-39, 3 = 40-70)
C4: Marca de la bebida rehidratante (1 = Gator, 2 = Spor, 3 = Fit, 4 = Energy, 5 = Power, 6 = Light).
C5: Sabor de la bebida rehidratante (1 = mandarina, 2 = uva, 3 = piña, 4 = lima limón, 5 = otros).
C6: Frecuencia de compra (1 = diariamente, 2 = semanalmente, 3 = quincenalmente, 4 = mensualmente, 5 = ocasionalmente).

Los datos de esta pregunta se encuentran en el archivo Bebidas.MTW. Utilice este archivo para completar los siguientes espacios en blanco:

Solución:

a. La hipótesis de que la edad del consumidor y la marca de bebida rehidratante son independientes tiene P-value igual a _____

b. El valor de la prueba estadística de la hipótesis de que todas las marcas tienen la misma proporción del mercado es _____

c. El valor de la prueba estadística para probar la hipótesis de que el género y la frecuencia del consumidor no tienen relación es _____

d. Si se juntan los consumidores de bebidas rehidratantes que compran diariamente con aquellos que compran semanalmente, entonces la hipótesis que supere el 40% tendrá un valor observado de la correspondiente prueba estadística igual a _____

17. Antes de un cambio de imagen, los dueños de una cadena de supermercados conocían que la distribución de sus clientes era de la siguiente manera:

Sucursal	Camacho	San Isidro	Óvalo Gutiérrez	San Miguel	Otros
Porcentaje	10	35	10	20	25

Para evaluar el efecto del cambio de imagen, los dueños de la cadena seleccionaron una muestra aleatoria de sus clientes durante el último fin de semana y registraron la asistencia de estos a las sucursales de la tabla anterior. Los datos obtenidos son:

Sucursal	Camacho	San Isidro	Óvalo Gutiérrez	San Miguel	Otros
N.º de clientes	20	75	32	35	38

¿Puede afirmarse que la distribución de clientes de la empresa es la misma que antes del cambio de imagen? Use $\alpha = 0.05$.

Solución:

a. Las hipótesis por probar son:

H_0: La distribución de los clientes es $\pi_1 = 0.1$, $\pi_2 = 0.35$, $\pi_3 = 0.1$, $\pi_4 = 0.2$, $\pi_5 = 0.25$.

H_1: La distribución de los clientes NO es $\pi_1 = 0.1$, $\pi_2 = 0.35$, $\pi_3 = 0.1$, $\pi_4 = 0.2$, $\pi_5 = 0.25$.

b. $\alpha = 0.05$

c. La prueba estadística es $\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(4)}$

p = 0 porque no se necesita estimar ningún parámetro.

d. Región crítica: $RC = \langle \chi^2_{(1-\alpha,k-p-1)}, \infty \rangle$

$\chi^2_{(1-\alpha,k-p-1)} = \chi^2_{(1-0.05,5-0-1)} = \chi^2_{(0.95,4)} = 9.48773$

Por lo tanto, $RC = \langle 9.48773, \infty \rangle$

e. Valor de la estadística de prueba: $\chi_0^2 = 11.06214$ con P-value = 0.025875.

20	0.1	20	0
75	0.35	70	0.3571429
32	0.1	20	7.2000000
35	0.2	40	0.6250000
38	0.25	50	2.8800000
200	**1**		**11.0621429**

f. Se rechaza la H_0, la distribución de los clientes de la empresa no es la misma que antes del cambio de imagen con $\alpha = 0.05$.

18. Suponga que usted selecciona al azar una muestra aleatoria de transportistas y registra las siguientes variables:

X1: Monto total de las infracciones.
X2: Monto total mensual del consumo de petróleo.
X3: Empresa de transportes.

El reporte de los montos de las infracciones de las empresas de Emtsa y Huarochi es:

Descriptive Statistics: x1

Variable	Count	Mean	StDev
EMTSA	60	2002.4	138.9
Huarochi	60	1984.5	149.0

¿Puede afirmarse que los promedios de los montos de las infracciones de las empresas Emtsa y Huarochi son iguales? Use $\alpha = 0.04$.
Nota: Presente el procedimiento completo de la técnica estadística correspondiente.

Solución:

a. Para responder a la pregunta se deben formular las hipótesis:

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

b. $\alpha = 0.04$ y los montos de las infracciones se distribuyen normalmente. Las muestras son aleatorias e independientes.

c. La estadística por usar es la t para muestras independientes, pero como no se sabe si las varianzas son iguales o diferentes se debe hacer la prueba de varianzas. En efecto:

1. $H_0 : \sigma_1^2 = \sigma_2^2$

   $H_1 : \sigma_1^2 \neq \sigma_2^2$

2. Región crítica: $RC = \langle -\infty, 0.58261 \rangle \cup \langle 1.71642, \infty \rangle$

3. $F_0 = \dfrac{s_1^2}{s_2^2} \sim F_{(59,59)}$ y el valor de la estadística es: $F_0 = 0.87$.

4. No se rechaza la H_0. Las varianzas son homogéneas y se usa la estadística t con varianzas iguales, es decir:

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{(n_1 + n_2 - 2)}$$

d. Región crítica: $RC = \langle -\infty, -2.0767 \rangle \cup \langle 2.0767, \infty \rangle$

e. El valor de la estadística de prueba es: $t_0 = 0.68$

f. No se rechaza la H_0. Los promedios son iguales.

19. Diversas entidades financieras de Lima están tratando de colocar sus tarjetas de crédito, para esta finalidad hacen uso fundamentalmente de las bases de datos de Corinfo; en esta base de datos figuran todos los consumidores que alguna vez hicieron uso de algún crédito y los datos registrados corresponden a una serie de variables que describen el perfil crediticio de los consumidores, tales como ingreso mensual, monto del crédito, deudas pendientes, morosidad, estado civil, tipo de ocupación, edad, etcétera. Money's Card es una entidad financiera cuyo rubro fundamental son los negocios a través de tarjetas de crédito, y por encargo de su gerente de operaciones se ha seleccionado a todos los clientes cuyo perfil les ha resultado interesante, obteniéndose una base de datos de 9.000 clientes. De esta base de datos se ha extraído una muestra aleatoria de 750 clientes.

Pregunta 1. La siguiente información corresponde a la variable ingreso mensual de los clientes por tipo de empleo.

Variable	t_ocupacion	Count	Mean	SE Mean	StDev	Variance
ingres	DEPENDIENTE	477	3184.4	22.4	489.5	239572.3
	INDEPENDIENTE	243	3150.0	31.1	484.5	234733.2
	MIXTO	30	3242.1	93.4	511.6	261737.3

Considerando un nivel de significación de 0.04, ¿se puede concluir que el ingreso medio de los clientes con empleo de tipo dependiente es inferior al ingreso medio de los clientes de tipo de empleo mixto? Suponga varianzas desconocidas pero iguales.

Solución:

En este caso se utilizará la ayuda de Minitab para responder a la pregunta, en efecto:

a. $H_0 : \mu_1 = \mu_2$ (los ingresos medios en ambos tipos de empleo son iguales)
   $H_1 : \mu_1 < \mu_2$ (el ingreso medio en los dependientes es menor que en los mixtos)

b. La corrida del Minitab produce la siguiente salida:

Two-Sample T-Test and CI

Sample	N	Mean	StDev	SE Mean
1	477	3184	490	22
2	30	3242	12	93

```
Difference = mu (1) - mu (2)
Estimate for difference:  -57.7000
96% upper bound for difference:  104.3572
T-Test of difference=0 (vs<): T-Value = -0.62
P-Value = 0.266 DF =505
Both use Pooled StDev = 490.7960
```

Decisión y conclusión: Como el P-value (0.266) es mayor que el valor de α (0.04), entonces no se rechaza H_0. Se concluye que el ingreso medio de los clientes con empleo dependiente no es inferior al ingreso medio de los clientes de tipo de empleo mixto.

Pregunta 2. La siguiente información corresponde a la variable número de créditos obtenidos en los dos últimos meses por cada cliente clasificados por tipo de empleo y nivel del cliente (A y B).

Nivel: nivel asociado a cada cliente de acuerdo a su perfil crediticio.

Results for nivel = A

Rows: t_ocupacion Columns: n_credit

	0	1	2	3	4	All
Dependiente	11	17	12	3	1	44
Independiente	10	7	6	1	0	24
Mixto	0	1	2	0	1	4
All	21	25	20	4	2	72

Results for nivel = B

Rows: t_ocupacion Columns: n_credit

	0	1	2	3	4	All
Dependiente	3	29	12	10	0	54
Independiente	8	15	5	2	0	30
Mixto	0	2	1	0	0	3
All	11	46	18	12	0	87

¿Se puede afirmar con $\alpha = 0.05$ que, en el nivel A la proporción de clientes con empleo de tipo dependiente supera a la correspondiente proporción de clientes con empleo de tipo dependiente en el nivel B?

Solución:

Haciendo uso de Minitab para responder a la pregunta, se tiene:

a. $H_0 : \pi_1 = \pi_2$ (En el nivel A la proporción de clientes con empleo de tipo dependiente no supera a la correspondiente proporción de clientes con empleo de tipo dependiente en el nivel B).

$H_1 : \pi_1 > \pi_2$ (En el nivel A la proporción de clientes con empleo de tipo dependiente supera a la correspondiente proporción de clientes con empleo de tipo dependiente en el nivel B).

b. La salida de Minitab produce:

Test and CI for Two Proportions

Sample	X	N	Sample p
1	44	72	0.611111
2	54	87	0.620690

Difference = p (1) - p (2)
Estimate for difference: -0.00957854
95% lower bound for difference: -0.137062
Test for difference = 0 (vs > 0): Z = -0.12 P-Value = 0.549

Decisión y conclusión. No se rechaza la H_0 porque el P-value (0.549) es mayor que el valor α (0.05), por lo tanto, en el nivel A la proporción de clientes con empleo de tipo dependiente no supera a la correspondiente proporción de clientes con empleo de tipo dependiente en el nivel B.

Pregunta 3. La siguiente información corresponde a la variable número de créditos bancarios obtenidos por cada cliente en los dos últimos meses.

n_cred: número de créditos obtenidos en los dos últimos meses:

n_cred	0	1	2	3	4
Frecuencia	158	281	224	77	10

¿Se puede afirmar que el número de créditos bancarios obtenidos se ajusta a una distribución Poisson? Use $\alpha = 0.05$.

Solución:

Las hipótesis por probar son:

a. H_0: El número de créditos bancarios tiene distribución de Poisson.
   H_1: El número de créditos bancarios no tiene distribución de Poisson.

b. Con ayuda de Minitab se obtiene el siguiente reporte:

Goodness-of-Fit Test for Poisson Distribution

Poisson mean for X = 1.33333

X	Observed	Poisson Probability	Expected	Contribution to Chi-Sq
0	158	0.263597	197.698	7.9714
1	281	0.351463	263.597	1.1489
2	224	0.234309	175.731	13.2580
3	77	0.104137	78.103	0.0156
4	10	0.046494	34.871	17.7385

N	DF	Chi-Sq	P-Value
750	3	40.1324	0.000

Decisión y conclusión. El número de créditos bancarios no presenta una distribución de Poisson con $\lambda = 1.33333$, ya que el P-value (0) es menor que el $\alpha(0.05)$.

Pregunta 4. La siguiente información corresponde a las variables nivel crediticio y tipo de empleo.

Tabulated statistics: nivel, t_empleo

Rows: nivel Columns: t_empleo

	DEPENDIENTE	INDEPENDIENTE	MIXTO	All
A	44	24	4	72
B	186	87	12	285
C	193	102	11	306
D	54	30	3	87
All	477	243	30	750

¿Es el nivel crediticio de los clientes independiente del tipo de ocupación? Use $\alpha = 0.05$.

Solución:

Las hipótesis por probar son:

a. H_0: El nivel crediticio y el tipo de empleo son independientes.
   H_1: El nivel crediticio y el tipo de empleo no son independientes.

b. La salida de Minitab es:

Chi-Square Test: C3, C4, C5

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts
        C3          C4          C5          Total
1       44          24           4             72
        45.79       23.33        2.88
2      186          87          12            285
       181.26       92.34       11.40
3      193         102          11            306
       194.62       99.14       12.24
4       54          30           3             87
        55.33       28.19        3.48
Total  477         243          30            750
Chi-Sq = 1.425,     DF = 6,  P-Value = 0.964.
```

Decisión y conclusión: Como el P-value (0.964) es mayor que el valor de α (0.05), no se rechaza H_0. Es decir, el nivel crediticio y el tipo de empleo son independientes.

20. El diámetro interno de una pieza fabricada que debe acoplarse a otras piezas con cierta dimensión crítica es de 4.6 pulgadas. La variabilidad en la fabricación es $\sigma = 0.5$ pulgadas. El proceso es considerado bajo control y se continúa si produce esta pieza con una media de 4.6 pulgadas para la dimensión crítica. En otro caso, el proceso se considera fuera de control y es detenido. Para comprobar si el proceso está bajo control, el jefe de control de calidad de la fábrica decide tomar muestras aleatorias cada dos horas con $\alpha = 0.05$ y $n = 25$.

a. Formule las hipótesis apropiadas e indique los errores I y II.
b. Calcule los valores críticos para la media muestral, y con ellos establezca una regla de decisión.
c. Determine los valores de la curva OC y la curva de potencia para los siguientes valores del parámetro μ:

4.20 4.30 4.40 4.50 4.60 4.70 4.80 4.90 5.00

Solución:

a. $H_0 : \mu = 4.6$ (el proceso está bajo control)

   $H_1 : \mu \neq 4.6$ (el proceso está fuera de control)

   ETI: Afirmar erróneamente que el proceso está fuera de control.

   ETII: Afirmar erróneamente que el proceso está bajo control.

b. Se rechaza H_0 si: $|Z_0| > Z_{(0.975)} \Rightarrow |Z_0| > 1.95996 \Rightarrow \left|\dfrac{\bar{x} - 4.6}{0.5/\sqrt{5}}\right| > 1.95996$

de donde, no se rechaza H_0 si: $4.404 \leq \bar{x} \leq 4.796$ y se rechaza en otro caso.

Por lo tanto el criterio de decisión es:

i. Extraer una muestra aleatoria de tamaño 25 y calcular la media muestral.

ii. Si: $\bar{x} < 4.404$ *ó* $\bar{x} > 4.796$, entonces se rechaza H_0 y se concluye que el proceso está fuera de control.

Si $4.404 \leq \bar{x} \leq 4.796$, entonces no se rechaza H_0 y se concluye que el proceso está bajo control.

c. Con ayuda de Minitab se obtiene:

μ	4.20	4.30	4.40	4.50	4.60	4.70	4.80	4.90	5.0
FP	0.9793	0.8508	0.5160	0.1700	0.0500	0.1700	0.5160	0.8508	0.9793
OC	0.0206	0.1491	0.4839	0.8299	0.9500	0.8299	0.4839	0.1491	0.0206

PROBLEMAS PROPUESTOS

1. Error de tipo II es rechazar la hipótesis nula cuando realmente es verdadera. V F

2. Siempre el error de tipo I es 0.05. V F

3. Si en una prueba de diferencia de medias, las varianzas poblacionales son conocidas, entonces se usa la distribución Normal. V F

4. En la prueba de proporciones, el error estándar es $\sqrt{\dfrac{p(1-p)}{n}}$ V F

5. En una prueba de una varianza se usa la distribución F. V F

6. Para cada uno de los siguientes casos, identifique la hipótesis nula y el tipo de error, cuando sea apropiado:

a. Los televisores Sunglo requieren reparaciones en una mayor proporción que los televisores Zeta. Un estudio efectuado por Indecopi concluyó lo mismo.

b. Al realizar cierta tarea se comparó a dos trabajadores con respecto a su eficiencia; el trabajador B es realmente más eficiente que el A. Una evaluación concluyó que no hay diferencia entre las razones de eficiencia de A y de B.

7. Al evaluar solicitudes de crédito, el gerente de Banca Personal del banco Intrabank se encuentra con el problema de otorgar un préstamo a aquellos clientes que son buenos pagadores y negarlo a aquellos que no lo son. Estaría probando lo siguiente para cada cliente:

H_0: El solicitante es un buen pagador
H_1: El solicitante es un mal pagador

a. En términos de conceder o negar el préstamo, ¿cómo se cometería el error tipo I?; ¿cómo se cometería el error de tipo II?
b. Discuta la selección del nivel de significación en los siguientes casos:
   - Se tiene poco dinero para prestar, las tasas de interés son altas y hay un gran número de solicitudes.
   - Hay bastante dinero para ser prestado, las tasas de interés son moderadas y existe competencia para conseguir solicitudes de crédito.

8. En una prueba de hipótesis bilateral, la región de rechazo es:

a. A la izquierda del valor crítico.
b. A la derecha del valor crítico.
c. Entre los dos valores críticos.
d. Ninguno de los anteriores.

9. En una prueba de hipótesis donde $\alpha = 0.05$ y la potencia es 0.90, el error de tipo II es:

a. 0.05
b. 0.95
c. 0.10
d. 0.90
e. ninguno de los anteriores.

10. Suponga que se desea probar que $H_0 : \mu = 5$ frente a la $H_1 : \mu \neq 5$, $\sigma = 3$ y $\alpha = 0.05$, ¿cuál es el valor crítico que debe usarse?

a. 1.646985
b. –1.64698
c. 1.95996
d. $\pm$1.95996
e. ninguno.

11. Sea P-value el valor probabilístico de una prueba de hipótesis ($H_1 : \mu > \mu_0$) y α el nivel de significancia. No se rechaza H_0 si:

a. $P\text{-value} \geq \alpha$
b. $P\text{-value} < \alpha$
c. $P\text{-value} > \alpha / 2$

12. El Gobierno indica que la canasta familiar es de 500 nuevos soles, los candidatos para ocupar la Presidencia de la República sospechan que es mayor a 500. Se eligió una muestra de 50 familias, resultando un promedio de 600 nuevos soles y una desviación estándar de 400 nuevos soles. Efectúe una prueba de hipótesis para determinar quién tiene razón, con un nivel de significancia del 4%.

13. La gerencia general de una cadena de supermercados contempla la posibilidad de abrir una tienda en cierta zona de la ciudad si encuentra pruebas de que el gasto promedio mensual en consumo por familia es superior a S/.1.000. La decisión se tomará sobre la base de una encuesta aplicada a 500 familias del sector.

a. Formule convenientemente las hipótesis nula y alternativa e indique, en términos del enunciado, en qué consisten los errores de tipo I y tipo II.
b. Si como resultado de la encuesta se obtiene un gasto promedio de 1.280 nuevos soles, con una desviación estándar de 320 nuevos soles, ¿qué decisión debe adoptar la gerencia? Use $\alpha = 0.02$.

14. Un encargado de envasar arroz afirma que el contenido de las bolsas que vende la empresa por término medio es de por lo menos 995 gramos. En años anteriores la distribución del peso del arroz seguía una distribución Normal, con desviación típica de 9 gramos. Se eligió una muestra aleatoria de 16 bolsas, obteniendo un peso medio de 985 gramos. ¿Está en lo cierto lo que indica el encargado con un nivel de significancia del 5%?

15. Una empresa peruana que se dedica a la venta de franquicias afirma que por término medio obtiene, durante el primer año, un rendimiento del 12%. Una muestra aleatoria de 9 franquicias presentó los siguientes rendimientos porcentuales:

Rendimiento	8.1	9.2	15	11.6	8.7	3.9	16	11	14

Suponiendo que los rendimientos tienen distribución Normal, contrastar la afirmación de la compañía con un α del 10%.

16. La siguiente información representa el número de horas de sueño ganadas por 10 pacientes con un cierto somnífero. ¿Justifican estos datos admitir que el somnífero aumenta las horas de sueño? Use $\alpha = 0.01$.

Pacientes	1	2	3	4	5	6	7	8	9	10
Horas ganadas	1.2	-1.3	0.7	0.2	3.4	0.8	3.1	1.8	2.0	3.1

17. Royal Kola S.A., fabricante de bebidas gaseosas, afirma que su producto es superior al de su competidor porque en una muestra de 100 personas que probaron ambas bebidas con los ojos vendados, 54 de ellas indicaron que lo preferían al de este.

a. ¿Se justifica la afirmación del fabricante al 5% de significación?
b. Si el verdadero valor de π es 0.52, calcule β si $\alpha = 0.05$.
c. ¿Cuál es la potencia de la prueba cuando $\pi = 0.55$ con $\alpha = 0.05$?

18. De una muestra aleatoria de 134 clientes, 63 recordaron el precio correcto de un artículo. La cajera afirma que al menos la mitad de los compradores recuerdan el precio correcto.

a. Con un nivel de significancia del 5%, pruebe la hipótesis correspondiente.
b. Halle el P-value.
c. Halle la potencia de la prueba con un nivel del 5% si, en realidad, el 45% de los clientes son capaces de recordar el precio correcto.

19. Un ingeniero especialista en control de calidad desea probar que el porcentaje de polos de tipo exportación tienen menos fallas que el porcentaje de polos de tipo nacional. Se tomaron dos muestras de 60 y 50 y se encontraron 4 y 6 fallas, respectivamente. Con $\alpha = 0.02$, ¿el ingeniero de control de calidad tiene razón en su apreciación?

20. La información que se detalla a continuación corresponde a un experimento de medir la producción de leche y calcio de 12 vacas durante 15 días.

Día	Producción (L)	Calcio (%)	Raza
1	358.70	1.40	A
2	311.4	0.90	B
3	360.6	0.88	B
4	380.9	0.99	A
5	406.5	0.99	A
6	410.0	0.77	A
7	370.2	1.35	B
8	385.6	1.05	B
9	350.4	0.46	A
10	390.6	0.59	A
11	380.4	0.72	B
12	395.5	0.93	A
13	405.6	1.02	A
14	370.4	1.10	B
15	410.6	0.75	A

a. Considere que la producción de leche sigue una distribución Normal con varianza de 4.900. Pruebe la hipótesis de que el promedio de producción de leche diaria de las 12 vacas es, a lo más, 390 litros (L), con un nivel de significancia del 5%. Calcule el P-value. ¿Cuál es la potencia de la prueba si el promedio de producción diaria es de 400 litros?
b. Pruebe la hipótesis de que el promedio de calcio en la leche es de, por lo menos, 1.05%, con un nivel de significancia del 5%.
c. Pruebe la hipótesis de que el promedio de producción de leche de la raza A es mayor que el promedio de producción de leche de la raza B, con un error del 1%.
d. Pruebe la hipótesis de que la varianza del calcio de la leche de las vacas de A es mayor que la varianza del calcio de la leche de las vacas del tipo B, con un nivel de significancia del 5%.

21. En una de las academias situadas alrededor de un instituto educativo, que ofrecen cursos de preparación antes de los exámenes parciales, se eligieron dos muestras aleatorias y se registraron las notas obtenidas del curso de Matemáticas I. La primera muestra corresponde a los alumnos que asistieron a esos cursos de preparación y la otra a aquellos que no asistieron. Los resultados fueron los siguientes:

Asistieron	No asistieron
13	15
12	14
10	10
08	13
11	12
05	11
13	08
12	10

Suponiendo que las notas siguen una distribución Normal, pruebe la hipótesis de que el promedio de notas de los alumnos que no asistiron es mayor que el promedio de notas de los que asistieron, con un nivel de significancia del 5%.

Nota: Suponga que las varianzas son iguales.

22. Un diario peruano informó que el país deberá demostrar ante la Organización Mundial de Comercio los graves daños ocasionados a la industria textil que motivaron la aplicación de las salvaguardias provisionales a las confecciones extranjeras. En este contexto, suponga que usted recibe el encargo de analizar las ventas mensuales (en miles de dólares) de las empresas del sector textil. Para este fin, selecciona al azar 14 empresas y registra sus ventas sin la aplicación de las salvaguardias provisionales. De la misma manera, selecciona al azar 13 empresas y registra sus ventas con la aplicación de las salvaguardias provisionales. Los datos fueron los siguientes:

Sin salvaguardias	18	12	14	23	35	67	43	53	23	12	34	32	23	32
Con salvaguardias	32	34	35	43	41	42	40	48	21	23	54	34	42	

Use $\alpha = 0.06$ en sus cálculos. Suponga que las ventas mensuales se distribuyen normalmente.

a. ¿Son las varianzas homogéneas? Señale las hipótesis, la prueba estadística, la región crítica y su conclusión.
b. ¿Son las ventas con salvaguardias superiores a las ventas sin salvaguardias? Señale las hipótesis, la prueba estadística, la región crítica y su conclusión.
c. ¿Puede usted afirmar que el promedio de las ventas mensuales con salvaguardias es superior a los 35.000 dólares? Señale las hipótesis, la prueba estadística, la región crítica y su conclusión.

23. Una empresa que se dedica al negocio de comida rápida debe pagar por derechos de franquicia a la casa matriz. Se eligió una muestra aleatoria de 14 días, como a continuación se detalla:

Regalías diarias	Ventas diarias	Día de la semana
237	4.761	Lunes
203	4.070	Miércoles
210	4.577	Jueves
220	4.672	Jueves
200	4.000	Miércoles
305	6.112	Sábado
205	5.320	Martes
220	5.500	Viernes
310	6.150	Sábado
213	4.600	Lunes
290	4.900	Miércoles
305	5.300	Sábado
208	3.900	Martes
190	2.300	Jueves

a. Si las ventas diarias de comida rápida sigue una distribución Normal, con varianza de 81.000, pruebe la hipótesis de que el promedio de ventas de comida rápida es menor a 4.900, con un nivel de significancia del 5%.
b. ¿Es la venta promedio de los sábados mayor que la venta promedio de los miércoles, con un nivel de significancia del 1%? Suponga que las varianzas son iguales.
c. ¿Es la variabilidad de la regalía del día lunes igual a la variabilidad de la regalía del día miércoles, con un nivel del 10%?

24. Una conocida empresa de la capital se dedica a la comercialización de artículos de limpieza. Para comparar el efecto de la publicidad por radio y por televisión, se registraron las ventas diarias (en miles de nuevos soles) de sus locales de Mesa Redonda (MR), Los Olivos (LO), Ceres (C) y San Juan de Miraflores (SJM). A continuación se presenta una tabla con los valores de las ventas registradas por local y tipo de publicidad:

Local	Radio									Televisión			
MR	36	32	30	38	40	28	29	15	36	12	18	19	48
LO	28	29	20	27	30	26	15	18	19	46	48	49	55
C	10	40	23	12	16	18	42	36	45	8	63	36	10
SJM	23	22	21	22	23	22	21	20	23	22	21	22	22

Nota: Use $\alpha = 0.04$ en sus cálculos y suponga que las ventas se distribuyen normalmente. En los casos b) y c) presente el procedimiento completo de la técnica correspondiente.

a. Sea μ_3 el promedio de las ventas correspondientes a la publicidad por radio del local de Ceres. Se desea probar $H_0 : \mu_3 = 28$ versus $H_1 : \mu_3 < 28$.

i. Aplicando la siguiente regla de decisión: si el promedio muestral correspondiente es menor que 27 se rechaza la H_0, calcule el error tipo I, suponiendo $\sigma = 2$.

ii. Calcule el error tipo II si el verdadero valor de la media es 26.5. Suponga $\sigma = 2$.

b. En los datos correspondientes a la publicidad por radio, ¿puede afirmarse que el promedio de las ventas del local de San Juan de Miraflores fue mayor que el promedio de las ventas del local de Los Olivos?

c. ¿Puede afirmarse que el 50% de todas las ventas de la empresa es superior a 30.000 nuevos soles?

25. Un empresario tiene dos grifos de combustible. Sus registros indican que la media del número de galones de gasolina que vende a sus clientes es igual a cuatro galones en cada grifo. Además, el empresario sabe que las ventas de gasolina a sus clientes tienen distribución Normal. Para analizar el comportamiento de las ventas, el empresario seleccionó muestras aleatorias de las ventas de cada grifo obteniendo los siguientes resultados:

```
Descriptive Statistics: grifo 1, grifo 2
Variable   N     Mean        SE Mean
grifo 1    17    3.838       0.205
grifo 2    15    3.490       0.370
```

Use un nivel de significación de 0.05 en sus cálculos.

a. Si se desea probar que las varianzas poblacionales son iguales, señale:
   1. Hipótesis nula.
      Hipótesis alternante.
   2. La estadística de prueba y su valor.
   3. Regla de decisión y valores críticos.
   4. Decisión y conclusión.

b. Si se desea probar que ha bajado el consumo promedio de gasolina en el grifo 2, señale:
   1. Hipótesis nula.
      Hipótesis alternante.
   2. La estadística de prueba y su valor.
   3. Regla de decisión y valores críticos.
   4. Decisión y conclusión.

26. Un laboratorio farmacéutico usa un estabilizador de voltaje de 220 voltios. El ingeniero de planta debe controlar que el voltaje medio sea 220 voltios y que la desviación estándar no supere los cinco voltios. Si no se cumple una de las dos condiciones debe cambiar el equipo por otro perfectamente regulado. Para ello, al final de cada jornada laboral recibe un informe con 16 lecturas de voltaje realizadas cada media hora, para decidir si el aparato requiere ser cambiado o no. El ingeniero acaba de recibir la siguiente información:

```
Descriptive Statistics: voltaje
Variable   N      Mean    SE Mean   StDev    Variance
voltaje   16    219.88       1.81    7.25       52.52
```

Suponga que la distribución de la población es Normal y $\alpha = 0.05$.

a. ¿Es el voltaje medio igual a 220 voltios? Señale:
   Hipótesis nula.
   Hipótesis alternante.
   La estadística de prueba y su valor.
   Regla de decisión y valores críticos.
   Decisión y conclusión.
b. ¿Es la desviación estándar mayor que cinco voltios? Señale
   Hipótesis nula.
   Hipótesis alternante.
   La estadística de prueba y su valor.
   Regla de decisión y valores críticos.
   Decisión y conclusión.
c. ¿Es necesario cambiar el estabilizador de voltaje?, ¿por qué?

27. Suponga que el gerente de Lima Plaza desea probar la hipótesis nula de que un cliente que asiste a este centro comercial gasta en promedio S/.250 versus la hipótesis alternante de que gasta más de S/.250. Suponga que el gerente está dispuesto a cometer un error de 0.08 de rechazar la hipótesis nula cuando esta es verdadera en la población, y también suponga que la desviación estándar poblacional de los gastos es igual a S/.50. Si el gerente decide seleccionar al azar los gastos de 100 clientes, complete la siguiente tabla con los valores correspondientes de la potencia de la prueba y del error tipo II para cada uno de los valores de μ que se indican.

μ	Potencia de la prueba	β
255		
256		
257		

28. Una conocida empresa dedicada a la comercialización de pollos a la brasa, parrilladas y comida criolla brinda sus servicios en sus cuatro locales estratégicamente distribuidos en la capital. La venta de sus productos se realiza tanto en sus propios establecimientos como a domicilio (*delivery*). Con la finalidad de conocer el comportamiento del negocio, el gerente de la empresa decidió registrar al azar diversas variables correspondientes a las ventas realizadas durante el mes pasado. Algunas de estas variables son:

X_1: Monto de la factura en nuevos soles
X_2: Local (1, 2, 3, 4)
X_3: Entrega a domicilio (Sí, No)
X_4: Tipo de comida servida (pollos a la brasa, parrilladas o comida criolla)

Después de procesar la información con Minitab se formularon las siguientes preguntas:

Pregunta 1: El reporte de Minitab para comparar los montos de las ventas de comida criolla versus las ventas de pollos a la brasa es:

Two-Sample T-Test and CI: Monto_Comida Criolla; Monto_Pollos a la brasa

```
Two-sample T for Monto_Comida Criolla vs Monto_Pollos a la brasa
     N                      Mean           SE Mean
Monto_Comida Criolla          99    192.9           6.5
Monto_Pollos a la brasa      146    210.0           9.2
Difference = mu (Monto_Comida Criolla) - mu (Monto_Pollos a la brasa)
Estimate for difference: [          ]
T-Test of difference = 0 (vs <): T-Value = [          ] P-Value = [          ]
DF = [          ]
Both use Pooled StDev = 95.0928
```

a. Complete los espacios en blanco e indique sus conclusiones. Use $\alpha = 0.03$.
b. ¿Qué valores debe asumir α para rechazar la H_0?

Pregunta 2: El reporte de Minitab para comparar los montos de las ventas de los diferentes locales de la empresa es:

Descriptive Statistics: Monto

Variable Local	Count	Mean	Variance
La Molina	91	202.94	5002.60
Miraflores	121	193.67	4934.17
San Isidro	89	190.23	955.90
Surco	99	192.16	4428.17

Con $\alpha = 0.03$, y presentando el procedimiento completo de la prueba de hipótesis correspondiente, determine si el local de San Isidro vende más que el de Surco.

Pregunta 3: La distribución de frecuencias de las ventas del local de Miraflores es:

60-140	21
140-220	44
220-300	26
300-380	16
380-460	14

Si se desea probar que los montos de las ventas del local de Miraflores tienen distribución Normal con $\sigma^2 = 4900$.
Nota: Use $\alpha = 0.06$ en sus cálculos.

a. ¿Cuál es la regla de decisión para probar la hipótesis nula correspondiente?

b. Presente sus cálculos para obtener el valor de la prueba estadística correspondiente.

Pregunta 4: A continuación se presenta la tabla de contingencia de local versus tipo de comida.

Tabulated statistics: Local; Comida

Rows: Local Columns: Comida

Comida	Criolla	Parrillada	Pollo a la brasa	All
La Molina	19	30	91	42
Miraflores	34	40	47	121
San Isidro	20	31	38	89
Surco	26	33	40	9
All	99	146	155	400

Nota: Use $\alpha = 0.07$ en sus cálculos.
Presente el procedimiento completo de la prueba de hipótesis correspondiente para determinar si la variable local es independiente de la variable tipo de comida.

29. Cheese S.A. es una empresa importadora de quesos de España, Suiza, Inglaterra y Francia. Sus principales quesos son: cheddar, mozzarella, roquefort, parmesano y gorgonzola. El gerente de la empresa seleccionó al azar una muestra aleatoria de las ventas realizadas durante los últimos meses y registró la siguiente información:

C1: País de procedencia.
C2: Tipo de queso.
C3: Kilogramos vendidos.
C4: Destino del queso (repostería, comida, bufé).
Usando el archivo de datos Cheese.MTW conteste las siguientes preguntas:

Pregunta 1: Si la desviación estándar poblacional de los kilogramos vendidos es 3 y $\alpha = 0.06$.

a. ¿Puede el gerente afirmar que el promedio de kilogramos vendidos es mayor que 15? Señale:
   1. Hipótesis
      H_0
      H_1
   2. Prueba estadística y su valor:
   3. Región crítica y regla de decisión:
   4. Decisión y conclusión:

b. En a, ¿cuál es el error tipo I (α) si la región de rechazo es $\bar{x} > 15.2$?

c. En a, ¿cuál es el error tipo II (β) si el verdadero promedio es 15.3?

d. ¿Cuál es el tamaño de muestra necesario para probar H_0: $\mu = 15$ vs H_1: $\mu > 15$ con una probabilidad de aceptar la H_0 cuando es falsa de 0.04? Suponga que el verdadero promedio es igual a 15.2 kilogramos.

Pregunta 2: ¿Es el promedio de kilogramos vendidos de queso mozzarella menor que el peso promedio de queso roquefort? Con $\alpha = 0.03$, señale:

a. Hipótesis
   H_0:
   H_1:
b. Prueba estadística y su valor:
c. Región crítica y regla de decisión:
d. Decisión y conclusión:

Pregunta 3: ¿Es la proporción de quesos procedentes de España utilizados en repostería menor que la proporción de quesos procedentes de Francia utilizados en repostería? Con $\alpha = 0.04$, señale:

a. Hipótesis
   H_0:
   H_1:
b. Prueba estadística y su valor:
c. Región crítica y regla de decisión:
d. Decisión y conclusión:

30. Un inspector del Ministerio de Trabajo recibió el encargo de analizar el tiempo de trabajo diario (en horas) que realizan los miembros de dos conocidas compañías de seguridad: Top S.A. y Vipsa. Para este fin, el inspector seleccionó al azar muestras aleatorias de los tiempos de trabajo de los trabajadores de seguridad de ambas compañías y luego de utilizar Minitab obtuvo los siguientes reportes:

Nota importante: utilice en sus cálculos $\alpha = 0.02$.

Reporte I

Test for Equal Variances: Top S.A.; Vipsa

	N	Lower	StDev	Upper
Top S.A.	10	1.19426	1.82574	3.67641
Vipsa	9	0.53406	0.83333	1.77654

F-Test (Normal distribution)
Test statistic = [ ]; P-value = [ ]

a. Complete los espacios en blanco.
b. Conclusión.

Reporte II

Two-Sample T-Test and CI: Top S.A.; Vipsa

Two-sample T for Top S.A. vs Vipsa
Difference = mu (Top S.A.) - mu (Vipsa)
Estimate for difference: 2.22222
T-Test of difference=0 (vs >):T-Value = [ ]
P-Value = [ ] DF = [ ]

a. Complete los espacios en blanco. Utilice los resultados del reporte I.
b. Decisión y conclusión.

31. En una encuesta aplicada durante el 2005 a empresarios nacionales acerca del impacto del Tratado de Libre Comercio (TLC) con Estados Unidos en los diferentes sectores económicos se registraron las siguientes variables:

C1: Sector económico al que pertenece el empresario.
C2: Servicio o producto principal que ofrece el empresario.
C3: Ubicación de la empresa (centro, norte, oriente, sur).
C4: Ingreso estimado para el año 2005 (millones de dólares).
C5: Ingreso estimado del año 2006 en caso de firmarse el TLC (millones de dólares).
C6: Ingreso estimado del año 2006 en caso de NO firmarse el TLC (millones de dólares).
C7: Años de funcionamiento de la empresa.
Usando los datos que aparecen en el archivo TLC.MTW responda lo siguiente:

Pregunta 1. ¿Cuál es el tamaño de muestra para estimar el promedio de los ingresos estimados del año 2006 en caso de no firmarse el TLC con un error máximo de estimación de 0.2 millones de dólares con 0.98 de probabilidad y $\sigma = 0.5$ millones de dólares?

Pregunta 2. Según un instituto de estudios económicos, la contribución porcentual de cada sector económico al PBI es la siguiente:

PBI según sectores económicos	
Agropecuario	15
Pesca	10
Minería e hidrocarburos	25
Manufactura	5
Electricidad y agua	5
Construcción	18
Comercio	10
Otros servicios	12

¿Se ajustan los datos de la variable C1 al tipo de distribución que se señala en el instituto? Usando $\alpha = 0.03$, responda lo siguiente:

a. Hipótesis
   H_0:
   H_1:
b. Valor de la prueba estadística.
c. Valor(es) crítico(s) y regla de decisión.
d. P-value.
e. Decisión y conclusión.

Pregunta 3. ¿Puede afirmarse que existen diferencias significativas entre los promedios poblacionales de los ingresos del año 2006 en caso de NO firmarse el TLC de los sectores de manufactura y comercio? Use $\alpha = 0.02$ y los datos del archivo TLC.MTW.

Capítulo

4 Análisis de regresión

En este capítulo trataremos los siguientes temas:

- *Definiciones básicas*
- *Tipos de relaciones*
- *Tipos de modelos de regresión*
- *Análisis de regresión lineal simple*
- *Análisis de regresión lineal múltiple*

El análisis de regresión tiene por objetivos: primero, establecer la relación funcional entre una variable dependiente (respuesta) y un conjunto de variables independientes (predictoras o explicativas), y segundo, realizar pronósticos, tema de estudio abordado en este capítulo. También trata acerca de los tipos de relaciones y los tipos de modelo de regresión. Además, se presenta un detallado estudio del análisis de variancia de la regresión lineal simple y de la regresión lineal múltiple. Al finalizar, el lector será capaz de realizar un análisis de regresión lineal simple y múltiple haciendo uso de Excel y de Minitab.

1. Introducción

La técnica del análisis de regresión tiene por objetivo estudiar los modelos que mejor expliquen la relación estocástica cuantitativa entre una variable de interés y un conjunto de variables explicativas. Estos modelos son muy utilizados y se aplican en diferentes campos de la ciencia; su estudio conforma un área de investigación clásica dentro de la disciplina de la estadística desde hace muchos años.

Diversos investigadores han trabajado en este campo, el primero fue Sir Francis Galton (1822-1917); sus trabajos más importantes derivan de sus dos grandes aficiones: el estudio de la herencia y la expresión matemática de los fenómenos vinculados a ella. Galton asignó un número a un conjunto de variables, y de esta forma logró obtener una medida del grado de relación existente entre ellas. Sostenía la idea de que las personas excepcionalmente altas solían tener hijos de estatura menor a la de sus progenitores, mientras que personas muy bajas solían tener hijos más altos que sus padres, hecho que Galton enunció como la regresión a la mediocridad, aplicable a las tallas de una generación respecto de las siguientes. Este principio se considera la primera falacia sobre la Teoría de la Regresión. La justificación que se da hoy día a este hecho es que los valores extremos de una distribución se deben en gran parte al azar.

2. Definición

Cuando se sabe que existe alguna relación estocástica entre las variables, y hay necesidad de analizar este conjunto de variables, el método que proporciona la estadística es el análisis de regresión, el cual se define como un método de análisis de datos que sirve para poner en evidencia las relaciones estocásticas que existen entre diversas variables.

Ahora bien, la variable que influye sobre otra se denomina variable independiente (variable estímulo) y generalmente se denota por X, mientras que la varia-

ble que es influenciada se denomina variable respuesta (variable dependiente) y se representa por Y.

Los objetivos del análisis de regresión son:

- Obtener una ecuación que permita "predecir" el valor de Y una vez conocidos los valores de $X_1, X_2,..., X_k$, a estos modelos se les conoce como modelos predictivos.
- Conocer la relación funcional entre $X_1, X_2,..., X_k$, y la variable Y con el fin de conocer o explicar mejor los mecanismos de esa relación.

3. Tipos de relaciones

Las relaciones *entre variables* pueden presentarse en cualquiera de los tres casos siguientes:

Una variable que influye sobre otra:

En este caso se tiene que una variable X puede influir directamente sobre otra variable Y.

Ejemplo 1:

- La edad influye en el desarrollo mental del niño.
- La cantidad de proteína de la harina influye en el volumen del pan.
- El nivel de las lluvias influyen en la cantidad de la cosecha.

Variables influenciadas entre sí:

Puede presentarse el caso en que dos variables pueden estar influenciadas entre sí; es decir, existe una influencia recíproca.

Ejemplo 2:

- Precio y nivel de producción de un artículo.
- Peso y volumen de un producto perecedero.
- Peso y altura de las personas.

Variables no relacionadas influenciadas por otra variable:

En este caso dos variables pueden estar relacionadas entre sí en forma indirecta sin estar influenciadas directamente, por estar ambas influenciadas por una tercera variable asociada.

Ejemplo 3:

- El peso de los hermanos y el peso de las hermanas están relacionados por la influencia de la variable genética de los padres.

- Los precios del pan y de la leche están relacionados por la influencia del costo de vida a través de los años.
- Las notas de los cursos de Química y Bioquímica están relacionadas por la inclinación de los alumnos por los cursos de ciencias.

4. Tipos de modelo de regresión

Existen dos tipos de modelos de regresión: por la forma de influencia y por el número de variables independientes que influyen en la variable respuesta.

4.1 Por la forma de influencia

Existen los siguientes tipos de regresión:

a. Las variaciones de la variable independiente pueden provocar variaciones proporcionales en la variable respuesta. Este tipo de relación puede ser estudiado por los modelos de regresión lineal, siendo su representación matemática una recta. Ejemplo: el peso de un niño al nacer y su peso a los tres meses de edad.
b. Las variaciones de la variable independiente pueden provocar variaciones no proporcionales en la variable respuesta.
   Este tipo de relación debe analizarse a través de los modelos de regresión no lineal. Ejemplo: la precipitación pluvial de una zona y el rendimiento de los cultivos de dicha zona.

4.2 Por el número de variables independientes que influyen en la variable respuesta

Se pueden presentar los siguientes tipos de regresión:

a. Una sola variable independiente. Si solamente una variable independiente influye sobre la variable respuesta se tiene un modelo de regresión simple. Ejemplo: la cantidad de gluten de la harina de trigo y el volumen de pan.
b. Varias variables independientes. Si son varias las variables independientes que influyen en la variable respuesta se tiene un modelo de regresión múltiple. Ejemplo: donde la edad y la altura de las personas influyen sobre su peso.

5. Análisis de regresión lineal simple

El análisis de regresión lineal simple (RLS) permite estudiar una relación estocástica entre dos variables X e Y, donde los valores posibles de Y se pueden asociar con cualquier valor de X.

5.1 Metodología para la formulación de un modelo de regresión simple

Para cumplir con el objetivo de modelar la relación entre variables, se sugiere el siguiente procedimiento:

- *Especificación del modelo.*- Se refiere a la identificación de las variables que deben considerarse en el modelo, las relaciones que las unen, las propiedades estocásticas de las variables y las restricciones que afectan algunos de los parámetros.
- *Estimación*.- Consiste en estimar en forma satisfactoria, mediante operaciones matemáticas, los parámetros del modelo propuesto.
- *Verificación*.- Una vez realizada la estimación del modelo, es necesario comprobar la validez de este y de las estimaciones. Esto se realiza mediante una serie de pruebas estadísticas, que a su vez son comparadas con la realidad. De la confirmación o refutación de ellas depende la revisión o la reformulación del modelo.
- *Predicción*.- Si el modelo logra pasar el proceso de verificación, entonces está listo para ser utilizado. Los modelos de regresión usualmente son de utilidad para determinar el valor esperado de la variable respuesta sobre la base de valores conocidos de las variables independientes.

5.2 Especificación del modelo de regresión lineal simple

Especificación del modelo poblacional de la regresión lineal simple. Se propone el siguiente modelo poblacional:

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

Donde:

β_0, β_1 : Coeficientes de regresión a estimarse; β_0 es denominado intercepto y β_1 es denominado pendiente.
Y_i : Variable respuesta, explicada, variable pronosticada para la i-ésima observación.
X_i : Variable independiente, explicativa, predictora, regresora, etcétera.
U_i : Variable aleatoria no observable que puede tomar cualquier valor, se le conoce como variable perturbadora o error estadístico. Esta variable representa a las demás variables no consideradas en el modelo, a los errores de muestreo y cualquier otro aspecto no especificado en el modelo.

El supuesto es que la variable aleatoria Y está formada por una parte predecible la cual es función lineal de X y una parte no predecible que es el error aleatorio, este error aleatorio (U_i) incluye efectos de todos los otros factores no considerados en el modelo. En un estudio de regresión se pueden alcanzar varios objetivos:

- Los datos pueden utilizarse para estimar la magnitud de variabilidad o incertidumbre de la ecuación propuesta.
- Como los datos corresponden a una muestra, pueden utilizarse para estimar los valores verdaderos poblacionales de los parámetros considerados en la regresión.
- La ecuación de predicción se puede utilizar para predecir un rango razonable de valores futuros de la variable respuesta.

5.2.1 Supuestos básicos del modelo de regresión lineal simple

Se debe tener en cuenta que la variable explicativa X debe ser considerada como fija, es decir, X es una variable matemática medida sin error.

Un modelo de regresión lineal simple debe satisfacer los supuestos que se presentan a continuación:

- Supuesto N° 1
  En promedio el valor esperado de los errores U_i es cero (0) es decir, hay errores por exceso y por defecto que en promedio se anulan.

$$E(U_i \mid X_i) = 0$$

- Supuesto N° 2
  El error cometido en la i-ésima observación no depende del error cometido en la j-ésima observación, cuando esta suposición no es satisfecha se tiene un problema de autocorrelación.

$$\text{Covarianza} \begin{cases} Cov(U_i\ ,\ U_j) = E\left[(U_i - E(U_i))(U_j - E(U_j))\right] \\ E(U_i U_j) = 0 \qquad i \neq j \end{cases}$$

  Esto quiere decir que U_i y U_j no están correlacionados. También se conoce como la independencia de las observaciones.

- Supuesto N° 3
  La varianza de los errores U_i para cada X_i es un número constante σ^2; representa el supuesto de homocedasticidad o igual dispersión, es decir, que las poblaciones tienen igual varianza. Esto es:

$$V(U_i \mid X_i) = E\left[U_i - E(U_i)\right]^2 = E\left[U_i\right]^2 = \sigma^2$$

5.3 Estimación de parámetros en un modelo de regresión lineal simple

En situaciones prácticas, lo que está al alcance del investigador es una muestra de valores de Y correspondiente a $X's$ fijos, por consiguiente la tarea es la estimación de los parámetros β_0 y β_1 utilizando información muestral.

El modelo de regresión muestral: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$ (1)

Donde:

$\hat{\beta}_0$: Término constante, es la ordenada en el origen o intercepto y se interpreta como el valor estimado o predicho de Y cuando X es 0.

$\hat{\beta}_1$: Pendiente, es el cambio pronosticado en Y cuando hay un cambio unitario en X.

e_i : Término residual.

Si se define $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, se tiene que: $Y_i = \hat{Y}_i + e_i$.

La expresión $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ es conocida como la ecuación de regresión estimada.

Para la estimación de los parámetros de regresión β_0 y β_1 se emplea el método de los mínimos cuadrados ordinarios (MCO), que consiste en minimizar las sumas de los cuadrados de los residuales.

Se sabe que:

$$Y_i = \hat{Y}_i + e_i \Rightarrow e_i = Y_i - \hat{Y}_i$$

Esto quiere decir que el error residual es la diferencia entre el valor observado y el valor estimado.

Con el método de mínimos cuadrados ordinarios se desea minimizar la siguiente expresión:

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \qquad (2)$$

Derivando la expresión (2) respecto a $\hat{\beta}_0$ y $\hat{\beta}_1$ se obtiene:

$$\frac{\partial Q}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \qquad (3)$$

$$\frac{\partial Q}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} (X_i Y_i - X_i \beta_0 - \beta_1 X_i^2) \qquad (4)$$

Igualando a cero las expresiones (3) y (4), y luego de ordenar se obtienen las siguientes ecuaciones consideradas como "ecuaciones normales":

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \qquad (5)$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \qquad (6)$$

Al despejar los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ de las expresiones (5) y (6) se tiene:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \qquad \hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$

Para verificar si los valores hallados son mínimos, se obtiene la segunda derivada:

$$\frac{\partial^2 Q}{\partial {\beta_0}^2} = 2n > 0$$

$$\frac{\partial^2 Q}{\partial \beta_1^2} = 2\sum_{i=1}^{n} X_i^2 > 0$$

Como los valores son siempre positivos, entonces los valores de los estimadores hallados son mínimos.

De acuerdo con el teorema de Gauss-Markov, el cual se fundamenta en los supuestos del modelo clásico de regresión lineal, se puede concluir que los estimadores mínimos cuadráticos hallados son óptimos o de mínima varianza dentro de la clase de estimadores insesgados que son funciones lineales de las observaciones (uniformes).

Ejemplo 4:

Una empresa que se dedica a la venta de pizzas a domicilio desea determinar si existe una relación entre los gastos en publicidad y las ventas semanales. La tabla adjunta muestra la información de las últimas ocho semanas:

Gastos en publicidad	0	100	250	350	450	500	600	700
Ventas semanales (unidades)	120	350	500	550	550	650	800	1100

Solución:

El primer paso en la determinación del modelo es verificar el tipo de relación existente entre las dos variables en estudio; esto es, al examinar la gráfica de dispersión que a continuación se presenta.

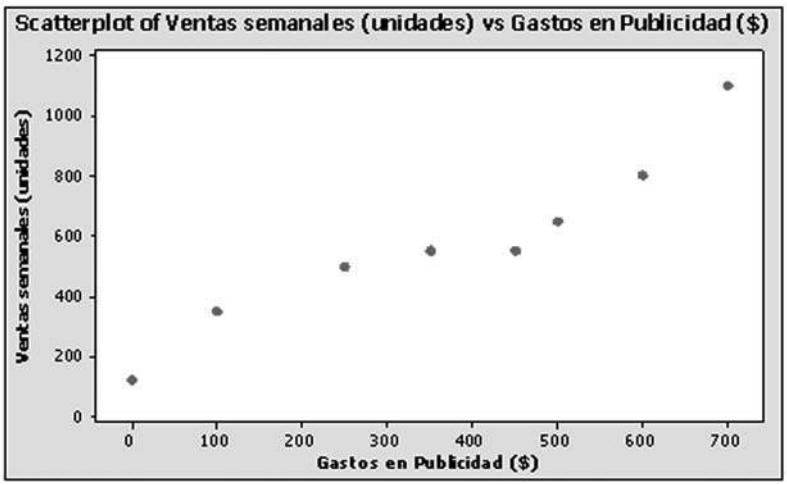


Figura 1. Diagrama de dispersión.

Se observa un patrón creciente entre las variables, es decir que a medida que aumentan los gastos de publicidad aumentan las ventas semanales de pizza.

Por consiguiente, el modelo poblacional que se propone es el modelo lineal:

$$\text{Ventas}_i = \beta_0 + \beta_1 * \text{Gasto en publicidad}_i + u_i \; ; \; i = 1, 2, \cdots, 8$$

Donde $\text{Ventas}_i = Y_i$ y $\text{Gasto en publicidad}_i = X_i$

- Estimación de los parámetros.

  Se procede a la estimación de los parámetros β_0 y β_1 para reemplazarlos en el modelo ajustado o estimado:

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{8(2175000) - 2950(4620)}{8(1497500) - 8702500} = \frac{3771000}{3277500} = 1.150572$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 577.5 - 1.1505721(368.75) = 153.226538$$

  La ecuación de regresión estimada es:

$$\text{Ventas}_i = 153.2265 + 1.15057 * \text{Gasto en publicidad}_i$$

- Interpretación:
  En promedio, las ventas semanales de pizzas son de 153 unidades cuando no hay gastos en publicidad. Cuando los gastos en publicidad aumentan en $1, las ventas aumentan en promedio 1.15 pizzas.

  El reporte del software Minitab es:

```
Regression Analysis: Ventas Semanales versus Gasto Publicidad
The regression equation is
Ventas Semanales = 153 + 1.15 Gasto Publicidad

Predictor            Coef
Constant           153.23
Gasto Publicidad   1.1506
```

5.3.1 Varianza de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$

Las siguientes fórmulas permiten obtener la varianza y desviación estándar de los estimadores:

$$V(\hat{\beta}_0) = \frac{\sigma_e^2}{n} + \bar{x}^2 \left(\frac{\sigma_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \qquad \textit{Desv.Estándar} \quad (\hat{\beta}_0) = \sqrt{V(\hat{\beta}_0)}$$

$$V(\hat{\beta}_1) = \frac{\sigma_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \textit{Desv.Estándar} \ (\hat{\beta}_1) = \sqrt{V(\hat{\beta}_1)}$$

El propósito de obtener la varianza de estos estimadores es la construcción de los intervalos de confianza, en forma práctica se trabaja con los estimadores de la varianza, es decir con la varianza muestral.

$$V(\hat{\beta}_0) = \frac{s_e^2}{n} + \bar{x}^2 \left(\frac{s_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) = s_e^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

$$V(\hat{\beta}_1) = \frac{s_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_e^2}{\left(\sum_{i=1}^{n} (x_i)^2 \right) - n\bar{x}^2}$$

Donde:

$$s_e^2 = \frac{\sum_{i=1}^{n} (e_i)^2}{n-2} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2}{n-2} = \frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i}{n-2}$$

En el caso de las pizzas (ejemplo 4), al reemplazar los valores se obtiene la siguiente tabla con algunos de los cálculos realizados:

Tabla de cálculos para el ejemplo 4.

Ventas semanales Y	Gasto publicidad X	$\hat{Y}$	e	e^2	X^2
120	0	153.227	-33.227	1104.003	0
350	100	268.284	81.716	6677.545	10000
500	250	440.87	59.13	3496.408	62500
550	350	555.927	-5.927	35.12664	122500
550	450	670.984	-120.984	14637.12	202500
650	500	728.513	-78.513	6164.226	250000
800	600	843.57	-43.57	1898.327	360000
1100	700	958.627	141.373	19986.32	490000
				53999.08	1497500

- $s_e^2 = \frac{\sum_{i=1}^{n} (e_i)^2}{n-2} = \frac{53999.08}{6} = 8999.8467$

- $V(\hat{\beta}_0) = s_e^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) = 8999.8467 \left(\frac{1}{8} + \frac{135976.56}{409687.5} \right) = 4112.058048$

  Entonces, el error estándar $s(\hat{\beta}_0) = \sqrt{V(\hat{\beta}_0)} = \sqrt{4112.058048} = 64.125331$

- $V(\hat{\beta}_1) = \frac{s_e^2}{\left(\sum_{i=1}^{n} (x_i)^2 \right) - n\bar{x}^2} = \frac{8999.8467}{409687.5} = 0.021968$

Entonces, el error estándar $s(\hat{\beta}_1) = \sqrt{V(\hat{\beta}_1)} = \sqrt{0.021968} = 0.148215$

Usando el software Minitab:

Regression Analysis: Ventas Semanales versus Gasto Publicidad

```
Predictor            SE Coef
Constant               64.13
Gasto Publicidad      0.1482
```

5.3.2 Intervalos de confianza para los parámetros

Los intervalos de confianza para los parámetros vienen dados por las siguientes expresiones:

- Intervalo de confianza para el término constante:

$$\beta_0 \in \left\langle \hat{\beta}_0 \mp t_{(1-\alpha/2,n-2)} s(\hat{\beta}_0) \right\rangle$$

- Intervalo de confianza para la pendiente: $\beta_1 \in \left\langle \hat{\beta}_1 \mp t_{(1-\alpha/2,n-2)} s(\hat{\beta}_1) \right\rangle$

  Para el ejemplo 4, se tiene:

- Intervalo de confianza para el término constante.

$$\beta_0 \in \left\langle \hat{\beta}_0 \mp t_{(1-\alpha/2,n-2)} s(\hat{\beta}_0) \right\rangle$$

$$\beta_0 \in \left\langle 153.226538 \mp t_{(0.975,6)} (64.125331) \right\rangle$$

Inverse Cumulative Distribution Function

```
Student's t distribution with 6 DF
P( X <= x )          x
       0.975    2.44691
```

$$\beta_0 \in \left\langle 153.226538 \mp 2.44691 (64.125331) \right\rangle$$

$$\beta_0 \in \langle -3.682376, 310.135452 \rangle$$

- Intervalo de confianza para la pendiente.

$$\beta_1 \in \left\langle \hat{\beta}_1 \mp t_{(1-\alpha/2,n-2)} s(\hat{\beta}_1) \right\rangle$$

$$\beta_1 \in \left\langle 1.150572 \mp t_{(0.975,6)} (0.148215) \right\rangle$$

$$\beta_1 \in \left\langle 1.150572 \mp 2.44691 (0.148215) \right\rangle$$

$$\beta_1 \in \langle 0.787903, 1.513241 \rangle$$

5.4 Tabla de análisis de varianza (Anova)

Con la finalidad de saber que tan bien predice la variable estímulo a la variable respuesta, se estudiará la variación de la variable Y. La variación total de los valores observados de Y alrededor de su media puede ser dividida en dos: una atribuible al modelo de regresión (variación explicada) y la otra a factores aleatorios (variación no explicada), tal como se presenta en la siguiente expresión:

$$\underbrace{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}_{\text{Variación total}\ (SCT)} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2}_{\text{Variación explicada}\ (SCR)} + \underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y})^2}_{\text{variación no explicada}\ (SCE)}$$

donde:

- $\sum_{i=1}^{n}(Y_i-\overline{Y})^2$: Suma de Cuadrados Total (SCT). Expresa las desviaciones de las observaciones respecto al promedio total. Si SCT tiende al valor cero, se concluye que no existe variabilidad en la variable respuesta.

- $\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2$: Suma de Cuadrados de la Regresión (SCR). Expresa las desviaciones de los valores ajustados respecto al promedio de los valores de Y. Si el valor de SCR se aproxima al valor de SCT, se concluye que el modelo propuesto es adecuado.

- $\sum_{i=1}^{n}(Y_i-\hat{Y})^2$: Suma de Cuadrados del Error (SCE). Expresa las desviaciones de los valores observados respecto a los valores ajustados. Si SCE tienden a cero, entonces todas las observaciones caen en la línea de regresión, por consiguiente el modelo es adecuado.

Esta partición puede ser representada en una tabla llamada Tabla de análisis de varianza (Anova o Anva):

Tabla de análisis de varianza — Regresión lineal simple

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F_0
Debido a la regresión	1	$SCR=\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2$	$CMR=\frac{SCR}{1}$	$F_0=\frac{CMR}{CME}$
Debido al error	n – 2	$SCE=\sum_{i=1}^{n}(Y_i-\hat{Y})^2$	$CME=\frac{SCE}{n-2}=s_e^2$	
Total	n – 1	$SCT=\sum_{i=1}^{n}(Y_i-\overline{Y})^2$		

Nótese que $CME=s_e^2$.

5.5 Verificación del modelo

Para verificar la validez estadística del modelo de regresión lineal simple propuesto se utiliza lo siguiente:

- Coeficiente de determinación: R^2.
- Coeficiente de correlación lineal simple: r.
- Pruebas de significación: Pruebas T y F.

5.5.1 *Coeficiente de determinación* (R^2)

El coeficiente de determinación indica en qué porcentaje la variable estímulo explica a la variable respuesta. Este coeficiente expresa la relación entre dos tipos de variación:

- V1 = Variación de los valores de Y alrededor de la línea de regresión.
- V2 = Variación de los valores de Y alrededor de su propia media.

$$R^2 * 100\% = \frac{SCR}{SCT} * 100\% = \frac{\text{Variación alrededor de la línea}}{\text{Variación alrededor de su media}} * 100\% = \frac{V1}{V2} * 100\%$$

Por consiguiente R^2, expresado en porcentaje, mide la variación total en Y explicada por el modelo de regresión. Por ser R^2 cociente entre dos sumas de cuadrados, luego de multiplicarlo por 100%, el mínimo valor que puede tomar es 0 y el máximo valor que puede tomar es 100%.

Ejemplo 5:

Para el ejemplo 4, usando software Minitab:

```
Analysis of Variance
Source          DF        SS        MS      F       P
Regression       1    542351    542351  60.26   0.000
Residual Error   6     53999      9000
Total            7    596350
```

Entonces, se tiene que: $R^2 * 100\% = \frac{542351}{596350} * 100\% = 90.9451\%$

Se puede concluir que la variación en gastos de publicidad explica el 90.9451% de la variación de ventas semanales de pizzas.

5.5.2 *Coeficiente de correlación lineal simple (r)*

El coeficiente de correlación lineal simple es una medida que indica el grado de asociación lineal entre dos variables; el coeficiente de correlación lineal simple se obtiene de la siguiente expresión:

$$r = \frac{\sum_{i=i}^{n} X_i Y_i - \frac{\sum_{i=i}^{n} X_i \sum_{i=i}^{n} Y_i}{n}}{\sqrt{\sum_{i=i}^{n} X_i^2 - \frac{\left(\sum_{i=i}^{n} X_i\right)^2}{n}} \sqrt{\sum_{i=i}^{n} Y_i^2 - \frac{\left(\sum_{i=i}^{n} Y_i\right)^2}{n}}}$$

En el caso de la regresión lineal simple, se cumple que $r = \pm\sqrt{R^2}$. En un modelo de regresión lineal simple, el signo del coeficiente de correlación corresponde al signo de la pendiente $\hat{\beta}_1$. El rango de r es: $-1 \le r \le 1$.

Si el coeficiente de correlación es positivo y tiende a 1, se dice que hay una relación directa y significativa entre las variables, si el coeficiente de correlación es negativo y tiende a -1 se dice que hay una relación inversa y significativa entre las variables, si el coeficiente de correlación es cero no existe relación entre las variables.

Ejemplo 6:

Para el ejemplo 4, se tiene que: $r = \sqrt{R^2} = \sqrt{0.909451} = 0.953651$

El grado de asociación entre las variables ventas semanales y gastos en publicidad es de 0.953651; por lo cual hay una relación directa y significativa entre dichas variables.

5.5.3 Pruebas de significación de las variables. Prueba T

Las pruebas individuales o pruebas *T* son independientes para cada parámetro del modelo de regresión lineal simple.

Procedimiento:

i. Hipótesis.

$H_0 : \beta_i = 0$ (la variable Xi no es significativa en el modelo).

$H_1 : \beta_i \neq 0$ (la variable Xi sí es significativa en el modelo).

ii. Especificación del nivel de significación o riesgo: α y suposiciones.

iii. Obtención de la estadística de prueba: $t_0 = \frac{\hat{\beta}_i - \beta_i}{s(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$, tal que H0 es verdadera.
Donde:

$\hat{\beta}_i$: Estimador.

$s(\hat{\beta}_i)$: Error estándar del estimador.

iv. Región crítica y regla de decisión.

$\text{RC} = \langle -\infty, t_{(\alpha/2, n-2)} \rangle \cup \langle t_{(1-\alpha/2, n-2)}, \infty \rangle$

Rechazar H_0 si t_0 RC, es decir, si: $t_0 < t_{(\alpha/2, n-2)}$ ó $t_0 > t_{(1-\alpha/2, n-2)}$

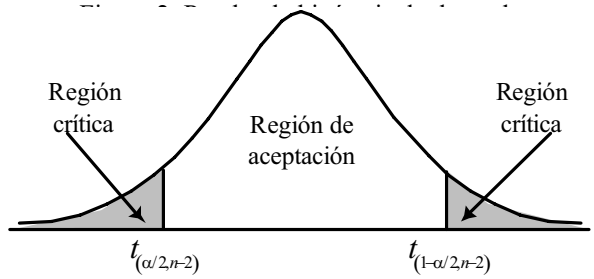


Figura 2. Prueba de hipótesis de dos colas.

También se puede calcular y utilizar el $\text{P-value} = 2 * P(t_{(n-2)} > t_0)$

v. Si t_0 , valor de la estadística de prueba pertenece a la región crítica se rechaza la hipótesis nula; en caso contrario no se rechaza.
Si $\text{P-value} < \alpha$, entonces se rechaza la hipótesis nula (H_0).

Ejemplo 7:

Para el ejemplo 4, se tiene que:

- Prueba con respecto al intercepto:
  i. Hipótesis por probar.

  $H_0 : \beta_0 = 0$ (el intercepto no es significativo).

  $H_1 : \beta_0 \neq 0$ (el intercepto sí es significativo).

  ii. Nivel de significación $\alpha = 0.05$ y las ventas semanales se distribuyen normalmente.

  iii. Estadístico de prueba: $t_0 = \dfrac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)} = \dfrac{153.226538 - 0}{64.125331} = 2.389485$

  iv. Región crítica: Como los puntos críticos son $t_{(0.025,6)} = -2.447$ y $t_{(0.975,6)} = 2.447$ por consiguiente la región crítica es
  $\text{RC} = \langle -\infty, -2.447 \rangle \cup \langle 2.447, \infty \rangle$.

  v. Decisión y conclusión: Como $2.3894 < 2.447$, entonces t_0 no pertenece a la región crítica (pertenece a la región de aceptación), y por lo tanto no se rechaza H_0. Es decir, la recta pasa por el origen.

- Prueba con respecto a la pendiente:
  i. Hipótesis por probar.

  $H_0 : \beta_1 = 0$ (la variable X_1 no es significativa en el modelo).

  $H_1 : \beta_1 \neq 0$ (la variable X_1 sí es significativa en el modelo).

  ii. Nivel de significación $\alpha = 0.05$ y las ventas semanales se distribuyen normalmente.

  iii. Estadística de prueba: $t_0 = \dfrac{\hat{\beta}_1 - \beta_1}{s\left(\hat{\beta}_1\right)} = \dfrac{1.150572 - 0}{0.148215} = 7.762858$

  iv. Región crítica: Como los puntos críticos son $t_{(0.025,6)} = -2.4469$ y $t_{(0.975,6)} = 2.4469$; por consiguiente la región crítica es
  $\text{RC} = \langle -\infty, -2.4469 \rangle \cup \langle 2.4469, \infty \rangle$

v. Decisión y conclusión: Como $t_0 = 7.7628 > 2.4469$ se rechaza la hipótesis nula, la variable gastos en publicidad influye en las ventas semanales de las pizzas.

$$\text{P-value} = 2 * P(t > 7.7629) = 2 * (1 - P(t \leq 7.7629)) = 2 * (1 - 0.99988) = 0.00024$$

Trabajando con tres valores decimales se puede considerar, para este caso, que el valor del P-value es igual a cero.
Usando el software Minitab para hallar la estadística de prueba t_0:

```
Predictor                  Coef  SE Coef     T      P
Constant                153.23    64.13  2.39  0.054
Gastos en Publicidad ($) 1.1506   0.1482  7.76  0.000
```

5.5.4 *Prueba de significación del modelo. Prueba F*

La presente prueba permite determinar si el modelo de regresión lineal es apropiado o aceptable para explicar la relación entre las variables en estudio.

Procedimiento:

i. Hipótesis por probar.
$H_0 : \beta_1 = 0$ (el modelo no es apropiado).
$H_1 : \beta_1 \neq 0$ (el modelo es apropiado).

ii. Nivel de significancia o riesgo: α y suposiciones.

iii. Estadística de prueba: $F_0 = \dfrac{CMR}{CME} \square F_{(1,n-2)}$, tal que H_0 es verdadera.

iv. Región crítica: $RC = \langle F_{(1-\alpha,1,n-2)}, \infty \rangle$ (prueba de una cola a la derecha)

F crítico: $F_{(1-\alpha,1,n-2)}$

También se puede calcular y utilizar el P-value: $P(F_{(1,n-2)} > F_0)$.

v. Si $F_0 > F$ crítico se rechaza la hipótesis nula, en caso contrario no se rechaza.
Si $\text{P-value} < \alpha$, entonces se rechaza la hipótesis nula (H_0).

Ejemplo 8:

En el ejemplo 4, se tiene que:

i. Hipótesis por probar.
$H_0 : \beta_1 = 0$ (el modelo no es apropiado).
$H_1 : \beta_1 \neq 0$ (el modelo es apropiado).

ii. Nivel de significación $\alpha = 0.05$.

iii. Estadística de prueba: $F_0 = \dfrac{CMR}{CME} = \dfrac{542351}{9000} = 60.261222$ y las ventas semanales se distribuyen normalmente.

Tabla de Análisis de Varianza (Anova)

Hallando la estadística de prueba F_0 haciendo uso del software Minitab:

```
Analysis of Variance
Source          DF        SS      MS      F
Regression       1    542351  542351  60.26
Residual Error   6     53999    9000
Total            7    596350
```

iv. Punto crítico $F_{(1-\alpha,1,6)} = 5.98738$.

v. Decisión y conclusión: Como $F_0 > F$ crítico se rechaza la hipótesis nula, entonces el modelo es apropiado.

$\text{P-value} = P(F > 60.26) = 0.00024$. Nótese que $0.00024 < 0.05$, lo que confirma la decisión de rechazar H_0.

Ejemplo 9:

Se desea determinar la influencia de la edad sobre el peso (g) en una muestra de 30 pollos desde 1 hasta 10 semanas de nacidos en 3 granjas. Los resultados se presentan a continuación:

Edad en semanas	Peso de pollos (g)		
	Granja 1	Granja 2	Granja 3
1	36	50	24
2	50	81	74
3	35	75	75
4	90	130	80
5	100	100	50
6	150	150	100
7	180	200	140
8	180	150	100
9	100	100	200
10	200	100	150

Usando software Minitab:

a. Para observar la gráfica de la dispersión de los puntos se ingresa a la opción Graph / Scaterplot, tal como se presenta en la figura 3.

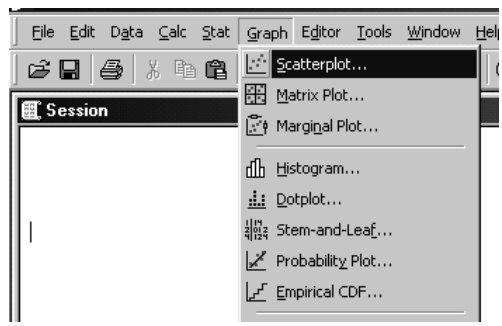


Figura 3. Secuencia de comandos.

El gráfico resultante aparece en la figura 4.

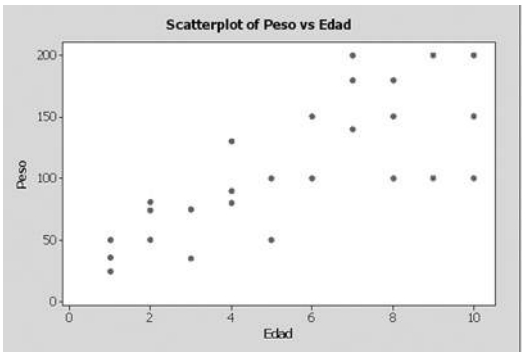


Figura 4. Diagrama de dispersión.

Interpretación: Se observa un patrón creciente entre la edad y el peso de los pollos.

b. Para obtener los resultados de la regresión lineal simple se procede como se indica en las figuras 5 y 6.

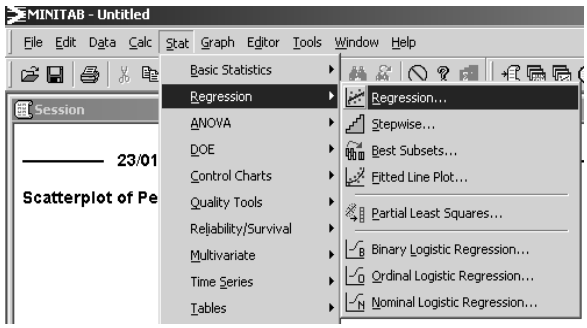


Figura 5. Secuencia para el análisis de regresión

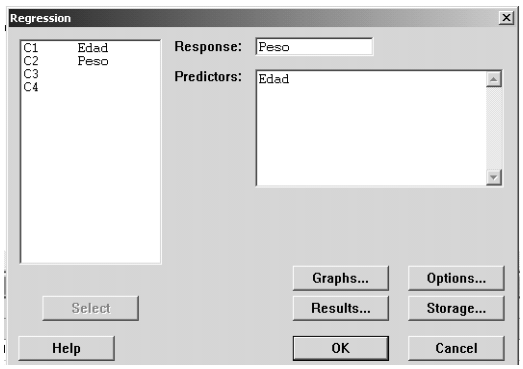


Figura 6. Ingreso de variables

Considerando intercepto en el modelo:

Regression Analysis: Peso versus Edad

The regression equation is

Peso = 36.6 + 13.1 Edad

Interpretación de los coeficientes estimados:

- Del modelo estimado se observa que cuando los pollos nacen estos pesan en promedio 36.6 gramos.
- De la estimación de la pendiente, por cada semana adicional en la edad, el peso de los pollos se incrementa en un promedio de 13.1 gramos (debido a que el signo de 13.1 es positivo, la relación entre edad y peso es directa).

Pruebas de significación:

Para observar la prueba de significancia para el intercepto y el coeficiente:

```
Predictor    Coef    SE Coef       T       P
Constant    36.56      14.03    2.60   0.015
Edad       13.051      2.262    5.77   0.000
```

- Con respecto a las pruebas de significación del intercepto y del coeficiente asociada a la variable edad:
  Intercepto:

  $H_0 : \beta_0 = 0$ (el intercepto no es significativo).

  $H_1 : \beta_0 \neq 0$ (el intercepto sí es significativo).

  Coeficiente asociada a la variable edad:

  $H_0 : \beta_1 = 0$ (la variable X1 no es significativa en el modelo).

  $H_1 : \beta_1 \neq 0$ (la variable X1 sí es significativa en el modelo).

  Se observa que para cada uno de ellos el P-value es menor que $\alpha = 0.05$, por consiguiente se rechaza la hipótesis nula. Es decir, el intercepto y la variable edad influyen en el modelo.

Prueba de validación:

```
S = 35.5807   R-Sq = 54.3%   R-Sq(adj) = 52.7%
```

- De acuerdo con el coeficiente de determinación, el 54.3% de la variación total del peso se debe a la edad de los pollos.

Prueba de significación del modelo:

$H_0 : \beta_1 = 0$ (el modelo no es apropiado)

$H_1 : \beta_1 \neq 0$ (el modelo es apropiado)

```
Analysis of Variance
Source           DF       SS      MS      F      P
Regression        1    42153   42153  33.30  0.000
Residual Error   28    35448    1266
Total            29    77601
```

- Con respecto a la prueba F de idoneidad del modelo se tiene un P-value igual a 0 y, por lo tanto menor que $\alpha = 0.05$, lo que implica que se rechaza la hipótesis nula y por consiguiente el modelo es apropiado.

Ejemplo 10:

En un artículo sobre pavimentos se indica que existe una relación entre el porcentaje de porosidad y el peso unitario (lb/pie^3) en muestras de concreto. Se tomaron 15 muestras, la información registrada se presenta en la siguiente tabla:

Muestra	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Peso unitario	99	101	102.6	103.1	105.6	107.2	108.7	110.9	113.1	113.2	113.5	113.6	115.2	115.5	120.1
% de porosidad	28.9	27.8	27.1	25.3	22.9	21.7	20.9	19.7	17	18.6	16.1	16.8	13.1	13.5	10.9

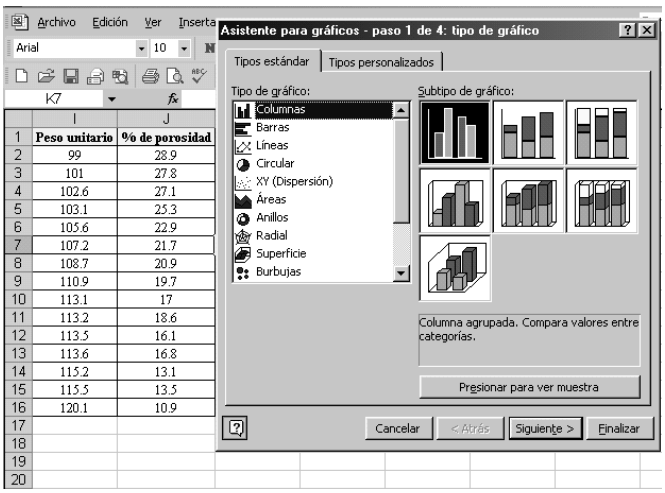
a. Haciendo uso de Excel represente la información de las variables mediante un diagrama de dispersión.
b. Determine la ecuación ajustada de mínimos cuadrados.
c. Determine el coeficiente de correlación e interprete los resultados.

Solución:

Empleando Excel:

a. Diagrama de dispersión

Seleccionar el tipo de gráfico en el icono de la barra superior  , elegir siguiente, elegir dispersión y luego siguiente (véase la figura 7).



	I	J
1	Peso unitario	% de porosidad
2	99	28.9
3	101	27.8
4	102.6	27.1
5	103.1	25.3
6	105.6	22.9
7	107.2	21.7
8	108.7	20.9
9	110.9	19.7
10	113.1	17
11	113.2	18.6
12	113.5	16.1
13	113.6	16.8
14	115.2	13.1
15	115.5	13.5
16	120.1	10.9
17		
18		
19		
20		

Figura 7. Gráfico en Excel.

En la pestaña Rango de datos se selecciona el rango en el que están contenidos los datos, incluyendo los nombres de las variables y se indica si los datos están en filas o en columnas; para nuestro caso columna seleccione la pestaña serie para comprobar si las variables corresponden a los valores de X e Y, y luego elija <Siguiente> (véase la figura 8).

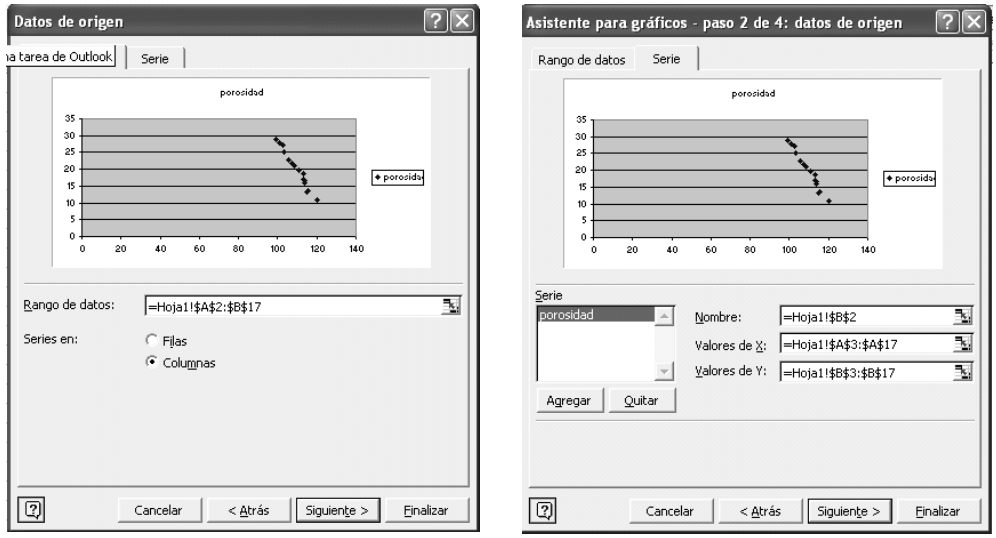


Figura 8. Elección del rango para el gráfico.

Se puede colocar el título del gráfico y los nombres para los ejes, etcétera. Seleccionar <Siguiente> y luego en <Finalizar> (véase la figura 9).

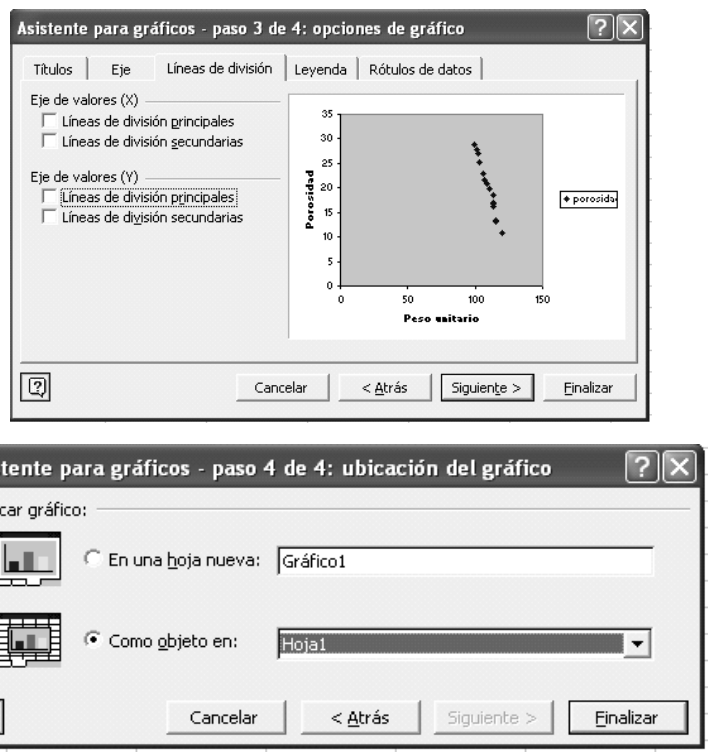


Figura 9. Nombre para los ejes.

b. Ecuación de mínimos cuadrados

En el diagrama de dispersión mostrado en la figura 9 se puede agregar la línea de tendencia, la ecuación de la recta y el coeficiente de correlación de la siguiente manera:

En el gráfico de dispersión se selecciona cualquier observación. Al presionar el botón derecho del *mouse* aparece una ventana, seleccionar <Agregar línea de tendencia> (véase la figura 10).

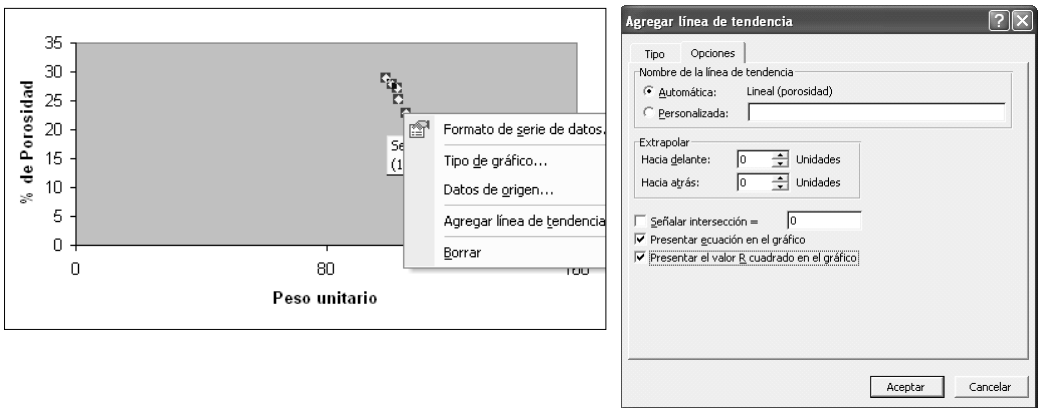


Figura 10. Selección del gráfico y opciones para línea de tendencia.

Seleccionar la opción <lineal> y en la pestaña de <Opciones> de la herramienta principal seleccionar <Presentar ecuación en el gráfico> y <Presentar el valor R cuadrado en el gráfico> (véase la figura 11).

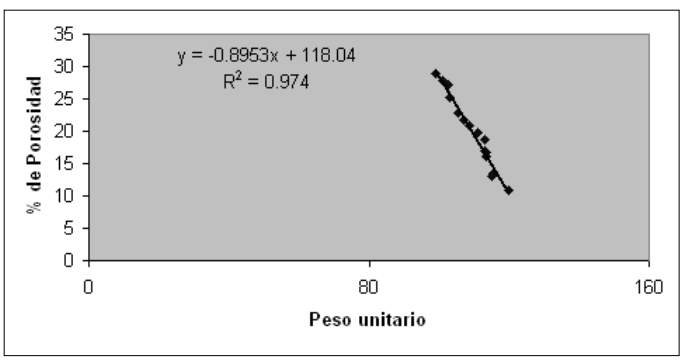


Figura 11. Gráfico con línea de tendencia.

La ecuación de mínimos cuadrados ajustada es:
Porcentaje de porosidad = 118.04 -0.8953*Peso unitario. Nótese que $\hat{\beta}_1 = -0.8953$, esto significa que si el peso unitario aumenta en una unidad el porcentaje de porosidad disminuye en 0.8953.

c. El coeficiente de correlación

El coeficiente de correlación es obtenido de la raíz cuadrada del coeficiente de determinación: $r = -\sqrt{R^2} = -\sqrt{0.974} = -0.986914$, por consiguiente existe una relación inversa y significativa entre las dos variables, es decir cuando el peso unitario aumenta el porcentaje de porosidad disminuye. Obsérvese que se considera el signo negativo por la relación inversa existente entre peso y porosidad.

6. Análisis de regresión lineal múltiple (RLM)

En este modelo de regresión, la variable dependiente se encuentra relacionada en forma lineal con dos o más variables regresoras o independientes.

6.1 Especificación del modelo de RLM

El modelo de RLM con k variables regresoras se puede representar de la siguiente manera:

Modelo de regresión poblacional:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + U_i \text{ , donde: } i = 1, 2, \cdots, n$$

Donde:

β_0 : Intercepto.
β_1, …, β_k : Coeficientes de regresión.
n : tamaño de la muestra.

La función de regresión poblacional se debe interpretar como la media o valor esperado de Y condicionado a los valores fijos de X. Como se considera una muestra de n observaciones, donde cada observación considera a las k variables regresoras se obtiene entonces un conjunto de ecuaciones lineales, como a continuación se detalla:

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + ... + \beta_k X_{1k} + U_1$$

$$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + ... + \beta_k X_{2k} + U_i$$

$$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_k X_{nk} + U_n$$

En forma abreviada:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + U_i \quad \text{donde: } i = 1, 2, \cdots, n$$

Este sistema de ecuaciones se puede expresar usando matrices.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{1K} \\ 1 & X_{21} & X_{22} & X_{2k} \\ \vdots & & & \\ 1 & X_{n1} & X_{n2} & X_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}$$

Donde:
Y: Vector columna de observaciones de la variable respuesta de orden (nx1).
X: Matriz nx(k+1) de las variables explicativas.
β: Vector (k+1)x1 de parámetros desconocidos.
U: Vector columna de variables de perturbación de orden nx1.

En forma abreviada: $Y_{n\text{x}1} = X_{n\text{x}(k+1)}\beta_{(k+1)\text{x}1} + U_{n\text{x}1}$

6.1.1 Supuestos básicos del modelo de RLM

Un modelo de RLM debe satisfacer los siguientes supuestos básicos:

- Supuesto N.º 1
  En promedio el valor esperado de U es 0; es decir, hay errores por exceso y por defecto, que en promedio se anulan
  $$E(U_i) = 0$$
- Supuesto N.º 2
  Los errores U_1, U_2, …, U_n son independientes y tienen varianza constante.
  $$Var(U) = E(UU') = \sigma^2 I$$
  La matriz V(U) se denomina matriz de Varianza - Covarianza
- Supuesto N.º 3
  La matriz $X_{n\text{x}(k+1)}$ es no estocástica, lo cual implica que está formada por números fijos.
- Supuesto N.º 4
  La matriz X tiene un rango igual al número de columnas de la matriz, en este caso es $k + 1$. Esto significa que tiene $k + 1$ columnas linealmente independientes; es decir, que no existe una relación lineal exacta entre las variables X.

6.2 Tabla de análisis de varianza (Anova)

Como se ha mencionado anteriormente, la variabilidad de los valores observados de Y alrededor de su media, puede ser atribuida a dos causas, una atribuida a la regresión y la otra a factores aleatorios inherentes al error y cuya variabilidad es no explicada. Esta variación es expresada de la siguiente manera:

$$\underbrace{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}_{\text{Variabilidad alrededor de la media}} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{Variabilidad debido al error}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}_{\text{Variabilidad debido a la regresión}}$$

Tabla de análisis de varianza — Regresión lineal múltiple

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F_0	P-value
Debido a la regresión	k	$SCR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$	$CMR = \frac{SCR}{k}$	$F_0 = \frac{CMR}{CME}$	$P(F > F_0)$
Debido al error	$n - k - 1$	$SCE = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$	$CME = \frac{SCE}{n-k-1}$		
Total	$n - 1$	$SCT = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$			

6.3 Obtención de estimadores en un modelo de regresión lineal múltiple

La información que está al alcance del investigador es una muestra de valores de Y correspondiente a $X's$ fijos, por consiguiente la tarea es la estimación de los parámetros basándose en la información muestral.

Modelo de regresión muestral: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_k X_{ik} + e_i$ donde: $i = 1, 2, \cdots, n$

Además se tiene que: $Y_i = \hat{Y}_i + e_i$.

Por consiguiente, la ecuación de regresión estimada es:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_k X_{ik}$$

De forma abreviada:

$$\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j X_{ij}$$

Entonces se deduce:

$$e_i = (Y_i - \hat{Y}_i) \quad i = 1, 2, ..., n$$

En forma matricial, el modelo de regresión estimada es:

$$Y = X\hat{\beta} + e$$

Donde:

$\hat{\beta}$: Vector de estimadores del vector de parámetros $(k+1)$x1.

e : Vector de residuales o errores.

El objetivo es determinar los valores del vector $\hat{\beta}$ de tal manera que los residuales sean lo más pequeños posibles, el método más adecuado para lograrlo es el de mínimos cuadrados. Luego de la aplicación del método se obtiene la siguiente expresión matricial para los estimadores:

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Donde:

$\hat{\beta}$: Vector de estimadores de mínimos cuadrados.

$(X'X)$: Matriz simétrica, para que tenga inversa su determinante debe ser diferente de cero.

Nota: El método de máxima verosimilitud también puede ser utilizado pero se necesita que el vector $U \sim N(0,\sigma^2 I)$:

Ejemplo 11:

La siguiente información se refiere a las notas obtenidas en una evaluación del curso de estadística (escala vigesimal), el tiempo en horas dedicado al estudio del curso antes de la evaluación y el número de veces que el alumno está llevando el curso (número de matrículas en el mismo curso) para una muestra de 10 alumnos.
Se desea establecer un modelo de regresión que explique el rendimiento del alumno en función del tiempo dedicado al estudio y el número de veces que el alumno lleva el curso.

Nota del curso	Horas de estudio	Número de veces que lleva el curso
12	1	1
14	3	2
15	5	2
11	4	3
16	5	1
17	4	1
10	1	1
8	2	2
18	5	1
19	6	2

Solución:

Modelo muestral:

$$Nota = \hat{\beta}_0 + \hat{\beta}_1 (horas\ de\ estudio) + \hat{\beta}_2 (número\ de\ veces\ que\ lleva\ el\ curso) + e_i$$

Para obtener los estimadores por el método de mínimos cuadrados, haciendo uso de la matriz X y la inversa de $X'X$ se tiene lo siguiente:

$$X=\begin{bmatrix}1&1&1\\1&3&2\\1&5&2\\1&4&3\\1&5&1\\1&4&1\\1&1&1\\1&2&2\\1&5&1\\1&6&2\end{bmatrix};\quad X'Y=\begin{bmatrix}140\\551\\218\end{bmatrix};\ (X'X)^{-1}=\begin{bmatrix}0.956376&-0.100671&-0.308725\\-0.100671&0.036913&-0.020134\\-0.308725&-0.020134&0.238255\end{bmatrix}$$

Por lo tanto, se tiene que el vector de coeficientes estimados es:

$$\hat{\beta}=(X'X)^{-1}X'Y=\begin{bmatrix}0.956376&-0.100671&-0.308725\\-0.100671&0.036913&-0.020134\\-0.308725&-0.020134&0.238255\end{bmatrix}\begin{bmatrix}140\\551\\218\end{bmatrix}=\begin{bmatrix}11.12081\\1.85570\\-2.37584\end{bmatrix}$$

Formulación del modelo estimado:

$$Nota=\hat{\beta}_0+\hat{\beta}_1(horas\ de\ estudio)+\hat{\beta}_2(número\ de\ veces\ que\ lleva\ el\ curso)$$

Haciendo uso del software Minitab:

Regression Analysis: Nota del cur versus Horas de est, N° de veces

```
The regression equation is
Nota del curso = 11.1 + 1.86 Horas de estudio
                 - 2.38 N° de veces que lleva el curso
```

El modelo es:

$$Nota=11.1+1.86(horas\ de\ estudio)-2.38(número\ de\ veces\ que\ lleva\ el\ curso)$$

6.3.1 Propiedades de los estimadores

Las propiedades de los estimadores obtenidos por el método de mínimos cuadrados, según el teorema de Gauss-Markov son:

- Insesgamiento, esto es; $E(\hat{\beta})=\beta$
- Varianza mínima, esto es $V(\hat{\beta})=\sigma^2(X'X)^{-1}$ es la más pequeña

La varianza poblacional σ^2 se estima con la varianza muestral $s_e^2=CME$, definida como:

$$s_e^2=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-k-1}=\frac{\sum_{i=1}^{n}e_i^2}{n-k-1}=\frac{SCE}{n-k-1}=CME\ \text{(Cuadrado Medio del Error)}$$

El valor de los errores estándar de estimadores con notación $s(\hat{\beta})$ permite obtener las estimaciones por intervalos.

Ejemplo 12:

Utilizando los datos del ejemplo 11 con $n = 10$ alumnos, y $k = 2$ variables regresoras.
Para calcular el resultado correspondiente a la suma de cuadrados debido al error (SCE), se deben calcular los valores estimados para, de esa forma, obtener la suma de cuadrados de los residuales (valor real – valor estimado):

$$\sum_{i=1}^{n} e_i^2 = 1.958 + 4.259 + 0.419 + 0.173 + 4.095 + 0.693 + 0.361 + 4.328 + 0.001 + 2.240 = 18.527$$

Es decir, $SCE = 18.527$; entonces:

$$s_e^2 = CME = \frac{SCE}{n-k-1} = \frac{18.527}{10-2-1} = \frac{18.527}{7} = 2.647$$

6.3.2 Intervalos de confianza de los estimadores — RLM

Los intervalos de confianza para los parámetros del modelo de regresión lineal múltiple vienen dados por las siguientes expresiones:

$$\hat{\beta}_i - t_{(1-\alpha/2,\,n-k-1)} s(\hat{\beta}_i) < \beta_i < \hat{\beta}_i + t_{(-\alpha/2,\,n-k-1)} s(\hat{\beta}_i)$$

$$\text{ó} \quad \beta_i \in \left\langle \hat{\beta}_i \mp t_{\left(1-\alpha/2,\,n-k-1\right)} s(\hat{\beta}_i) \right\rangle$$

Ejemplo 13:

Continuando con el ejemplo 11, haciendo uso de los cálculos realizados inicialmente y utilizando el valor de CME = 2.647 como una estimación de σ^2 , se procede a estimar la varianza de cada uno de los estimadores:

$$V(\hat{\beta}) = \hat{\sigma}^2 \left(X'X\right)^{-1}$$

$$V(\hat{\beta}) = 2.647 \left[\begin{array}{ccc} 0.956376 & -0.100671 & -0.308725 \\ -0.100671 & 0.036913 & -0.020134 \\ -0.308725 & -0.020134 & 0.238255 \end{array}\right]$$

$$V(\hat{\beta}) = \left[\begin{array}{ccc} 2.531527 & -0.266477 & -0.817195 \\ -0.266477 & 0.097708 & -0.053295 \\ -0.0.817195 & -0.053295 & 0.630661 \end{array}\right]$$

Entonces se tienen los siguientes valores para las varianzas y desviaciones estándar de los estimadores:

$V(\hat{\beta}_0) = 2.5315$ y $s(\hat{\beta}_0) = 1.5911$

$V(\hat{\beta}_1) = 0.0977\,\frac{1}{n}$ y $s(\hat{\beta}_1) = 0.3126$

$V(\hat{\beta}_2) = 0.6307$ y $s(\hat{\beta}_2) = 0.7941$

Haciendo uso del software Minitab:

Tabla de coeficientes:

Predictor	Coef	SE Coef	T	P
Constant	11.121	1.591	6.99	0.000
Horas de estudio	1.8557	0.3126	5.94	0.001
N° de veces que lleva el curso	-2.3758	0.7941	-2.99	0.020

Por lo tanto, con un nivel de significancia de $\alpha = 0.05$, los intervalos de confianza son:

- Intervalo de confianza para el término constante, β_0 :

$$\hat{\beta}_0 - t_{(1-\alpha/2,\,n-k-1)}\, s(\hat{\beta}_0) < \beta_0 < \hat{\beta}_0 + t_{(1-\alpha/2\,n-k-1)}\, s(\hat{\beta}_0)$$

$$11.121 - t_{(0.025,7)}(1.5911) < \beta_0 < 11.121 + t_{(0.975,7)}(1.5911)$$

Inverse Cumulative Distribution Function

Student's t distribution with 7 DF

P(X <= x)	x
0.975	2.36462

$$11.121 - 2.36462(1.591) < \beta_0 < 11.121 + 2.36462(1.591)$$

$$7.358889 < \beta_0 < 14.883110$$

Interpretación: Con un 95% de confianza el verdadero valor de β_0 se encuentra dentro del intervalo $\langle 7.3589,\ 14.8831 \rangle$.

- Intervalo de confianza para el parámetro β_1

$$\hat{\beta}_1 - t_{(1-\alpha/2,\,n-k-1)}\, s(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t_{(1-\alpha/2\,n-k-1)}\, s(\hat{\beta}_1)$$

$$1.8557 - t_{(0.025,7)}(0.3126) < \beta_1 < 1.8557 + t_{(0.975,7)}(0.3126)$$

$$1.8557 - 2.36462(0.3126) < \beta_1 < 1.8557 + 2.36462(0.312\ 6)$$

$$1.116519 < \beta_1 < 2.594880$$

Interpretación: con un 95% de confianza el verdadero valor del parámetro β_1 se encuentra dentro del intervalo $\langle 1.1165,\ 2.5949 \rangle$.

- Intervalo de confianza para el parámetro β_2

$$\hat{\beta}_2 - t_{(1-\alpha/2,\,n-k-1)}\, s(\hat{\beta}_2) < \beta_2 < \hat{\beta}_2 + t_{(1-\alpha/2\,n-k-1)}\, s(\hat{\beta}_2)$$

$$-2.3758 - t_{(0.025,7)}(0.7941) < \beta_2 < -2.3758 + t_{(0.975,7)}(0.7941)$$

$$-2.3758 - 2.36462(0.7941) < \beta_2 < -2.3758 + 2.36462(0.7941)$$

$$-4.253545 < \beta_2 < -0.498055$$

Interpretación: con un 95% de confianza el verdadero valor del parámetro β_2 se encuentra dentro del intervalo $\langle -4.2535,\ -0.4981 \rangle$.

6.4 Pruebas de verificación

Son las pruebas que se pueden realizar con la finalidad de verificar la validez estadística del modelo de regresión lineal múltiple propuesto, las principales pruebas son:

- Coeficiente de determinación múltiple: R^2
- Prueba global o prueba del modelo de regresión lineal múltiple – Prueba F
- Prueba individual o prueba de cada coeficiente β_i – Prueba T

6.4.1 Coeficiente de determinación múltiple (R^2)

El coeficiente de determinación múltiple se denota por R^2 y se define como el porcentaje de la variación total de los valores de la variable respuesta Y, que es explicada por el conjunto de variables $X_1, X_2, X_3, \cdots, X_k$.

$$R^2 * 100\% = \frac{SCR}{SCT} * 100\% = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2} * 100\%$$

Como se sabe, $0 \leq R^2 \leq 1$; cuando tome valores cercanos a cero (0) peor será el ajuste del plano de regresión a los datos; cuanto más se acerca a la unidad, o al 100% en caso de R^2*100%, mejor será el ajuste.

Ejemplo 14:

Para el ejemplo 11, haciendo uso del software Minitab:

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	101.473	50.737	19.17	0.001
Residual Error	7	18.527	2.647		
Total	9	120.000			

Entonces:

$SCR = 101.4732$ y $SCT = 120$

Luego:

$$R^2 * 100\% = \frac{SCR}{SCT} * 100\% = \frac{101.4732}{120} * 100\% = 0.84561 * 100\% = 84.561\%$$

La variación en las notas de los alumnos es explicada en un 84.56% por la variación del tiempo que dedicaron los alumnos a estudiar el curso antes de la práctica y el número de veces que han llevado el curso.

6.4.2 Prueba de significación del modelo — Prueba F

La prueba de significación del modelo sirve para determinar si el modelo de regresión lineal múltiple, con las variables independientes utilizadas, es apropiado o no.

Procedimiento:

i. Las hipótesis por probar:
$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ (el modelo no es apropiado).

$H_1 : \text{Al menos un } \beta_i \neq 0$ (el modelo es apropiado). $i = 1, ..., k$

ii. Nivel de significación α y suposiciones de la distribución de la variable.

iii. Estadística de prueba: $F_0 = \dfrac{CMR}{CME} \sim F_{(k,n-k-1)}$ si H_0 es verdadera.

iv. Región crítica y regla de decisión: $\text{RC} = \langle F_{(1-\alpha,k,n-k-1)}, \infty \rangle$, siendo el valor crítico $F_{(1-\alpha,k,n-k-1)}$

También se puede calcular y utilizar el P-value: $P(F_{(k,n-k-1)} > F_0)$

v. Valor del estadístico (empleando Anova) y regla de decisión
Si $F_0 > F$ crítico se rechaza la hipótesis nula (H_0).
Si $\text{P-value} < \alpha$, entonces se rechaza la hipótesis nula (H_0).

Ejemplo 15:

Para el ejemplo 11, se tiene que:

i. Las hipótesis por probar:
$H_0 : \beta_1 = \beta_2 = 0$ (el modelo no es apropiado).

$H_1 : Al\ menos\ un\ \beta_i \neq 0$ (el modelo es apropiado).

ii. Nivel de significación $\alpha = 0.05$ y las notas del curso tienen una distribución normal.

iii. Estadística de prueba: $F_0 = \dfrac{CMR}{CME} \sim F_{(2,7)}$, tal que H_0 es verdadera.

iv. Valor crítico $F_{(1-0.05,2,10-2-1)} = F_{(0.95,2,7)} = 4.73741$

Región crítica: $\text{RC} = \langle 4.73741, \infty \rangle$

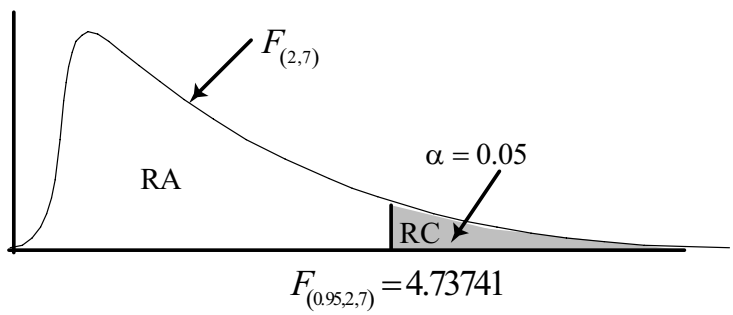


Figura 12. Prueba de hipótesis de una cola.

v. Valor del estadístico (empleando Anova) y regla de decisión

$$F_0 = \frac{50.737}{2.647} = 19.17$$

Tabla Anova de la regresión lineal múltiple

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F_0	P-value
Debido a la regresión	2	SCR = 101.473	*CMR = 50.737*	19.17	0.001
Debido al error	7	SCE = 18.527	CME = 2.647		
Total	9	SCT = 120.000			

Calculando el valor del P-value:

$$P(F_{(2,7)} > 19.17) = 1 - P(F_{(2,7)} < 19.17) = 1 - 0.998554 = 0.001446 \approx 0.001$$

Decisión y conclusión: Como $F_0 = 19.17$, entonces $F_0 > F$ (la estadística de prueba es mayor que el valor crítico), por lo tanto se rechaza la hipótesis nula; es decir, el modelo es apropiado.

6.4.3 Prueba individual de las variables — Prueba T

La prueba individual conocida también como la prueba de significación de las variables tiene el siguiente procedimiento:

Procedimiento:

i. Hipótesis

$H_0 : \beta_i = 0$ (la variable X_i no influye en el modelo).

$H_1 : \beta_i \neq 0$ (la variable X_i sí influye en el modelo).

ii. Nivel de significación o riesgo α y suposición de la distribución de la variable.

iii. Estadística de prueba: $t_0 = \frac{\hat{\beta}_i - \beta_i}{s(\hat{\beta}_i)} \sim t_{(n-k-1)}$, tal que H_0 es verdadera.

iv. Región crítica: $\text{RC} = \langle -\infty, t_{(\alpha/2)} \rangle \cup \langle t_{(1-\alpha/2)}, \infty \rangle$

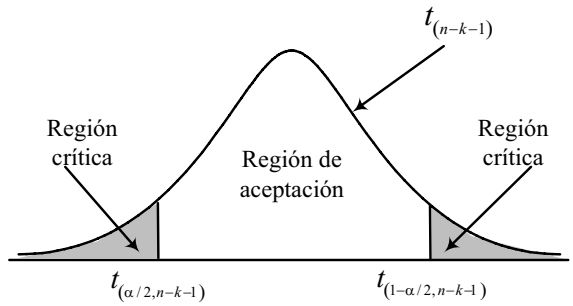


Figura 13. Prueba de hipótesis de dos colas.

También se puede calcular y utilizar el $\text{P-value} = 2 * P(t_{(n-k-1)} > t_0)$

v. Regla de decisión: Se rechaza H_0 si $t_0 < t_{(\alpha/2,n-k-1)}$ ó $t_0 > t_{(1-\alpha/2,n-k-1)}$

Si $\text{P-value} < \alpha$, entonces se rechaza la hipótesis nula (H_0)

Ejemplo 16:

Para el ejemplo 11, se tiene que:

Prueba de significación:

- Prueba con respecto al intercepto:
  i. Hipótesis por probar

  $H_0 : \beta_0 = 0$ (la recta pasa por el origen)

  $H_1 : \beta_0 \neq 0$ (la recta pasa fuera del origen)

  ii. $\alpha = 0.05$ y las notas del curso se distribuyen normalmente.

  iii. Estadística de prueba: $t_0 = \dfrac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)} = \dfrac{\hat{\beta}_0}{s(\hat{\beta}_0)}$

  iv. Región crítica: $\text{RC} = \langle -\infty, -2.36462 \rangle \cup \langle 2.36462, \infty \rangle$

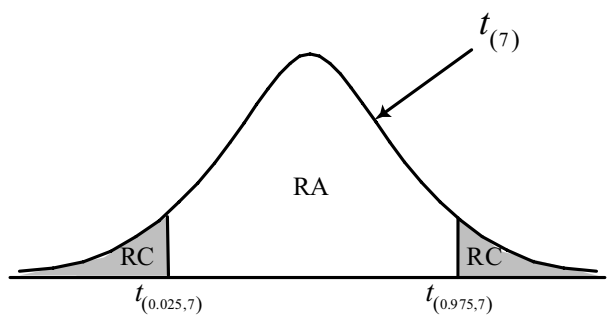


Figura 14. Prueba de hipótesis de dos colas.

v. Valor de la estadística de prueba y regla de decisión

$$t_0 = \frac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)} = \frac{11.12081 - 0}{1.5910} = 6.9898$$

Como $t_0 = 6.9898 > 2.36462$, se rechaza la hipótesis nula, indicando que la recta pasa fuera del origen.

- Prueba para las pendientes:

Para β_1

i. Hipótesis por probar

$H_0 : \beta_1 = 0$ (la variable X_1 no influye en el modelo).

$H_1 : \beta_1 \neq 0$ (la variable X_1 sí influye en el modelo).

ii. $\alpha = 0.05$ y las notas del curso se distribuyen normalmente.

iii. Estadística de prueba: $t_0 = \dfrac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} = \dfrac{\hat{\beta}_1}{s(\hat{\beta}_1)}$

iv. Región crítica: $\text{RC} = \langle -\infty, -2.36462 \rangle \cup \langle 2.36462, \infty \rangle$

v. Valor de la estadística de prueba: $t_0 = \dfrac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} = \dfrac{1.85570 - 0}{0.3126} = 5.9363$

Decisión y conclusión: como $t_0 = 5.9363 > 2.36462$, se rechaza la hipótesis nula, la variable X_1 influye en el modelo.

Para β_2

i. Hipótesis por probar

$H_0 : \beta_2 = 0$ (la variable X_2 no influye en el modelo).

$H_1 : \beta_2 \neq 0$ (la variable X_2 sí influye en el modelo).

ii. $\alpha = 0.05$ y las notas del curso se distribuyen normalmente.

iii. Estadística de prueba: $t_0 = \dfrac{\hat{\beta}_2 - \beta_2}{s(\hat{\beta}_2)} = \dfrac{\hat{\beta}_2}{s(\hat{\beta}_2)}$

iv. Región crítica: $\text{RC} = \langle -\infty, -2.36462 \rangle \cup \langle 2.36462, \infty \rangle$

v. Valor de la estadística de prueba: $t_0 = \dfrac{\hat{\beta}_2 - \beta_2}{s(\hat{\beta}_2)} = \dfrac{-2.37584 - 0}{0.7941} = -2.9913$

vi. Decisión y conclusión: como $t_0 = -2.9913 < -2.36462$, se rechaza la hipótesis nula, la variable X_2 influye en el modelo.

PROBLEMAS RESUELTOS

1. La siguiente información resultó de un estudio realizado para examinar la relación entre una medida de la corrosión de hierro (Y) y la concentración de Na PO_4 (X, en ppm)

X	2.5	5.03	7.6	11.6	13	19.6	26.2	33	40	50	55
Y	7.68	6.95	6.3	5.75	5.01	1.43	0.93	0.72	0.68	0.65	0.56

a. Construya una gráfica de dispersión de los datos. ¿Parece razonable el modelo de regresión lineal?

Solución:

Ingresar a la opción Graph / Scatterplot..., seleccionar el gráfico <Simple>, luego en <OK> (véase la figura 15).

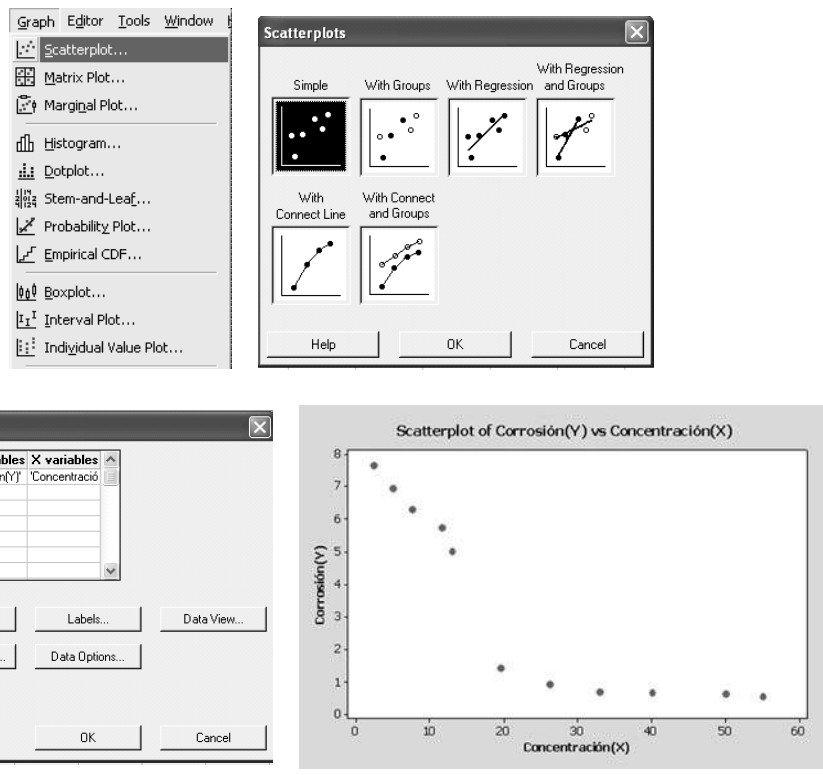


Figura 15. Secuencia para gráfico de dispersión.

Observando el gráfico parece razonable pensar en una línea recta con pendiente negativa.

b. Calcule la recta de regresión estimada, utilícela para pronosticar el valor de la rapidez de corrosión que se observaría para una concentración de 33 ppm.

Solución:

Ingresar a la opción Stat / Regression / Regression..., en el cuadro de diálogo seleccionar Corrosión en <Response> y Concentración en <Predictors>, luego en <Ok> (véase la figura 16).

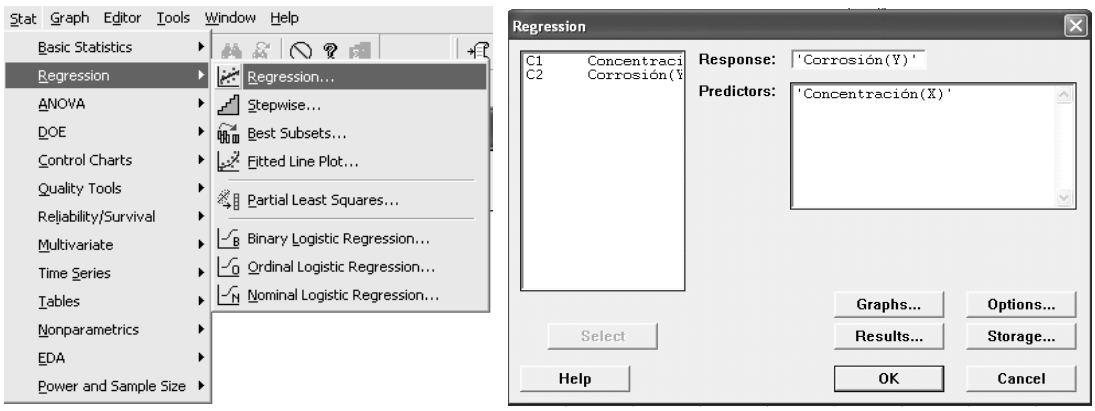


Figura 16. Secuencia para diagrama de dispersión.

Se obtiene el siguiente reporte:

Regression Analysis: Corrosión(Y) versus Concentración(X)

```
The regression equation is
Corrosión(Y) = 6.74 - 0.142 Concentración(X)
Predictor             Coef  SE Coef      T      P
Constant            6.7351   0.7573   8.89  0.000
Concentración(X)  -0.14202  0.02553  -5.56  0.000

S = 1.48088   R-Sq = 77.5%   R-Sq(adj) = 75.0%
Analysis of Variance
Source          DF      SS      MS      F      P
Regression       1  67.849  67.849  30.94  0.000
Residual Error   9  19.737   2.193
Total           10  87.586
```

La regresión estimada es: Corrosión (Y) = 6.7351 - 0.14202 Concentración(X).

Interpretación: en promedio, el índice de corrosión es de 6.7351 unidades cuando no hay concentración de Na PO_4. Cuando la concentración de Na PO_4 aumenta en 1 ppm, el índice de corrosión disminuye en promedio 0.14202 unidades.

Para obtener el pronóstico para una concentración de 33 ppm se debe seguir el siguiente procedimiento:

En la pestaña del cuadro de regresión elegir <Options>, colocar el valor 33 en el recuadro <Prediction intervals for new observations>: luego <Ok>, finalmente <Ok>.

Los pasos indicados se pueden apreciar en la figura 17.

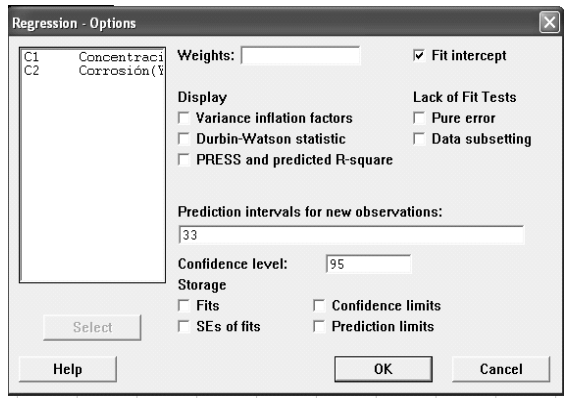


Figura 17. Opciones para intervalo de nuevas observaciones.

Se obtiene el siguiente reporte:

```
Predicted Values for New Observations
New
Obs     Fit   SE Fit       95% CI             95% PI
  1   2.048    0.503   (0.911, 3.186)   (-1.489, 5.586)
Values of Predictors for New Observations
New
Obs   Concentración(X)
  1                33.0
```

La corrosión estimada es de: 2.048 (véase la columna con encabezado Fit).

c. ¿Qué porcentaje de la variación de la corrosión se puede atribuir a la concentración?

Solución:

El valor obtenido en el coeficiente de determinación indica que el 77.5% de la variación de la corrosión se puede atribuir a la concentración.

d. Estime con 95% de confianza el promedio de Y cuando $X = 45$ ppm.

Solución:

Estimación interválica para nuevas observaciones:

```
Predicted Values for New Observations
New Obs    Fit     SE Fit        95% CI            95% PI
   1     0.344    0.699  (-1.236, 1.925)  (-3.360, 4.048)
 Values of Predictors for New Observations
New
Obs  Concentración(X)
 1              45.0
```

El intervalo de confianza para el promedio Y cuando $X = 45$ ppm es: $\langle -1.236, 1.925 \rangle$

Interpretación: Con un 95% de confianza el promedio de Y, cuando $X = 45$ ppm, se encuentra dentro del intervalo $\langle -1.236, 1.925 \rangle$.

2. Una empresa de transporte interprovincial recolectó los datos del recorrido (X) de sus vehículos en miles de millas, y el costo total de mantenimiento (Y) en miles de nuevos soles.

A continuación se presenta la información recopilada:

Observación	Costo total (miles de S/.)	Millas recorridas (miles)	Observación	Costo total (miles de S/.)	Millas recorridas (miles)
1	234.5	3776	13	252.3	4251
2	205.9	3232	14	224.4	3844
3	202.7	3141	15	215.3	3276
4	198.5	2928	16	202.5	3184
5	195.6	3063	17	200.7	3037
6	200.4	3096	18	201.8	3142
7	200.1	3096	19	202.1	3159
8	201.5	3158	20	200.4	3139
9	213.2	3338	21	209.3	3203
10	219.5	3492	22	213.9	3307
11	243.7	4019	23	227	3585
12	262.3	4394	24	246.4	4073

Empleando Excel

a. Determine la ecuación de mínimos cuadrados. Use un nivel de significancia del 5%.

Solución:

Ingresar los datos en las columnas A y B.

Para determinar la ecuación de mínimos cuadrados ingresar a la opción Herramientas / Análisis de datos, en la ventana que aparece seleccionar <Regresión>, luego seleccionar rango de las variables, activar <En una hoja nueva> y elegir <Aceptar> (véanse las figuras 18 y 19).

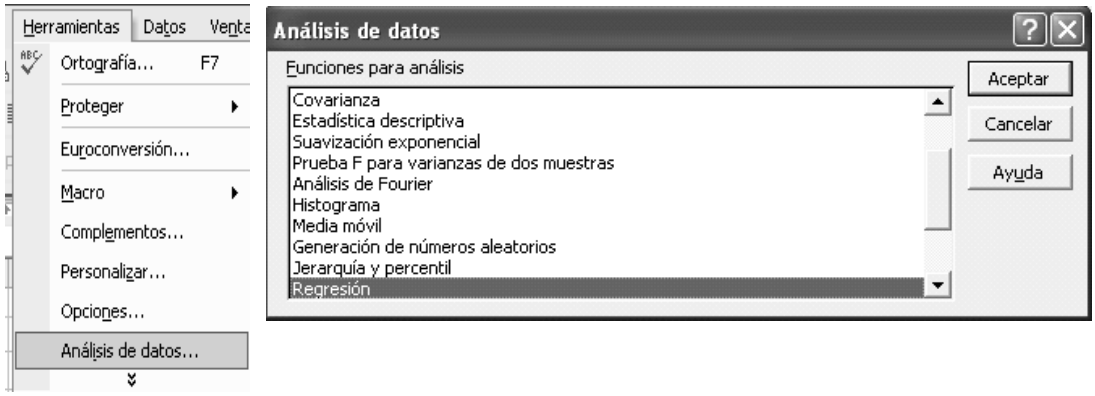


Figura 18. Secuencia para la ecuación de regresión.

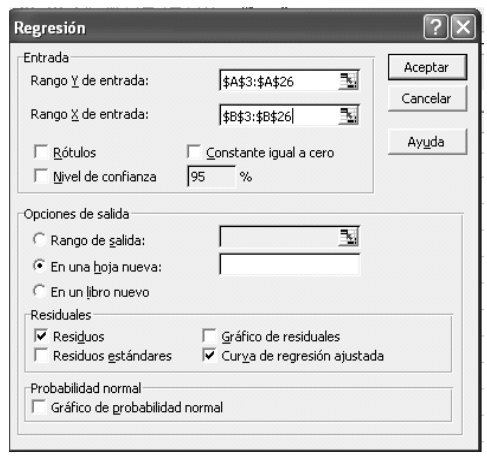


Figura 19. Rango donde se encuentran los valores.

El reporte es el siguiente:

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.9837
Coeficiente de determinación R^2	0.9676
R^2 ajustado	0.9661
Error típico	3.5408
Observaciones	24

ANÁLISIS DE VARIANZA

Fuente de variación	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	1	8238.8134	8238.8134	657.1456	0
Residuos	22	275.8199	12.5373		
Total	23	8514.6333			

	Coeficientes	Error típico	Estadístico t	Probabilidad (P-value)	Inferior 95%	Superior 95%
Intercepción	61.3757	6.0588	10.1299	9.54E-10	48.8105	73.9409
Variable X 1	0.0452	0.0018	25.6349	7.04E-18	0.0415	0.0488

La ecuación es:

$$\text{Costo total} = 61.3757 + 0.0452 \cdot (\text{Millas recorridas})$$

Interpretación: En promedio, el costo total de mantenimiento de un vehículo es de 61.3757 miles de nuevos soles cuando no existen millas por recorrer. Cuando las millas recorridas aumentan en 1 millar, el costo total de mantenimiento de un vehículo aumenta en promedio 0.0452 miles de nuevos soles.

b. Interprete el coeficiente de correlación.

Solución:

El coeficiente de correlación es 0.9837, este valor indica que existe una fuerte relación directa entre las variables costo total y millas recorridas.

c. ¿La recta pasa por el origen?

Solución:

Prueba para determinar la contribución del intercepto:

i. Formulación de la hipótesis:

$H_0 : \beta_0 = 0$

$H_1 : \beta_0 \neq 0$

ii. Nivel de significación:
$\alpha = 0.05$ y el costo total tiene distribución normal.

iii. Estadística de prueba:
Del reporte del Excel; $t_0 = 10.1299$ y P-value $= 0.000$

iv. Región crítica: $\text{RC} = \langle -\infty, -2.07387 \rangle \cup \langle 2.07387, \infty \rangle$

Si P-value $< \alpha$, se rechaza la hipótesis nula (H_0)

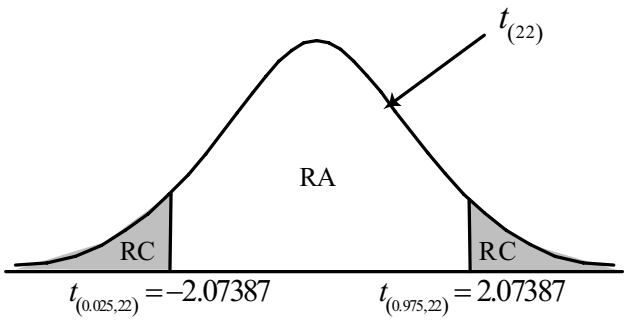


Figura 20. Prueba de hipótesis de dos colas.

v. Regla de decisión:
El P-value = 0.000 es menor que el nivel de significación (0.05), se rechaza H_0; se concluye que la recta pasa fuera del origen.

d. ¿β_1 contribuye significativamente al modelo?

Solución:

i. Prueba de hipótesis:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

ii. Nivel de significación:
$\alpha = 0.05$ y el costo total tiene distribución normal.

iii. Estadística de prueba:

$t_0 = 25.6349$

iv. Región crítica:
P-value = 0.000

v. Regla de decisión.

vi. Si P-value $< \alpha$, se rechaza la hipótesis nula (H_0).
El P-value = 0 es menor que el nivel de significación, se rechaza H_0; se concluye β_1 contribuye significativamente al modelo.

e. ¿El modelo es adecuado?

Solución:

i. Formulación de la hipótesis:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

ii. Nivel de significación:
$\alpha = 0.05$ y el costo total tiene distribución normal.

iii. Estadística de prueba:

$F_0 = 657.1456$; P-value = 0.000

iv. Regla de decisión.
Si P-value $< \alpha$, se rechaza la hipótesis nula (H_0)

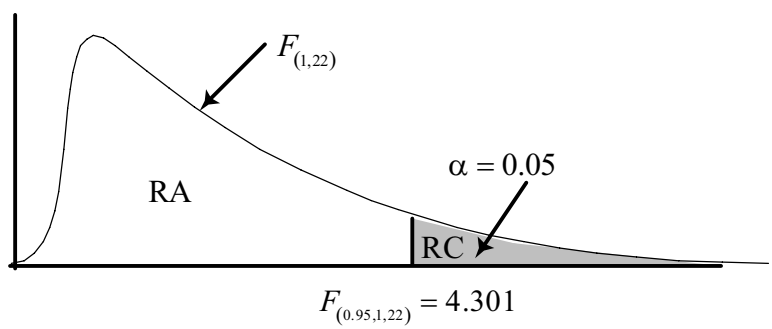


Figura 21. Prueba de hipótesis de una cola.

v. Decisión y conclusión:
El P-value = 0.000 es menor que el nivel de significación (0.05), se rechaza H0; se concluye que el modelo es adecuado.

f. Calcule un intervalo de confianza para β_1 con una confianza del 95%.

Solución:

Obtención del intervalo de confianza para β_1 :

$$\beta_1 \in \left\langle \hat{\beta}_1 \pm t_{(1-\alpha/2,n-2)} s\left(\hat{\beta}_1\right) \right\rangle$$

$$\beta_1 \in \left\langle 0.0452 \pm t_{(0.975,22)} \left(0.0018\right) \right\rangle$$

$$\beta_1 \in \left\langle 0.0452 \pm 2.0739\left(0.0018\right) \right\rangle = \left\langle 0.0415, 0.0488 \right\rangle$$

Decisión y conclusión: el intervalo de confianza 0.0415, 0.0488 significa que se tiene 95% de confianza que β_1 se encuentre dentro de ese intervalo.

3. La siguiente información representa el monto por cobro de comisiones de 15 sucursales bancarias elegidas al azar y los beneficios obtenidos por las mismas sucursales: use un nivel de significancia del 10%.

Sucursales	Montos por cobro de comisiones (miles $)	Beneficios (miles $)
1	29.2	84.1
2	22.8	51.3
3	11.0	33.5
4	23.3	95.0
5	10.1	10.05
6	3.9	23.0
7	7.3	24.2
8	16.9	53.2
9	19.4	52.3
10	7.5	35.0
11	20.5	60.2
12	19.8	50.1
13	23.1	90.2
14	6.3	32.0
15	4.3	38.1

a. Determine la ecuación de mínimos cuadrados.

Solución:

Haciendo uso del software Minitab:

Regression Analysis: Montos versus Beneficios

```
The regression equation is
Montos = 1.67 + 0.274 Beneficios

Predictor      Coef   SE Coef      T       P
Constant      1.671     2.685   0.62   0.544
Beneficios  0.27358   0.04924   5.56   0.000
```

La ecuación de regresión ajustada es:

$$\text{Montos por cobros de comisiones} = 1.671 + 0.27358(\text{Beneficios})$$

b. Interprete el coeficiente de determinación.

Solución:

Haciendo uso del software Minitab:

```
S = 4.63437   R-Sq = 70.4%   R-Sq(adj) = 68.1%
```

El coeficiente de determinación es de 70.4%, es decir la variabilidad de los montos por cobro de comisiones es explicado en un 70.4% por el modelo.

c. ¿La variable X contribuye significativamente al modelo?

Solución:

Prueba para determinar la contribución de la variable X_1:

i. Formulación de la hipótesis:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

ii. Nivel de significación: $\alpha = 0.10$

iii. Estadística de prueba:

Haciendo uso del software Minitab:

```
Predictor       Coef   SE Coef      T       P
Constant       1.671     2.685   0.62   0.544
Beneficios   0.27358   0.04924   5.56   0.000
```

Entonces:

$t_0 = 5.56$; P-value = 0.000

iv. Regla de decisión.

Si P-value $< \alpha$, se rechaza la hipótesis nula (H_0).

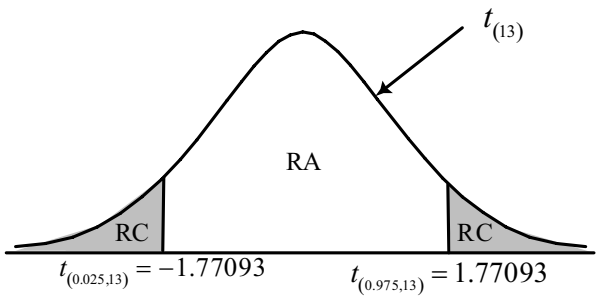


Figura 22. Prueba de hipótesis de dos colas.

v. Decisión y conclusión:
El P-value es menor que el nivel de significancia (0.10), entonces se rechaza H_0; se concluye entonces que β_1 asociado a la variable X_1 contribuye significativamente al modelo.

d. Calcule el intervalo de confianza para β_0 con una confianza del 90%.

Solución:

Intervalo de confianza para β_0 es:

$$\beta_0 \in \left\langle \hat{\beta}_0 \pm t_{(1-\alpha/2,n-2)} s(\hat{\beta}_0) \right\rangle$$

$$\beta_0 \in \left\langle 1.671 \pm t_{(0.95,13)}(2.685) \right\rangle$$

$$\beta_0 \in \left\langle 1.671 \pm 1.77093\,(2.685) \right\rangle = \langle -3.083947, 6.425947 \rangle$$

El valor de β_0 se encuentra entre $\langle -3.0839, 6.42594 \rangle$ con 90% de confianza.

4. Los siguientes tiempos fueron registrados en ocho competencias olímpicas en carrera masculina para diferentes distancias.

Tiempo (seg) Y	9.8	19.9	44.26	103.6	215.9	806.3	1659	7798
Distancia (m) X	100	200	400	800	1500	5000	10000	42196

a. Estime la ecuación de regresión lineal simple.

Solución:

Haciendo uso del software Minitab:

Regression Analysis: Tiempo(seg) versus Distancia (mts.)

```
The regression equation is
Tiempo(seg) = - 63.1 + 0.185 Distancia (mts.
Predictor            Coef   SE Coef       T      P
Constant          -63.10     27.98   -2.26  0.065
Distancia (mts.  0.185420  0.001811  102.36  0.000
```

La ecuación es:

$$\text{Tiempos (Seg.)} = -63.10 + 0.185420(\text{Distancia (mts)})$$

b. Interprete el valor de la pendiente.

Solución:

Si se incrementa la distancia en 1 metro, el tiempo registrado se incrementa en 0.185420 segundos.

c. Calcule el intervalo de confianza para la pendiente β_1 con un nivel de significancia del 5%.

Solución:

$$\beta_1 \in \left\langle \hat{\beta}_1 \pm t_{(1-\alpha/2,n-2)} s(\hat{\beta}_1) \right\rangle$$

$$\beta_1 \in \left\langle 0.185420 \pm t_{(0.975,6)} (0.001811) \right\rangle$$

$$\beta_1 \in \left\langle 0.185420 \pm 2.44691(0.001811) \right\rangle = \langle 0.180989, 0.189851 \rangle$$

Decisión y conclusión: se tiene un 95% de confianza que β_1 se encuentre dentro del inter-valo 0.1809, 0.1899.

5. Un experimento consiste en determinar la relación entre el alargamiento (mm) producido en un plástico y la fuerza a la que es sometido en toneladas por centímetro cuadrado. El experimento se repitió 10 veces y los resultados se reportan a continuación. Utilice un nivel de significancia del 5%.

Reportes del Minitab:

Regression Analysis: Alargamiento (Y) versus Fuerza(X)

```
The regression equation is
Alargamiento(Y) = 4.38 + 56.2 Fuerza(X)
Predictor     Coef  SE Coef     T      P
Constant     4.380    7.670  0.57  0.584
Fuerza(X)   56.182    6.976  8.05  0.000
S = 10.5536   R-Sq = 89.0%   R-Sq(adj) = 87.6%
Analysis of Variance
Source          DF      SS      MS      F      P
Regression       1  7225.0  7225.0  64.87  0.000
Residual Error   8   891.0   111.4
Total            9  8116.0
```

a. Interprete los coeficientes de la recta estimada.

Solución:

La ecuación estimada es:

$$\text{Alargamiento} = 4.380 + 56.182(\text{Fuerza})$$

Es decir:

$\hat{\beta}_0 = 4.380$, el alargamiento del plástico es de 4.38 mm cuando no es sometido a ninguna fuerza.

$\hat{\beta}_1 = 56.182$, por cada tonelada de fuerza aplicada el alargamiento aumenta en 56.182 mm.

b. ¿Cuál o cuáles de los parámetros del modelo no son significativos?

Solución:

Prueba de hipótesis para los coeficientes de regresión.

i. **Prueba de hipótesis para β_0.**

$H_0 : \beta_0 = 0$

$H_1 : \beta_0 \neq 0$

ii. Nivel de significación: $\alpha = 0.05$

iii. Estadística de prueba:

$$t_0 = \frac{4.38}{7.67} = 0.57$$

Observando el reporte de Minitab:

$t_0 = 0.57$; P-value = 0.584

iv. Regla de decisión:
Si P-value $< \alpha$, se rechaza la hipótesis nula (H_0)

v. Decisión y conclusión:
El valor de la estadística de prueba (0.57) es pequeño, comparando el valor crítico (2.306). Por consiguiente, no se rechaza H_0 y se concluye que β_0 no es significativa en el modelo. Obsérvese que el valor del P-value = 0.584 es mayor que $\alpha = 0.05$, lo que confirma la conclusión antes mencionada.

i. **Prueba de hipótesis para β_1:**

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

ii. Nivel de significación: $\alpha = 0.05$

iii. Estadística de prueba:

$$t_0 = \frac{56.182}{6.976} = 8.05$$

Observando el reporte de Minitab:

$t_0 = 8.05$; P-value = 0.000

iv. Regla de decisión.
Si P-value $< \alpha$, se rechaza la hipótesis nula (H_0)

v. Decisión y conclusión:
El valor del P-value = 0.000 es menor que $\alpha = 0.05$, por lo tanto se rechaza H_0. Se concluye que β_1 es significativa en el modelo, es decir, la variable fuerza aplicada es apropiada para explicar el comportamiento del alargamiento.

c. En términos del enunciado interprete el valor de R^2.

Solución:

El 89% de la variación total en el alargamiento es explicado por la fuerza.

6. El gerente de una cadena de restaurantes ha registrado datos sobre el número de sus menús semanales y los impuestos anuales en nuevos soles de 24 locales.

Local	Número de menús	Impuestos
1	296	5028
2	259	4918
3	279	455
4	259	4557
5	299	5058
6	299	3890
7	309	5860
8	289	5603
9	360	5828
10	315	5300
11	310	6217
12	309	5960
13	300	5050
14	369	8230
15	419	6670
16	405	7782
17	439	9038
18	376	5960
19	445	8798
20	380	6082
21	389	8368
22	370	8140
23	458	9150
24	410	7790

a. Suponga que desea utilizar un modelo de regresión lineal simple para explicar la relación entre el monto de los impuestos con el número de menús, obtenga la ecuación de ajuste de mínimos cuadrados que relacione el número de menús y los impuestos pagados.

Solución:

Haciendo uso del software Minitab

Regression Analysis: Impuestos versus Número de menús (semanal)

```
The regression equation is
Impuestos = - 2843 + 26.1 Número de menús (semanal)
Predictor                    Coef  SE Coef      T      P
Constant                    -2843     1425  -1.99  0.059
Número de menús (semanal)  26.125    4.041   6.46  0.000
```

La ecuación estimada es:

$$\text{Impuesto} = -2843 + 26.125(\text{Número de menús})$$

b. Interprete el valor de $\hat{\beta}_1$.

Solución:

Interpretación: por cada menú que se incremente en la semana, el impuesto aumenta en 26.125 nuevos soles.

c. Determine los impuestos pagados cuando se venden 500 menús semanales.

Solución:

Se sabe que el modelo es:

$$\text{Impuesto} = -2843 + 26.125(\text{Número de menús})$$

Entonces la estimación es:

$$\text{Impuesto} = -2843 + 26.125(500) = 10219.5$$ nuevos soles.

7. La Secretaría General de Transportes reportó información sobre 10 personas víctimas de accidentes de tránsito, de las cuales se tomó información acerca de sus edades y y de la medición de la concentración de alcohol en la sangre después del accidente. Use un nivel de significancia del 3%.

Edad	17	43	31	37	21	20	46	22	20	29
Alcohol	0.22	0.19	0.18	0.20	0.26	0.26	0.19	0.25	0.26	0.25

a. Formule la ecuación de regresión estimada que mejor describa estos datos.

Solución:

Haciendo uso del software Minitab.

Regression Analysis: Alcohol versus Edad

```
The regression equation is
Alcohol = 0.297 - 0.00247 Edad
Predictor        Coef    SE Coef      T      P
Constant      0.29662    0.02205  13.45  0.000
Edad       -0.0024691  0.0007289  -3.39  0.010
```

La ecuación estimada es:

$$\text{Concentración de Alcohol} = 0.29662 - 0.0024691(\text{Edad})$$

b. ¿Es el modelo de regresión lineal apropiado para explicar la relación de la edad del accidentado y el nivel de alcohol?

Solución:

Haciendo uso del software Minitab

```
Analysis of Variance
Source          DF         SS          MS      F      P
Regression       1  0.0059159  0.0059159  11.48  0.010
Residual Error   8  0.0041241  0.0005155
Total            9  0.0100400
```

Prueba para determinar la adecuación o confiabilidad del modelo.

i. Formulación de la hipótesis:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

ii. Nivel de significación: $\alpha = 0.03$

iii. Estadística de prueba:

$$F_0 = \frac{CMR}{CME} = \frac{0.0059159}{0.0005155} = 11.48 \text{ ; P-value} = 0.010$$

iv. Regla de decisión:
Si P-value $< \alpha$, se rechaza la hipótesis nula (H_0)

v. Decisión y conclusión:
El P-value = 0.010 es menor que el nivel de significación (0.03), por lo tanto se rechaza H_0; se concluye que el modelo es adecuado.

8. La siguiente información corresponde al modelo de regresión lineal simple de la tasa mensual de asesinatos cometidos y la cantidad de armas vendidas durante 10 meses.

Regression Analysis: Tasa de asesinatos versus Armas vendidas

```
The regression equation is
Tasa de asesinatos = 3.00 + 0.000961 Armas vendidas
Predictor            Coef    SE Coef
Constant           2.9995     0.9951
Armas vendidas  0.0009605  0.0001784
S = 1.89974
```

Fuentes de variabilidad	GL	SC	CM	F	P-value
Regresión					
Error					
Total		133.44			

a. Complete la tabla del análisis de varianza (Anova).

Solución:

Del reporte de Minitab: el número de variables regresoras es 1, por consiguiente los grados de libertad asociados a la regresión es 1.

Los grados de libertad del total son 9, debido a que se tienen (n-1) grados de libertad.

Los grados de libertad del error son 8, que se obtienen por diferencia.

Se tiene como dato que la desviación estándar de los errores s = 1.89974, pero se sabe que s^2 = CME, pero también se sabe que CME = SCE / GL (error).

Por lo tanto: s^2 = SCE / 8 SCE = 8*(1.89974^2) = 28.8721

La suma de cuadrados de la regresión es SCR = SCT – SCE, entonces SCR = 133.44 – 28.8721 = 104.5679.

El cuadrado medio de la regresión es CMR = SCR / GL (regresión), entonces CMR = 104.5679 / 1 = 104.5679.

El cuadrado medio del error es CME = SCE / GL (error), entonces CME = 28.8720 / 8 = 3.6090.

El estadístico es F = CMR / CME,

entonces F = 104.5679 / 3.6090 = 28.9741

El valor del P-value es:

$$P-value = P(F_{(k,n-k-1)} > F_0) = P(F_{(1,8)} > 28.9741) = \left[1 - P(F_{(1,8)} \leq 28.9741)\right]$$

$$P-value = \left(1 - 0.999341\right) = 0.000659$$

Entonces la tabla del Anova es:

Fuentes de variabilidad	GL	SC	CM	F	P-value
Regresión	1	104.5679	104.5679	28.9741	0.000659
Error	8	28.8721	3.609		
Total	9	133.44			

b. ¿Cuál es el valor del P-value correspondiente a la hipótesis $H_0: \beta_0 = 0$?

Solución:

Se tiene como información:

```
Predictor             Coef     SE Coef     T
Constant            2.9995      0.9951  3.0143
Armas vendidas  0.0009605   0.0001784  5.3839
```

Las estadísticas de pruebas son halladas de la siguiente manera:

Constante : t_0 = Coef. / SE Coef., entonces 2.9995 / 0.9951 = 3.0143
Armas: t_0 = 0.0009605 / 0.0001784 = 5.3839

Hallando el P-value de la constante:

$$P-value = 2P(t_{(n-2)} > t_0) = 2P(t_{(8)} > 3.0143) = 2\left[1 - P(t_{(8)} \leq 3.0143)\right]$$

$$P-value = 2(1-0.991648) = 2(0.008352) = 0.016704$$

Nota: Observe que se ha multiplicado por dos a la probabilidad ya que la prueba es de dos colas.

c. ¿Es el modelo de regresión lineal simple apropiado? Use $\alpha = 0.06$

Solución:

Se sabe que:

Fuente de variación	GL	SC	CM	F	P-value
Regresión	1	104.5679	104.5679	28.9741	0.000659
Error	8	28.8721	3.6090		
Total	9	133.44			

Prueba para determinar la adecuación o confiabilidad del modelo.

i. Formulación de hipótesis:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

ii. Nivel de significancia: $\alpha = 0.06$

iii. Estadística de prueba:

$$F_0 = \frac{104.5679}{3.6090} = 28.9741 \text{ ;} \qquad \text{P-value} = 0.000659$$

iv. Regla de decisión.
Si P-value $< \alpha$, se rechaza la hipótesis nula (H_0)

v. Decisión y conclusión:
El P-value = 0.000659 es menor que el nivel de significación (0.06), por lo tanto se rechaza H_0; se concluye que el modelo es adecuado.

d. ¿Cuál es la tasa mensual de asesinatos si se venden 1.300 armas?

Solución:

Del reporte:

$$\text{Tasa mensual de asesinatos} = 2.9995 + 0.0009605(\text{Armas vendidas})$$

Entonces la tasa mensual de asesinatos es cuando se venden 1.300 armas es:

$$\text{Tasa mensual de asesinatos} = 2.9995 + 0.0009605(1300) = 4.2482$$

9. Un fabricante de vehículos hizo pruebas antes de lanzar su producto al mercado, para ello registró el peso (libras) y el rendimiento de los automóviles (millas/galón). Se estimó el siguiente modelo de regresión a partir de una muestra de 10 vehículos (la desviación estándar de cada uno de los coeficientes aparece entre los paréntesis).

$$\hat{Y} = 37.5 - 0.00261 Peso \qquad R^2 = 56.6\%$$
$$(2.522) \quad (0.00081)$$

a. Interpretar el coeficiente estimado para X_1 (peso).

Solución:

El coeficiente se interpreta como que por cada libra que se incrementa en el peso del vehículo el rendimiento disminuye en 0.00261 millas/galón.

b. Probar la hipótesis nula de que el verdadero coeficiente de X_1 es 0, frente a la alternativa de que es diferente de cero con un nivel de significancia del 5%.

Solución:

i. Formulación de hipótesis:

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

ii. Nivel de significancia: $\alpha = 0.05$

iii. Estadística de prueba:

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} = \frac{-0.00261 - 0}{0.00081} = -3.2222;$$

iv. Regla de decisión:

Si $t_0 < t_{(\alpha/2,n-2)}$ ó $t_0 > t_{(1-\alpha/2,n-2)}$, entonces se rechaza la H_0

Haciendo uso del software de Minitab:

```
Inverse Cumulative Distribution Function
Student's t distribution with 8 DF
P( X <= x )          x
      0.025   -2.30600
      0.975    2.30600
```

Entonces la regla de decisión para este problema es:

Si $t_0 < -2.3060$ ó $t_0 > 2.3060$, entonces se rechaza la H_0

v. Decisión y conclusión:
El estadístico t_0 es menor a –2.30600, por consiguiente se rechaza la hipótesis nula. Esto significa que la variable peso sí es apropiada para explicar el rendimiento de automóviles.

c. Interpretar el coeficiente de determinación.

Solución:

El coeficiente de determinación indica que el 56.6% de las variaciones del rendimiento (millas/galón) se deben al peso del vehículo.

d. Encontrar e interpretar un intervalo de confianza del 95% para el coeficiente de X_1.

Solución:

Obtención del intervalo de confianza para la pendiente:

$$\beta_1 \in \left\langle \hat{\beta}_1 \pm t_{(1-\alpha/2,n-2)} s(\hat{\beta}_1) \right\rangle$$

$$\beta_1 \in \left\langle -0.00261 \pm t_{(0.975,8)} (0.00081) \right\rangle$$

$$\beta_1 \in \left\langle -0.00261 \pm 2.3060 (0.00081) \right\rangle = \left\langle -0.004478, -0.000742 \right\rangle$$

Interpretación: se tiene 95% de confianza de que el valor poblacional de β_1 se encuentre en el intervalo $\langle -0.004478, -0.000742 \rangle$.

10. Se hizo un estudio de mercado sobre el consumo de helados (Kg per cápita por semana) durante la primavera y el verano, el precio/kg del helado (dólares), el ingreso familiar de los consumidores (dólares) y la temperatura (grados Fahrenheit).

Consumo(Y)	Precio(X1)	Ingreso(X2)	Temperatura(X3)
0.387	1.33	359	63
0.375	1.37	358	61
0.394	1.32	360	65
0.428	1.30	370	69
0.407	1.32	366	68
0.345	1.37	357	55
0.328	1.38	357	47
0.289	1.39	352	42
0.269	1.41	343	32
0.258	1.42	343	23

a. Encuentre la recta de regresión que exprese el consumo en término del precio.

Solución:

Haciendo uso de software de Minitab.

Regression Analysis: Consumo(Y) versus Precio(X1)

```
The regression equation is
Consumo(Y) = 2.26 - 1.40 Precio(X1)
Predictor      Coef   SE Coef        T        P
Constant     2.2573    0.1902    11.87    0.000
Precio(X1)  -1.4029    0.1397   -10.04    0.000
```

Entonces el modelo es: $\text{Consumo} = 2.2573 - 1.4029 * (\text{Precio})$

b. Determine si existe correlación entre el consumo y el precio.

Solución:

Haciendo uso de software de Minitab

```
S = 0.0172709   R-Sq = 92.7%   R-Sq(adj) = 91.7%
```

El coeficiente de determinación R^2*100% = 92.7%, es decir, $R^2 = 0.927$, entonces el coeficiente de correlación es: $r = \sqrt{R^2} = \sqrt{0.927} = 0.9628$, este valor significa que las variables consumo y precio tienen una fuerte asociación directa.

c. Encuentre la línea de regresión que exprese el consumo en términos del ingreso.

Solución:

Regression Analysis: Consumo(Y) versus Ingreso(X2)

```
The regression equation is
Consumo(Y) = - 2.00 + 0.00660 Ingreso(X2)
Predictor         Coef     SE Coef       T       P
Constant       -2.0033      0.2635   -7.60   0.000
Ingreso(X2)  0.0065954   0.0007389    8.93   0.000
```

Entonces el modelo es: $\text{Consumo} = -2.0033 + 0.006595(\text{Ingreso})$

d. ¿Existe correlación entre el consumo y el ingreso?

Solución:

Haciendo uso de software de Minitab

```
S = 0.0192466   R-Sq = 90.9%   R-Sq(adj) = 89.7%
```

De igual manera, $r = \sqrt{R^2} = \sqrt{0.909} = 0.9534$, por consiguiente hay una fuerte relación directa entre las variables consumo e ingreso.

e. Encuentre el modelo de regresión entre consumo, precio, ingreso y temperatura.

Solución:

Haciendo uso de software de Minitab

Regression Analysis: Consumo(Y) versus Precio(X1), Ingreso(X2), ...

```
The regression equation is
Consumo(Y) = 0.530 - 0.461 Precio(X1) + 0.00094 Ingreso(X2)
             + 0.00209 Temperatura(X3)
Predictor             Coef    SE Coef      T      P
Constant            0.5296     0.7168   0.74  0.488
Precio(X1)         -0.4609     0.2726  -1.69  0.142
Ingreso(X2)       0.000942   0.001357   0.69  0.514
Temperatura(X3)  0.0020923  0.0007708   2.71  0.035
```

Entonces el modelo es:

$$\text{Consumo} = 0.5296 - 0.4609(\text{Precio}) + 0.000942(\text{Ingreso}) + 0.002092(\text{Temperatura})$$

f. Estime el consumo cuando el precio/Kg es 1.29 dólares, el ingreso es de 400 dólares y la temperatura es de 25 grados Fahrenheit.

Solución:

$$\text{Consumo} = 0.5296 - 0.4609(1.29) + 0.000942(400) + 0.002092(25)$$

$$\text{Consumo} = 0.3641 \text{ Kg.}$$

El consumo estimado es de 0.3641 Kg por semana.

11. Se realizó un estudio sobre el empuje de un motor de turbina (hp), velocidad de rotación primaria (rpm), rapidez del flujo de combustible (m/seg), presión (bar) y temperatura de escape (^{o}C) .Use un $\alpha = 0.05$.

Vehículo	Empuje del motor (Y)	Velocidad (X_1)	Rapidez (X_2)	Presión (X_3)	Temperatura (X_4)
1	45435	2139	30255	206	1735
2	3655	1676	29330	163	1599
3	3201	1475	28961	143	1543
4	4835	2122	29835	210	1670
5	4340	1702	29099	175	1587
6	3819	2107	27310	155	1580
7	4444	1674	20521	190	1689
8	4189	2165	17980	180	1598
9	3982	2048	19780	139	1684
10	3623	1658	19021	208	1478
11	3126	2062	20681	199	1561
12	4561	1929	18950	197	1668
13	3630	1596	18700	145	1729
14	4330	1400	19681	153	1691
15	4119	2048	17870	173	1568

a. Interprete los coeficientes estimados del modelo de regresión lineal múltiple.

Solución:

Haciendo uso de software de Minitab

Regression Analysis: Empuje del m versus Velocidad (X, Rapidez (X2), ...

```
The regression equation is
Empuje del motor (Y) = -147657 + 6.1 Velocidad(X1) + 0.754 Rapidez(X2)
+ 131 Presión(X3) + 63.1 Temperatura(X4)
Predictor            Coef  SE Coef      T      P
Constant          -147657    58438  -2.53  0.030
Velocidad (X1)       6.07    10.13   0.60  0.562
Rapidez (X2)       0.7540   0.4799   1.57  0.147
Presión (X3)        131.4    107.1   1.23  0.248
Temperatura (X4)    63.12    32.83   1.92  0.083
```

Interpretación de los coeficientes:

$\hat{\beta}_1$: Por cada revolución por minuto en que se incrementa en la velocidad, el empuje del motor de turbina se incrementa en 6.07 unidades hp, permaneciendo constantes las variables rapidez, presión y temperatura.

$\hat{\beta}_2$: Por cada m/seg que se incrementa en la rapidez, el empuje del motor de turbina se incrementa en 0.7540 hp, permaneciendo constantes las variables velocidad, presión y temperatura.

$\hat{\beta}_3$: Por cada unidad adicional que se incrementa en la presión, el empuje del motor de turbina se incrementa en 131.4 hp, permaneciendo constantes las variables velocidad, rapidez y temperatura.

$\hat{\beta}_4$: Por cada grado que se incrementa en la temperatura, el empuje del motor de turbina se incrementa en 63.12 hp, permaneciendo constantes las variables velocidad, rapidez y presión.

b. ¿Es el modelo lineal múltiple apropiado?

Solución:

Haciendo uso de software de Minitab

```
Analysis of Variance
Source          DF          SS          MS      F      P
Regression       4   767747080  191936770  2.29  0.132
Residual Error  10   838776923   83877692
Total           14  1606524004
```

Prueba F. Confiabilidad del modelo

i. Formulación de la hipótesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$H_1 : \text{Al menos un } \beta_i \text{ es diferente de cero} \quad \beta_i \neq 0$

ii. Nivel de significancia: $\alpha = 0.05$

iii. Estadística de prueba:

$$F_0 = \frac{CMR}{CME} = \frac{191936770}{83877692} = 2.29 \; ; \quad \text{P-value} = 0.132.$$

iv. Regla de decisión:
Si P-value $< \alpha$, entonces se rechaza H_0

v. Decisión y conclusión:
El P-value = 0.132 es mayor al nivel de significancia (0.05), se concluye que no se rechaza la hipótesis nula, es decir las variables velocidad, rapidez, presión y temperatura en forma conjunta no son apropiadas para explicar el empuje del motor.

c. ¿Contribuye la velocidad (X_1) a explicar significativamente el empuje del motor?

Solución:

i. Prueba de hipótesis de β_1

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

ii. Nivel de significancia: $\alpha = 0.05$

iii. Estadística de prueba:

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} = \frac{6.071 - 0}{10.13} = 0.60 \qquad \text{P-value} = 0.562$$

iv. Regla de decisión:
Si P-value $< \alpha$, entonces se rechaza H_0.

v. Decisión y conclusión:

El P-value = 0.562 es mayor que el nivel de significación (0.05), es decir no se rechaza H_0, esto significa que X_1 no es una variable significativa en el modelo.

12. Los siguientes datos fueron registrados por una compañía de mudanzas sobre los pesos (miles de libras), la distancia de traslado (miles de km) y los daños sufridos (en dólares):

Peso (X_1)	4.0	3.0	1.6	1.2	3.4	4.8	3.5	2.0	1.6	3.5	4.0	1.0
Distancia (X_2)	1.5	2.2	1.0	2.0	0.8	1.6	1.4	2.6	1.2	2.4	2.4	1.0
Daño (Y)	160	112	69	90	123	186	148	105	74	156	175	140

a. Probar si el modelo lineal múltiple es adecuado.

Solución:

Haciendo uso de software de Minitab.

Regression Analysis: Daño(Y) versus Peso (X1), Distancia (X2)

```
The regression equation is
Daño(Y) = 54.1 + 24.2 Peso (X1) + 3.9 Distancia (X2)
Predictor         Coef  SE Coef     T      P
Constant         54.06    25.69  2.10  0.065
Peso (X1)       24.157    6.247  3.87  0.004
Distancia (X2)    3.86    12.63  0.31  0.767
S = 25.6641   R-Sq = 64.2%   R-Sq(adj) = 56.3%
Analysis of Variance
Source          DF       SS      MS     F      P
Regression       2  10647.8  5323.9  8.08  0.010
Residual Error   9   5927.8   658.6
Total           11  16575.7
```

Para probar si el modelo es adecuado se trabajará con la prueba F:

i. Formulación de la hipótesis:

$H_0 : \beta_1 = \beta_2 = 0$ (el modelo no es apropiado).

$H_1 : \text{Al menos } \beta_i \neq 0$ (el modelo sí es apropiado).

ii. Nivel de significación: $\alpha = 0.05$

iii. Estadística de prueba:

$$F_0 = \frac{5323.9}{658.6} = 8.08 \; ; \qquad \text{P-value} = 0.010$$

iv. Regla de decisión:
Si P-value $< \alpha$, entonces se rechaza H_0.

v. Decisión y conclusión:
Como el valor P-value = 0.010 es menor que el nivel de significación (0.05), entonces se rechaza H_0 y se concluye que las variables peso y distancia son apropiados para explicar el comportamiento del daño.

b. ¿La variable X_2 contribuye significativamente al modelo?

Solución:

Para probar que la variable X_2 contribuye al modelo se debe plantear la prueba T:

i. Prueba de hipótesis:

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

ii. Nivel de significación: $\alpha = 0.05$

iii. Estadística de prueba:

$$t_0 = \frac{3.86}{12.63} = 0.31 \; ; \qquad \text{P-value} = 0.767$$

iv. Regla de decisión:
Si P-value $< \alpha$, entonces se rechaza H_0.

v. Decisión y conclusión:
Como el P-value = 0.767 es mayor que el nivel de significación, no se rechaza la hipótesis nula; es decir, X_2 no contribuye significativamente a explicar el daño.

c. Interprete los coeficientes estimados $\hat{\beta}_1$ y $\hat{\beta}_2$.

Solución:

Interpretación de $\hat{\beta}_1$ y $\hat{\beta}_2$

$\hat{\beta}_1 = 24.157$; si se mantiene constante la distancia, entonces, en promedio, el daño se incrementa en 24.157 dólares cuando el peso se incrementa en mil libras.

$\hat{\beta}_2 = 3.86$ si se mantiene constante el peso, entonces el daño se incrementa en 3.86 dólares cuando la distancia se incrementa en mil kilómetros.

d. Estime los daños de un cargamento que pesa 2.400 libras y se trasladó a 1.200 Km.

Solución:

Se sabe que el modelo es:
$\text{Daño} = 54.06 + 24.2(\text{Peso}) + 3.86(\text{Distancia})$

Entonces:
$\text{Daño} = 54.06 + 24.2(2.4) + 3.86(1.2) = 116.772$ dólares.

13. El gerente de un banco está interesado en obtener una mejor "percepción" de las características de las familias que pagan sus gastos mensuales con cheques. El gerente considera que el número de cheques girados por mes (Y) está relacionada con las siguientes variables:

X_1: Edad del jefe de la familia (en años).

X_2: Ingreso familiar (en miles de dólares).

X_3: Número de miembros de la familia.

Para analizar esta relación, se seleccionó una muestra aleatoria de 28 clientes registrándose información de las variables indicadas. El reporte de Minitab del análisis de regresión de los datos registrados fue:

```
Regression Analysis
Predictor          Coef         T
Constant         -17.154     -2.46
      X1          0.1141      0.63
      X2          1.8451      6.08
      X3          2.7930      2.35
Se = 8.811
Analysis of Variance
Source         DF       SS           MS          F
Regression     3    5645.4       1881.35     25.29
Error          24   1785.5         74.40
Total          27   7430.9
```

Nota: Formule las hipótesis correspondientes, la prueba estadística y su valor, la regla de decisión y sus conclusiones.

a. Interprete los valores de los estimadores de los coeficientes de regresión β_1 y β_3.

Solución:

Interpretación de los coeficientes:

$\hat{\beta}_1 = 0.1141$ significa que por cada año de edad adicional que tenga el jefe de familia, el número de cheques girados mensualmente aumenta en 0.1141, manteniendo constantes el ingreso familiar y el número de miembros de la familia.

$\hat{\beta}_3 = 2.793$ significa que por cada miembro adicional que tenga la familia, el número de cheques girados mensualmente aumenta en 2.793, manteniendo constantes la edad del jefe de familia y el ingreso familiar.

b. ¿Es el modelo significativo con $\alpha = 0.01$?

Solución:

Hipótesis de confiabilidad del modelo.

i. Prueba de hipótesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

$H_1 : \text{Al menos un } \beta_i \neq 0$

ii. Nivel de significación: $\alpha = 0.01$

iii. Estadística de prueba:

$$F_0 = \frac{1881.35}{74.40} = 25.29$$

iv. Regla de decisión:

Si $F_0 > F_{(1-\alpha,k,n-k-1)}$, entonces se rechaza H_0, en caso contrario no se rechaza.

Haciendo uso del software de Minitab

Inverse Cumulative Distribution Function

```
F distribution with 3 DF in numerator and 24 DF in
denominator
P( X <= x )          x
        0.99   4.71805
```

Entonces:
Se rechaza H_0 si $F_0 > F_{(0.99,3,24)} \Rightarrow 25.59 > 4.71805$

Decisión y conclusión:
Se rechaza H_0. El modelo es apropiado.

c. Calcule e interprete el coeficiente de determinación.

Solución:

$$R^2 = \frac{5645.4}{7430.9}(100) = 75.97\%$$

El 75.97% de la variación total es explicada por las variables consideradas en el modelo.

d. Calcular un intervalo de confianza para el coeficiente de regresión de X_2. Use $\alpha = 0.05$.

Solución:

La desviación estándar del coeficiente de regresión es

$$s(\hat{\beta}_2) = \frac{\hat{\beta}_2}{t_0} = \frac{1.8451}{6.08} = 0.303470$$

Por otro lado, se sabe que:

$$\beta_2 \in \left\langle \hat{\beta}_2 \pm t_{(1-\alpha/2,n-k-1)} s(\hat{\beta}_2) \right\rangle$$

$$\beta_2 \in \left\langle 1.8451 \pm t_{(0.975,24)} (0.30347) \right\rangle = \left\langle 1.8451 \pm 2.0639 (0.30347) \right\rangle = \langle 1.218768 , 2.471432 \rangle$$

Interpretación: se tiene 95% de confianza de que β_2 se encuentre dentro del intervalo $\langle 1.2187, 2.4714 \rangle$.

14. La Sunat desea estimar el importe mensual (Y) de los impuestos no pagados (en millones de dólares) detectados por su departamento de auditoría. Actualmente se posee información de los últimos 10 meses de las siguientes variables:

X_1: Horas de trabajo en auditoría de campo (en cientos).

X_2: Número de horas que la computadora tarda en detectar impuestos no pagados.

Suponiendo un modelo lineal múltiple de dos variables y usando Minitab, se obtuvieron los siguientes resultados:

```
The regression equation is
Y = - 13.8 + 0.564 X1 + 1.10 X2
Predictor      Coef    StDev       T       P
Constant    -13.82    13.32   -1.04   0.334
   X1       0.5637   0.3033    1.86   0.105
   X2       1.0995   0.3131    3.51   0.010
S = 1.071       R-Sq = 72.9%
```

a. Interprete los coeficientes de la recta estimada.

Solución:

Coeficientes de la recta

$\hat{\beta}_1 = 0.5637$: Si se mantiene constante el número de horas que trabaja la computadora, los impuestos no pagados se incrementan en \$563.700 por cada 100 horas que aumente la auditoría de campo.

$\hat{\beta}_2 = 1.0995$: Si se mantiene constante el número de horas de auditoría de campo, los impuestos no pagados se incrementan en \$1.099.500 por cada hora adicional de tiempo de la computadora.

b. ¿Cuál o cuáles de los parámetros del modelo no son significativos? Justifique estadísticamente su respuesta planteando sus hipótesis adecuadas. Suponga $\alpha = 0.05$.

Solución:

Formulación de la hipótesis:

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

Como: $t_0 = -1.04$ y P-value = 0.334 el cual es mayor que $\alpha = 0.05$, se concluye que no se rechaza H_0, es decir la constante no es significativa en el modelo.

Prueba de hipótesis de β_1:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Se tiene como información que: $t_0 = 1.86$ y P-value = 0.105

Como P-value = 0.105 es mayor que $\alpha = 0.05$, se concluye que no se rechaza H0, es decir la variable X_1 no es significativa en el modelo.

Prueba de hipótesis de β_2:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Se tiene como información que: $t_0 = 3.51$ y P-value = 0.010

Como P-value = 0.010 es menor que $\alpha = 0.05$, se concluye que se rechaza H_0, es decir la variable X_2 es significativa en el modelo.

Analizando el modelo, la gerencia de la Sunat considera que el modelo no es lo suficientemente adecuado para explicar la variabilidad de los impuestos no pagados, por lo que considera conveniente agregar una nueva variable:

X_3: Recompensa (premio) a los informantes (miles de nuevos soles).

Al usar nuevamente Minitab se obtiene el siguiente reporte:

The regression equation is

$Y = -45.8 + 0.597\ X_1 + 1.18\ X_2 + 0.405\ X_3$

Predictor	Coef	StDev	T	P
Constant	-45.796	4.878	-9.39	0.000
X1	0.59697	0.08112	7.36	0.000
X2	1.17684	0.08407	14.00	0.000
X3	0.40511	0.04223	9.59	0.000

$S_e = 0.2861$ R-Sq = 98.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regresión	3	29.1088	9.7029	118.52	0.000
Error	6	0.4912	0.0819		
Total	9	29.6000			

c. ¿Qué efecto tiene haber añadido la variable X_3 al modelo?

Solución:

Efecto de X_3:

El efecto es significativo, en el sentido que incrementa el valor de R^2*100% en 25.4% (98.3% - 72.9% = 25.4%) por lo que se puede concluir que el modelo es apropiado para predecir el monto de los impuestos no pagados.

Se puede notar, además, que todos los coeficientes son significativos en el modelo ya que los P-value son aproximadamente iguales a cero.

d. ¿Qué hipótesis H_0 plantearía para usar el valor P-value de la tabla Anova? ¿Cuál sería su conclusión?

Solución:

Formulación de hipótesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (el modelo no es significativo).

$H_1 : \text{Al menos un } \beta_i \neq 0$ (el modelo es significativo).

Decisión y conclusión:

Como el valor P-value es aproximadamente igual a cero se rechaza H_0 y se concluye que el modelo es significativo.

e. En términos del enunciado interprete el valor de R^2 del modelo 2.

Solución:

La variación total en impuestos no pagados es explicado en 98.3% por las tres variables independientes (horas de trabajo de auditoría de campo, número de horas que trabaja la computadora y recompensa a los informantes).

15. El gerente general de una empresa que ofrece servicios de "software a medida" supone que el costo de producción (Y en miles de nuevos soles) de un programa está relacionado a las siguientes variables:

X_1 = Tiempo de demora (en horas) para elaborar el programa.

X_2 = Número de programadores que intervienen en la construcción del programa.

X_3 = Gastos operativos (en miles de nuevos soles) para la construcción del programa.

Para probar esta suposición el gerente registró la información apropiada y luego de procesarla con Minitab obtuvo el siguiente reporte:

Regression Analysis: Y versus X1; X2; X3

```
Predictor      Coef      T
Constant    -10.829   -3.00
X1          0.25946    2.83
X2           1.1228    2.75
X3            6.698    1.23
Analysis of Variance
Source           DF      SS
Regression
Residual Error    6    1.64
Total                476.40
```

Use en sus cálculos.

a. ¿Es la variable X_3 significativa? Señale las hipótesis, la prueba estadística y la región crítica correspondiente $\alpha = 0.04$.

Solución:

Primero se obtiene la desviación estándar de los coeficientes:

$$s\left(\hat{\beta}_0\right) = \frac{\hat{\beta}_0}{t_0} = \frac{-10.829}{-3.00} = 3.61 \qquad s\left(\hat{\beta}_1\right) = \frac{\hat{\beta}_1}{t_0} = \frac{0.25946}{2.83} = 0.09$$

$$s\left(\hat{\beta}_2\right) = \frac{\hat{\beta}_2}{t_0} = \frac{1.1228}{2.75} = 0.41 \qquad s\left(\hat{\beta}_3\right) = \frac{\hat{\beta}_3}{t_0} = \frac{6.698}{1.23} = 5.45$$

Prueba de significancia para el X_3

i. Formulación de hipótesis:

$H_0 : \beta_3 = 0$ (la variable X_3 no influye en el modelo).

$H_1 : \beta_3 \neq 0$ (la variable X_3 influye en el modelo).

ii. Estadística de prueba:

Del reporte: $t_0 = \frac{6.698}{5.45} = 1.23$

iii. Regla de decisión:

Si $t_0 < t_{(\alpha/2)}$ ó $t_0 > t_{(1-\alpha/2)}$ se rechaza la hipótesis nula.

Reporte de Minitab

Inverse Cumulative Distribution Function

```
Student's t distribution with 6 DF
P( X <= x )          x
        0.02   -2.61224
        0.98    2.61224
```

Entonces:

Si $t_0 < -2.61224$ ó $t_0 > 2.61224$ se rechaza la hipótesis nula.

iv. Región crítica:

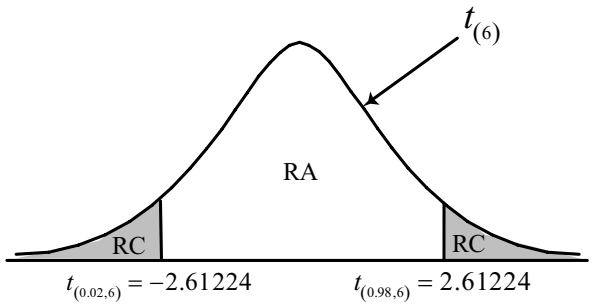


Figura 23. Prueba de dos colas para β_3.

v. Decisión y conclusión:
No se rechaza la hipótesis nula, debido a que t_0 pertenece a la región de aceptación, por consiguiente X_3 no influye en el modelo.

b. ¿Cuál es el intervalo de confianza para β_2? Interprete.

Solución:

Se sabe que: $\left\langle \hat{\beta}_2 \pm t_{(1-\alpha/2,n-k-1)} s\left(\hat{\beta}_2\right) \right\rangle$

Entonces:

$\left\langle 1.1228 \pm t_{(1-0.04/2,10-3-1)}(0.41) \right\rangle$

$\left\langle 1.1228 \pm t_{(0.98,6)}(0.41) \right\rangle$

Reporte de Minitab:

```
Inverse Cumulative Distribution Function
Student's t distribution with 6 DF
P( X <= x )          x
        0.98   2.61224
```

$\left\langle 1.1228 \pm 2.61224(0.41) \right\rangle = \langle 0.0562, 2.1894 \rangle$

Se tiene 96% de confianza de que $\hat{\beta}_2$ se encuentre en el intervalo $\langle 0.0562,\ 2.1894 \rangle$.

c. ¿Es factible afirmar que el costo de producción aumenta en 250 nuevos soles por cada hora adicional de demora en la construcción del programa? Señale las hipótesis, la prueba estadística y la región crítica correspondiente.

Solución:

i. Formulación de la hipótesis:

$H_0 : \beta_1 = 0.25$

$H_1 : \beta_1 \neq 0.25$

ii. Estadística de prueba: $t_0 = \dfrac{\hat{\beta}_1 - \beta_1}{s\left(\hat{\beta}_1\right)} = \dfrac{0.25946 - 0.25}{0.09} = \dfrac{0.00946}{0.09} = 0.11$

iii. Regla de decisión:

Si $t_0 < t_{(\alpha/2)}$ ó $t_0 > t_{(1-\alpha/2)}$ se rechaza la hipótesis nula.

iv. Región crítica:

Reporte de Minitab

Inverse Cumulative Distribution Function

```
Student's t distribution with 6 DF
P( X <= x )          x
        0.02  -2.61224
        0.98   2.61224
```

Entonces:

Si $t_0 < -2.61224$ ó $t_0 > 2.61224$ se rechaza la hipótesis nula.

v. Decisión y conclusión:

No se rechaza la hipótesis nula, t_0 pertenece a la región de aceptación, es decir que si se puede afirmar que el costo aumenta en 250 nuevos soles por cada hora adicional de demanda en la construcción del programa.

d. Si el tiempo de demora en construir un programa es de 83 horas con 6 programadores y un gasto total operativo de 950 nuevos soles, ¿cuál es su pronóstico del costo de producción del programa?

Solución:

Por el reporte brindado se sabe que:

$$\text{Costo de producción} = -10.829 + 0.25946\left(\text{Tiempo d e demora}\right) +$$

$$+1.1228\left(\text{Número de programadores}\right) + 6.698\left(\text{Gastos operativos}\right)$$

Entonces:

$$\text{Costo de Producción} = -10.829 + 0.25946(83) + 1.1228(6) + 6.698(950) = 6380.5429$$

e. ¿Cuál es el valor de la prueba estadística para probar la hipótesis de que el modelo es apropiado?

Solución:

Las hipótesis adecuadas son:

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (el modelo no es significativo)

$H_1 : \text{Al menos un } \beta_i \neq 0$ (el modelo es significativo)

Se tiene que:

```
Analysis of Variance
Source           DF       SS
Regression
Residual Error    6     1.64
Total                 476.40
```

Completando la tabla Anova:
Se sabe que el número de variables regresoras es 3, por consiguiente el número de grados de libertad asociados a la regresión es de 3.

Como los grados de libertad del error son 6, entonces $n - k$ -1 = 6 y por lo tanto $n = 10$.

Como los grados de libertad del total es $(n - 1)$ entonces su valor es 9.

La SCR = SCT – SCE SCR = 476.40 – 1.64 = 474.76
El CMR = SCR / GL CMR = 474.76 / 3 = 158.25
El CME = SCE / GL CME = 1.64 / 6 = 0.2733
El F = CMR / CME F = 158.25 / 0.2733 = 579.0340

Entonces:

Analysis of Variance				
Source	DF	SS	**MS**	**F**
Regression	**3**	**474.76**	**158.25**	**579.0340**
Residual Error	6	1.64	**0.2733**	
Total	**9**	476.40		

Luego, la estadística de prueba es: $F_0 = 579.0340$

16. La Sociedad Peruana de Cardiología reúne datos sobre el riesgo de un ataque. Un estudio que duró 10 años proporcionó datos acerca de cómo se relacionan la edad (en años), la presión sanguínea (en mmHg) y los hábitos de fumar (sí fuma = 1, no fuma = 0) con el riesgo de ataque cardíaco. La información que se presenta a continuación es una salida de Minitab con parte de los datos del estudio. El riesgo de un ataque cardíaco (en porcentaje) se define como la probabilidad, multiplicada por 100, de que el paciente sufra un ataque cardíaco en los próximos 10 años.

La ecuación de regresión es:
Riesgo = –9 2.1 + 1.08 Edad + 0.254 Presion + 8.49 Fumador

Predictor	Coef	SE Coef	T	P
Constant	-92.06		-6.35	0.0000
Edad	1.0801		6.83	0.0000
Presión	0.25364		5.89	0.0000
Fumador	8.495		2.97	0.0009

Source	DF	SS	MS	F
Regresión				
Residual error				
Total	19	4119.0		

Source	DF	Seq SS
Edad	1	1758.2
Presión	1	1613.9
Fumador	1	265.5

Use $\alpha = 0.03$ en sus cálculos.

a. Complete los espacios en blanco.

Solución:

Primero, se obtiene la desviación estándar de los coeficientes:

$$s\left(\hat{\beta}_0\right) = \frac{\hat{\beta}_0}{t_0} = \frac{-92.06}{-6.35} = 14.4976 \qquad s\left(\hat{\beta}_1\right) = \frac{\hat{\beta}_1}{t_0} = \frac{1.0801}{6.83} = 0.1581$$

$$s\left(\hat{\beta}_2\right) = \frac{\hat{\beta}_2}{t_0} = \frac{0.25364}{5.89} = 0.0431 \qquad s\left(\hat{\beta}_3\right) = \frac{\hat{\beta}_3}{t_0} = \frac{8.495}{2.97} = 2.8603$$

Por consiguiente la tabla quedaría:

Predictor	Coef	SE Coef	T	P
Constant	-92.06	14.4976	-6.35	0.0000
Edad	1.0801	0.1581	6.83	0.0000
Presion	0.25364	0.0431	5.89	0.0000
Fumador	8.495	2.8603	2.97	0.0009

Completando la tabla Anova.

Como el número de variables regresoras es 3, los grados de libertad asociados a la regresión son de 3.

Los grados de libertad del error se obtienen por la diferencia de GL(total) – GL(regresión), es decir (19 – 3 = 16)

Se tiene entonces que:

Source	DF	Seq SS
Edad	1	1758.2
Presión	1	1613.9
Fumador	1	265.5

Se sabe que la suma de cuadrados de la regresión está dado por:
SCR = SC (Edad) + SC (Presión) + SC (Fumador)
SCR = 1758.2 + 1613.9 + 265.5 = 3637.6
SCE = SCT – SCR lo que implica que SCE = 4119.0 – 3637.6 = 481.4
El CMR = SCR / GL $\Rightarrow$ CMR = 36373.6 / 3 = 1212.53
El CME = SCE / GL $\Rightarrow$ CME = 481.4 / 16 = 30.0875
La F = CMR / CME $\Rightarrow$ F = 1212.53 / 30.0875 = 40.3001
El valor del P-value es:

$$P-value = P\left(F_{(k,n-k-1)} > F_0\right) = P\left(F_{(3,16)} > 40.3001\right) = \left[1 - P\left(F_{(3,16)} \leq 40.3001\right)\right]$$

Reporte de Minitab:

Cumulative Distribution Function

```
F distribution with 3 DF in numerator and 16 DF in
denominator
       x       P( X <= x )
40.3001       1.00000
```

Por consiguiente: P-value = (1 – 1) = 0
La tabla quedaría:

Source	DF	SS	MS	F	P
Regresión	3	3637.6	1212.53	40.3001	0.000
Residual error	16	481.4	30.0875		
Total	19	4119			

b. Calcule e interprete el coeficiente de determinación.

Solución:

El coeficiente de determinación está dado por:

$$R^2 = \frac{SCR}{SCT} = \frac{3637.6}{4119.0} = 0.8831$$

Interpretación. La variabilidad del riesgo de que un paciente sufra un ataque cardíaco es explicado en un 88.31% por su edad, su presión y el hecho de que sea fumador o no.

c. ¿Es el modelo propuesto adecuado? Señale las hipótesis, la región crítica y sus conclusiones.

Solución:

i. Formulación de la hipótesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (el modelo no es adecuado)

$H_1 : \text{Al menos un } \beta_i \neq 0$ (el modelo es adecuado)

ii. Nivel de significancia: $\alpha = 0.03$

iii. Estadística de prueba:

$F_0 = 40.3001$

iv. Regla de decisión:

Si $F_0 > F_{(1-\alpha,k,n-k-1)}$ se rechaza la hipótesis nula.

Reporte de Minitab:

Inverse Cumulative Distribution Function

F distribution with 3 DF in numerator and 16 DF in denominator

P(X <= x)	x
0.97	3.85003

Como la estadística de prueba 40.3001 es mayor que el valor crítico 3.85003 entonces se rechaza la hipótesis nula.

v. Decisión y conclusión:

Se concluye entonces que el modelo propuesto es adecuado (significativo).

d. Calcule e interprete un intervalo de confianza para β_2.

Solución:

Se sabe que:

$$\left\langle \hat{\beta}_2 \pm t_{(1-\alpha/2,n-k-1)} s\left(\hat{\beta}_2\right) \right\rangle$$

$$\left\langle 0.25364 \pm t_{(1-0.03/2,20-3-1)}\left(0.0431\right) \right\rangle = \left\langle 0.25364 \pm t_{(0.985,16)}\left(0.0431\right) \right\rangle$$

Reporte de Minitab:

Inverse Cumulative Distribution Function

Student's t distribution with 16 DF

P(X <= x)	x
0.985	2.38155

$$\left\langle 0.25364 \pm 2.38155\left(0.0431\right) \right\rangle = \left\langle 0.150995, 0.356285 \right\rangle$$

Es decir, β_2 se encuentra con un 97% de confianza en el intervalo $\langle 0.150995,\ 0.356285 \rangle$

e. El señor Juan Pérez es un fumador con 58 años de edad y con presión sanguínea 175, ¿cuál es el riesgo porcentual de que el señor Pérez sufra un ataque cardíaco en los próximos 10 años?

Solución:

Se sabe que el modelo es:

$$\text{Riesgo} = -92.06 + 1.0801(\text{Edad}) + 0.25364(\text{Presión}) + 8.495(\text{Fumador})$$

Entonces:

$$\text{Riesgo} = -92.06 + 1.0801(58) + 0.25364(175) + 8.495(1) = 23.4678$$

Es decir, el señor Pérez tiene 23.46% de riesgo de sufrir un ataque cardíaco en los próximos 10 años.

17. Utilice el siguiente reporte de Minitab para responder las preguntas que se plantean a continuación.

Regression Analysis: Ingreso versus N° Hijos, Edad escolar, Padres alcohólicos

```
Ingreso = 258.84 - 3.305 N°Hijos - 5.467 Edad escolar
+ 2.99 Padres alcohol
Predictor        Coef   SE Coef
Constant       258.84     17.14
N° Hijos       -3.305     3.688
Edad escolar   -5.467     6.832
Padres alcoh     2.99     10.73
S = 166.481
Analysis of Variance
```

Fuentes de variabilidad	GL	SC	CM	F	P-value
Regresión					
Error					
Total	999	27647616			

a. Complete la tabla del análisis de varianza.

Solución:

Como k = 3 (número de variables regresoras) ⇒ los grados de libertad son 3.
Los GL (error) son: 999 – 3 = 996 (diferencia del GL (total) – GL (regresión))
Se sabe que s = 166.481, que s^2 = CME y que CME = SCE / GL (error)
Por lo tanto: S^2 = SCE / 996 ⇒ SCE = 996*(166.481^2) = 27605059.67
La SCR = SCT – SCE ⇒ SCR = 27647616 – 27605059.67 = 42556.33
El CMR = SCR / GL (regresión) ⇒ CMR = 42556.33 / 3 = 14185.44
El CME = SCE / GL (Error) ⇒ CME = 27605059.67 / 996 = 27715.92
El F = CMR / CME ⇒ F = 14185.44 / 27715.92 = 0.5119

El valor de P-value es:

$$P-value = P\left(F_{(k,n-k-1)} > F_0\right) = P\left(F_{(3,996)} > 0.5119\right) = \left[1 - P\left(F_{(3,996)} \leq 0.5119\right)\right]$$

Reporte de Minitab:

Cumulative Distribution Function

```
F distribution with 3 DF in numerator and 996 DF in
denominator
      x      P( X <= x )
0.5119      0.325850
```

Reemplazando.
P-value = (1 – 0.32585) = 0.67415

Completando la tabla:

Fuentes de variabilidad	GL	SC	CM	F	P-value
Regresión	3	42556.33	14185.44	0.5119	0.6741
Error	996	27605059.67	27715.92		
Total	999	27647616			

b. ¿Cuál es el valor del P-value correspondiente a la hipótesis H_0: $\beta_3 = 0$?

Solución:

La estadística de prueba está dada por:

$$t_0 = \frac{\hat{\beta}_3 - \beta_3}{s\left(\hat{\beta}_3\right)} = \frac{2.99 - 0}{10.73} = 0.2787$$

El valor del P-value está dado por:

$$\text{P-value} = 2P\left(t_{(n-k-1)} > t_0\right) = 2P\left(t_{(996)} > 0.2787\right) = 2\left[1 - P\left(t_{(996)} \leq 0.2787\right)\right]$$

Reporte de Minitab:

Cumulative Distribution Function

```
Student's t distribution with 996 DF
      X      P( X <= x )
0.2787      0.609734
```

$$P-value = 2\left(1 - 0.609734\right) = 2\left(0.390266\right) = 0.780532$$

c. ¿Cuál es la regla de decisión para juzgar si el modelo es apropiado? Use $\alpha = 0.07$.

Solución:

i. Formulación de la hipótesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (el modelo no es adecuado)

$H_1 : \text{Al menos un } \beta_i \neq 0$ (el modelo es adecuado)

ii. Nivel de significancia:
$\alpha = 0.07$

iii. Estadística de prueba:

$F_0 = 0.5119$

iv. Regla de decisión:

Si $F_0 > F_{(1-\alpha,k,n-k-1)}$ se rechaza la hipótesis nula.

Reporte de Minitab:

Inverse Cumulative Distribution Function

```
F distribution with 3 DF in numerator and 996 DF in
denominator
P( X <= x )          x
        0.93   2.36061
```

Entonces:

$0.5119 < F_{(0.93,3,996)} \Rightarrow 0.5119 < 2.3606$ (no se rechaza la hipótesis nula).

v. Decisión y conclusión:

No se rechaza la hipótesis, es decir el modelo no es significativo.

d. ¿Cuál es el ingreso de una familia con 5 hijos, 2 de ellos en edad escolar y el padre alcohólico?

Solución:

Se sabe que el modelo es:

$$\text{Ingreso} = 258.84 - 3.305(\text{N° hijos}) - 5.467(\text{Edad escolar}) + 2.99(\text{Padre alcohólico})$$

Entonces:

$$\text{Ingreso} = 258.84 - 3.305(5) - 5.467(2) + 2.99(1) = 234.371$$

18. La siguiente información corresponde a una muestra aleatoria de 35 pacientes de quienes se conoce su edad y una medición de su tensión sistólica. Se desea estudiar la variación en la tensión sistólica en función de la edad del individuo.

Regression Analysis: Tensión Sistólica versus Edad

```
Predictor           Coef   SE Coef
Constant         119.576     6.163
Edad              0.4346    0.1593
Analysis of Variance
Source                DF        SS
Regression             1     970.7
Residual Error        33    4304.0
Total                 34    5274.7
```

a. Indique el valor de la prueba estadística para probar la H_0: $\beta_1 = 0.7$.

Solución:

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{s\left(\hat{\beta}_1\right)} = \frac{0.4346 - 0.7}{0.1593} = \frac{-0.2654}{0.1593} = -1.66$$

b. Obtenga el valor del coeficiente de correlación lineal simple.

Solución:

Téngase presente que únicamente en la regresión lineal simple $r = \sqrt{R^2}$, luego:

$$r = \sqrt{R^2} = \sqrt{\frac{SCR}{SCT}} = \sqrt{\frac{970.7}{5274.7}} = \sqrt{0.184029} = 0.428987$$

Este valor significa que existe una pobre relación directa entre la edad y la tensión sistólica.

c. Indique cuál es el valor de la suma de cuadrados de las desviaciones de las observaciones con respecto a la línea de ajuste.

Solución:

Por definición $SCE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right) = 4304.0$

Respuesta: 4304.0

d. Obtenga el P-value de la H_0: El modelo no es apropiado.

Solución:

Hallando F_0

$$F_0 = \frac{CMR}{CME} = \frac{SCR/GL}{SCE/GL} = \frac{970.7/1}{4304/33} = \frac{970.7}{130.43} = 7.44$$

Obteniendo el P-value:

$$P - value = P\left(F_{(k,n-k-1)} > F_0\right) = P\left(F_{(1,33)} > 7.44\right) = \left[1 - P\left(F_{(1,33)} \leq 7.44\right)\right]$$

Reporte de Minitab:

Cumulative Distribution Function

```
F distribution with 1 DF in numerator and 33 DF in
denominator
    x   P( X <= x )
7.44      0.989860
P-value = (1 - 0.98986) = 0.01014
```

e. Obtenga el intervalo de confianza del 94% para β_1.

Solución:

Se sabe que:

$$\left\langle \hat{\beta}_1 \pm t_{(1-\alpha/2,n-k-1)} s\left(\hat{\beta}_1\right) \right\rangle$$

$$\left\langle 0.4346 \pm t_{(1-0.06/2,35-1-1)}\left(0.1593\right) \right\rangle = \left\langle 0.4346 \pm t_{(0.97,33)}\left(0.1593\right) \right\rangle$$

Reporte de Minitab:

Inverse Cumulative Distribution Function

```
Student's t distribution with 33 DF
P( X <= x )          x
        0.97   1.94770
```

$$\left\langle 0.4346 \pm 1.9477\left(0.1593\right) \right\rangle = \left\langle 0.124331,\ 0.744869 \right\rangle$$

Se tiene 94% de confianza de que β_1 se encuentre en este intervalo.

19. En una encuesta aplicada a empresarios nacionales acerca de políticas económicas se registró la siguiente información:

C1: Sector económico al que pertenece el empresario.
C2: Servicio o producto principal que ofrece el empresario.
C3: Ubicación de la empresa (centro, norte, oriente, sur).
C4: Ingreso estimado para el año 2005 (millones de dólares).
C5: Ingreso estimado del año 2006 en caso de firmarse el TLC (millones de dólares).
C6: Ingreso estimado del año 2006 en caso de no firmarse el TLC (millones de dólares).
C7: Años de funcionamiento de la empresa.

Los datos aparecen en el archivo "TLC.MTW".

Ejecute la regresión de C5 (Y) versus C4 (X_1) y C7 (X_2) para responder las siguientes preguntas:

a. ¿Cuál es el P-value de la H_0: $\beta_2 = 0.02$ versus H_1: $\beta_1 > 0.02$?

Solución:

Haciendo uso del Minitab

Regression Analysis: 2006-Si TLC versus Ingreso 2005, Años de func.

```
The regression equation is
2006-Si TLC = - 0.251 + 1.15 Ingreso 2005 + 0.0141 Años de func
Predictor         Coef  SE Coef      T      P
Constant       -0.2506   0.1598  -1.57  0.118
Ingreso 2005   1.14645  0.01276  89.84  0.000
Años de func.  0.01408  0.01523   0.92  0.356
```

Del reporte se sabe que:

$\hat{\beta}_2 = 0.01408$; $s\left(\hat{\beta}_2\right) = 0.01523$.

Hallando el P-value para la hipótesis $H_0 : \beta_2 = 0.02$ versus $H_1 : \beta_2 > 0.02$:
La estadística de prueba:

$$t_0 = \frac{\hat{\beta}_2 - \beta_2}{s\left(\hat{\beta}_2\right)} = \frac{0.01408 - 0.02}{0.01523} = \frac{-0.00592}{0.01523} = -0.3887$$

El P-value es:

$$\text{P-value} = P\left(t_{(n-k-1)} > t_0\right) = P\left(t_{(240-2-1)} > -0.3887\right) = \left[1 - P\left(t_{(237)} \leq -0.3887\right)\right]$$

Reporte de Minitab:

Cumulative Distribution Function

```
Student's t distribution with 237 DF
       x  P( X <= x )
-0.3887      0.348924
P-value = (1 - 0.348924) = 0.651076
```

b. ¿Cuál es el valor del coeficiente de correlación entre Y y X_2? ¿Qué significa?

Solución:

Coeficiente de correlación:

A partir del software Minitab en el módulo de correlación se obtiene el valor de $R_{Y,X_2} = -0.017$

Interpretación: Como el valor del coeficiente de correlación es muy pequeño (cercano a cero) se concluye que no existe correlación lineal entre los ingresos estimados del 2006 en caso de firmarse el TLC y los años de funcionamiento.

c. ¿Cuál es el intervalo de confianza del 98% para el ingreso promedio estimado del año 2006 en caso de firmarse el TLC de las empresas con $X_1 = 0.8$ y $X_2 = 8$? Interprete.

Solución:

Intervalo de confianza para el ingreso promedio estimado del año 2006.

$\mu_{Y|X_1=0.8;X_2=8} = (0.6634,\ 0.8951)$

Interpretación: hay un 98% de confianza de que el ingreso promedio estimado del 2006 con $X_1 = 0.8$ y $X_2 = 8$ se encuentre entre $\langle 0.6634,\ 0.8951 \rangle$.

d. ¿Cuál es el valor de la suma de cuadrados de los residuales (error)?

Solución:

Haciendo uso de software de Minitab.

```
Analysis of Variance
Source           DF       SS         MS        F      P
Regression        2   890.36     945.18  4037.16  0.000
Residual Error  237    55.49       0.23
Total           239  1945.85
```

Entonces, suma de cuadrados de los residuales es SC Error = 55.49.

PROBLEMAS PROPUESTOS

1. Un valor de r^2 cercano a cero indica que existe una fuerte correlación. V F

2. Suponga que la pendiente de una ecuación es positiva, entonces el valor de r debe ser la raíz cuadrada positiva de r^2. V F

3. El agregar variables adicionales a una regresión lineal múltiple siempre reducirá el error estándar de la estimación. V F

4. Para determinar si una regresión es significativa como un todo, se calcula el valor de la estadística de prueba F y se compara con un valor obtenido en la tabla F. V F

5. Una empresa de embutidos está interesada en medir el efecto del precio de jamón sobre la cantidad vendida.

Precio (S/./Kg)	20	18	25	19	21	19	20	21
Ventas (Kg)	50	60	35	70	45	65	55	50

a. Trace un diagrama de dispersión.
b. Desarrolle la ecuación de estimación que mejor describa estos datos.
c. Pronostique las ventas, cuando el precio del jamón es S/.22.

6. La siguiente información corresponde al consumo de gasolina de automóviles y el número de cilindros del motor.

Consumo (S/./Km)	0.18	0.3	0.7	0.45	0.68	0.25
Número de cilindros	4	4	8	6	8	4
Consumo (S/./Km)	0.2	0.5	0.4	0.42	0.6	0.35
Número de cilindros	4	6	6	6	8	4

a. Encuentre la línea de regresión de mínimos cuadrados del consumo de gasolina como función del número de cilindros.
b. Calcule el coeficiente de correlación lineal simple e interprete el resultado.
c Pruebe la hipótesis nula de que la pendiente de la recta de regresión poblacional es igual a cero (0) $\alpha = 0.05$.

7. Se cree que existe una relación de las holguras entre dos placas con soldadura (Y) y el número de placas en transistores (X).

Holgura (CMS)	0.746	0.252	0.541	0.519	0.357
Número de placas	40	20	25	23	21
Holgura	1.741	0.378	0.501	0.45	0.777
Número de placas	48	22	24	20	30

a. Obtenga la recta de regresión ajustada de mínimos cuadrados.
b. Pruebe la hipótesis $H_o : \beta_1 = 0$ $(\alpha = 0.05)$.
c. Construya un intervalo de confianza para la pendiente $(\alpha = 0.05)$.
d. Determine si el modelo es apropiado $(\alpha = 0.05)$.

8. Una compañía de productos químicos desea estudiar los efectos del tiempo sobre la eficiencia de una operación de extracción, para este fin se obtienen los siguientes datos:

Tiempo (minutos X)	27	45	41	19	35	39	19	49	15	31
Eficiencia (% Y)	47	84	80	46	62	72	52	87	37	68

El ingeniero encargado del estudio usa el software Minitab para procesar los datos anteriores y obtiene el siguiente reporte:

Regression Analysis: Eficiencia versus Tiempo

```
The regression equation is
Eficiencia = 18.1 + 1.42 Tiempo
Predictor    Coef  SE Coef     T      P
Constant   18.060    5.163  3.50  0.008
Tiempo     1.4200   0.1523  9.32  0.000
S = 5.38516   R-Sq = 91.6%   R-Sq(adj) = 90.5%
```

a. Interprete el valor de cada uno de los coeficientes de la recta anterior.
b. ¿Qué indica el valor del coeficiente de determinación?
c. ¿Puede usted afirmar que la pendiente de la recta es significativamente distinta de cero, con $(\alpha = 0.05)$?

9. El gerente de producción de una fábrica desea desarrollar un modelo con el fin de predecir el tiempo (Y) para realizar una tarea manual de montaje, basada en el tiempo de capacitación (X). A continuación se presentan las sumatorias correspondientes al tiempo de capacitación en horas y el tiempo de realización de la tarea en minutos. Correspondientes a una muestra aleatoria de 18 empleados.

$$\sum X = 354 \quad \sum Y = 288 \quad \sum X^2 = 7392 \quad \sum Y^2 = 4784 \quad \sum XY = 5833$$

a. Calcule e interprete los coeficientes de la ecuación de regresión.
b. ¿Es la pendiente de la recta significativamente distinta de 0? $\alpha = 0.05$.
c. Estime con 95% de confianza el tiempo de eficiencia promedio de los empleados que recibieron 20 horas de capacitación.
d. Calcule e interprete los coeficientes de correlación y de determinación.
e. Construya un intervalo para predecir el tiempo de eficiencia de un empleado que recibió 15 horas de capacitación.

10. Una consultora ha impartido una serie de cursos de dirección financiera para ejecutivos. Al finalizar dichos cursos se les pidió a los participantes que dieran una valoración de estos. Se estimó el siguiente modelo de regresión a partir de una muestra de 25 cursos (en paréntesis aparece la desviación estándar de cada uno de los coeficientes).

$$\hat{Y} = \underset{}{42.97} + \underset{(0,29)}{0.38X_1} + \underset{(0,21)}{0.52X_2} + \underset{(0,11)}{0.08X_3} + \underset{(3,59)}{6,21X_4} \qquad R^2 = 0.675$$

Donde:
Y : Valoración promedio de los participantes de dicho curso.
X_1: Porcentaje del tiempo que se dedicó al trabajo en grupo sobre la duración del curso.
X_2: Dinero invertido en material del curso, en dólares por miembro del grupo
X_3: Dinero invertido en comida y bebida, en dólares por miembro del grupo
X_4: Variable dummy que toma el valor 1 si hubo charla de un profesor invitado y 0 en otros casos.

a. Interpretar el coeficiente estimado asociado a X_4.
b. Probar la hipótesis nula de que el verdadero coeficiente de X_4 es 0, frente a la alternativa de que es positivo.
c. Interpretar el coeficiente de determinación.
d. Encontrar e interpretar un intervalo de confianza del 95% para el coeficiente de X_2.

11. La resistencia a la tensión de una fibra sintética se ve afectada por el tiempo de secado (X_1), la temperatura de secado (X_2) y el porcentaje de algodón en la fibra (X_3). En la siguiente tabla se presentan los datos de una muestra aleatoria de 11 observaciones:

Y	213	220	216	234	230	235	238	230	236	231	243
X_1	2.0	2.3	2.3	2.5	3.0	3.4	3.4	3.4	4.0	4.0	4.1
X_2	145	140	140	146	138	135	135	135	141	141	145
X_3	13	15	15	18	20	19	19	19	16	16	17

a. Ajuste un modelo de regresión múltiple a los datos.
b. ¿Cuál variable o cuáles variables son las más significativas en el modelo?
c. ¿Cuál será la resistencia de la fibra cuando $X_1 = 2.5$, $X_2 = 140$ y $X_3 = 16$?
d. ¿Qué le indica el valor del coeficiente de determinación?

12. A partir de una muestra de 27 recién graduados de una universidad se obtuvieron los siguientes datos sobre: la nota promedio de la carrera (Y), el número de horas a la semana dedicadas al estudio (X_1), el número promedio de horas dedicadas a la preparación de exámenes (X_2), el número de horas a la semana dedicadas a divertirse (X_3), si los estudiantes subrayaban el texto o tomaban nota a partir de este ($X_4 = 1$) si es "sí", 0 si "no") y el número medio efectivo de horas de clase que tuvieron por semestre (X_5). Los datos procesados se presentan a continuación (en paréntesis aparece el error estándar de cada coeficiente) $\alpha = 0.05$:

$$Y = 2.55 + 0.0071X_1 + 0.0227X_2 - 0.231X_3 + 0.277X_4 + 0.141X_5$$
$$(1.117)\ (0.0134)\ (0.04982)\ (0.05857)\ (0.2156)\ (0.06404)$$

s = 0.4454 R - sq = 50.0%

Análisis de varianza

Source	DF	SS	MS	F	P
Regression	5	4.1907	0.8381	4.21	0.008
Residual Error	21	4.1856	0.1993		
Total	26	8.3763			

a. Interprete el coeficiente estimado para X_4.
b. Pruebe la hipótesis nula de que el verdadero coeficiente de X_4 es 0, frente a la alternativa de que es positivo.
c. Interprete el coeficiente de determinación.
d. Encuentre e interprete un intervalo de confianza del 95% para el coeficiente de X_3.
e. ¿Cuál variable o cuáles variables no son significativas al modelo?
f. ¿A qué conclusión llega con el cuadro de análisis de varianza?
g. ¿Cómo interpretaría el coeficiente estimado de X_3?

13. Se sospecha que la potencia eléctrica consumida en el mes por una planta química (Y) está relacionada con la temperatura ambiente promedio (X_1), el número de días del mes (X_2), la pureza promedio del producto (X_3) y las to-

neladas producidas (X_4). Los datos correspondientes a los 12 últimos meses son los siguientes:

Y	240	236	290	274	301	316	300	296	267	276	288	261
X_1	25	31	45	60	65	72	80	84	75	60	50	38
X_2	24	21	24	25	25	26	25	25	24	25	25	23
X_3	91	90	88	87	91	94	87	86	88	91	90	89
X_4	100	95	110	88	94	99	97	96	110	105	100	98

Formule su modelo de regresión lineal múltiple y estime puntualmente y por intervalos el consumo de potencia para un mes en el que $X_1 = 75^{\circ} F$, $X_2 = 24$ días, $X_3 = 90\%$, $X_4 = 98$ toneladas. $\gamma = 0.975$.

14. La siguiente información se refiere a la demanda de PC, el precio (S/.) y el ingreso familiar (S/.).

Demanda	Precio	Ingreso
40	1.900	2.000
60	1.500	2.500
50	1.450	1.750
55	1.200	1.750
60	1.300	2.100
65	1.100	1.800
70	800	1.400
65	900	1.200
75	980	1.600
75	1.000	1.800

a. Encuentre la línea de regresión.
b. Establezca e interprete el coeficiente de determinación múltiple.
c. ¿Qué valor de demanda estima, si el precio de una PC es S/.1.150 y el ingreso familiar es de S/.1.800?
d. Contraste a un nivel de significancia del 5% la hipótesis de que, conjuntamente, estas dos variables no influyen en la demanda.

15. Simón Pérez, gerente comercial de una distribuidora, desea conocer el comportamiento de las ventas de las cámaras digitales. Simón considera que la publicidad y el precio son los factores determinantes de la demanda. La información recabada es la siguiente:

Demanda (unidades)	Publicidad (número de anuncios)	Precio ($)
33	3	125
61	6	115
70	10	140
82	13	130
17	9	145
24	6	140

a. Obtenga la ecuación de mínimos cuadrados ajustada para predecir la demanda sobre la base de la publicidad y el precio. Interprete los coeficientes de regresión.
b. Si el número de anuncios es 7 y el precio es $132, ¿cuál es su pronóstico de la demanda?

16. La consultora Magnum ha decidido realizar un estudio para una auditora, en la cual el estudio le va a permitir conocer la relación del monto mensual de los impuestos no pagados (Y) y las variables: horas de trabajo en auditoría de campo (X_1) y número de empresas evasoras (X_2). Para este fin, registra la información que se presenta en la tabla siguiente:

Y (millones de dólares)	30	23	28	30	27	29	32	34	33	26
X_1	320	310	290	280	270	500	470	480	400	260
X_2	71	70	72	75	70	76	80	82	80	66

a. ¿Es el modelo de regresión lineal múltiple apropiado para explicar la relación de Y versus X_1 y X_2?
b. ¿Cuál es la interpretación del coeficiente de regresión correspondiente a X_2?
c. Si en un mes determinado se dedican 350 horas de trabajo en auditoría y el número de empresas evasoras es 74, ¿entre qué valores se encontrará el promedio del monto mensual de los impuestos no pagados?
d. ¿Cuál(es) variable(s) independiente(s) no es(son) significativa(s)?
e. Calcule e interprete el coeficiente de determinación.
f. ¿Puede afirmarse que por cada empresa adicional evasora que se identifique, el monto mensual de impuestos no pagados aumentará en 800.000 dólares?

17. El contador de una empresa que alquila departamentos ha tomado una muestra aleatoria de inquilinos y presenta la siguiente tabla (anexada). De la información brindada por el contador se desea predecir el alquiler (en dólares por mes) basándose en el tamaño del departamento (número de habitaciones) y su distancia del centro de la ciudad (en kilómetros).

Renta ($)	Número de habitaciones	Distancia desde el centro de la ciudad
230	2	1
880	6	1
300	3	2
340	4	3
200	2	10
190	1	4

a. Calcule la ecuación de mínimos cuadrados que relacione mejor estas tres variables. Interprete los coeficientes de regresión.
b. Si alguien está buscando un departamento de dos habitaciones situado a dos kilómetros del centro de la ciudad, ¿qué renta pagaría aproximadamente?

18. Si se tienen los siguientes datos, use el software Minitab para calcular la ecuación de regresión del mejor ajuste y conteste lo siguiente:

a. ¿Cuál es la ecuación de regresión? ¿Cuál es el error estándar de estimación?
b. ¿Cuál es el R^2 de esta regresión?
c. Proponga un intervalo de confianza de aproximadamente 95% para el valor de Y cuando los valores de X_1, X_2, X_3 y X_4 son 115.6, 71.8, 93.4 y 0.7, respectivamente.

X_1	X_2	X_3	X_4	Y
112.4	92.6	91.2	-0.2	8.22
115.7	70.4	92.4	0.5	27.39
114.8	81.8	89.6	0.2	19.46
111.8	101.4	90.9	-0.4	2.91
117.6	62.2	92.1	0.8	38.55
119.9	51.6	90.3	1	70.32

19. Se desea predecir la demanda anual (unidades vendidas) de un artefacto utilizando las siguientes variables independientes:

Precio = precio del artefacto (P en dólares)
Ingreso = ingresos del consumidor (I en miles de dólares)
Sustituto = precio de una mercancía sustituta (S en dólares)

Demanda	P	I	S
65	18	1.100	17
75	25	1.200	22
75	25	1.300	19
80	25	1.400	20
100	23	1.500	23
90	24	1.600	18
95	23	1.700	24
85	24	1.800	21

a. Determine la ecuación de regresión del mejor ajuste para estos datos. Interprete los coeficientes de regresión.
b. ¿Corresponden a lo esperado los signos (+ o –) de los coeficientes de regresión de las variables independientes? Explique su respuesta brevemente.
c. Establezca e interprete el coeficiente de determinación para este problema.
d. Establezca e interprete el error estándar de estimación para este problema.
e. Aplicando la ecuación, ¿qué pronosticaría (puntual y por intervalo) usted para la variable demanda si el precio del artefacto fuera de $6, los ingresos del consumidor fueran de $1.200 y el precio de la mercancía sustituta fuera de $17?

20. Juan Castañeda está pensando en vender su casa; para fijar el precio que debe pedir ha reunido los datos de cinco cierres de ventas recientes en los cuales se considera: los precios de venta (en miles de dólares), las superficies de las casas (en metros cuadrados), el número de pisos, el número de baños y la antigüedad de la casa (en años).

Precios de venta	Metros cuadrados	Pisos	Baños	Edad
49.65	89	1.0	1.0	2
67.95	95	1.0	1.0	6
81.15	126	2.0	1.5	11
95.25	176	1.0	1.0	17
100.35	200	2.0	1.5	12

a. Determine la ecuación de regresión del mejor ajuste para estos datos. Interprete los coeficientes de regresión.
b. ¿Cuál es el R^2 de esta ecuación? ¿Qué mide este número?
c. Si la casa del señor Castañeda tiene 180 m^2, un piso, dos baños y seis años de antigüedad, ¿qué precio de venta puede pedir el señor Castañeda?

21. Una aerolínea cuya sede está en Lima ha realizado una encuesta al azar en sus siete oficinas, obteniendo los siguientes datos correspondientes al mes pasado.

Ventas ($)	Promoción ($)	Competidores	Gratis
79.3	2.5	10	3
200.1	5.5	8	6
163.2	6.0	12	9
200.1	7.9	7	16
146	5.2	8	15
177.7	7.6	12	9
30.9	2.0	12	8

Donde:
- Ventas: Ingreso total basado en el número de boletos vendidos (miles de dólares).
- Promoción: Cantidad gastada en la promoción de la aerolínea en el área (miles de dólares).
- Competidores: Número de aerolíneas competidoras en el área.
- Gratis: El porcentaje de pasajeros que viaja gratuitamente (por diversas razones).

a. Determine la ecuación de regresión del mejor ajuste. Interprete los coeficientes de regresión.
b. ¿Disminuyen significativamente las ventas si los pasajeros viajan gratis? Formule y pruebe la hipótesis apropiada. Utilice $\alpha = 0.05$.

c. ¿Con un incremento en las promociones de $1.000 cambian las ventas en $28.000, o es el cambio significativamente diferente de $28.000? Formule y pruebe las hipótesis apropiadas. Utilice $\alpha = 0.10$.
d. Construya un intervalo de confianza de 90% para el coeficiente de la variable competidores.

22. Los siguientes datos muestrales han sido obtenidos al azar de ocho hospitales de una gran ciudad. Un analista financiero intenta ver si existe una relación entre el número de camas que tiene un hospital, sus admisiones y gastos.

Hospital	Camas X_1	Admisiones de pacientes (miles) X_2	Gastos (miles de dólares) Y
N° 1	1082	31.0	268
N° 2	713	29.4	177
N° 3	489	13.4	101
N° 4	464	19.2	120
N° 5	451	16.3	108
N° 6	444	14.7	135

a. Ajuste una ecuación de regresión lineal múltiple a estos datos.
b. Utilice la ecuación obtenida en el inciso (a) para estimar los gastos de un hospital cuando el número de camas es igual a 650 y las admisiones son de 30.500 mil pacientes.

23. La compra y venta de agua mineral en la temporada es una oportunidad de negocios para muchos empresarios. Un estudio realizado en enero del año pasado en Lima Metropolitana basado en una muestra aleatoria de personas consumidoras de agua mineral permitió registrar los valores de las siguientes variables:

C1 = Edad de la persona consumidora de agua mineral (en años).
C2 = Ocupación.
C3 = Gasto semanal en agua mineral (en nuevos soles).
C4 = Consumo semanal de agua mineral (en litros).
C5 = Marca de agua mineral.

Los datos se encuentran en el archivo Agua Mineral.MTW.

Ejecute la regresión C4 (Y) versus C1(X_1) y C3(X_2) para responder las siguientes preguntas:

a. La señorita Ángeles, de 18 años de edad, gasta 15 nuevos soles semanalmente en agua mineral, estime la cantidad de litros de agua mineral que consume.
b. Obtenga el intervalo de confianza del 96% para el consumo promedio semanal de agua mineral para los consumidores de 20 años de edad que gastan semanalmente 16 nuevos soles en agua mineral.
c. El porcentaje de la variabilidad total explicada por el modelo es:_____.

24. El trabajo de Carlos, estudiante de ingeniería que labora para la empresa Libertad Asociados, consiste en analizar la información correspondiente de 102 egresados universitarios contratados por la compañía. La información de cada egresado contratado es la siguiente:

- X_1: Sexo (F = Femenino, M = Masculino).
- X_2: Edad del egresado (en años).
- X_3: Promedio ponderado acumulativo en la universidad.
- X_4: Categoría de la universidad donde estudió (1 = Excelente, 2 = Buena y 3 = Regular).
- X_5: Puntaje en una prueba de ingreso para el personal de la empresa.
- X_6: Puntaje en una prueba de rendimiento al segundo año. El resultado de esta prueba es una puntuación numérica que asume valores desde 0 (muy malo) hasta 100 (excelente). Suponga que una puntuación inferior a 50 representa un rendimiento insatisfactorio (I), 50-69 representa un rendimiento satisfactorio (S), 70-89 representa un rendimiento por encima del promedio (P), y más de 89 un rendimiento excelente (E).

El siguiente reporte de Minitab corresponde a los resultados de la regresión de los puntajes del rendimiento obtenidos en el segundo año (X_6) versus la edad (X_2), el promedio ponderado acumulativo en la universidad (X_3) y el puntaje obtenido en la prueba de ingreso para el personal de la empresa (X_5).

Regression Analysis: Puntaje segundo versus Edad, Promedio, Puntaje
La ecuación de regresión es:
Puntaje segundo = -59.3 + 0.106 Edad + 8.71 Promedio + 0.088 Puntaje

$s(\hat{\beta}_i)$	(0.3977)	(0.9063)	(0.1194)

Nota: Los valores entre paréntesis son las desviaciones estándar de cada coeficiente estimado.

Si se sabe además que: R-Sq = 67.9%

Colabore usted con Carlos en su análisis estadístico de la información de los 102 egresados respondiendo las siguientes preguntas.

Use $\alpha = 0.02$ en sus cálculos.

a. Complete la siguiente tabla:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	Valor F	P-value
Regresión					
Error					
Total		24421.00	--------	--------	--------

b. ¿Es el modelo en conjunto apropiado? Ejecute el procedimiento completo de la técnica estadística correspondiente.
c. Si se desea probar $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 > 0$, ¿cuál es el valor del P-value para esta prueba?

25. Un nuevo paquete de software se diseñó con el objetivo de lograr acceso y mantenimiento de los datos a gran escala. La eficiencia se mide en términos del número de operaciones de entrada/salida (E/S) de disco (llamadas bloques) necesarias para acceder al conjunto de datos y darle mantenimiento, cuanto menor sea el número de bloques leídos, con mayor rapidez se efectuará la operación. A fin de evaluar el nuevo software se registró el número de operaciones de E/S necesarias para acceder a un conjunto de datos a gran escala y el tamaño del conjunto (el tamaño se mide como el número de registros del conjunto). Los datos fueron los siguientes:

Número de registros X (miles)	Número de E/S a disco Y (miles)
350	36
200	20
450	45
50	5
400	40
150	18
350	38
300	32
150	21
500	54
100	11
400	43
200	19
50	7
250	26

a. Calcule e interprete, en términos del enunciado, los coeficientes de la recta de regresión empleada para estimar Y en función de X.
b. En términos del enunciado, ¿cómo interpreta usted los coeficientes de correlación y de determinación?
c. Con 95% de confianza, en cuánto estima el número de E/S de disco para un conjunto de datos que tiene 375 mil registros?
d. ¿Estaría usted en condiciones de afirmar que al aumentar X en una unidad Y aumenta más de 0.12? Haga la comprobación con $\alpha = 0.04$. Formule y pruebe las hipótesis.
e. ¿Es el modelo lineal adecuado para hacer estimaciones sobre la variable Y? Formule las hipótesis correspondientes. Use $\alpha = 0.04$.

26. En la tabla adjunta se presentan el número de páginas y el precio, en dólares, de doce libros técnicos:

páginas	310	300	280	310
precio	35	35	32	73
páginas	400	170	430	230
precio	80	18	70	32
páginas	420	610	420	450
precio	25	50	54	37

a. Calcule e interprete los coeficientes de la recta empleada para estimar el precio del libro en función del número de páginas.
b. Interprete el valor del coeficiente de determinación.
c. Estime con 95% de confianza el precio promedio de los libros que tienen 350 páginas.
d. ¿Es el coeficiente de regresión del número de páginas significativamente distinto de cero? Use $\alpha = 0.05$.
e. ¿Es razonable afirmar que el modelo planteado es adecuado en su conjunto? Use $\alpha = 0.05$.

27. Se está estudiando el efecto de cierta fricción constante sobre el grosor de láminas metálicas de la misma aleación. Se han efectuado 10 mediciones cuyos resultados son:

X	3	5	6	8	9	10	11	12	14	15
Y	9.52	9.06	8.81	8.54	8.28	8.03	7.82	7.6	7.35	7.06

X: tiempo de fricción (hrs)
Y: grosor de la lámina (mm)

a. Calcule e interprete, en términos del enunciado, los coeficientes de la recta de regresión.
b. ¿Cómo interpreta, en términos del enunciado, los coeficientes de correlación y de determinación?
c. Construya un intervalo de confianza del 96% para la pendiente de la recta.
d. Con 99% de confianza, estime (puntualmente y por intervalo) el grosor promedio de la lámina cuando el tiempo de fricción es de 13 horas.
e. ¿Es el modelo lineal, adecuado para estimar el grosor de la lámina? Use $\alpha = 0.01$. Indique la región crítica y la estadística de prueba.

28. A continuación se ofrecen los resultados de un estudio que se realizó en una empresa. En este estudio se extrajo una muestra de 10 empleados, de cada empleado se registró el número de unidades por hora que había producido (Y). También se registraron los resultados que obtuvo cada empleado en una prueba de aptitud (X_1). Los resultados del estudio son los siguientes:

```
The regression equation is Y = - 7.50 + 0.244 X1
Predictor       Coef      Stdev   t-ratio       p
Constant      -7.495      8.576     -0.87   0.408
X1           0.24369    0.06154      3.96   0.004
s = 6.415       R-sq = 66.2%
```

a. ¿Cómo interpreta los coeficientes de la ecuación de regresión y el coeficiente de determinación?
b. Suponga que en el estudio anterior se añade una segunda variable, años de experiencia (X_2).

   Ahora los resultados son:

```
The regression equation is  Y = - 13.8 + 0.212 X1 + 2.00 X2
Predictor     Coef      Stdev     t-ratio       p
Constant   -13.825      1.795       -7.70   0.000
      X1   0.21217    0.01266       16.76   0.000
      X2    1.9995     0.1456       13.73   0.000
s = 1.298    R-sq = 98.8%
```

   b.1 ¿Cómo interpreta los coeficientes de la ecuación de regresión?
   b.2 ¿Qué efecto tiene el haber añadido X_2 sobre la producción?

Capítulo

5 *Diseño de experimentos*

En este capítulo trataremos los siguientes temas:

- *Definiciónes básicas*
- *Tipos de variabilidad*
- *Etapas de un diseño de experimento*
- *Definiciones importantes*
- *Principios básicos de un diseño experimental*
- *Tipos de diseños experimentales*
- *Diseño completamente aleatorio*

Después de realizar una breve introducción al diseño de experimentos se explican los conceptos de tipos de variabilidad, las etapas de un diseño de experimento, las definiciones y los principios básicos de un diseño experimental. Por otro lado, se señalan los principales tipos de diseños experimentales, sobre todo el diseño completamente aleatorio, desarrollado con la ayuda de una diversidad de ejercicios solucionados y propuestos. La herramienta informática utilizada en este capítulo es el Minitab.

1. Introducción

El diseño de experimentos tuvo sus inicios en 1919, con Sir Ronald A. Fisher, quien desarrolló las bases de la teoría de diseños experimentales empleando conceptos como bloqueo, aleatorización, replicación, experimentación factorial y confusión, entre otros.

El objetivo de Fisher fue explicar la variabilidad de los resultados que obtenía en sus diferentes experimentos en la estación agrícola de *Rothamsted Experimental Station* en Inglaterra. En 1935 Fisher publicó su obra *The design of experiments*.

Estas ideas fueron rápidamente aplicadas a otras áreas, como la medicina, la biología, las ciencias sociales, la industria, etcétera.

En los últimos años, la teoría y las aplicaciones del diseño de experimentos se han consolidado y expandido, permitiendo que las personas tengan una mejor calidad de vida con los diversos productos y servicios que ofrecen las empresas.

2. Definición

Un diseño experimental es un método estructurado y organizado para analizar los cambios deliberados en las variables de entrada de un proceso o sistema, de tal manera que sea posible observar e identificar las causas de los cambios que se producen en la respuesta de salida.

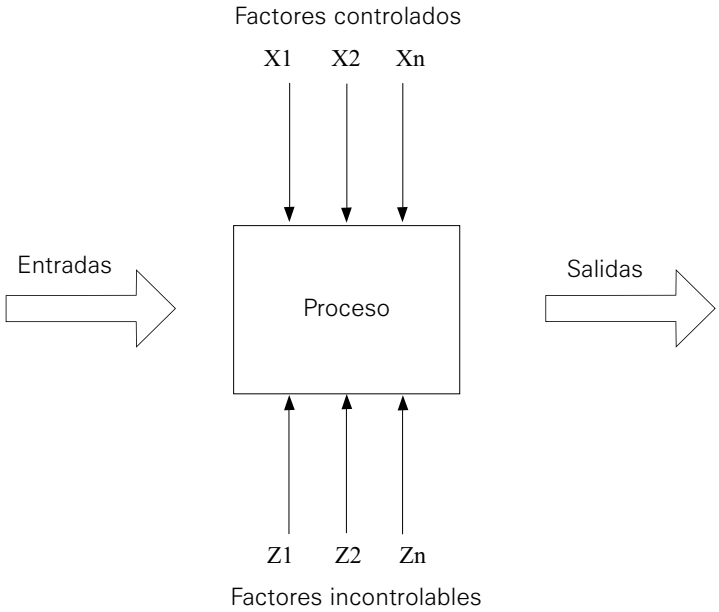


Figura 1. Esquema de un proceso o sistema.

Aunque los diseños experimentales difieren unos de otros en muchos aspectos, existen diseños que se utilizan con mucha frecuencia. Algunos de los más usados son los siguientes: diseños completamente aleatorios, diseños en bloques o con un factor bloque, diseños con dos o más factores bloque. Los modelos de diseños de experimentos tienen como objetivo determinar si uno o más factores influyen en la variable de interés.

Ejemplo 1:

Analizar el rendimiento de los alumnos de una determinada asignatura, para lo cual se desea estudiar el efecto de diferentes factores, como el profesor que imparte la asignatura, el método de enseñanza usado, la edad del alumno, el género del alumno, etcétera.

Ejemplo 2:

Una empresa de telecomunicaciones está interesada en conocer la influencia de varios factores en la variable tiempo de duración de una llamada telefónica; los factores en este estudio son: día de la semana en que se realiza la llamada, ciudad de la que se realiza la llamada; hora en que se produce la llamada; tipo de teléfono (fijo, móvil) que se utiliza y edad de la persona que hace la llamada.

La metodología del diseño de experimentos se basa en la experimentación. La experimentación se realiza en un laboratorio, donde la mayoría de las causas de variabilidad están controladas con el fin de minimizar el error experimental.

3. Tipos de variabilidad

Existen tres tipos de variabilidad: variabilidad sistemática, variabilidad típica y variabilidad sistemática no planificada.

- Variabilidad sistemática. La variabilidad de los resultados es debida a diferencias estructurales que son impuestas en el diseño por el experimentador. Por ejemplo, cinco diferentes marcas de jarabes para la tos.
- Variabilidad típica. Se le conoce con el nombre de ruido aleatorio, se denomina también error de medida; es una variable no controlable.

  Esta variabilidad es inevitable, pero si el experimento ha sido bien planificado es posible estimar o medir su valor, lo que es de gran importancia para obtener conclusiones y realizar predicciones.

  Por ejemplo, el tiempo de recuperación de una persona que padece tos no depende únicamente del jarabe, pues hay otras variables que influyen, como la edad, el peso, etcétera.
- Variabilidad sistemática no planificada. Se debe a causas desconocidas, los resultados están siendo sesgados sistemáticamente por causas desconocidas. La presencia de esta variabilidad supone la principal causa de conclusiones erróneas y estudios incorrectos al ajustar un modelo estadístico. Por ejemplo: cuando existen factores que dependen de otros factores o de las interrelaciones entre estos.

4. Etapas de un diseño de experimento

Las etapas de un diseño de experimento son las siguientes:

- Planteamiento del problema. En esta etapa se deben definir claramente los objetivos que persigue el experimento.
- Identificación de todas las posibles fuentes de variación, incluyendo la unidad experimental, los factores y el tratamiento y sus niveles.
- Elección de las unidades experimentales según las condiciones de estudio.
- Especificación de la medida de la variable respuesta y el procedimiento a utilizar para registrar los valores.
- Elección del diseño experimental y realización del experimento. La elección del diseño depende de las unidades experimentales y de la precisión deseada, una vez elegido el diseño se lleva a cabo el experimento de acuerdo al diseño seleccionado.
- Análisis de los datos. Cualquier diseño experimental implica un modelo lineal que refleje todos los factores o variables considerados en el diseño. Al aplicar la técnica estadística correspondiente a este modelo para evaluar las posibles fuentes de variación y la descomposición de la variabilidad total en sus partes componentes, de acuerdo con el modelo lineal en estudio, se determina qué efectos son significativos.
- Conclusiones. Después de haber analizado los datos, el experimentador debe formular conclusiones, recomendaciones y sugerencias.

5. Definiciones importantes

A continuación se presentan algunas definiciones importantes para el diseño de experimentos.

5.1. Unidad experimental

Es el sujeto u objeto, intervalo de espacio o tiempo, sobre el que se experimenta o se aplica el tratamiento.

Ejemplo 3:

En el campo industrial: el trabajador, una máquina, un lote de material, etcétera.
En el campo informático: una computadora, una página web, un buscador de internet, etcétera.
En medicina: un paciente, un centro hospitalario, etcétera.

5.2 Factor

Es una variable independiente de interés del experimentador, en la cual se desea estudiar su efecto sobre la variable respuesta. En la gran mayoría de las investigaciones de tipo cualitativo o cuantitativo, se trabaja con más de una variable independiente.

- Factor cualitativo:
  - Tipos de variedad de trigo que se desea comparar.
  - Tipos de dietas de animales.
  - Marcas de fármacos utilizados para el tratamiento de una enfermedad.
- Factor cuantitativo:
  - Cantidad de nutrientes de un tipo de alimento en diferentes cantidades.
  - Cantidad de megabytes de memoria en las computadoras.
  - Niveles de temperatura de una variable de interés.

5.3 Niveles de un factor

Son los diferentes tipos o grados específicos del factor que se tendrán en cuenta en la realización del experimento. Los niveles de un factor recibieron el nombre de "tratamientos".

Ejemplo 4:

- En variedades de trigo:
  Niveles: Buck Ponch, Buck Mataco, Klein estrella.
- En fármacos para la relajación muscular:
  Niveles: Innovar, Droperidol, Fentanyl.
- En temperatura:
  Niveles: 30°C, 50°C, 70°C.

5.4 Tratamientos

Un tratamiento es un efecto que se desea estudiar. Implica el nivel particular de un factor que deben imponerse a una unidad experimental dentro del marco del diseño seleccionado.

Ejemplo 5:

- Utilizar en el cultivo de trigo la variedad Buck Ponch.
- Utilizar el Droperidol como fármaco para la relajación muscular.

6. Principios básicos de un diseño experimental

Existen tres principios básicos:

- Repetición del experimento.
- Aleatoriedad.
- Formación de bloques.

6.1 Repetición del experimento

Este principio tiene varias implicancias, una de ellas suministra la estimación del error experimental e incrementa la precisión reduciendo los errores estándares. La repetición del experimento permite observar el comportamiento de la variable en estudio en el tiempo pero bajo las mismas condiciones de experimentación.

6.2 Aleatoriedad

Es la asignación aleatoria de los tratamientos a las unidades experimentales, una suposición frecuente es que las observaciones o errores en ellas son independientes, la aleatoriedad hace válido este supuesto.

6.3. Formación de bloques

Un bloque es un conjunto de unidades experimentales lo más homogéneas posible. También se dice que es la forma de asignar los tratamientos a las unidades experimentales de modo que las observaciones realizadas en cada bloque se hagan bajo condiciones experimentales lo más parecidas posible.

Cuando un bloque recibe todos los tratamientos se denomina bloque completo; en caso contrario se denomina bloque incompleto.

7. Tipos de diseños experimentales

Los más utilizados son los siguientes:

- Diseño completamente aleatorio.
- Diseño en bloques o con un factor bloque.
- Diseño cuadrado latino.
- Experimentos factoriales.

En este capítulo solo se desarrollará el diseño completamente aleatorio o DCA. A continuación se presenta una breve explicación de cada uno de los tipos de diseños experimentales.

7.1 Diseño completamente aleatorio

Un diseño aleatorio es aquel en el que los tratamientos o factores son asignados aleatoriamente a las unidades experimentales. El modelo matemático de este diseño tiene la forma:

$$\text{Respuesta} = \text{Constante} + \text{Efecto Tratamiento} + \text{Error}$$

7.2 Diseño en bloques o con un factor bloque

En este diseño el experimentador agrupa las unidades experimentales en bloques, a continuación se determina la distribución de los tratamientos en cada bloque y, por último, se asignan al azar las unidades experimentales a los tratamientos dentro de cada bloque. El modelo matemático de este diseño es:

$$\text{Respuesta} = \text{Constante} + \text{Efecto Bloque} + \text{Efecto Tratamiento} + \text{Error}$$

El diseño en bloques más simple es el denominado *diseño en bloques completos*, en el que en cada tratamiento se observa el mismo número de veces en cada bloque.

El diseño en bloques completos con una única observación por cada tratamiento se denomina diseño en bloques completamente aleatorizado o diseño en bloques aleatorizado.

Cuando el tamaño del bloque es inferior al número de tratamientos no es posible observar la totalidad de tratamientos en cada bloque y se habla entonces de diseño en bloques incompletos.

7.3 Diseño cuadrado latino

En ocasiones hay dos o más fuentes de variación lo suficientemente importantes como para ser designadas factores de bloqueo. En tal caso, ambos factores bloque pueden ser cruzados o anidados.

Los factores bloque están cruzados cuando existen unidades experimentales en todas las combinaciones posibles de los niveles de los factores bloques.

8. Diseño completamente aleatorio

Este diseño también es denominado modelo de clasificación en un solo sentido y permite comparar más de dos tratamientos o niveles de un factor único.

El modelo matemático de este diseño es: $Y_{ij} = \mu + \tau_i + e_{ij}$

Si $\mu_i = \mu + \tau_i$, el modelo puede expresarse como:

$$Y_{ij} = \mu_i + e_{ij}$$

Donde:

Y_{ij} : Es la j-ésima respuesta correspondiente al efecto del i-ésimo tratamiento.

μ : Es una constante común a todas las observaciones, denominado media global o promedio general.

τ_i : Efecto del i-ésimo tratamiento.

e_{ij}: El componente aleatorio asociado al modelo.

μ_i : Representa la media del i-ésimo tratamiento.

La idea básica en este modelo es que cada tratamiento es asignado al azar a cada unidad experimental. Las suposiciones son:

a. Tiene una distribución de probabilidad normal.
b. Cada distribución de probabilidad tiene la misma varianza.
c. Las observaciones o la respuesta que se observa en cada uno de los tratamientos son variables aleatorias independientes.

Lo que se desea es probar la igualdad de las medias μ_i de los i tratamientos, para ello se supone que los errores experimentales son variables aleatorias independientes con distribución normal con media cero y varianza constante σ^2. El modelo descrito conjuntamente con las suposiciones anteriores puede ser:

- **Modelo de efectos fijos**. Son aquellos en los cuales los tratamientos son seleccionados por el experimentador, en esta situación se desea probar hipótesis respecto a las medias de los tratamientos. Las conclusiones solo se aplican a los tratamientos seleccionados.
- **Modelo de efectos aleatorios**. Los tratamientos considerados forman parte de una muestra aleatoria y se desean generalizar las conclusiones a todos los tratamientos de la población; en este caso los efectos (τ_i) son variables aleatorias y en esta situación se prueban hipótesis con referencia a la variabilidad de estos efectos.

En este libro se desarrolla únicamente el modelo de efectos fijos.

8.1 Modelo de efectos fijos

El modelo de efectos fijos es el siguiente: $Y_{ij} = \mu + \tau_i + e_{ij}$

Donde: $i = 1,\ldots,a$ $\quad j = 1,\ldots,n$ $\quad N = an$

Nota: Obsérvese la diferencia entre N y n. N representa el total de los datos mientras que n representa el número de repeticiones por tratamiento.

El objetivo principal es la prueba de hipótesis de las medias de los tratamientos, considerándose que el número de repeticiones por tratamiento es igual para todos.

Procedimiento de la prueba de hipótesis para las medias de los tratamientos:

a. Formulación de las hipótesis.

$H_0 : \mu_1 = \mu_2 = ... = \mu_a$

$H_1 :$ Al menos un $\mu_i \neq \mu_j$ donde $i \neq j$

Si la hipótesis nula es cierta, todos los tratamientos tienen la misma media común μ constante.

Otra forma de expresar las hipótesis anteriores en términos de los efectos de los tratamientos es:

$H_0 : \tau_1 = \tau_2 = ... = \tau_a = 0$

$H_1 :$ al menos un $\tau_i \neq 0.$

Es igual probar la igualdad de medias de tratamientos o probar que los efectos de los tratamientos son cero.

b. Selección del nivel de significación α.
c. Cálculo de la estadística de prueba utilizando la tabla 1 de análisis de varianza.

Tabla 1 Tabla del análisis de varianza.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medio	F_0	P-value
Debido a los tratamientos	$SCTrat = \sum_{i=1}^{a} \frac{y_{i\square}^2}{n} - \frac{y_{\square\square}^2}{N}$	$a-1$	$CMTrat = \frac{SCTrat}{a-1}$	$F = \frac{CMR}{CME}$	$P(F > F_0)$
Debido al error	$SCE = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\square})^2$	$a(n-1)$	$CME = \frac{SCE}{a(n-1)}$		
Total	$SCT = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\square\square}^2}{N}$	$an-1$			

Donde:

$y_{i\square} = \sum_{j=1}^{n} y_{ij}$: Total de observaciones del i-ésimo tratamiento

$\bar{y}_{i\square} = \frac{\sum_{j=1}^{n} y_{ij}}{n}$: Promedio de las observaciones bajo el i-ésimo tratamiento

$y_{\square} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}$: Suma total de observaciones

$\bar{y}_{\square} = \frac{y_{\square}}{N} = \frac{\sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}}{N}$: Promedio general

d. Determinación de la región crítica; es decir, el conjunto de valores de la distribución muestral que indican que la hipótesis nula debe ser rechazada.
e. Conclusión. Si la estadística de prueba es mayor que $F_{(1-\alpha, a-1, a(n-1))}$ se rechaza la hipótesis nula; en caso contrario no se rechaza.

Ejemplo 7:

Una nutricionista ha realizado un estudio con tres tipos de dieta, para lo cual seleccionó 18 pacientes a los que se les asignó un tipo de dieta y, después de un mes, se registraron las pérdidas de peso en cada paciente (véase la tabla 2). Al nivel de significación de 0.05, ¿se puede concluir que existe una diferencia en el promedio de pérdida de peso de las tres dietas?

Tabla 2. Datos de pérdidas de peso del ejemplo 7

Dieta 1	Dieta 2	Dieta 3
5	2	6
6	3	5
4	3	4
7	4	3
8	3	5
6	2	6

Solución:

Modelo estadístico: $Y_{ij} = \mu + \tau_i + e_{ij} \quad i = 1,\ldots,3 \quad j = 1,\ldots,6 \quad N = 18$

Donde: Y_{ij} : Pérdida de peso en kilogramos registrado usando el i-ésimo tipo de dieta, en el j-ésimo paciente.

a. Formulación de las hipótesis

$H_0 : \mu_1 = \mu_2 = \mu_3$

H_1 : Al menos un $\mu_i \neq \mu_j$ donde $i \neq j$

b. Nivel de significación: $\alpha = 0.05$
c. Obtención de la estadística de prueba por usar mediante la siguiente expresión:

$F = \frac{CMTrat}{CME} \quad \square \; F_{(2,15)} \big/ H_0$ es verdadera

Se procede a construir la tabla Anova (tabla de análisis de varianza), para lo cual se obtienen las siguientes sumas de cuadrados:

Suma de cuadrados entre tratamientos, cuya fórmula es la siguiente:

$\sum_{i=1}^{a} \frac{y_{i\square}^2}{n} - \frac{y_{\square}^2}{N}$

Desarrollando cada expresión de la fórmula:

$$y_{1\square} = \sum_{j=1}^{6} y_{1j} = 5+6+...+6 = 36 \; ; \quad y_{1\square}^2 = 1296$$

$$y_{2\square} = \sum_{j=1}^{6} y_{2j} = 2+3+...+2 = 17 \; ; \quad y_{2\square}^2 = 289$$

$$y_{3\square} = \sum_{j=1}^{6} y_{3j} = 6+5+...+6 = 29 \; ; \quad y_{3\square}^2 = 841$$

$$\sum_{i=1}^{a} \frac{y_{i\square}^2}{n} = \frac{1296+289+841}{6} = 404.333$$

$$\frac{y_{\square\square}^2}{N} = \frac{(36+17+29)^2}{18} = \frac{6724}{18} = 373.556$$

Reemplazando en la fórmula se tiene:

$$\sum_{i=1}^{a} \frac{y_{i\square}^2}{n} - \frac{y_{\square\square}^2}{N} = 404.333 - 373.556 = 30.777$$

Calculando la suma de cuadrados del total (SCT):

$$\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\square\square}^2}{N}$$

$$\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 = 5^2+6^2+...+6^2+2^2+3^2+...+2^2+6^2+5^2+...+6^2 = 424$$

Reemplazando en la fórmula: $SCT = 424 + 373.556 = 50.444$

Suma de cuadrados del error (SCE), que se obtiene mediante:

$$SCE = SCT - SCTrat$$

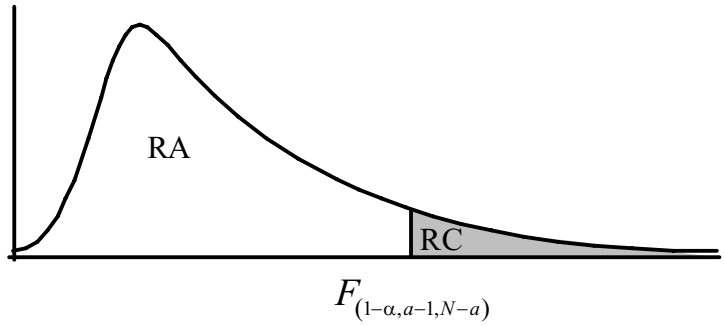
Suma de cuadrados del error $(SCE) = \left(\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\square\square}^2}{N}\right) - \left(\sum_{i=1}^{a} \frac{y_{i\square}^2}{n} - \frac{y_{\square\square}^2}{N}\right)$

Por consiguiente la $SCE = 50.444 - 30.777 = 19.667$

Reemplazando los valores calculados en la tabla del Anova se tiene:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medio	F_0	P-value
Debido a los tratamientos	30.777	2	$\frac{30.777}{2} = 15.3887$	$\frac{15.3887}{1.111} = 13.85$	$P(F > 13.85) = 0.000391$
Debido al error	19.6667	15	$\frac{19.667}{15} = 1.111$		
Total	50.444	17			

d. Obtención del valor crítico.



Haciendo uso del software Minitab

Inverse Cumulative Distribution Function

```
F distribution with 2 DF in numerator and 15 DF in
denominator
P( X <= x )          x
        0.95   3.68232
```

e. Regla de decisión.

Como el valor del estadístico de prueba $F_0 = 13.85$ es mayor que el valor crítico $F_c = F_{(0.95,2,15)} = 3.68232$ se rechaza la hipótesis nula, concluyendo que hay diferencia de pérdida en el peso promedio de los tres tipos de dietas. Es decir, las dietas en estudio producen diferentes pérdidas de peso promedio.

8.2 Estimación de los parámetros del modelo

Se utiliza el método de mínimos cuadrados, que consiste en minimizar la suma de cuadrados de los errores respecto a μ y los efectos fijos τ_i y se obtiene lo siguiente:

$$\hat{\mu} = \bar{y}_{\square} \qquad \hat{\mu}_i = \bar{y}_{i\square} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} \qquad \hat{\tau}_i = \bar{y}_{i\square} - \bar{y}_{\square}$$

Estos estimadores son insesgados y de varianza mínima.

Ejemplo 8:

En el caso anterior, estime la media general y los efectos de los tratamientos.

Media general: $\hat{\mu} = \bar{y}_{\square} = \frac{82}{18} = 4.556$

Efectos de los tratamientos:

$$\hat{\tau}_1 = \bar{y}_{1\square} - \bar{y}_{\square\square} = \frac{36}{6} - 4.556 = 1.444$$

$$\hat{\tau}_2 = \bar{y}_{2\square} - \bar{y}_{\square\square} = \frac{17}{6} - 4.556 = -1.723$$

$$\hat{\tau}_3 = \bar{y}_{3\square} - \bar{y}_{\square\square} = \frac{29}{6} - 4.556 = 0.277$$

Como puede apreciarse en los resultados, las dietas 1 y 3 producen una pérdida de peso mayor que la dieta 2.

8.3. Intervalo de confianza para los parámetros del modelo

Los intervalos de confianza para los parámetros son:

a. El intervalo de confianza para μ_i con un nivel de confianza $(1 - \alpha)$ es :

$$\mu_i \in \left\langle \bar{y}_{i\square} \pm t_{\left(1-\alpha/2,\, a(n-1)\right)} \sqrt{\frac{CMError}{n}} \right\rangle$$

Ejemplo 9:

Utilizando los datos del ejemplo 7, determine el intervalo del 95% de confianza para la media de pérdida de peso de la dieta 1.

$$\left\langle \bar{y}_{1\square} \pm t_{\left(1-\alpha/2,\, a(n-1)\right)} \sqrt{\frac{CMError}{n}} \right\rangle = \left\langle 6 \pm t_{(0.975,15)} \sqrt{\frac{1.111}{6}} \right\rangle = \langle 6 \pm 0.917 \rangle = \langle 5.083; 6.917 \rangle$$

b. El intervalo de confianza para $\mu_i - \mu_j$ con nivel de confianza $(1 - \alpha)$ es:

Si el número de repeticiones de cada tratamiento es igual:

$$(\mu_i - \mu_k) \in \left\langle \bar{y}_{i\square} - \bar{y}_{k\square} \pm t_{\left(1-\alpha/2,\, a(n-1)\right)} \sqrt{\frac{2CMError}{n}} \right\rangle$$

Si el número de repeticiones de cada tratamiento es diferente.

$$(\mu_i - \mu_k) \in \left\langle \bar{y}_{i\square} - \bar{y}_{k\square} \pm t_{\left(1-\alpha/2,\, a(n-1)\right)} \sqrt{CMError\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \right\rangle$$

Ejemplo 10:

Continuando con el ejemplo 7, ¿existe diferencia en el promedio de pérdida de peso de la dieta 1 con respecto al promedio de pérdida de peso de la dieta 2, con un 95% de confianza?

Se utiliza la fórmula: $(\mu_1 - \mu_2) \in \left\langle \bar{y}_{1\square} - \bar{y}_{2\square} \pm t_{\left(1-\alpha/2,\, a(n-1)\right)} \sqrt{\frac{2CMError}{n}} \right\rangle$

Porque el número de repeticiones $n = 6$ es el mismo para las dietas 1 y 2

$$\left\langle (6-2.83) \pm 2.13145\sqrt{\frac{2(1.111)}{6}} \right\rangle = \langle 3.17 \pm 1.297 \rangle = \langle 1.873, 4.467 \rangle$$

El intervalo $\langle 1.873; 4.467 \rangle$ indica que hay diferencias significativas respecto a la pérdida de peso promedio en ambos tipos de dieta. La pérdida de peso con la dieta 1 es mayor que con la dieta 2 con 95% de confianza.

PROBLEMAS RESUELTOS

1. Un ingeniero electrónico está interesado en el efecto de cinco diferentes tipos de recubrimiento en la conductividad en tubos de ensayo catódicos utilizados en un dispositivo de telecomunicaciones. Se obtuvieron los siguientes datos de conductividad:

Tipo de recubrimiento	Conductividad			
1	143	141	150	146
2	152	149	137	143
3	134	133	132	127
4	129	127	132	129
5	147	148	144	142

a. Identifique la unidad experimental. ¿Hay alguna diferencia en la conductividad debida a los diferentes tipos de recubrimiento?

Solución:

Unidad experimental: El tubo de ensayo catódico.
Modelo estadístico:

$Y_{ij} = \mu + \tau_i + e_{ij}$, donde: $i = 1,\ldots,5$ $\quad j = 1,\ldots,4$ $\quad a = 5$
$n = 4$ $\quad N = 20$

donde:
Y_{ij} : Conductividad usando el i-ésimo tipo de recubrimiento en el j-ésimo tubo de ensayo.

- Formulación de hipótesis:
  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
  H_1 : Al menos un $\mu_i \neq \mu_j$ donde $i \neq j$
- Nivel de significación: $\alpha = 0.05$
- Cálculo de la estadística de prueba: $F_0 = \dfrac{CMTrat}{CME}$

  Suma de cuadrados entre tratamientos: $SCTrat = \sum_{i=1}^{a} \dfrac{y_{i\square}^2}{n} - \dfrac{y_{\square\square}^2}{N}$

$$SCTrat = \left(\frac{(143+\ldots146)^2+\ldots+(147+\ldots142)^2}{4} \right) -$$

$$\left(\frac{(143+\ldots146+\ldots+147+\ldots+142)^2}{20} \right)$$

$$SCTrat = (388871.75 - 387811.25) = 1060.5$$

Suma de cuadrados del total: $SCT = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\square}^2}{N}$

$$SCT = (143^2+\ldots+142^2) - \left(\frac{(143+\ldots+142)^2}{20} \right) = (389115 - 387811.25) = 1303.75$$

Suma de cuadrados del error: $SCE = \sum_{i=1}^{a}\sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{i\square} \right)^2$ ó

$$SCE = SCT - SCTrat$$

$$SCE = (1303.75 - 1060.5) = 243.25$$

Obteniendo los cuadrados medios:

Cuadrado medio entre tratamientos: $CMTrat = \frac{SCTrat}{a-1}$

Reemplazando valores: $CMTrat = \frac{1060.5}{5-1} = 265.125$

Cuadrado medio del error: $CME = \frac{SCE}{a(n-1)}$

Reemplazando valores: $CME = \frac{243.25}{5(4-1)} = \frac{243.25}{15} = 16.216667$

Por consiguiente, el valor de la estadística de prueba es:

$$F_0 = \frac{CMTrat}{CME} = \frac{265.125}{16.216667} = 16.348921$$

- Obteniendo el valor crítico:

  $F_{(1-\alpha,a-1,N-a)} = F_{(1-0.05,5-1,20-5)} = F_{(0.95,4,15)} = 3.05557$

  Haciendo uso del software Minitab

  Inverse Cumulative Distribution Function

  ```
  F distribution with 4 DF in numerator and 15 DF in
  denominator
  P( X <= x )          x
          0.95   3.05557
  ```

- Cálculo del P-value.

  $P(F > 16.3489221) = 0.000024$

- Conclusión: Como $F_0 > F_{(1-\alpha,a-1,N-a)}$ .es decir $16.348921 > 3.05557$, se rechaza la hipótesis nula; se puede afirmar entonces que hay diferencia en la conductividad debido al tipo de recubrimiento, con un nivel de confianza del 95%.

Desarrollando con el software Minitab

En el menú Stat, tal como se aprecia en la figura 2, seleccione Anova / One-Way (Unstacked); se elige esta opción porque la información se encuentra en varias columnas. Seleccione las columnas del visor, luego en <Ok>.

El resultado se muestra a continuación:

One-way Anova: R1, R2, R3, R4, R5

```
Source      DF       SS      MS       F      P
Factor       4   1060.5   265.1   16.35  0.000
Error       15    243.3    16.2
Total       19   1303.8
S = 4.027    R-Sq = 81.34%    R-Sq(adj) = 76.37%
```

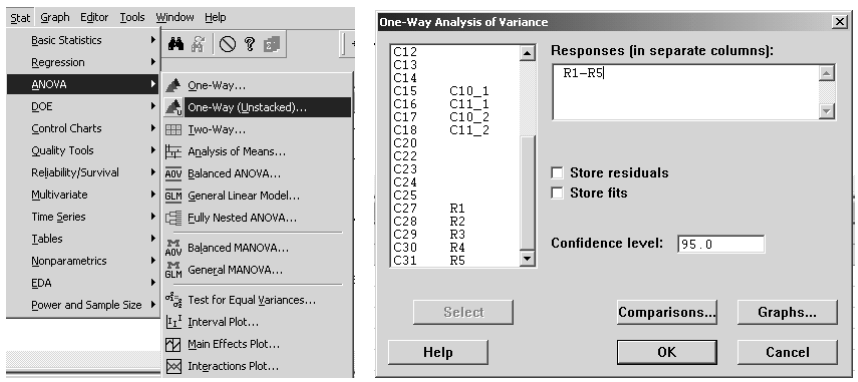


Figura 2. Secuencia para análisis de varianza e ingreso de datos.

Como el valor crítico $F_0 = 16.35$ tiene un P-value aproximadamente igual a cero, se rechaza la hipótesis nula. Se concluye, entonces, que existen diferencias en la conductividad debido al tipo de recubrimiento.

b. Estimación de la media general y los efectos de los tratamientos.

Solución:

Estimación de la media general: $\hat{\mu} = \overline{y}_{\square\square}$
Obtención del promedio:

$$\overline{y}_{\square\square} = \frac{y_{\square\square}}{N} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}}{N} = \frac{(143+141+\ldots+144+142)}{20} = \frac{2785}{20} = 139.25$$

Estimación de los efectos para cada uno de los tratamientos: Como: $\hat{\tau}_i = \overline{y}_{i\square} - \overline{y}_{\square\square}$

$$\hat{\tau}_1 = \bar{y}_{1\square} - \bar{y}_{\square\square} = \frac{(143+\ldots+146)}{4} - 139.25 = 145 - 139.25 = 5.75$$

$$\hat{\tau}_2 = \bar{y}_{2\square} - \bar{y}_{\square\square} = \frac{(152+\ldots+143)}{4} - 139.25 = 145.25 - 139.25 = 6$$

$$\hat{\tau}_3 = \bar{y}_{3\square} - \bar{y}_{\square\square} = \frac{(134+\ldots+127)}{4} - 139.25 = 131.5 - 139.25 = -7.75$$

$$\hat{\tau}_4 = \bar{y}_{4\square} - \bar{y}_{\square\square} = \frac{(129+\ldots+129)}{4} - 139.25 = 129.25 - 139.25 = -10$$

$$\hat{\tau}_5 = \bar{y}_{5\square} - \bar{y}_{\square\square} = \frac{(147+\ldots+142)}{4} - 139.25 = 145.25 - 139.25 = 6$$

Como puede apreciarse, los tipos de recubrimiento 1, 2 y 5 tienen una conductividad superior al promedio, mientras que el tercero y el cuarto tipos de recubrimiento tienen una conductividad menor.

Desarrollando con el software Minitab

El Minitab presenta una opción donde se guarda la media y los valores de los efectos de los tratamientos excepto 1; en este caso, se puede utilizar la opción "Modelo General". Para ello la información debe estar en dos columnas: en una de ellas se considera la variable respuesta y en la otra la identificación del nivel.

En el ejemplo considerado, la información original está en cinco columnas, para colocar en dos columnas como se requiere, se ingresa a la opción Data / Stack / Columns…

En el cuadro de diálogo <Stack the following columns:> se seleccionan las cinco columnas, se activa la opción <Column of current worksheet> y se selecciona "Conductividad", que es una columna vacía donde guardará los valores de la variable respuesta/; y en <Store subscripts in> se selecciona "Nivel", que es la otra columna vacía que guardará la identificación del nivel.

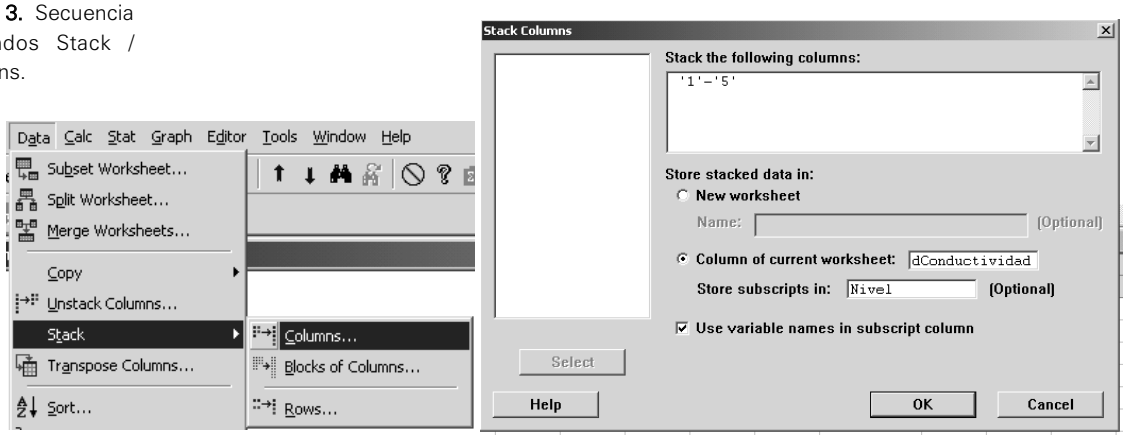


Figura 3. Secuencia comandos Stack / Columns.

De lo anterior se obtienen los siguientes resultados en la hoja de *worksheet*:

	C1	C2	C3	C4	C5	C6	C7-T
	1	2	3	4	5	Conductividad	Nivel
1	143	152	134	129	147	143	1
2	141	149	133	127	148	141	1
3	150	137	132	132	144	150	1
4	146	143	127	129	142	146	1
5						152	2
6						149	2
7						137	2
8						143	2
9						134	3
10						133	3
11						132	3
12						127	3
13						129	4
14						127	4
15						132	4
16						129	4
17						147	5
18						148	5
19						144	5
20						142	5

Figura 4. Resultados de la función.

Se ingresa a la opción: Stat / Anova / General Linear Model…
En el cuadro de diálogo, en <Responses>, seleccione "Conductividad", y en <Model>, seleccione "Nivel"; luego pulse en <Storage>.
Todo esto se aprecia en las figuras 5 y 6.

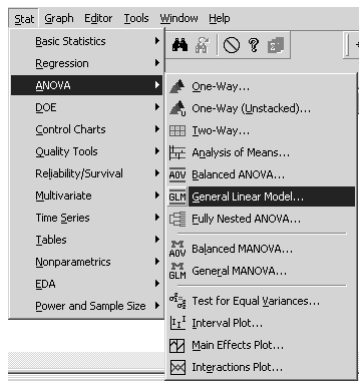


Figura 5. Secuencia del análisis de varianza.

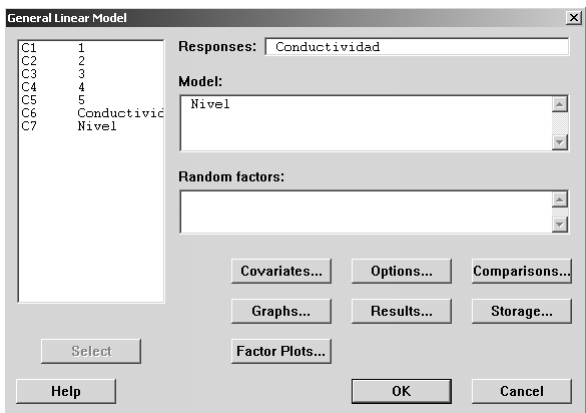


Figura 6. Ingreso de las variables.

Activar la opción <Coefficients>, tal como se presenta en la figura 7.

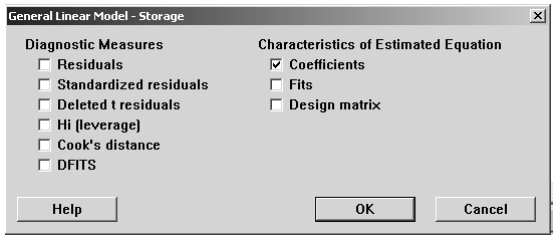


Figura 7. Opción para los coeficientes.

Se obtiene como resultado lo siguiente:

General Linear Model: Conductividad versus Nivel

```
Factor  Type   Levels  Values
Nivel   fixed       5  1, 2, 3, 4, 5
Analysis of Variance for Conductividad, using Adjusted SS for Tests
Source  DF   Seq SS   Adj SS  Adj MS      F      P
Nivel    4  1060.50  1060.50  265.13  16.35  0.000
Error   15   243.25   243.25   16.22
Total   19  1303.75
S = 4.02699   R-Sq = 81.34%   R-Sq(adj) = 76.37%
Unusual Observations for Conductividad
Obs  Conductividad      Fit  SE Fit  Residual  St Resid
  7        137.000  145.250   2.013    -8.250     -2.37 R
```

Worksheet 2 ***

↓	C1	C2	C3	C4	C5	C6	C7-T	C8
	1	2	3	4	5	Conductividad	Nivel	COEF1
1	143	152	134	129	147	143	1	139.25
2	141	149	133	127	148	141	1	5.75
3	150	137	132	132	144	150	1	6.00
4	146	143	127	129	142	146	1	-7.75
5						152	2	-10.00
6						149	2	
7						137	2	
8						143	2	
9						134	3	
10						133	3	
11						132	3	

Figura 8. Coeficientes en la hoja de datos.

En la figura 8 se presentan los resultados de aplicar el comando General Lineal Model.

En la columna C8 de la hoja Worksheet 2 la primera celda corresponde al estimado del promedio general; desde la segunda hasta la cuarta celda son los estimados de los efectos de los tratamientos.

c. Obtenga una estimación de intervalo del 95% del recubrimiento medio de tipo 1.

Solución:

Se sabe que el intervalo de confianza para el parámetro μ_i es:

$$\left\langle \bar{y}_{i\square} \pm t_{\left(1-\alpha/2,a(n-1)\right)}\sqrt{\frac{CMError}{n}} \right\rangle$$

Para el recubrimiento del tipo 1:

$$\left\langle \bar{y}_{1\square} \pm t_{(0.975,15)}\sqrt{\frac{CME}{4}} \right\rangle = \left\langle 145 \pm t_{(0.975,15)}\sqrt{\frac{16.216667}{4}} \right\rangle$$

Haciendo uso del software Minitab

Inverse Cumulative Distribution Function

```
Student's t distribution with 15 DF
P( X <= x )          x
       0.975   2.13145
```

Obteniendo:

$$\langle 145 \pm 2.13145(2.013496)\rangle = \langle 145 \pm 4.291666\rangle = \langle 140.708334, 149.291666\rangle$$

Este intervalo contiene al parámetro μ_i con un 95% de confianza.

d. Obtener una estimación de intervalo de 95% de la diferencia media entre los tipos de recubrimiento 1 y 4.

Solución:

La expresión que se usar para obtener la estimación por intervalos para la diferencia entre el recubrimiento 1 y 4:

$$\left\langle \bar{y}_{1\square} - \bar{y}_{4\square} \pm t_{\left(1-\alpha/2,a(n-1)\right)}\sqrt{\frac{2\text{CME}}{n}} \right\rangle = \left\langle 145 - 129.25 \pm t_{(0.975,15)}\sqrt{\frac{2(16.216667)}{4}} \right\rangle =$$

$$= \langle 15.75 \pm 2.13145(4.054167)\rangle = \langle 9.681, 21.819\rangle$$

El $\mu_1 - \mu_4$ se encuentra en este intervalo con un 95% de confianza.

2. Para estudiar el efecto de la temperatura sobre el rendimiento de un proceso químico se produjeron cinco lotes con cada uno de tres valores de temperatura. Los resultados aparecen en la tabla siguiente.

Nivel de temperatura	Rendimiento				
50°C	34	24	36	39	32
60°C	30	31	34	23	27
70°C	23	28	28	30	31

a. Indique la unidad experimental, el factor y sus niveles. ¿El nivel de temperatura tiene efecto sobre la media del rendimiento del proceso químico? Utilice $\alpha = 0.05$

Solución:

Unidad experimental: Un lote.

Factor: Temperatura (°C).

Niveles: 50°C, 60°C, 70°C

Modelo estadístico:

$$Y_{ij} = \mu + \tau_i + e_{ij}, \qquad i = 1,\ldots,3 \qquad j = 1,\ldots,5 \qquad a = 3$$
$$n = 5 \qquad N = 15$$

donde:

Y_{ij}: Rendimiento del proceso químico usando el i-ésimo nivel de temperatura en el j-ésimo lote de estudio.

- Formulación de hipótesis:

  $H_0 : \mu_1 = \mu_2 = \mu_3$

  H_1 : Al menos un $\mu_i \neq \mu_j$ donde $i \neq j$

- Nivel de significación: $\alpha = 0.05$
- Calculando la estadística de prueba: $F_0 = \dfrac{CMTrat}{CME}$

  Suma de cuadrados de tratamientos: $SCTrat = \sum_{i=1}^{a} \dfrac{y_{i\square}^2}{n} - \dfrac{y_{\square\square}^2}{N}$

$$SCTrat = \left(\frac{(34+\ldots+32)^2 + \ldots + (23+\ldots+31)^2}{5} \right) - \left(\frac{(34+\ldots+32+\ldots+23+\ldots+31)^2}{15} \right)$$

$$SCTrat = (13570 - 13500) = 70$$

Suma de cuadrados del total: $SCT = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \dfrac{y_{\square\square}^2}{N}$

$$SCT = \left(32^2 + \ldots + 31^2\right) - \left(\frac{(34+\ldots+31)^2}{15} \right) = (13806 - 13500) = 306$$

Suma de cuadrados del error: $SCE = \sum_{i=1}^{a}\sum_{j=1}^{n} \left(y_{ij} - \overline{y}_{i\square}\right)^2$ ó

$SCE = SCT - SCTrat$

$$SCE = (306 - 70) = 236$$

Cuadrado medio de tratamientos: $CMTrat = \dfrac{SCTrat}{a-1}$

$$CMTrat = \frac{70}{3-1} = 35$$

Cuadrado medio del error: $CME = \dfrac{SCE}{a(n-1)}$

$$CME = \frac{236}{3(5-1)} = \frac{236}{12} = 19.667$$

Entonces:

Estadístico de prueba: $F_0 = \dfrac{\text{CMTrat}}{\text{CME}} = \dfrac{35}{19.667} = 1.7797$

- Calculando el valor crítico:

$$F_{(1-\alpha,a-1,N-a)} = F_{(1-0.05,3-1,15-3)} = F_{(0.95,2,12)} = 3.88529$$

Haciendo uso del software Minitab

Inverse Cumulative Distribution Function

```
F distribution with 2 DF in numerator and 12 DF
in denominator
P( X <= x )          x
        0.95   3.88529
```

- Cálculo del P-value.

$$P(F > 1.7797) = 1 - P(F \le 1.7797) = 1 - 0.789559 = 0.210441$$

- Conclusión: Como $1.7797 < 3.88529$ entonces no se rechaza la hipótesis nula; es decir, el nivel de temperatura no tiene efecto sobre la media del rendimiento del proceso químico con 5% de significación.

Desarrollando con el software Minitab

One-way Anova: 50 °C, 60 °C, 70 °C

```
Source    DF      SS      MS       F      P
Factor     2    70.0     5.0    1.78      0.210
Error     12   236.0    19.7
Total     14   306.0
S = 4.435    R-Sq = 22.88%    R-Sq(adj) = 10.02%
```

Presenta un P-value de 0.21 que es mayor que 0.05 (nivel de significación), por lo tanto no se rechaza la hipótesis nula. Se concluye, entonces, que no hay diferencia en el rendimiento del proceso químico cuando se aplican las temperaturas de 50°C, 60°C y 70°C con 5% de significación.

b. Estimar la media general y los efectos de cada tratamiento.

Solución:

El estimador de la media general es: $\hat{\mu} = \overline{y}_{\square\square}$
Obtención del promedio:

$$\overline{y}_{\square\square} = \frac{y_{\square\square}}{N} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}}{N} = \frac{(34+24+\ldots+20+31)}{15} = \frac{450}{15} = 30$$

Estimación de los efectos para cada uno de los tratamientos:

$\hat{\tau}_i = \overline{y}_{i\square} - \overline{y}_{\square\square}$

$$\hat{\tau}_1 = \overline{y}_{1\square} - \overline{y}_{\square\square} = \frac{(34+\ldots+32)}{5} - 30 = 33 - 30 = 3$$

$$\hat{\tau}_2 = \overline{y}_{2\square} - \overline{y}_{\square\square} = \frac{(30+\ldots+27)}{5} - 30 = 29 - 30 = -1$$

$$\hat{\tau}_3 = \overline{y}_{3\square} - \overline{y}_{\square\square} = \frac{(23+\ldots+31)}{5} - 30 = 28 - 30 = -2$$

Como puede apreciarse, el primer nivel de temperatura (50°C) tiene una media superior al promedio, mientras que el segundo y el tercer niveles de temperatura tienen un nivel menor al promedio.

c. Obtener el intervalo de confianza para la media de cada nivel de temperatura. Utilice un $\alpha = 0.05$

Solución:

Se sabe que: $\mu_i \in \left\langle \overline{y}_{i\square} \pm t_{(1-\alpha/2,\,a(n-1))}\sqrt{\frac{CMError}{n}} \right\rangle$

Haciendo uso del software Minitab

Inverse Cumulative Distribution Function

```
Student's t distribution with 15 DF
P( X <= x )          x
       0.975   2.13145
```

Para cada temperatura:

$$\left\langle 33 \pm t_{(0.975,12)}\sqrt{\frac{19.667}{5}} \right\rangle = \langle 28.678846,\ 37.321154 \rangle \text{ para } \mu_1$$

$$\left\langle 29 \pm t_{(0.975,12)}\sqrt{\frac{19.667}{5}} \right\rangle = \langle 24.678846,\ 33.321154 \rangle \text{ para } \mu_2$$

$$\left\langle 28 \pm t_{(0.975,12)}\sqrt{\frac{19.667}{5}} \right\rangle = \langle 23.678846,\ 32.321154 \rangle \text{ para } \mu_3$$

3. Una empresa consultora recibió el encargo de evaluar cuatro marcas de automóviles respecto al rendimiento de combustible. En la tabla siguiente se presentan los resultados obtenidos, en kilómetros por galón.

Marca de automóvil	A	B	C	D
Rendimiento	19	19	24	18
	21	20	26	17
	20	22	23	16
	19	21	25	19
	21	23	27	20

a. Formulación del modelo estadístico:

Solución:

Unidad experimental: Combustible.

Factor: Automóvil.

Tratamiento: Cuatro marcas de automóviles.

Modelo estadístico para determinar si existe diferencia en el rendimiento de las marcas de automóviles respecto al combustible:

$$Y_{ij} = \mu + \tau_i + e_{ij} \quad , \qquad \begin{matrix} i = 1,\ldots,4 & j = 1,\ldots,5 & a = 4 \\ n = 5 & N = 20 \end{matrix}$$

donde:

Y_{ij}: Rendimiento en kilómetros por galón de la i-ésima marca de automóvil en la j-ésima repetición de combustible.

Nota: Se supone que los automóviles son semejantes en sus características.

b. ¿Existen diferencias significativas entre las marcas de los carros respecto al rendimiento de combustible? Utilice $\alpha = 0.05$

Solución:

- Formulación de hipótesis:

  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

  H_1 : Al menos $\mu_i \neq \mu_j$ donde $i \neq j$

- Nivel de significación: $\alpha = 0.05$

  Estadística de prueba: $F_0 = \dfrac{\text{CMTrat}}{\text{CME}}$

  Haciendo uso del software Minitab

One-way Anova: A, B, C, D

```
Source   DF       SS     MS       F       P
Factor    3   130.00  43.33   20.39   0.000
Error    16    34.00   2.13
Total    19   164.00
S = 1.458   R-Sq = 79.27%   R-Sq(adj) = 75.38%
```

- Conclusión:
  El valor de $F_0 = 20.39$ y el P-value es aproximadamente igual a 0, este último es menor que 0.05 (nivel de significación), por lo tanto, se rechaza la hipótesis nula. Es decir, hay diferencias en el rendimiento de las cuatro marcas de automóvil cuando se utiliza el mismo tipo de combustible.

c. Compare el intervalo de confianza de la diferencia de las medias poblacionales de las marcas A y C versus el intervalo de confianza de las diferencias de los tratamientos de las mismas marcas de automóviles. Utilice $\alpha = 0.05$

Solución:

Obteniendo los intervalos de confianza:
Intervalo de confianza de la diferencia de las medias de las marcas A y C.

Haciendo uso del software Minitab:

Two-Sample T-Test and CI: A, C

```
Two-sample T for A vs C
   N   Mean  StDev  SE Mean
A  5  20.00   1.00     0.45
C  5  25.00   1.58     0.71
Difference = mu (A) - mu (C)
Estimate for difference:  -5.00000
95% CI for difference:  (-7.04723, -2.95277)
T-Test of difference = 0 (vs not =): T-Value = -5.98
P-Value = 0.001  DF = 6
```

Intervalo de confianza de la diferencia de los tratamientos de las marcas A y C.

Haciendo uso del software Minitab

```
Fisher 95% Individual Confidence Intervals
All Pairwise Comparisons
Simultaneous confidence level = 81.11%
A subtracted from:
    Lower  Center   Upper  --------+---------+---------+---------+-
B  -0.954   1.000   2.954                  (---*---)
C   3.046   5.000   6.954                           (---*---)
D  -3.954  -2.000  -0.046            (---*---)
                          --------+---------+---------+---------+-
                               -5.0       0.0       5.0      10.0
```

Una vez obtenidos los dos intervalos de confianzas se realizan las comparaciones:

El intervalo de confianza de las diferencias de medias poblacionales de las marcas A y C es de $\langle -7.04723, -2.95277 \rangle$ y el intervalo de confianza de las diferencias de las medias de los tratamientos de las marcas A y C es de $\langle -6.954, -3.046 \rangle$. Se puede observar que la diferencia entre los intervalos de confianza para el último intervalo de confianza se usa una t con 19 grados de libertad y el factor de corrección $\sqrt{\frac{2CME}{n}}$ es menor que $\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$.

4. La fábrica de pantalones CK tiene cuatro trabajadores que se dedican exclusivamente a coser pantalones previamente cortados. El gerente de CK sospecha que los trabajadores no están trabajando a un mismo nivel. Para probar su sospecha, el gerente decide registrar en forma aleatoria el tiempo (en minutos) que demora cada trabajador en coser un pantalón. Cinco observaciones fueron registradas para cada trabajador, obteniéndose la siguiente información:

Trabajador	Tiempo				
1	28	25	29	30	28
2	27	28	30	28	25
3	29	30	32	35	37
4	27	28	26	28	27

Solución:

Unidad experimental: Un pantalón.
Factor: Trabajador.
Tratamiento: Trabajador 1, trabajador 2, trabajador 3 y trabajador 4.
Modelo estadístico para determinar si existe diferencia en el tiempo que cada trabajador se demora en coser un pantalón es:

$$Y_{ij} = \mu + \tau_i + e_{ij} \qquad i = 1,\ldots,4 \qquad j = 1,\ldots,5 \qquad a = 4$$
$$n = 5 \qquad N = 20$$

Donde:

Y_{ij}: Tiempo del i-ésimo trabajador en la j-ésima costura de pantalón.

- Formulación de hipótesis:

  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
  $H_1 :$ Al menos un $\mu_i \neq \mu_j$ donde $i \neq j$

- Nivel de significación: $\alpha = 0.05$
- Estadística de prueba: $F_0 = \dfrac{CMTrat}{CME}$

Haciendo uso del software Minitab

One-way Anova: 1, 2, 3, 4

```
Source  DF      SS      MS      F       P
Factor   3   95.35   31.78   6.76   0.004
Error   16   75.20    4.70
Total   19  170.55
S = 2.168   R-Sq = 55.91%   R-Sq(adj) = 47.64%
```

- Conclusión:

  El valor de $F_0 = 6.76$ y el P-value es 0.004, este último es menor que 0.05 (nivel de significación), por lo tanto se rechaza la hipótesis nula. Es decir, hay diferencias entre los tiempos promedios que demoran los trabajadores en coser un pantalón.

b. Suponga que usted observa al trabajador 2, ¿cuánto es el tiempo mínimo que se demora en coser un pantalón con 95% de confianza? ¿Cuál es el tiempo máximo?

Solución:

Se conoce que: $\left\langle \bar{y}_{i\square} \pm t_{\left(1-\alpha/2,\, a(n-1)\right)} \sqrt{\frac{CMError}{n}} \right\rangle$

Para el trabajador 2: $\left\langle \bar{y}_{2\square} \pm t_{(0.975,16)} \sqrt{\frac{CME}{n}} \right\rangle = \left\langle 27.6 \pm t_{(0.975,16)} \sqrt{\frac{4.70}{5}} \right\rangle$

Haciendo uso del software Minitab

Inverse Cumulative Distribution Function

```
Student's t distribution with 16 DF
P( X <= x )          x
       0.975   2.11991
```

Obteniendo:

$\left\langle 27.6 \pm 2.11991\left(2.055329\right) \right\rangle = \left\langle 25.544671,\ 29.655329 \right\rangle$

Por lo tanto:
El tiempo mínimo que se demora el trabajador 2 en coser un pantalón es de 25.5 minutos y el tiempo máximo es de 29.7 minutos, con un 95% de confianza.

c. ¿Puede usted afirmar que el trabajador 3 es más rápido que el trabajador 4?

Solución:

Se conoce que: $\left\langle \bar{y}_{i\square} - \bar{y}_{k\square} \pm t_{\left(1-\alpha/2,\, a(n-1)\right)} \sqrt{\frac{2CME}{n}} \right\rangle$

Para la diferencia entre el trabajador 3 y 4:

$$\left\langle \bar{y}_{3\square} - \bar{y}_{4\square} \pm t_{\left(1-\alpha/2,\, a(n-1)\right)} \sqrt{\frac{2\text{CME}}{n}} \right\rangle = \left\langle 32.6 - 27.2 \pm t_{(0.975,16)} \sqrt{\frac{2(4.7)}{5}} \right\rangle$$

$$\left\langle 5.4 \pm 2.11991(1.88) \right\rangle = \left\langle 2.493326,\ 8.306674 \right\rangle$$

El intervalo de confianza contiene al parámetro $\mu_3 - \mu_4$ con 95% de confianza.

Entonces:

Como en el intervalo contiene valores positivos para la diferencia $\mu_3 - \mu_4$, no se puede afirmar que el trabajador 3 es más rápido que el trabajador 4.

5. Se realizan cuatro mezclas experimentales para medir la resistencia de concreto y se sometieron a cargas de compresión hasta romperse. ¿Hay evidencia estadística que permita concluir que los tipos de mezcla influyen en la resistencia del concreto? Use un nivel de significación del 4%.

Mezcla A	Mezcla B	Mezcla C	Mezcla D
2.3	2.2	2.15	2.25
2.2	2.1	2.15	2.15
2.25	2.2	2.2	2.25

Solución:

Unidad experimental: Concreto.
Factor: Mezcla.
Tratamiento: mezclas A, B, C y D.
Modelo estadístico para determinar si existe diferencia en la resistencia de concreto respecto al tipo de mezcla:

$$Y_{ij} = \mu + \tau_i + e_{ij} \qquad i = 1,\ldots,4 \qquad j = 1,\ldots,3 \qquad a = 4 \qquad n = 3 \qquad N = 12$$

donde:
Y_{ij}: Resistencia del i-ésimo tipo de mezcla en el j-ésimo concreto.

- Formulación de hipótesis:

  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

  $H_1 :$ Al menos un $\mu_i \neq \mu_j$ donde $i \neq j$

- Nivel de significación: $\alpha = 0.04$

- Estadística de prueba: $F_0 = \dfrac{CMTrat}{CME}$

Haciendo uso del software Minitab

One-way Anova: Mezcla A, Mezcla B, Mezcla C, Mezcla D

```
Source  DF        SS       MS      F       P
Factor   3   0.01500  0.00500   2.00   0.193
Error    8   0.02000  0.00250
Total   11   0.03500
S = 0.05    R-Sq = 42.86%    R-Sq(adj) = 21.43%
```

- Conclusión:
  El valor de $F_0 = 2$ y el P-value es 0.193, este último es mayor que 0.04 (nivel de significación). Por lo tanto, no se rechaza la hipótesis nula. Es decir, no hay diferencias en la resistencia del concreto en los distintos tipos de mezclas.

6. Se tomó información de muestras aleatorias de agua, durante cuatro días, en cinco lugares del río Mantaro para medir si la cantidad de oxígeno varía de un lugar a otro. Los lugares 1 y 2 se escogieron antes de pasar una mina, uno cerca de la orilla y el otro a mitad del río; el lugar 3 se tomó en el sitio donde la mina descarga el agua industrial, el cuarto se tomó río abajo pero en la mitad y el quinto se tomó río abajo pero en la orilla. ¿Proporcionan los datos suficiente evidencia para indicar que existe diferencia entre las cantidades medias de oxígeno de los cinco lugares? Use un nivel de significación del 3%.

Muestras	Lugar 1	Lugar 2	Lugar 3	Lugar 4	Lugar 5
	5.9	6.1	6.3	6.1	6
	6.3	6.6	6.4	6.4	6.5
	4.8	4.3	5	4.7	5.1
	6	6.2	6.1	5.8	5.9

Unidad experimental: Muestra de agua.

Factor: Lugar.

Tratamiento: lugar 1, …, lugar 5 de agua.

Modelo matemático para determinar si existe diferencia en las medidas de oxígeno en las muestras tomadas respecto a los lugares:

$$Y_{ij} = \mu + \tau_i + e_{ij}\,, \quad i = 1,\ldots,5 \quad j = 1,\ldots,4 \quad a = 5 \quad n = 4 \quad N = 25$$

donde:
Y_{ij}: Cantidad de oxígeno en el i-ésimo lugar de la j-ésima muestra.

- Formulación de hipótesis:

  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

  $H_1 : \text{Al menos un } \mu_i \neq \mu_j \qquad \text{donde } i \neq j$

- Nivel de significación: $\alpha = 0.03$

- Estadística de prueba: $F_0 = \dfrac{CMTrat}{CME}$

  Haciendo uso del software Minitab

One-way Anova: Muestra 1; Muestra 2; Muestra 3; Muestra 4; Muestra 5

```
Source  DF      SS      MS      F       P
Factor   4   0.120   0.030   0.05   0.994
Error   15   8.338   0.556
Total   19   8.458
S = 0.7455    R-Sq = 1.42%    R-Sq(adj) = 0.00%
```

- Conclusión:

  El valor de $F_0 = 0.05$ y el P-value es 0.994, es mayor que 0.03 (nivel de significación). Por lo tanto, se acepta la hipótesis nula. Es decir, no hay diferencias entre las cantidades de oxígeno de los cinco lugares.

7. Un ingeniero consultor ha recibido el encargo de evaluar los bronceadores de piel que se comercializan en el mercado nacional. Para tal fin, ha decidido registrar el tiempo (en minutos) que los usuarios necesitan para alcanzar el bronceado deseado. Muestras aleatorias de supermercados y farmacias de cinco conocidas marcas de bronceadores fueron seleccionadas, obteniéndose los siguientes resultados:

Rayo de Sol (1)	Brillo Total (2)	Sunrise (3)	Sunset (4)	Sun (5)
25	35	42	33	26
26	34	43	32	28
24	36	50	30	29
20	37	45	32	24
30	38	48	32	18
35	25	43	30	42

Use $\alpha = 0.03$

a. ¿Hay diferencias significativas entre los bronceadores?

Solución:

Unidad experimental: Una persona.

Factor: Bronceador.

Tratamiento: Rayo de Sol, Brillo Total, Sunrise, Sunset y Sun.

Modelo estadístico para determinar si existe diferencia entre las marcas de bronceadores en cuanto al tiempo que necesitan los usuarios para alcanzar el bronceado deseado:

$$Y_{ij} = \mu + \tau_i + e_{ij}\ , \qquad i = 1,\ldots,5 \qquad j = 1,\ldots,6 \qquad a = 5$$
$$n = 6 \qquad N = 30$$

donde:

Y_{ij}: Tiempo para el bronceado deseado del i-ésimo bronceador en la j-ésima persona.

- Formulación de hipótesis:

  $H_0 : \mu_1 = \mu_2 = \ldots = \mu_5$

  $H_1 :$ Al menos un $\mu_i \neq \mu_j$ donde $i \neq j$

- Nivel de significación: $\alpha = 0.03$

- Estadística de prueba: $F_0 = \dfrac{CMTrat}{CME}$

  Haciendo uso del software Minitab

  One-way Anova: Rayo de Sol, Brillo Total, Sunrise, Sunset, Sun

```
Source  DF       SS      MS       F       P
Factor   4   1310.5   327.6   13.18   0.000
Error   25    621.3    24.9
Total   29   1931.9
S = 4.985    R-Sq = 67.84%    R-Sq(adj) = 62.69%
```

- Conclusión:
  El valor de $F_0 = 13.18$ y el P-value es aproximadamente igual a 0, este último es menor que 0.03 (nivel de significación), por lo tanto se rechaza la hipótesis nula. Es decir, hay diferencias en los tiempos para alcanzar el bronceado deseado entre las diferentes marcas de bronceadores.

b. ¿Puede afirmarse que Sunrise demora más en broncear que Sunset?

Solución:

Se conoce que con tamaños de muestra iguales la fórmula es:

$$\left\langle (\bar{y}_{i\square} - \bar{y}_{k\square}) \pm t_{\left(1-\alpha/2, a(n-1)\right)} \sqrt{\frac{2CME}{n}} \right\rangle$$

Para la diferencia entre las marcas de bronceador Sunrise y Sunset:

$$\left\langle (\bar{y}_{3\square} - \bar{y}_{4\square}) \pm t_{\left(1-\alpha/2, a(n-1)\right)} \sqrt{\frac{2CME}{n}} \right\rangle = \left\langle (45.166667 - 31.5) \pm t_{(0.985,25))} \sqrt{\frac{2(24.9)}{6}} \right\rangle$$

$$\langle 13.666667 \pm 2.30113(2.880972) \rangle = \langle 7.037175,\ 20.296158 \rangle$$

Entonces:

Como en el intervalo contiene valores positivos para el parámetro $\mu_3 - \mu_4$, se puede afirmar que Sunrise (3) demora más en alcanzar el bronceado deseado que Sunset (4) con 97% de confianza.

c. ¿Cuál es el tiempo máximo que necesita Rayo de Sol para broncear en la forma deseada a un usuario y cuál es el tiempo mínimo?

Solución:

Se conoce: $\left\langle \bar{y}_{i\square} \pm t_{\left(1-\alpha/2,a(n-1)\right)}\sqrt{\frac{CMError}{n}} \right\rangle$

Se sabe que: $\alpha = 0.03$

Para el bronceador Rayo de Sol (1):

$$\left\langle \bar{y}_{1\square} \pm t_{(0.985,16)}\sqrt{\frac{CME}{n}} \right\rangle = \left\langle 26.666667 \pm t_{(0.985,25)}\sqrt{\frac{24.9}{6}} \right\rangle$$

Haciendo uso del software Minitab

Inverse Cumulative Distribution Function

```
Student's t distribution with 25 DF
P( X <= x )          x
       0.985   2.30113
```

Obteniendo:

$$\left\langle 26.666667 \pm 2.30113\left(2.037155\right)\right\rangle = \left\langle 21.978908,\ 31.354425\right\rangle$$

Entonces:

El tiempo mínimo que demora el bronceador Rayo de Sol en broncear en la forma deseada por el usuario es de 21.9 minutos y el tiempo máximo es de 31.3 minutos.

d. Realice y presente los cálculos necesarios para obtener la suma de cuadrados de tratamientos.

Solución:

Calculando la suma de cuadrados entre tratamientos:

$$SCTrat = \sum_{i=1}^{a} \frac{y_{i\square}^2}{n} - \frac{y_{\square\square}^2}{N}$$

$$SCTrat = \left(\frac{\left(25+\ldots+35\right)^2+\ldots+\left(26+\ldots+42\right)^2}{6} \right) - \left(\frac{\left(25+\ldots+35+\ldots+26+\ldots+42\right)^2}{30} \right)$$

$$SCTrat = \left(34112.6667 - 32802.1333\right) = 1310.5333$$

8. El reporte de Minitab del análisis de varianza del diseño completamente al azar correspondiente al ingreso familiar en ocho departamentos en estudio es:

One-way ANOVA: Ingreso_Amaz, Ingreso_Apur, Ingreso_Ayac, Ingreso_Huan, ...

Fuentes de Variabilidad	GL	SC	CM	F
Factor	7	71541	10220.1429	0.47675248
Error	104	2229448	21437	
Total	111	2300989		

Level	N	Mean	StDev
Ingreso_Amazonas	14	244.2	102.6
Ingreso_Apurimac	14	238.3	165.7
Ingreso_Ayacucho	14	238.4	133.6
Ingreso_Huancave	14	290.9	171.3
Ingreso_Lambayeq	14	247.0	105.7
Ingreso_Otros	14	266.0	164.1
Ingreso_Tacna	14	281.2	177.1
Ingreso_Tumbes	14	206.3	129.8

a. ¿Cuál es el estimador puntual de μ_5?

Solución:

De acuerdo con el reporte del software Minitab, se tiene que el estimador puntual de μ_5 es el valor del promedio de los ingresos asociados al quinto departamento, Lambayeque $(i = 5)$, es: $\bar{y}_{5\square} = 247.0$

b. ¿Cuál es el P-value de la tabla del análisis de varianza?

Solución:

Para obtener el P-value se tiene que calcular la probabilidad de que la estadística de prueba sea mayor que el valor observado, es decir se tiene que calcular la siguiente expresión:

$P(F_{(a-1,N-a)} > F_0)$

Reemplazando valores se tiene:

$$P(F_{(7,104)} > 0.47675248) = 1 - P(F_{(7,104)} \leq 0.47675248)$$

Haciendo uso del software Minitab

Cumulative Distribution Function

F distribution with 7 DF in numerator and 104 DF in denominator

x	P(X <= x)
0.476752	0.150488

Reemplazando el valor obtenido se tiene:

$$P(F_{(7,104)} > 0.47675248) = 1 - 0.150488 = 0.849512$$

Entonces, el P-value es 0.849512

c. ¿Es el ingreso familiar en Ayacucho mayor que el ingreso familiar en Tumbes?

Solución:

Se conoce: $(\mu_i - \mu_k) \in \left\langle \overline{y}_{i\square} - \overline{y}_{k\square} \pm t_{\left(1-\alpha/2, a(n-1)\right)} \sqrt{\frac{2CME}{n}} \right\rangle$

Considerando: $\alpha = 0.05$

Entonces:

$$\left\langle \overline{y}_{Ayacucho} - \overline{y}_{Tumbes} \pm t_{(0.975,8(14-1))} \sqrt{\frac{2(21437)}{14}} \right\rangle =$$

$$\left\langle \overline{y}_{Ayacucho} - \overline{y}_{Tumbes} \pm t_{(0.975,104)} \sqrt{\frac{2(21437)}{14}} \right\rangle$$

Haciendo uso del software Minitab

Inverse Cumulative Distribution Function

```
Student's t distribution with 104 DF
P( X <= x )          x
        0.98   2.07983
```

Reemplazando el valor obtenido:

$$\left\langle \overline{y}_{Ayacucho} - \overline{y}_{Tumbes} \pm (2.07983) \sqrt{\frac{2(21437)}{14}} \right\rangle = \langle -82.9961568,\ 147.196157 \rangle$$

- Conclusión.
  No es posible afirmar que el ingreso familiar en Ayacucho es mayor que el ingreso familiar en Tumbes. Con 95% de confianza.

9. A fin de analizar el efecto del color del envase en un producto se seleccionaron 30 establecimientos comerciales de similares características. A ocho de ellos tomados al azar se les envió una partida del producto con envases rojos (1); a otros siete, el envase verde (2); otros siete el envase anaranjado (3) y al resto el envase azul (4). Al cabo de un mes se recogieron los datos siguientes referidos a las cantidades vendidas en cada establecimiento.

Ventas ($)	Color	Ventas	Color
70	1	25	3
50	1	26	3
65	1	30	3
77	1	28	3
62	1	29	3
63	1	33	3
55	1	32	3
52	1	78	4
49	2	77	4
48	2	80	4
47	2	85	4
43	2	85	4
44	2	89	4
42	2	84	4
40	2	90	4

Al nivel del 5% de significación, ¿existe diferencia en el número promedio de ventas por el color del envase?

Solución:

Nota: En este ejemplo se ilustra el uso del diseño completamente al azar con diferente número de repeticiones por tratamiento.
Unidad experimental: Un establecimiento comercial.
Factor: Color.
Tratamientos: rojo, verde, anaranjado y azul.
Modelo estadístico para determinar si existe diferencia en los promedios de ventas debido a los cuatro tipos de colores de envase:

$Y_{ij} = \mu + \tau_i + e_{ij}$, $i = 1,\ldots,4$ $j = 1,\ldots,n_i$ $n_1 = 8$ $n_2 = 7$ $n_3 = 7$ $n_4 = 8$

donde:
Y_{ij} : Promedio de ventas del i-ésimo tipo de color de envase en el j-ésimo

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
H_1 : Al menos un $\mu_i \neq \mu_j$ donde $i \neq j$

- Nivel de significación: $\alpha = 0.05$
- Estadística de prueba: $F_0 = \dfrac{CMTrat}{CME}$

Haciendo uso del software Minitab

One-way Anova: Ventas($) versus Color

```
Source  DF       SS       MS       F       P
Color    3  12308.9  4103.0  122.21  0.000
Error   26    872.9    33.6
Total   29  13181.9
S = 5.794   R-Sq = 93.38%   R-Sq(adj) = 92.61%
```

- Conclusión:
  El valor de $F_0 = 122.21$ y el P-value es 0 que es menor que 0.05 (nivel de significación), por consiguiente se rechaza la hipótesis nula. Se concluye entonces que hay diferencia en el promedio de ventas debido al color del envase.

10. Suponga que una compañía industrial ha adquirido tres máquinas nuevas de diferentes marcas y desea determinar si una de ellas es más rápida (tiempo en minutos) que las otras en la producción de cierto producto. Cinco productos de cada máquina son registradas al azar, obteniéndose los resultados que se presentan en el cuadro siguiente:

Marca 1	Marca 2	Marca 3
25	31	24
30	39	30
36	38	28
38	42	25
31	35	28

Al nivel del 5% de significación, ¿existe diferencia en los tiempos de producción por tipo de marca de la máquina?

Solución:

Modelo estadístico para determinar si existe diferencia en el tiempo de producción por tipo de marca de la máquina:

$$Y_{ij} = \mu + \tau_i + e_{ij}\ , \quad i = 1,\ldots,3 \qquad j = 1,\ldots,5$$

donde:
Y_{ij} : Cifras de producción del i-ésimo tipo de marca de la máquina en la j-ésima producto.

- Formulación de hipótesis

  $H_0 : \mu_1 = \mu_2 = \mu_3$
  $H_1 :$ Al menos un $\mu_i \neq \mu_j$ donde $i \neq j$

- Nivel de significación: $\alpha = 0.05$

- Estadística de prueba: $F_0 = \dfrac{CMTrat}{CME}$

  Haciendo uso del software Minitab

One-way Anova: Marca 1; Marca 2; Marca 3

```
Source  DF      SS      MS     F       P
Factor   2  250.0   125.0   7.50   0.008
Error   12  200.0    16.7
Total   14  450.0
S = 4.082   R-Sq = 55.56%   R-Sq(adj) = 48.15%
```

- Conclusión:
  El valor de $F_0 = 7.5$ y el P-value es 0.008, que es menor que 0.05 (nivel de significación); por consiguiente, se rechaza la hipótesis nula. Se concluye entonces que hay diferencia en los tiempos de producción debido al tipo de marca de la máquina. Con un nivel de significación del 5%.

11. Se sabe que el dióxido de carbono (CO_2) tiene un efecto crítico en el crecimiento microbiológico. Cantidades pequeñas de CO_2 estimulan el crecimiento de muchos microorganismos, mientras que altas concentraciones inhiben el crecimiento de la mayor parte de ellos. Este último efecto se utiliza comercialmente cuando se almacenan alimentos perecederos. Se realizó un estudio para investigar el efecto de CO_2 sobre la tasa de crecimiento del Pseudomona Fragi, un corruptor de alimentos. Se administró dióxido de carbono a cinco presiones atmosféricas diferentes. La respuesta anotada es el cambio porcentual en la masa celular después de un tiempo de crecimiento de una hora. Se utilizaron 10 cultivos en cada nivel y se registraron los siguientes datos:

Nivel del factor (Presión en atmósferas de CO_2)

	0.0	0.083	0.29	0.5	0.86
Cambio porcentual de masa celular	62.6	50.9	45.5	29.5	24.9
	59.6	44.3	41.4	22.8	17.2
	64.5	47.5	29.8	19.2	7.80
	59.3	49.5	38.3	20.6	10.5
	58.6	48.5	40.2	29.2	17.8
	64.6	50.4	38.5	24.1	22.1
	50.9	35.2	30.2	22.6	22.6
	56.2	49.9	27.0	32.7	16.8
	52.3	42.6	40.0	24.4	15.9
	62.8	41.6	33.9	29.6	8.80

Plantear las hipótesis y contrastarlas, usando un nivel del 5%.

Solución:

Unidad experimental: un cultivo de Pseudomona Fragi.
Modelo estadístico para determinar si existe diferencia en el cambio porcentual de masa celular por nivel de presión:

$$Y_{ij} = \mu + \tau_i + e_{ij}\ , \quad i = 1,\ldots,5 \qquad j = 1,\ldots,10$$

donde:
Y_{ij} : Cambio porcentual de masa celular de Pseudomona Fragi del i-ésimo nivel de presión en el j-ésimo cultivo analizado.

- Formulación de hipótesis

  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

  $H_1 : \text{Al menos un } \mu_i \neq \mu_j \qquad \text{donde } i \neq$

- Nivel de significación : $\alpha = 0.05$
- Estadística de prueba: $F_0 = \dfrac{CMTrat}{CME}$

  Haciendo uso del software Minitab

One-way Anova: 0.0; 0.083; 0.29; 0.5; 0.86

```
Source   DF        SS       MS        F       P
Factor    4  11274.2  2818.5  101.39  0.000
Error    45   1250.9    27.8
Total    49  12525.1
S = 5.272   R-Sq = 90.01%   R-Sq(adj) = 89.13%
```

- Conclusión:

  El valor de $F_0 = 101.39$ y tiene un P-value aproximadamente igual a 0 que es menor que 0.05 (nivel de significación), por consiguiente se rechaza la hipótesis nula. Se concluye, entonces, que hay diferencia en el cambio porcentual de masa celular por nivel de presión.

12. Una entidad científica con sede en Lima realizó un experimento en el que intervinieron cinco nuevas variedades de camotes: SR90.323, LM92.086, LM92.048, LM92.148 y JEWEL; con la finalidad de seleccionar las mejores variedades. El experimento fue ejecutado de la siguiente manera: tomar 200 gramos de camote y someterlo a 100°C por un lapso de 48 horas y luego pesarlo. El contenido de materia seca resultó del cociente entre el peso seco y el peso fresco; luego, dicho resultado es multiplicado por 100 para obtenerlo en forma porcentual. Este procedimiento descrito se realizó cuatro veces para cada tipo de variedad de camote.

SR90.323	LM92.086	LM92.048	LM92.148	JEWEL
40.5	33.4	37.1	34.7	27.3
40.23	33.3	34.8	35	29.3
40.6	35	35.4	37.1	29.2
39.8	31.1	35.2	30.5	27.2

Se desea saber si el contenido del peso seco varía de acuerdo a la variedad de camote. Use un nivel de significación de 5%.

Solución:

Unidad experimental: 200 gramos de camote.
Factor: Variedad de camote.
Modelo estadístico para determinar si existe diferencia en el contenido porcentual de materia seca por variedad de camote:

$$Y_{ij} = \mu + \tau_i + e_{ij}\ , \quad i = 1,\ldots,5 \qquad j = 1,\ldots,4$$

donde:

Y_{ij} : Contenido porcentual de materia seca de la i-ésima variedad de camote en la j-ésima planta cosechada.

- Formulación de hipótesis

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$
$$H_1 : \text{Al menos un } \mu_i \neq \mu_j \qquad \text{donde } i \neq j$$

- Nivel de significación: $\alpha = 0.05$
- Estadística de prueba: $F_0 = \dfrac{CMTrat}{CME}$

Haciendo uso del software Minitab

One-way Anova: SR90.323; LM92.086; LM92.048; LM92.148; JEWEL

```
Source  DF       SS      MS       F       P
Factor   4   301.41   75.35   29.66   0.000
Error   15    38.11    2.54
Total   19   339.52
S = 1.594   R-Sq = 88.78%   R-Sq(adj) = 85.78%
```

- Conclusión:

El valor de $F_0 = 29.66$ presenta un P-value aproximadamente igual a 0 que es menor que 0.05 (nivel de significación), por consiguiente se rechaza la hipótesis nula. Se concluye, entonces, que hay diferencia en el contenido porcentual de materia seca entre las cinco variedades de camote.

13. Una empresa transnacional que se dedica a la venta de productos de alta tecnología desea medir la eficiencia de sus vendedores. Se eligen a cuatro vendedores al azar y se registran sus niveles de ventas mensuales, en dólares.

V1	V2	V3	V4
18596	4692	1564	8694
12536	3047	2447	3120
21563	3313	2862	856
16785	1619	3451	1258
27564	3501	2944	1080
12563	2124	4374	3131
8569	1486	560	1120
16589	2879	1996	485

¿Hay alguna diferencia entre las ventas por tipo de vendedor? Use un nivel de significación del 1%.

Nota: Este ejemplo es un modelo de efectos aleatorios.

Solución:

Unidad experimental: Un mes de ventas de productos de alta tecnología.
Factor: Vendedor.
Modelo estadístico para determinar si existe diferencia en el nivel de venta mensual por tipo de vendedor:

$$Y_{ij} = \mu + \tau_i + e_{ij}\,, \qquad i = 1,\ldots,4 \qquad j = 1,\ldots,8$$

donde:
Y_{ij} : Nivel de venta mensual, en dólares, del i-ésimo tipo de vendedor en el j-ésimo mes de ventas de productos de alta tecnología.

- Formulación de hipótesis

  $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
  $H_1 :$ Al menos un $\mu_i \neq \mu_j$ donde $i \neq j$

- Nivel de significación: $\alpha = 0.01$
- Estadística de prueba: $F_0 = \dfrac{CMTrat}{CME}$

  Haciendo uso del software Minitab

One-way Anova: V1; V2; V3; V4

```
Source   DF          SS          MS       F       P
Factor    3  1216797379   405599126   36.09   0.000
Error    28   314660509    11237875
Total    31  1531457888
S = 3352   R-Sq = 79.45%   R-Sq(adj) = 77.25%
```

- Conclusión:

  El valor de $F_0 = 36.09$ presenta un P-value aproximadamente igual a 0, que es menor que 0.01 (nivel de significación); por consiguiente, se rechaza la hipótesis nula. Se concluye, entonces, que hay diferencia en el nivel de venta mensual entre los cuatro vendedores.

14. En una encuesta aplicada a empresarios nacionales acerca del impacto del Tratado de Libre Comercio (TLC) con Estados Unidos en los diferentes sectores económicos se registraron las siguientes variables:

C1: Sector económico al que pertenece el empresario.
C2: Servicio o producto principal que ofrece el empresario.
C3: Ubicación de la empresa (centro, norte, oriente, sur).
C4: Ingreso estimado para el año 2007 (millones de dólares).
C5: Ingreso estimado del año 2008 en caso de firmarse el TLC (millones de dólares).
C6: Ingreso estimado del año 2008 en caso de NO firmarse el TLC (millones de dólares).
C7: Años de funcionamiento de la empresa.

Los datos aparecen en el archivo "TLC.MTW".

¿Existen diferencias significativas entre los ingresos promedios estimados del año 2008 de los diferentes sectores económicos en caso de firmarse el TLC? Use $\alpha = 0.04$.

Solución:

Unidad experimental: Un empresario nacional.
Modelo estadístico para determinar si existe diferencia en el ingreso promedio estimado del año 2008 por tipo de sector económico:

$$Y_{ij} = \mu + \tau_i + e_{ij}$$

donde:
Y_{ij} : Ingreso promedio, millones de dólares, del i-ésimo sector económico

$H_0 : \mu_1 = \mu_2 = ... = \mu_8$

H_1 : Al menos un $\mu_i \neq \mu_j$ donde $i \neq j$

- Nivel de significación: $\alpha = 0.04$
- Estadística de prueba: $F_0 = \dfrac{CMTrat}{CME}$

  Haciendo uso del software Minitab

One-way Anova: 2008-Si TLC versus Sector

```
Source    DF        SS     MS      F       P
Sector     7     64.45   9.21   1.14   0.342
Error    232   1881.40   8.11
Total    239   1945.8
```

- Conclusión:
  El valor de $F_0 = 1.14$ presenta un P-value de 0.342 que es mayor que 0.04 (nivel de significación), por consiguiente no se rechaza la hipótesis nula. Se concluye, entonces, que no existen diferencias significativas entre los ingresos promedios estimados del año 2008 por tipo de sector económico.

PROBLEMAS PROPUESTOS

1. Una fábrica de hilados tiene un gran número de telares. Se supone que cada telar produce la misma cantidad de prendas por minuto. Para investigar esta hipótesis se eligen cinco telares, y se mide en tiempos distintos la cantidad de libras por minuto que producen. Con ello se obtienen los datos siguientes:

Telar	Salida				
1	4	4.1	4.2	4.0	4.1
2	3.9	3.8	3.9	4.0	4.0
3	4.1	4.2	4.1	4.0	3.9
4	3.6	3.8	4.0	3.9	3.7
5	3.8	3.6	3.9	3.8	4.0

a. ¿Hay alguna diferencia en la producción de los telares? ¿Cuál es la unidad experimental? Determine el factor y los niveles.
b. Estime la media general y los efectos de tratamiento.
c. Calcule IC para la media de cada tratamiento.

2. La ciudad de Trujillo tiene cuatro locales de comida rápida. Las cantidades de hamburguesas vendidas en los establecimientos en 10 semanas elegidas al azar se presentan a continuación. Al nivel del 1% de significación, ¿existe diferencia en el número promedio vendido entre los cuatro restaurantes?

R1	R2	R3	R4
70	65	80	100
75	70	85	105
72	69	99	110
71	73	100	120
80	90	120	140
85	100	130	150
79	110	135	155
83	109	132	143
85	115	125	145
88	118	128	150

3. Tres cadenas de supermercados de la capital dicen tener los precios más bajos de un determinado producto. Primero se seleccionó una muestra al azar de 10 días y se registro el precio en cada uno de ellos. Al nivel del 6% de significación, ¿existe diferencia en los precios medios en los supermercados?

S1	S2	S3
3.6	2.2	2.3
3.4	2.1	2.5
3.5	2.2	2.5
3.6	2.2	2.3
3.4	2	2.3
3.3	2.1	2.1
3.4	2.2	2.6
3.3	2.3	2.4
3.4	2	2.5
3.4	2.4	2.6

4. La Comisión Latinoamericana de Aviación Civil (CLAC) es un organismo regional y una de sus funciones es recopilar la información de tráfico de origen y destino de pasajeros provenientes de estados miembros de la CLAC que notifican esta clase de información. Se recolectaron datos de tres líneas aéreas en la ruta Lima-Estados Unidos, como se muestra a continuación, ¿existe diferencia significativa en la cantidad de pasajeros en las tres líneas aéreas a un nivel de significación del 3%?

Línea1	Línea 2	Línea 3
10112	8584	25548
8651	7662	22963
13409	8076	29015
14420	8854	22347
14893	8817	28347
12390	6919	27871
17548	8065	33671
16203	9766	28733

5. Un investigador en psicología evolutiva ha estudiado los efectos de cuatro técnicas de estudio sobre el aprendizaje del idioma inglés en una muestra aleatoria de 20 niños. Los cuatro grupos del mismo tamaño se han formado aleatoriamente y han sido asignados a cada tratamiento al azar. Algunos de los resultados se muestran en la siguiente tabla:

F.V.	SC	G.L.	MS	F	P
Entre grupos			76.63		
Dentro de grupos					
Total	541.20	19			

a. Defina la variable y el tipo de diseño de Anova que se ha aplicado.
b. Complete la tabla.
c. ¿Qué decisión estadística es razonable tomar según los datos, con $\alpha = 0.10$?

6. La división de entrenamiento de un departamento de recursos humanos en una empresa de publicidad decide poner a prueba una serie de métodos tradicionalmente empleados en los programas de capacitación. Para ello elige al azar 30 empleados para que tomen el curso de Estrategias de Mercadeo y divide a dichos empleados en tres grupos de igual número de integrantes y en forma aleatoria, donde el primer grupo recibirá el curso mediante el uso de cintas de vídeo, el segundo a través de una conferencia y el tercero a través de estudios de caso. La evaluación de los participantes con respecto al rendimiento en el curso se hizo a través de una escala que va de 1 a 20 puntos.

A partir de los datos que se presentan a continuación, determine:

Si existe alguna diferencia significativa entre los diversos métodos de enseñanza utilizados, y en caso de que exista, defina la estrategia más efectiva.

Cintas de videos	Conferencias	E. Casos
12	12	15
15	14	17
10	13	16
8	12	15
13	10	14
15	8	16
14	10	18
12	11	16
11	12	17
12	13	17

7. La empresa SSD ha recibido el encargo de evaluar las ventas diarias (en miles de nuevos soles) de cinco centros comerciales. Para este fin, se seleccionaron al azar las ventas diarias realizadas por estos centros comerciales, obteniéndose los siguientes resultados:

Mega Plaza	Jockey Plaza	Rey de Gamarra	Plaza San Miguel	Polvos Azules
120	180	200	130	120
130	130	220	80	300
180	110	160	70	280
230	190	250	120	300
170	150	160	180	250
165	180	280	160	290

Use $\alpha = 0.03$ en sus cálculos. Suponga que las ventas diarias tienen distribución normal.

a. ¿Cuál es la unidad experimental?
b. ¿Cuál es el estimador puntual de μ_3? ¿Cuál es el estimador puntual de μ_5?
c. ¿Cuáles son sus conclusiones respecto a las ventas diarias de los cinco centros comerciales? Señale las hipótesis, la prueba estadística y la región crítica correspondiente.
d. ¿Es posible afirmar que Rey de Gamarra vende más que Jockey Plaza? Sustente su respuesta.
e. Si se reemplaza el valor de Y_{42} por 380 mil nuevos soles, ¿cuál es el intervalo de confianza para μ_4?

8. Se desea estudiar la influencia de las concentraciones de pectina sobre la textura de una mermelada, al 0.25%, 1% y 2% y medir las viscosidades con viscosímetro rotacional. Los resultados fueron los siguientes:

Pectina		
0.25%	1%	2%
32.5	33	35
28.9	31.2	40
33.6	38.3	40.5
33.4	34.3	40.8
33.6	32.5	42
35.3	33.8	43
29.5	32.5	46

a. A la vista de los resultados, ¿qué conclusiones se pueden obtener sobre la viscosidad en los tres tipos de pectina, con $\alpha = 0.05$?
b. ¿Es posible afirmar que la viscosidad es mayor cuando se coloca el 1% de pectina que cuando se utiliza el 0.25%? Use $\alpha = 0.04$.

9. Se analizaron muestras de cuatro tipos de cereales para determinar el contenido de tiamina (B1), una de las vitaminas del complejo B que ayudan a las células del organismo a convertir los carbohidratos en energía. La tiamina es esencial para el funcionamiento del corazón, músculos y sistema nervioso.

Trigo	Cebada	Arroz	Avena
5.1	6.4	4.3	8.2
4.8	7.8	4.8	6.0
5.9	8.2	5.2	7.9
6.2	6.3	6.1	7.1
6.1	7.6	1.6	5.6
6.6	6.0	5.9	7.1
5.7	5.7	4.2	7.4

a. Utilice esta información y un nivel de significación del 0.01 para probar la hipótesis nula de que no existe diferencia en el contenido de tiamina en los cuatro tipos de cereales.
b. Estime la media general y los efectos de tratamiento.
c. Calcule IC para la media de cada tratamiento con $\alpha = 0.99$.
d. ¿Es posible afirmar que el contenido de tiamina en la avena es mayor que en el trigo, con un nivel de confianza del 98%?

10. Un estudio del servicio de investigación agraria de Estados Unidos (ARS) confirma que el consumo de ácidos grasos 'trans' está asociado con una mayor cantidad en la sangre de lipoproteínas de baja densidad (LDL) o colesterol malo.

El estudio ha sido realizado en 36 voluntarios durante 35 días, utilizando cuatro diferentes tipos de margarinas; al finalizar el periodo de experimentación se midió el colesterol a los voluntarios. Los datos son los siguientes:

Dhora	Swiss	Dulce	Untarella
180	190	200	210
190	195	210	200
200	180	220	230
220	230	280	250
235	240	250	240
250	180	190	260
180	200	220	245
190	220	230	250
185	260	240	280

a. Indique la unidad experimental, el factor y el tratamiento para este experimento.
b. Determine los estimadores de la media general y del efecto de los tratamientos.
c. ¿Es el nivel de colesterol igual en los voluntarios utilizando las cuatro marcas de margarina? Use un nivel de significación del 5%.
d. ¿Es el nivel de colesterol de los voluntarios que utilizaron la margarina Dulce igual a aquellos que utilizaron Dhora con 7% de nivel de significación?

11. La información que se presenta corresponde al rendimiento del personal de una empresa y la categoría de la universidad de procedencia del evaluado (Excelente, Buena y Regular). Use un nivel de significación del 2%.

Reporte Minitab:

One-way Anova: Puntaje rendimiento segundo año versus Categoría de universidad

```
Source        SS
Categoría    952
Error      10240
Total      11192

Level       N      Mean    StDev
Excelente  34     72.38    11.15
Buena      34     67.47    10.09
Regular    34     64.07     6.18
```

a. ¿Hay diferencias entre los puntajes promedios de rendimiento de las distintas categorías de universidad?
   a.1 Estadística de prueba (con sus valores respectivos) y regla de decisión con valor(es) crítico(s):
   a.2 Conclusión:
b. ¿Es el puntaje promedio de rendimiento de los egresados de las universidades "excelentes" mayor que el puntaje promedio de rendimiento de los egresados de las universidades "buenas"?

12. Tres técnicas diferentes permiten analizar la consistencia de una pieza fabricada. Las cinco piezas seleccionadas, a las que se les aplicó cada una de las técnicas, dieron los siguientes resultados:

Técnica 1	2.5	6	5.5	4	6
Técnica 2	1.5	6.5	5	7.5	6.5
Técnica 3	7	5	4.5	6.5	3

Con un nivel de significación del 5%, y considerando satisfechas las suposiciones necesarias, ¿puede considerarse que las tres técnicas proporcionan resultados similares?

13. El control de calidad de una determinada pieza se realiza a través de tres pruebas, considerándose que los resultados proporcionados por las tres pruebas son igualmente significativos. Para contrastar dicha hipótesis, con un nivel de significación del 5%, se seleccionan tres muestras, que dan los siguientes resultados:

Prueba 1	55	46	61	50	42	69	48	56	62	59
Prueba 2	36	50	42	32	71	55	61	29	42	60
Prueba 3	56	58	40	31	63	45	50	60	49	52

A la vista de estas observaciones, y considerando ciertas las hipótesis necesarias, ¿puede mantenerse que las tres pruebas aportan resultados similares?

14. Mizuno Corporation es una empresa que produce pelotas de golf y ha fabricado cuatro diseños, las cuales se han llevado al campo de golf para probarlas, se ha registrado la distancia recorrida en metros.

D1	D2	D3	D4
207.01	210.07	218.00	225.07
207.01	209.99	217.99	223.81
207.01	210.05	217.98	224.26
206.99	209.96	217.99	225.54
207.00	209.98	218.01	225.77
207.00	210.03	218.00	225.47
207.00	210.07	218.00	226.14
206.99	209.91	218.02	226.21
207.01	210.00	218.01	226.47
207.01	210.06	218.00	226.91
207.02	210.00	218.01	223.16
206.99	210.06	217.99	226.29

a. Con un nivel de significación del 3%, ¿existe diferencia en las distancias entre los cuatro diseños?
b. Si en el inciso (a) hay un rechazo de la hipótesis nula determine qué pares de diseños son los distintos. Utilice un nivel de significación del 5%.

15. La Secretaría General de Transportes dispone de información de una muestra aleatoria de transportistas, la cual registra las siguientes variables:

X1: Monto total de las infracciones.
X2: Monto total mensual del consumo de petróleo.
X3: Empresa de transportes.
El reporte de los montos de consumo mensual de petróleo por empresa es:

One-way Anova: x2 versus x3

Source	DF	SS	MS	F	P
X3					
Error		2571483			
Total		2600029			

Level	N	Mean	StDev
27 de mayo	60	2201.2	89.9
Etusa	60	2210.9	96.0
Huarochiri	60	2196.1	101.5
Transa	60	2185.0	90.8
Veloz	60	2210.8	87.9

a. El valor crítico para probar la "$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$" con $\alpha = 0.04$ es: _____
b. Valor observado: _____
c. El intervalo de confianza del 96% para $\mu_4 - \mu_5$ es:
Límite inferior: _____ y Límite superior: _____
d. El estimador puntual de μ es: _____

16. La información que a continuación se detalla representa el precio de la gasolina de 84 octanos en tres provincias.

a. ¿Existe evidencia de una diferencia en el precio promedio en las tres provincias? Use un nivel de significación del 2%.
b. ¿En qué provincias el precio promedio de la gasolina es igual? Utilice un nivel de significación del 2%.

Lima	Trujillo	Arequipa
9.52	10.16	10.54
11.03	10.02	10.85
9.94	10.14	10.73
10.17	9.72	10.72
9.62	9.88	10.59
9.17	9.35	10.82
9.53	10.07	10.57
10.28	10.29	10.72
10.78	9.71	10.77
9.23	9.87	10.96

17. El jefe de producción de la fábrica Ropas S.A. está interesado en conocer si la resistencia a la ruptura de una fibra sintética está relacionada con el porcentaje de algodón de la fibra, la información se da en la siguiente tabla:

Porcentaje de algodón			
15%	20%	25%	30%
8	13	15	20
8	17	19	26
15	13	18	23
12	19	19	20
10	19	19	23

¿El porcentaje de algodón tiene algún efecto sobre la resistencia a la tensión de ruptura? Use un nivel de significación del 1%.

18. En la ladrillera "Rex" se utilizan cuatro temperaturas de cocción y se cree que esta afecta la densidad de cierto tipo de ladrillo.

Temperaturas			
100°F	125°F	150°F	175°F
22.8	21.6	21.9	22.9
22.9	21.4	22.8	22.8
21.7	21.6	21.8	21.9
21.6	21.6	21.6	22.8
21.6	21.8	21.5	22.9
21.5	20.2	22.5	21.9

a. Señale la unidad de análisis y la variable en estudio.
b. ¿Afecta la temperatura de cocción a la densidad de ladrillos? Utilice un nivel de significación de 0.03.

19. La tienda Jamoncito.com comercializa jamón ibérico de primera calidad. Para comparar el nivel de ventas de tres tipos de jamón se ha recolectado información de las ventas de seis días:

Jamón ibérico Bellota (kg)	Jamón ibérico Recebo (kg)	Jamón ibérico Reserva (kg)
50	55	53
52	58	51
49	60	52
50	65	55
49	66	52
47	65	51

Nota: El precio del jamón fue igual en los tres tipos.

a. Señale la unidad de análisis y la variable en estudio.
b. ¿Influye el tipo de jamón en las ventas? Considere un nivel de significación al 3%.

20. En una investigación se consideró cuatro tratamientos clínicos, asignándose aleatoriamente las personas a cada tratamiento, luego se hizo una evaluación del rendimiento del tratamiento con puntajes entre 0 a 150 puntos. Los datos son:

A	B	C	D
42	45	48	101
44	46	50	100
48	50	55	98
50	53	59	99
49	60	60	95
55	65	63	90

a. ¿Se puede concluir que el tipo de tratamiento clínico influye sobre el rendimiento con un nivel de significación del 2%?
b. ¿Cuál es el grupo que tiene mejor rendimiento con una confianza del 95%?

21. Un empresario de transporte de pasajeros tiene cinco ómnibus (U1, U2, U3, U4, U5) que cubren la ruta Lima-Huancayo-Lima, y desea estudiar el gasto en combustible de sus unidades. Para este fin, registró al azar los galones de petróleo que consumen los ómnibus, obteniendo los siguientes resultados:

Descriptve Statistics: U1,U2,U3,U4,U5

```
Variable    Count       Sum
      U1        7    618.00
      U2        7    673.00
      U3        7    669.00
      U4        7    585.00
      U5        7    760.00
```

$\sum y_{ij}^2 = 317285$

Nota: Use $\alpha = 0.03$ en sus cálculos.

a. Construya la tabla del análisis de varianza y pruebe la hipótesis correspondiente.
b. ¿Es el consumo promedio de combustible de la unidad U3 mayor que el consumo promedio de combustible de la unidad U1? Justifique su respuesta.

22. Los consumos de pescado (en toneladas) en cuatro distritos de Lima Metropolitana fueron registrados al azar en diferentes días, obteniéndose la siguiente información:

One-Way Anova: Ate; Comas; Breña; Lince

```
S = 3.654 R-Sq=36.53%
Level    N     Mean     StDev
Ate      5    15.60    3, 362
Comas    5    20.40    4, 827
Breña    5    19.00    3, 391
Lince    5    22.40    2, 702
```

a. ¿Son iguales en promedio los consumos de pescado en los cuatro distritos de Lima metropolitana? Use un $\alpha = 0.03$ en sus cálculos.
b. Calcule un intervalo de confianza del 97% para $\mu_{Ate} - \mu_{Comas}$ e interprete los resultados.

23. La compra y venta de agua mineral en la temporada de verano es una oportunidad de negocio para muchos empresarios. Un estudio realizado en Lima metropolitana basado en una muestra aleatoria de personas consumidoras de agua mineral permitió registrar los valores de las siguientes variables. Los datos se encuentran en el archivo "Agua Mineral.MTW".

C1 = Edad de la persona consumidora de agua mineral (en años).
C2 = Profesión/Ocupación.
C3 = Gasto semanal en agua mineral (en nuevos soles).
C4 = Consumo semanal de agua mineral (en litros).
C5 = Marca de agua mineral.

Utilice $\alpha = 0.04$ en sus cálculos.
a. Unidad de análisis, variable, tratamiento.
b. ¿Puede afirmarse que el gasto semanal en agua mineral es el mismo para todas las marcas?
c. El intervalo de confianza del 96% para la diferencia de las medias de los tratamientos de San Antonio y San Luis es: _____

24. Un representante de la oficina de Defensa del Consumidor registró al azar los precios de ventas de la gasolina en las diferentes cadenas de ventas de combustibles de Lima metropolitana. Los datos son los siguientes:

Oil	Rep	PFY	Rimax
14.29	15.5	14.8	15.49
14.59	15.6	14.9	15.6
15.29	15.4	15.3	15.55
15.49	15.3	15.2	16.4
15.55	15.8	15.6	16.19

a. Unidad experimental, tratamiento y factor.
b. ¿Puede afirmarse que no hay diferencia significativa entre los precios de venta de la gasolina de las diferentes cadenas de venta de combustible? Use $\alpha = 0.03$.
c. Calcule un intervalo de confianza del 98% para $\mu_{\text{PFY}} - \mu_{\text{Rep}}$

25. La distribuidora mayorista La Exclusiva S.A. es una compañía que cuenta con un número apreciable de vendedores para sus principales productos: café, té, avena y leche. Su mercado está compuesto por las panaderías y bodegas de la ciudad. Después de siete años de operaciones, el señor Contreras (gerente general de la empresa) considera que la compañía ha logrado conquistar una buena parte del mercado. Sin embargo, el señor Contreras piensa que es tiempo de hacer un análisis de la eficiencia de sus esfuerzos de ventas para tomar decisiones al respecto. Para este fin, recolectó la información de 148 puntos de venta elegidos al azar, de los cuales 124 eran bodegas y 24 panaderías. La siguiente tabla muestra la estructura de los datos.

Variable	Descripción	Valores
Café	Ventas de café	Monto en nuevos soles
Té	Ventas de té	Monto en nuevos soles
Avena	Ventas de avena	Monto en nuevos soles
Leche	Ventas de leche	Monto en nuevos soles
Tipo	Tipo de establecimiento	1 = panadería 2 = bodega
Zona	Zona de la ciudad	1 = norte 2 = centro 3 = sur 4 = este

Use $\alpha = 0.02$.

¿Existen diferencias significativas entre las ventas promedios de leche de cada zona?

a. Formulación de hipótesis:
   H_0: _____
   H_1: _____
b. Valor de la prueba estadística: _____
c. Valor(es) crítico(s) y regla de decisión: _____
d. P-value: _____
e. Conclusión: _____

26. En un tratamiento contra la hipertensión se seleccionaron 40 enfermos de características similares en el hospital Sabogal. A cada enfermo se le administró al azar uno de los fármacos P, A, B, AB, formando cuatro grupos de 10. Para valorar la eficacia de los tratamientos, se registró el descenso de la presión diastólica desde el estado basal (inicio del tratamiento) hasta una semana después de iniciado dicho tratamiento. Los resultados, después de registrarse algunos abandonos, fueron los siguientes:

P:	10	0	15	-20	0	15	-5	35	20
A:	20	25	33	25	30	18	27	0	35
B:	15	10	25	30	15	35	25	22	11
AB:	10	5	-5	15	20	20	0	10	25

a. La unidad experimental, tratamiento y factor.
b. ¿Existen diferencias significativas entre los cuatro tratamientos? Use $\alpha = 0.04$.

27. Se tomaron cuatro muestras al azar de cinco marcas de cerveza. Se determinó el valor calórico expresado en calorías por 100 ml, para saber si las marcas difieren en la variable de respuesta analizada. La información se presenta en el siguiente cuadro:

Back	Abbot Ale	Brahman	Stella Artois	Beck's
31.5	32.3	28.8	24.5	25.4
32.0	31.9	27.6	25.3	26.3
32.7	33.0	29.1	24.9	23.9
30.9	31.7	27.7	26.1	24.1
31.6	32.4	28.9	26.1	24.2

a. Plantear el modelo de análisis y supuestos.
b. Probar la hipótesis con un nivel de confianza del 95% de que las marcas no difieren en el valor calórico.
c. Si las dos primeras marcas corresponden a cervezas blancas y las siguientes tres a cervezas negras, plantear un contraste que permita ver si existen diferencias en el valor calórico entre ambos colores de cervezas.
d. Hallar un intervalo de confianza del 95% para la diferencia entre las marcas con mayor y menor valor calórico.

28. Se desea saber si cierto fertilizante produce aumentos en el rendimiento de árboles frutales y determinar cuál es la dosis adecuada. Para lograr estos objetivos se decide aplicar diferentes concentraciones del fertilizante (expresadas en mg de producto activo por litro de solución) y ver el rendimiento alcanzado por la planta (expresado en Kg).

dosis-25	dosis-50	dosis-75	dosis-100
28	28.1	29.1	28.9
27.9	29.2	29.2	29.3
27.2	31	29.3	27.4
26.9	29.2	29.5	29.5
26.8	28.4	30.2	29.6

a. Plantear el modelo de análisis y supuestos.
b. ¿Hay diferencia significativa en el promedio del rendimiento de los árboles frutales en los cuatro tipos de dosis con un nivel de significación del 5%?
c. Hallar un intervalo de confianza del promedio de la dosis más alta con un nivel de significación del 6%.

29. Una comercializadora de productos de cómputo realiza una encuesta a sus clientes sobre el rendimiento de cuatro marcas de impresoras.

One-way ANOVA: M1, M2, M3, M4

```
Source  DF      SS      MS      F      P
Factor
Error
Total   23  592783
S = 74.51   R-Sq = 81.27%   R-Sq(adj) = 78.46
                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level  N    Mean  StDev   -+---------+---------+---------+--------
M1     6   850.0   59.7      (----*---)
M2     6  1170.0  113.1                            (---*---)
M3     6   801.7   61.8   (---*----)
M4     6   955.0   45.1             (----*---)
                           -+---------+---------+---------+--------
                          750       900      1050      1200
Pooled StDev = 74.5
```

a. Complete la tabla de Anova.
b. ¿Hay diferencia significativa en el promedio del rendimiento en las cuatro marcas de impresoras? Use un nivel de significación del 7%.
c. ¿Qué marcas de impresoras produce igual promedio en el rendimiento con una confianza del 95%?

Respuestas a los problemas propuestos

Capítulo 1

2. a) 15; b) $E(\bar{x}) = 17$ *y* $V(\bar{x}) = 4.666667$; c) 1/15; d) 1/3.
3. La respuesta es la opción b).
4. a) $k = 13.3654$; b) 0.008454; c) 0.27142.
5. a) 0.610261; b) 0.449711; c) 0.987904; d) -0.691197;
   e) 0.039205; f) $a = 0.55643$; $b = 1.51256$; $c = 0.40612$
6. a) 0.430554; b) disminuye de 1.428869 a 0.5.
7. a) 0.68296; b) 0.0000003.
8. $n = 9$ observaciones.
9. 0.033095.
10. $b_1 = 0.369457$ y $b_2 = 1.879889$.
11. a) $p \sim N(0.37, 0.015268^2)$; b) 0.952057; c) 0.83529.
12. 0.051237.
13. a) 0.958368; b) 0.041366.
14. a) 0.9957
15. a) 0.004340 b) 0.2841
16. 0.077792.
17. b.1) 0.190787; b.2) 0.03855; c.1) 0.204374; c.2) 299.334; d) 385
18. a.ii) 0.4898; b) 27 medicamentos.
19. a) entre 0.392451 a 0.916332; b) 0.87.
20. a) 0.999791; b) 0.99168.
21. a) 0.03223; b) 0.104748.
22. a.i) 0.734014; a.ii) 0.02275; a.iii) 57 pacientes.
    b.) 0.924092; c) 0.9723; d) 0.963314; e) 0.0002034.

Capítulo 2

2. β^* es insesgado.

3. $\dfrac{\sigma^2}{n_1+n_2}$

5. $\hat{\beta}=\sqrt{\dfrac{\sum_{i=1}^{n} X_i^2}{n}}$

7. $\hat{\mu}=\dfrac{\sum_{i=1}^{n}\dfrac{X_i}{\sigma_i^2}}{\sum_{i=1}^{n}\dfrac{1}{\sigma_i^2}}$; es insesgado.

8. $\hat{\pi}=\dfrac{1}{\bar{x}}$

9. a) $P=\dfrac{18}{30}=0.6;$ b) 6.54814.

10. a) Cualitativa $(1-\alpha)\%$. b) Homogéneas. c) Margen de error.
    d) Dependientes. e) $\sqrt{\dfrac{\pi(1-\pi)}{n}}$
    f) una función de las observaciones de una muestra aleatoria.

11. a) Bajo costo y menor tiempo de análisis.
    b) El valor esperado del estimador coincide con el parámetro.
    c) Que para muestras grandes ($n > 30$) la distribución de cualquier estimador tiende a la normal, independientemente de la población de donde fue extraída.
    d) Proporciona la teoría necesaria para hacer afirmaciones válidas de una población sobre la base de una muestra extraída de ella.

12. ⟨1441.20, 1558.80⟩

13. ⟨2882400, 3117600⟩

14. ⟨423450; 526550⟩

15. $n_1 = 1.41985n_0$.

16. $\sigma = 10$ y el intervalo es: ⟨473.467, 486.533⟩ por lo tanto la duración media es diferente de 500

17. 374

18. a) ⟨0.86261, 2.17739⟩ b) ⟨1.00029, 2.03971⟩
    c) Aumenta la precisión, es decir la amplitud del intervalo disminuye.

19. a) ⟨1003.1, 1008.9⟩ b) ⟨2.8, 7.3⟩

20. a) 0.95864 b) 145

21. a) ⟨281028, 298972⟩ b) ⟨0.476266, 0.514334⟩ no. c) ⟨0.01467, 0.03895⟩

22. a) ⟨5.11674, 22.3897⟩ b) ⟨16.6935, 19.4247⟩

23. ⟨1.49; 3.97⟩
24. a) i) 0.066807 ii) 0.308779
25. a) ⟨1770.24, 1829.76⟩ b) ⟨0.224623, 0.375377⟩ c) 505
26. a) ⟨170, 7830⟩ b) Incrementar n.
27. a) 99 b) 24
28. a) ⟨0.126348, 0.198652⟩ b) 2401
29. ⟨7.7237, 14.2763⟩ Sí.
30. a) ⟨0.374246, 0.482897⟩. La conclusión del gerente es falsa.
    b) ⟨0.057442, 0.617198⟩. Si es correcta la afirmación del gerente.
    c) ⟨–33. 6611, –4.2709⟩. Es falsa la afirmación del gerente.
    d) 0.104762, –0.016497, 0.226021. La afirmación del gerente es correcta.
31. a) ⟨246.238, 251.305⟩ b) ⟨–0.0702662, 0.150660⟩. No es correcta la afirmación.
    c) La prueba de varianzas indica que son iguales y el intervalo para la diferencia de medias es: ⟨–0.84308, 19.34308). No es correcto afirmar.
32. ⟨–2.04265, 6.04265⟩. Si la resistencia es la misma en toda la plancha.
33. –5.900, ⟨–59.00692, 47.20692⟩, 160.1625
34. La prueba de varianzas indica que son diferentes y el intervalo para la diferencia de medias es: ⟨–3.17825, 12.35825⟩. La afirmación del gerente tiene sustento.
35. Las varianzas son diferentes, el intervalo para la diferencia de medias es: ⟨–2.4128, 9.3850⟩; por lo tanto la máquina 1 no tiene mayor tiempo de fabricación.
36. Las varianzas son iguales y el intervalo para la diferencia de medias es: ⟨–1.68971, 0.18971). No se puede decir que uno demora más que el otro.
37. Son muestras dependientes y el intervalo es: ⟨–2.4128, 9.3850⟩. No reduce el tiempo de fabricación.
38. ⟨–11.8483, 5.84827⟩. No se distinguen los resultados de ambos ayudantes.
39. a) ⟨20.9049, 27.6951⟩ b) ⟨4.4923, 10.5296⟩ c) Las varianzas son iguales y el intervalo para la diferencia de medias es: ⟨–6.75188, 4.15188⟩..
40. a) ⟨8.21686, 8.25114⟩ b) ⟨0.0169, 0.0469⟩.
41. ⟨0.718308, 1.00137⟩
42. ⟨0.0908980, 3.42793⟩ b) ⟨–0.199131, 0.999131⟩.
43. ⟨–0.144863, 0.224863⟩
44. ⟨–0.0132415, 0.0732415⟩ No existe.
45. $\bar{x}_1 = 200$; $\bar{x}_2 = 180$; $s_1 = 70$; $s_2 = 72$

    a) ⟨0.306089, 3.18716⟩ b) ⟨–19.7314, 59.7314⟩ c) ⟨–0.284903, 0.362284⟩
    d) ⟨16814, 23186⟩

Capítulo 3

1. Falso.
2. Falso.

3. Verdadero.
4. Falso.
5. Falso.
6. a) H_0 :La proporción de televisores que requieren reparación en Sunglo y Zeta es la misma. No hay error.

   b) H_0 Ambos trabajadores son igualmente eficientes. Se comete error de tipo II.
7. a) ETI: Negar el préstamo cuando el solicitante es un buen prospecto.

   ETII: Conceder el préstamo cuando el solicitante es un mal prospecto.

   b) i. α pequeño ii. α puede ser relativamente grande.
8. D.
9. C.
10. D.
11. A.
12. $H_0 = \mu =$ 500(el gobierno tiene razón).

    $t_0 = 1.77$; p value = 0.042, como $\alpha = 0.04$, entonces se acepta y por lo tanto el gobierno tiene razón.

13. a) H_0: El gasto promedio mensual en consumo por familia es igual a S/. 1000.
    H_1: El gasto promedio mensual en consumo por familia es superior a S/. 1000
    ETI: Se acepta que el gasto promedio mensual, en consumo por familia, sea igual a S/.100, siendo esto falso.
    ETII: Se rechaza que el gasto promedio mensual, en consumo por familia, sea superior a S/. 100, siendo esto verdadero.
    b) $t_0 = 19.57$; p value = 0; $t_C = 2.05913$. Se rechaza H_0
14. RH_0, El peso promedio del arroz es menos de 995 gramos. $Z_0 = -4.44$; $Z_C = 1.645$.
15. $t_0 = -0.91$; p value = 0.338; $t_c = -1$ $\text{R.A} = <-1.86; 1.86>$.

    Aceptar H_0, la venta de franquicias en promedio durante el primer año produce un rendimiento del 12%.
16. Rechazar H_0, por lo tanto el somnífero aumenta las horas de sueño. $t_0 = 3.2$; p value = 0.005; $t_C = 2.822$.
17. a) Aceptar H_0, por lo tanto el Royal Kola S.A. no tiene razón. $Z_0 = 0.8$;
    p value = 0.212.
    b) $\beta = 0.89359$ c) $1-\beta = 0.258459$
18. a) $Z_0 = -0.69$; $Z_C = -1.64485$. Aceptar H_0, por lo menos la mitad de los compradores recuerdan el precio correcto.

    b) p value = 0.245 c) Potencia de prueba = 0.312164
19. Aceptar H_0, el porcentaje de polos tipo exportación no tiene menos fallas que el porcentaje tipo nacional; $Z_0 = -0.97$; $Z_C = -2.05$

20. a) Aceptar H_0, la producción de leche es a lo más 390 litros.

p value $= 0.726$. $1-\beta = 0.14$

b) $t_0 = -1.89$; p value $= 0.04$; $t_C = -1.76131$. Rechazar H_0, el promedio de calcio en la leche es menos de 1.05%.

c) $t_0 = 2.1$; p value $= 0.028$; $t_C = 2.65031$. Aceptar H_0, el promedio de producción de
leche de la raza A no es mayor al promedio de producción de leche de la raza B.

d) $F_0 = 1.59244$; $F_C = 4.81832$. Aceptar H_0, la varianza del calcio de la leche de las vacas de A no es mayor que la varianza del calcio de la leche de las vacas del tipo B.

21. $t_0 = -0.88$; p value $= 0.197$; $t_C = -1.76131$. Aceptar H_0, el promedio de las notas de los alumnos que no asistieron a cursos de preparación no es mayor a los que asistieron.

22. a) $H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

$F_0 = 3.38$; RA $= \langle 0.3829, 2.6929 \rangle$.

Conclusión: Las varianzas son diferentes.

b) $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 < \mu_2$

$t_0 = -1.86$; $RC = \langle -\infty, -1.6076 \rangle$.

Conclusión: Se rechaza H_0, las ventas con salvaguardias son superiores a las sin salvaguardias.

c) $H_0 : \mu_2 = 35$

$H_1 : \mu_2 > 35$

$t_0 = 1.0065$; $RC = \langle 1.64865, \infty \rangle$.

Conclusión: se acepta H_0, las ventas promedio mensuales no son mayores de 35 000 dólares.

23. a) $Z_0 = -2.24$; p value $= 0.011$; $Z_C = -1.64485$. Rechazar H_0, las ventas en promedio es menos de 4900.

b) $t_0 = 3.74$; p value $= 0.01$; $t_0 = 3.74695$. Aceptar H_0, el promedio de ventas de los sábados no es mayor al promedio de las ventas de los miércoles.

c) $F_0 = 1.59244$; p value $= 0.318$; $RA = \langle 0.005, 18.51 \rangle$. Aceptar H_0, la variabilidad de las regalías en ambos días son homogéneas.

24. a) i) 0.066807 ii) 0.308779

b) Primero se hace la prueba de varianzas para determinar si son iguales o diferentes.

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

$F_0 = 0.04$; p value = 0; por lo tanto las varianzas son diferentes.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

$t_0 = -0.89$; p value = 0.799; se acepta H_0 y por lo tanto el promedio de ventas en el local de San Juan es igual al local de Los Olivos.

c) $H_0 : \pi_1 = 0.5$

$H_1 : \pi_1 > 0.5$

$Z_0 = -2.77$; p value = 0.997

Se acepta H_0 y no se puede afirmar que el 50% de todas las ventas de la empresa sea superior a 30 000 nuevos soles.

25. a) i) $H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

ii) $F \sim F_{(16.14)}$; y $F_0 = 0.347906$

iii) Se acepta H_0 si $F_0 \in \langle 0.35499, 2.92339 \rangle$

iv) Rechaza H_0, $F_0 \in RC$; por lo tanto las varianzas poblacionales son diferentes.

b) i) $H_0 : \mu_2 = 4$

$H_1 : \mu_2 < 4$

ii) $t \sim t_{(14)}$; y $t_0 = -1.29$

iii) Se rechaza H_0 si: $t_0 < t_{(0.05,15)} \Rightarrow t_0 < -1.76131$

iv) Se acepta H_0, el consumo promedio de gasolina es de cuatro galones.

26. a) i) $H_0 : \mu_2 = 220$

$H_1 : \mu_2 \neq 220$

ii) $t \sim t_{(15)}$, y $t_0 = -0.07$

iii) Se rechaza H_0 si: $|t_0| > t_{(0.025,15)} \Rightarrow |t_0| > 2.13145$

iv) Aceptar H_0, el estabilizador tiene un consumo promedio de 220 voltios.

b) i) $H_0 : \sigma^2 = 25$

$H_1 : \sigma^2 > 25$

ii) $\chi_0^2 \sim \chi_{(15)}^2$; $\chi_0^2 = 31.512$

iii) Se rechaza H_0 si: $\chi_0^2 > \chi_{(0.95,15)}^2 \Rightarrow \chi_0^2 > 24.9958$

iv) Se rechaza H_0, la desviación estándar es mayor de cinco voltios.

27.

μ	Potencia de la prueba	β
255	0.342712	0.657288
256	0.418758	0.581242
257	0.497977	0.502023

28. Pregunta 1:
a) –1 7.1; –1.38; 0.084; 243.
b) $\alpha > 0.084$

Pregunta 2:
Primero se hace la prueba de varianzas para determinar si son iguales o diferentes.

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

$F_0 = 0.22$; p value = 0; por lo tanto las varianzas son diferentes.

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

$t_0 = -0.26$; p value = 0.602.
Se acepta H_0, por lo tanto, en el local de San Isidro no se vende más en promedio que en el de Surco.

Pregunta 3:
a) Las hipótesis que se deben probar son:
H_0: Los montos de las ventas de Miraflores se distribuyen como normal.
H_1: Los montos de las ventas de Miraflores no se distribuyen como normal.
Se rechaza H_0 si: $\chi_0^2 > \chi_{(0.94,3)}^2 \Rightarrow \chi_0^2 > 7.40688$

b) Usando la tabla siguiente se tiene que: $\chi_0^2 = 91.9188$, por lo tanto se rechaza H_0, se concluye que las ventas en Miraflores no se distribuyen como normal.

[soles >	Oi	Pi	Ei
60 - 140	21	0.086	10.51
140 - 220	44	0.3368	40.7559
220 - 300	26	0.4029	48.7465
300 - 460	30	0.1659	20.0764

Nota: Las dos últimas categorías se juntaron en una sola por que la última frecuencia esperada era menor que 5.

Pregunta 4:

Las hipótesis que se deben probar son:

H_0: El local donde se vende comida es independiente del tipo de comida.

H_1: El local donde se vende comida no es independiente del tipo de comida.

Se rechaza H_0 si: $\chi_0^2 > \chi^2_{(0.93,6)} \Rightarrow \chi_0^2 > 11.6599$

$\chi_0^2 = 5.647$; $p - value = 0.464$; $GL = 6$

Se acepta H_0 y se concluye que el local es independiente del tipo de comida.

29. Pregunta 1:

a) a.1 $H_0 : \mu = 15$

$H_1 : \mu > 15$

a.2 $Z_0 = 0.7448$

a.3 Se rechaza H_0 si $Z_0 > 1.55477$.

a.4 Se acepta H_0, el promedio no es mayor que 15.

b) $\alpha = P(\bar{x} > 15.2 | \mu = 15) = 0.0296731$; donde $\bar{x} \sim N(15,\ 9/800)$

c) $\beta = P(\bar{x} < 15.1649 | \mu = 15.3) = 0.1013926$; donde $\bar{x} \sim N(15.3,\ 9/800)$

d) $n = 2459$

Pregunta 2:

Primero se hace la prueba de varianzas para determinar si son iguales o diferentes.

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

$F_0 = 0.9$; p value = 0.514; por lo tanto las varianzas son iguales.

a) $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 < \mu_2$

b) $t_0 = -1.57$

c) Se rechaza H_0 si $t_0 < t_{(0.03,356)}$, entonces: $t_0 < 1.88681$

d) Se acepta H_0 y por lo tanto los pesos promedio vendidos de ambos quesos son los mismos.

Pregunta 3:

a) $H_0 : \pi_1 = \pi_2$

$H_1 : \pi_1 < \pi_2$

b) $Z_0 = -1.15$

c) Región crítica: RC = $\langle -\infty, -1.75069 \rangle$

d) Se acepta H_0, y se concluye que la proporción de quesos de España usados en repostería es la misma que la de Francia.

30. Reporte I:

a) 4.8 , 0.03 b) Las varianzas son homogéneas

Reporte II:

a) 3.5442 , 0.002 , 17

b) Los guachimanes de la empresa Top S.A. trabajan en promedio más horas diarias que los guachimanes de la empresa Vipsa.

31. Pregunta 1:

$$n=\left(\frac{Z_{0.99}\sigma}{E}\right)^2=\left(\frac{2.32635(0.5)}{0.98}\right)^2=34$$

Pregunta 2:

a) Las hipótesis que se deben probar son:

H_0: Los datos se ajustan a lo que señala el instituto.

H_1: Los datos no se ajustan a lo que señala el instituto.

b) Usando la tabla siguiente se tiene que: $\chi_0^2=77.0833$.

Sector	Oi	Pi	Ei
Agropecuario	30	0.15	36
Pesca	30	0.10	24
Minería	30	0.25	60
Manufactura	30	0.05	12
Electricidad	30	0.05	12
Construcción	30	0.18	43.2
Comercio	30	0.10	24
Otros	30	0.12	28.8

c) Se rechaza H_0 si: $\chi_0^2>\chi^2_{(0.97,7)} \Rightarrow \chi_0^2>15.5091$

d) *p-value* = 0.00

e) Se rechaza H_0 y se concluye que los datos no se distribuyen como afirma el instituto.

Pregunta 3:

Primero se hace la prueba de varianzas para determinar si son iguales o diferentes.

$H_0:\sigma_1^2=\sigma_2^2$

$H_1:\sigma_1^2\neq\sigma_2^2$

$F_0=1.66$; p value = 0.178; por lo tanto las varianzas son iguales.

a) $H_0:\mu_1=\mu_2$

$H_1:\mu_1\neq\mu_2$

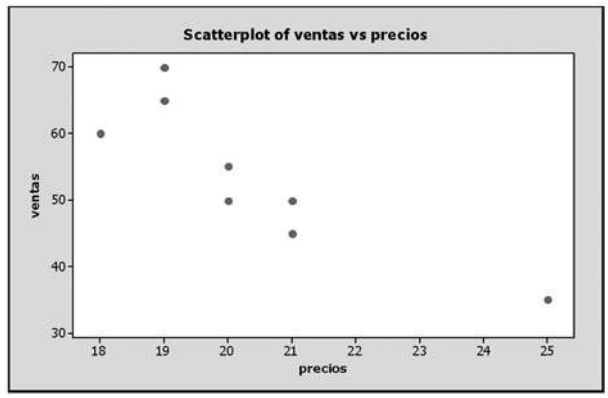
b) $t_0=0$; p value = 1; $s_p=3.5878$; G.L. = 58

c) Se rechaza H_0 si $|t_0|>t_{(0.99,58)} \Rightarrow |t_0|>2.39238$

d) Se acepta H_0 y por lo tanto no existen diferencias entre los promedios estimados del año 2006 en los sectores comercio y manufacturas en caso de no firmarse el TLC.

Capítulo 4

1. Falso.
2. Verdadero.
3. Falso.
4. Verdadero.
5. a)



b) Ventas = 147 – 4.59 Precio; c) Ventas = 46.02 kilos.

6. a) Consumo = –0.149 + 0.100 Cilindros;
   b) El grado de asociación entre consumo y el número de cilindros de los automóviles es 0.95.
   c) $t_0 = 9.88$, p value = 0, se rechaza la hipótesis nula, el número de cilindros influye en el consumo de gasolina (S/).

7. a) Holgura = - 0.4949 + 0.041067 número de placas.
   b) Se tiene $t_0 = 6.47$, p value = 0. Se rechaza la hipótesis nula, el número de placas influye en la holgura de estas.
   c) El intervalo de confianza para la pendiente: $(0.026 < \beta_1 < 0.057)$
   d) Se tiene $F_0 = 41.84$, p value = 0. Se rechaza la hipótesis nula al nivel del 5%.

8. a) Coeficientes:
   β_0: Cuando el tiempo permanece constante la eficiencia es de 18.1%.
   β_1: Por cada minuto que pase la eficiencia se incrementa en 1.42%.
   b) El 91.6% de la variación de la eficiencia se debe al tiempo de extracción.
   c) La pendiente de la recta es significativamente distinta de cero, con $\alpha = 0.05$, $t_0 = 9.32$y un p-value = 0.

9. a) $\beta_0 = 8.27054264$,si el tiempo de capacitación permanece constante el tiempo de eficiencia promedio será de 8.2705. $\beta_1 = 0.39302326$, por cada hora que se incremente en la capacitación la eficiencia se incrementa en 0.3930.
   b) $t_0 = 3.114$; RA = ⟨–2.11991, 2.11991⟩; se rechaza la hipótesis nula.

c) $\langle 14.818,\ 17.44\rangle$. d) Coeficiente de Correlación: $r = 0.6143$. Existe una relación entre el tiempo de eficiencia y el tiempo de capacitación.

Coeficiente de determinación: $R^2 = 0.3773$. Indica que el 37.73% de la variación de la eficiencia se debe al tiempo de capacitación.

e) $\langle 8.47,\ 19.87\rangle$

10. a) β_4: Cada vez que hay charla de un profesor, la valoración promedio se incrementa en 6.21

b) Prueba de hipótesis para β_4 $H_0 : \beta_4 = 0$

$H_1 : \beta_4 > 0$

$t_0 = 1.7298,\ RC = \langle 1.7247, \infty\rangle$

Se rechaza a 5% de nivel de significancia.

c) El 67.5% de las variaciones de la puntuación promedio se debe al porcentaje del tiempo que dedicó al trabajo en grupo sobre la duración del curso, dinero invertido en material del curso, en dólares por miembro del grupo, al dinero invertido en comida y bebida, si hubo una charla del profesor.

d) El dinero invertido en material del curso es: $0.08658 < \beta_2 < 0.9534$

11. a) La ecuación de la recta es: $Y = 42.7 + 8.09X_1 + 0.831X_2 + 2.67X_3$

b) Variables significativas

Predictor	Coef	SE Coef	T	P
Constant	42.65	55.59	0.77	0.468
x1	8.089	1.661	4.87	0.002
x2	0.8306	0.3462	2.40	0.048
x3	2.6665	0.6893	3.87	0.006

Al 5% X_1, X_2, X_3 son significantes.

c) La resistencia estimada es: 221.82.

d) El 89.6% de las variaciones de la resistencia a la tensión se debe al tiempo de secado, la temperatura de secado y al porcentaje de algodón.

12. a) Siendo constantes el número de horas a la semana que se dedicaron a estudiar, el número medio de horas que se dedicaron a prepararse para los exámenes, el número de horas a la semana que se fueron a divertirse y el número medio efectivo de clase que tuvieron por semestre. Por cada vez que el alumno subrayaba el texto la nota media se incrementa en 0.277.

b) Prueba de hipótesis para β_4

$H_0 : \beta_4 = 0$

$H_1 : \beta_4 > 0$

$t_0 = 1.2847,\ RC = \langle 1.7472,\ \infty\rangle$

Se acepta la hipótesis nula.

c) El 50% de la variación de la nota media se debe a las variaciones de número de horas que se dedicaron a estudiar, número medio de horas que se

dedicaron a preparar exámenes, número que se fueron a divertir, si subrayaban el texto o no, y el número medio efectivo de clases.

d) El número de horas de diversión al 95% de confianza es $-0.35317 < \beta_3 < -0.10883$

e) Las variables X_1, X_2, X_3, X_4 no son significantes al modelo a un 5% de nivel de significancia.

f) $F = 4.21$, p value = 0.008, se rechaza la hipótesis nula, el modelo es idóneo.

g) Si las variables independientes permanecen constantes, entonces por cada hora de diversión, el promedio de notas de la carrera disminuye en 0.231.

13. La ecuación de la regresión es:

P = – 1 03 + 0.605 Temperatura + 8.92 Número de días + 1.44 Pureza + 0.014 Toneladas del producto.

El consumo de potencia estimado es: 287.56; <234.88, 340.24>

14. a) La línea de regresión es: Demanda = 84.3 – 0.0397 Precio + 0.0142 Ingreso

b) El 87.6% de las variaciones de la demanda se debe a las variaciones del precio y del ingreso.

c) La demanda estimada es 64.

d) $F = 24.63$; p value = 0.001. Se rechaza la hipótesis nula, el modelo es idóneo.

15. a) La ecuación de la regresión es:

Ventas = 219 + 6.38 Publicidad (número de anuncios) – 1.67 Precio

β_0 Si la publicidad y precio permanece constante entonces las ventas en promedio es de 219.

β_1 Si el precio permanece constante, por cada anuncio, las ventas aumentan en 6.38

β_2 Si la publicidad permanece constante, por cada $ que se incremente en el precio de la grabadora las ventas disminuyen en 1.67.

b) Las ventas estimadas son de 43 grabadoras.

16. a) Prueba global

```
Analysis of Variance
Source           DF        SS        MS      F       P
Regression        2    77.455    38.727  11.23   0.007
Residual Error    7    24.145     3.449
Total             9   101.600
```

Se rechaza la hipótesis nula, el modelo en su conjunto es idóneo.

b) β_2: Si las horas de trabajo en auditoría de campo permanece constante; entonces, por cada empresa evasora el importe mensual de los impuestos no pagados se incrementa en 0.665.

c) El promedio de los importes mensuales de los impuestos no pagados se encuentra entre 27.732 y 30.525.

d) Variables significativas

Predictor	Coef	SE Coef	T	P
Constant	-17.40	12.10	-1.44	0.194
X1	-0.00773	0.01135	-0.68	0.518
X2	0.6653	0.2046	3.25	0.014

La variable X_1 no es significante.

e) El 76.2% de las variaciones de los importes no pagados se debe a las variaciones de horas de trabajo en auditoría de campo y número de empresas evasoras.

f) No aumentará en 800 000 dólares.

17. a) La ecuación de la regresión es:
Renta ($) = – 20 + 132 (número de habitaciones) – 5.0 (distancia desde el centro)
$\beta_1 = 132$ Si la distancia permanece constante, entonces por cada habitación adicional la renta aumenta en $ 132
$\beta_2 = -5$ Si el número de habitaciones permanece constante, entonces por cada kilómetro adicional que se aleje de la ciudad la renta disminuye en $5.

b) Esperaría pagar $232.7.

18. a) La ecuación de la regresión es de:

$$Y = -1635 + 15.0X_1 - 2.14X_2 + 1.48X_3 - 114X_4$$

El error de estimación es: S = 0.195766

b) R2 = 100%; c) Intervalo de confianza para Y: ⟨–5.2796, 8.2587⟩

19. a) La ecuación de regresión es de:

Demanda = 6.1 – 0.14 (Precio) + 0.0280 (Ingreso) + 1.94 (Sustituto)

β_1 : Si el ingreso y el sustituto son constantes, entonces por cada dólar que se incremente en los artefactos la demanda disminuye en 0.14.
β_2 : Si el precio y el substituto son constantes, entonces por cada dólar de incremento en los ingresos la demanda se incrementa en 0.028.
β_3 : Si el precio y el ingreso son constantes, entonces por cada dólar que se emplee en el substituto la demanda se incrementa en 1.94.

b) Corresponde a lo que se espera en las variables precio y el ingreso.

c) R^2 * 100% = 71.3% de las variaciones de la demanda se debe a las variaciones del precio, ingreso y del precio del substituto.

d) El error estándar de estimación es: = 8.24316.

e) La demanda estimada es 72.

20. a) La ecuación de regresión es:
Precios de ventas = 27.1 + 0.196 (metros cuadrados) + 5.90 (pisos) + 1.73 (edad)
β_1 : Si el número de pisos y la antigüedad permanecen constantes, entonces por cada metro cuadrado de incremento el precio de venta se incrementa en $196.
β_2 : Si la cantidad en metros cuadrados y antigüedad permanece cons-

tante, entonces por cada piso adicional de la vivienda el precio de venta se incrementa en $5900.

β_3 : Si los metros y el piso se mantienen constantes, entonces por cada año adicional el precio de venta se incrementa en $1,730.

b) $R^2 * 100\% = 96.4\%$ de las variaciones de los precios de ventas de las casas se debe a las variaciones de la superficie, pisos y antigüedad.

c) El precio de venta estimado es de $70 830.

21. a) Ecuación de regresión:

Ventas ($) = 138 + 28.4 Promoción ($) – 10.2 Competidores – 4.67 Gratis.

β_1: Si el número de aerolíneas y el porcentaje de pasajeros que viaja gratuitamente permanece constante; por cada dólar empleado en la promoción las ventas se incrementan en $28400.

β_2: Permaneciendo constantes la inversión en promoción y el porcentaje de pasajeros que viaja gratuitamente; por cada aerolínea que se incrementa, las ventas disminuirán en $10200.

β_3: Si la promoción y las aerolíneas competidoras permanecen constantes entonces, por cada punto porcentual que se incremente las ventas disminuyen en $ 4670.

b) Prueba de hipótesis para β_3:

$H_0 : \beta_3 = 0$

$H_1 : \beta_3 \neq 0$

$t_0 = -1.87$, p value $= 0.158$

Se acepta la hipótesis nula al 5% de nivel de significancia, esta variable no es significante.

c) Prueba de hipótesis para β_2

$H_0 : \beta_2 = 28$

$H_1 : \beta_2 \neq 28$

$t_0 = 0.07575 \quad RC = 1,2.9199 \quad \text{R.A} = \langle -2.353, 2.353 \rangle$

Se acepta la hipótesis nula, el cambio no es significativamente diferente de 28,000.

d) Intervalo de confianza para $\beta_2 = \langle -20.738, 0.33 \rangle$

22. a) Ecuación de regresión:

Gastos (miles de dólares) = 0.6 + 0.219 (camas) + 0.87 (admisiones)

b) Los gastos estimados son de $169.470.

23. a) La señorita Ángeles consume 11.8383 litros de agua mineral semanal mente.

b) Intervalo de confianza para consumo de agua mineral: 11.6286, 12.1366

c) Variabilidad explicada por el modelo es 0.1%.

24. a)

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	Valor F	P value
Regresión	3	16581.859	5527.28		
Error	95	7839.141	82.52	66.981	0
Total	98	24421.00	----------	----------	----------

b) $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
H_1 : Al menos un β_i es diferente de cero

$F_0 = 66.98$, p value = 0; el modelo en conjunto es apropiado.

c) $t_0 = 9.6105$; $P(t_{95} > 9.6105) = 1 - P(t_{95} < 9.6105) = 1 - 1 = 0$; p value = 0

25. a) El modelo de la regresión es:
$Y = 1.40 + 0.101X$
Por cada mil registros que se lea, el número de operaciones de entrada y salida se incrementa en 101.

b) El coeficiente de correlación:
El número de operaciones de entrada y salida está fuertemente asociada con el número de registros leídos con un valor de 0.99196.
Coeficiente de determinación:
El 98.4% de las variaciones del número de operaciones de entrada y salida se debe a las variaciones del número de registros leídos.

c) La estimación es: ⟨34.860, 43.706⟩

d) $H_0 : \beta_1 = 0.12$; $t_0 = \dfrac{0.101014 - 0.12}{0.003570} = -5.318$; $t_{(tabulada)} = -1.89887$
$H_1 : \beta_1 < 0.12$
Se rechaza la hipótesis nula, no se puede afirmar que al aumentar X en una unidad, Y aumenta más de 0.12.

e) $H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$
Si el modelo es adecuado $F_0 = 800.83$, p value = 0

26. a) El modelo de regresión es:
precio = 20.0 + 0.0695 páginas
El precio promedio es de 20 dólares, permaneciendo constante el número de páginas.
Por cada página que se incremente en el libro, el precio aumenta en 0.0695.

b) El 16.4% de las variaciones del precio de los libros se debe a las variaciones del número de páginas.

c) 95% CI ⟨31,84, 56.82⟩

d) $Ho : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$
$t_0 = 1.40$; p value = 0.191; el coeficiente de la regresión es igual a cero.

e) $H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

$F_0 = 1.97$; p value $= 0.191$

El modelo planteado en su conjunto no es adecuado al 0.05.

27. a) El modelo de regresión es:

$Y = 10.1 - 0.201X$

Si el tiempo de fricción permanece constante, el grosor de la lámina en promedio es de 10.
Por cada hora que se emplee en la fricción el grosor de la lámina disminuye en 0.201.

b) Coeficiente de correlación:
Hay una correlación entre el tiempo de fricción y el grosor de la lámina de 0.9975.
Coeficiente de determinación:
El 99.5% de las variaciones del grosor de lámina se debe a las variaciones del tiempo de fricción.

c) El intervalo es: $\langle -0.201 \mp 2.44898(0.004298) \rangle = \langle -0.2115, -0.1905 \rangle$

d) 99% CI $\langle 7.3765,\ 7.5493 \rangle$

e) $F_0 =$ 1665.37; $F_{tabulada} =$ 5.46712; como $F_0 > F_{tabulada}$, se rechaza la hipótesis nula.

28. a) β_0 no tiene interpretación ya que una producción no puede ser negativa.
β_1 (Pendiente): Por cada punto que se incremente en el resultado de la evaluación la producción se incrementa en 0.24369.
R^2 (Coeficiente de determinación): el 66.2% de las variaciones de la producción es explicada por las variaciones de los resultados de la prueba de aptitud.

b.1) El nuevo modelo de regresión es:

$Y = -13.8 + 0.212X_1 + 2.00X_2$

β_0 no tiene interpretación ya que una producción no puede ser negativa.

β_1: Por cada punto que se incremente en el resultado de la evaluación la producción se incrementa en 0.212, si los años de experiencia permanecen constantes;

β_2: Por cada año de experiencia que aumente, la producción se incrementa en 2, si el resultado de la evaluación permanece constante.

b.2) $H_0 : \beta_2 = 0$; $t_0 = 13.73$; p value $= 0$,
$H_1 : \beta_2 \neq 0$
años de experiencia es significativa al modelo de estimación; aumenta el R^2.

Capítulo 5

1. a) Sí hay diferencia en la producción de los telares: $F_0 = 5.77$;
   p value $= 0.003$
   Unidad experimental: telar.
   Factor: cantidad de libras por minuto.
   Niveles: cinco tipos de telares.

   b) $\bar{Y}_{..} = 3.936$ $\quad \hat{\tau}_1 = 0.144$ $\quad \hat{\tau}_3 = 0.124$ $\quad \hat{\tau}_5 = -0.116$

   $\hat{\tau}_2 = -0.016$ $\quad \hat{\tau}_4 = -0.136$

   c) $3.967 < \mu_1 < 4.193$ $\quad 3.947 < \mu_3 < 4.173$ $\quad 3.707 < \mu_5 < 3.933$

   $3.807 < \mu_2 < 4.033$ $\quad 3.687 < \mu_4 < 3.913$

2. Hay diferencia en el promedio vendido en los cuatro restaurantes con un nivel de significancia del 1%; $F_0 = 16.33$; p value $= 0$.

3. Existen diferencias en los precios en los supermercados. $F_0 = 256.59$; p value $= 0$.

4. Existen diferencias en el promedio de pasajeros en las tres líneas aéreas. $F_0 = 100.32$; p value $= 0$

5. a) Variable: Aprendizaje

   Diseño: Completamente aleatorio

   b) La tabla queda de la siguiente manera:

F.V.	G.L.	SC	MS	F	P
Entre grupos	3	229.89	76.63	3.938	0.0279
Dentro de grupos	16	311.31	19.4568		
Total	19	541.20			

   c) Se rechaza la hipótesis nula al 10% de nivel de significancia.
   Hay diferencia en el aprendizaje entre las cuatro técnicas empleadas.

6. Sí hay diferencias entre los métodos enseñados. $F_0 = 19.51$; p value $= 0$
   La estrategia más efectiva es estudios de casos, el promedio es de 16.100.

7. a) La unidad experimental es el centro comercial.

   b) $\bar{Y}_{3.} = 211.67$ y $\bar{Y}_{5.} = 256.67$

   c) Hipótesis:

   $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

   $H_1 :$ Al menos un μ_i es diferente

   $F_0 = 6.98$; $\quad RA = \langle 0.0891993, 3.80667 \rangle$

   Se rechaza la hipótesis nula; las ventas son distintas en los centros comerciales.

d) Las ventas de Rey de Gamarra no son mayores que las del Jockey Plaza, ambos tienen ventas muy parecidas $\langle -8.99, 118.99 \rangle$

e) $(111.87 < \mu_4 < 234.79)$

8. a) Se concluye que la viscosidad en los tres tipos de pectina es distinta.
   b) Es posible afirmar que viscosidad es igual cuando se coloca el 1% de pectina que cuando se usa el 0.25%. (–1.94; 4.454)

9. a) Se rechaza la hipótesis nula a un nivel de significancia de 0.01.
   b) $\bar{Y}_{..} = 6.06429$ $\hat{\tau}_1 = -1.47857$ $\hat{\tau}_3 = 0.79286$
   $\hat{\tau}_2 = 0.97857$ $\hat{\tau}_4 = 0.97871$
   c) $(4.647 < \mu_1 < 6.879)$ $(3.46225 < \mu_3 < 5.70975)$
   $(5.73325 < \mu_3 < 7.98075)$ $(5.91925 < \mu_5 < 8.16675)$
   d) El contenido de tiamina en ambos productos —trigo y avena— es el mismo $\langle -0.147, 2.689 \rangle$.

10. a) Unidad experimental: un voluntario
    Tratamiento: Tipos de mantequilla
    b) $\bar{Y}_{..} = 220.278$ $\hat{\tau}_1 = -16.994$ $\hat{\tau}_3 = -9.722$
    $\hat{\tau}_2 = 6.389$ $\hat{\tau}_4 = 20.282$
    c) No es el mismo. $F_0 = 5.36$; p value $= 0.025$.
    d) Sí, el intervalo es: $\langle -0.10, 46.76 \rangle$

11. a)

Source	Df	Ss	Ms	F
Categoría	2	952	476	4.602
Error	99	10240	103.43	
Total	101	11192		

Valor crítico: $F_0 = 4.07076$
Conclusión: se rechaza la hipótesis nula, el rendimiento del personal es distinto de la universidad de donde procede, con un error del 2%.

b)

Level	N	Mean	StDev
Excelente	34	72.38	11.15
Buena	34	67.47	10.09

Intervalo de confianza de excelentes y buenos

$$(72.38 - 67.47) \pm t_{(0.99,99)} \sqrt{\frac{2(103.43)}{34}} = 4.91 \pm 2.36461(2.4666)$$

$-0.9225 < \mu_{excel} - \mu_{Bueno} < 10.742$; el rendimiento en promedio es igual.

12. One-way ANOVA: TÉCNICA 1, TÉCNICA 2, TÉCNICA 3

Source	DF	SS	MS	F	P
Factor	2	0.93	0.47	0.13	0.876
Error	12	41.80	3.48		
Total	14	2.73			

Se comprueba que las tres técnicas en promedio *proporcionan resultados similares*, valor del p value = 0.876 es mayor que el nivel de significancia, por lo tanto se acepta la hipótesis nula.

13. One-way ANOVA: Prueba 1; Prueba 2; Prueba 3

Source	DF	SS	MS	F	P
Factor	2	250	125	1,05	0,362
Error	27	3208	119		
Total	29	3458			

Se comprueba que las tres pruebas aportan resultados similares, al 5% de error. El valor del p value = 0.362 mayor que el nivel de significancia.

14. a) One-way ANOVA: D1, D2, D3, D4

Source	DF	SS	MS	F	P
Factor	3	2477.141	825.714	2495.41	0.000
Error	44	14.559	0.331		
Total	47	2491.700			

Sí existe diferencia en los cuatro diseños p value = 0 es menor que el nivel de significancia.

b) Todos son distintos

15. a) 2.54074 b) 0.81869 c) LI: -46.2635 LS: 24.0635 d) 2200.8

16. a) One-way ANOVA: Lima, Trujillo, Arequipa

Source	DF	SS	MS	F	P
Factor	2	4.289	2.145	12.97	0.000
Error	27	4.466	0.165		
Total	29	8.755			

Sí existe diferencia en las tres provincias, el p value = 0 es menor que el nivel de significancia, por lo tanto se rechaza la hipótesis nula.

b) Los precios son similares en Trujillo vs Lima $\langle -0.4569, 0.4426 \rangle$

17. One-way ANOVA: 15%, 20%, 25%, 30%

Source	DF	SS	MS	F	P
Factor	3	358.00	119.33	17.48	0.000
Error	16	109.20	6.83		
Total	19	67.20			

S = 2.612 R-Sq = 76.63% R-Sq(adj) = 72.24%

F_0 = 17.48; p value = 0, se rechaza la hipótesis nula el porcentaje de algodón influye en la resistencia a la tensión de la fibra.

18. a) Unidad de análisis = un ladrillo; variable = densidad.

b) One-way ANOVA: 100°F, 125°F, 150°F, 175°F

Source	DF	SS	MS	F	P
Factor	3	2.435	0.812	3.25	0.045
Error	19	4.740	0.249		
Total	22	7.175			

F_0 = 3.25; p value = 0.045, la temperatura no afecta la densidad del ladrillo.

19. a) Unidad de análisis = un día de venta; variable = ventas.
b) F_0 = 28.06; p value = 0; se rechaza la hipótesis nula, el tipo de jamón influye en las ventas

20. a) F_0 = 90.74; p value = 0; se rechaza la hipótesis nula, el tipo de tratamiento clínico influye en el rendimiento de los pacientes.

b) $48 \mp 2.08596 * 2.37697 = 48 \mp 4.9583 = \langle 48.21, 58.13 \rangle$

$55.83 \mp 4.9583 = \langle 50.87,\ 60.79 \rangle$

$97.17 \mp 4.9583 = \langle 92.21,\ 102.13 \rangle$

El que tiene mejor rendimiento es el tratamiento "D"

21. a)

Fuente de variación	D.F	SS	MS	F
Factor	4	2519.143	629.785	7.0513
Error	30	2679.43	89.314	
Total	34	5198.57		

Se rechaza la hipótesis nula; $F_0 = 7.0513 > F = 3.10$

b) El promedio de ambos consumos son iguales $(-4.2247 < \mu_3 - \mu_1 < 18.7297)$

22. a) $F_0 = 3.0695$ $F_0 = 3.85003$ $F_0 < F$; no se rechaza la hipótesis nula, en promedio los consume de pescados son iguales en los cuatro distritos de Lima metropolitana con un α = 0.03.

b) $(15.60 - 20.40) \mp t_{(0.995;16)} \sqrt{\frac{2 * 13.3517}{5}} = -4.8 \mp 5.5037$

El intervalo de confianza es: $-10.3037 < \mu_{ate} - \mu_{Comas} < 0.7037$, el consumo promedio en Ate y Comas es igual al 3% de significancia.

23. a) Unidad de análisis = Una persona consumidora de agua mineral, Variable = Gasto semanales agua mineral, Tratamiento = marca de agua mineral

b) $H_o = \mu_1 = \mu_2 = \mu_3$

H_1 = al menos dos son diferentes

$F_0 = 0.91$; p value = 0.403

No se rechaza la H_0, el gasto promedio es el mismo para las tres marcas.

c) $\langle -0.432, 0.345 \rangle$

24. a) Unidad de análisis =1 cadena de venta, Tratamiento = Marcas de combustible.

b) One-way ANOVA: Móbil Oil, Repsol, YPF, Primax

Source	DF	SS	MS	F	P
Factor	3	1.994	0.665	4.17	0.023
Error	16	2.550	0.159		
Total	19	4.544			

Sí existe diferencia de precios en las cadenas; F_0 =4.17 y p_value = 0.023 menor que el nivel de significancia.

c) $-1.0123 < \mu_{YPF} - \mu_{\text{Re}psol} < 0.2923$

25. a) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
H_1: al menos una media es diferente

b) $F_0 = 1.55$

c) Si $F_0 > 3.3817$, se rechaza la H_0, en caso contrario no se rechaza.

d) p value = 0.204 e) No se rechaza la hipótesis nula con $\alpha \approx 0.02$.

26. a) Unidad experimental = Un paciente del hospital Sabogal, Tratamiento el tipo de fármaco

b) One-way ANOVA: P, A, B, AB

Source	DF	SS	MS	F	P
Factor	3	2492.6	830.9	8.53	0.000
Error	31	3020.9	97.4		
Total	34	5513.5			

Sí hay diferencia significativa en el tipo de fármaco F_0 = 8.53, p value = 0, menor que el nivel de significancia.

27. a) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
H_1: Al menos dos μ_i son diferentes

Poblaciones normales y muestra aleatoria independiente

b) way ANOVA: Backus, Abbot Ale, Brahman, Stella Artois, Beck's

Source	DF	SS	MS	F	P
Factor	4	241.066	60.266	108.31	0.000
Error	20	11.128	0.556		
Total	24	252.194			

Existen diferencias significativas en las ventas de las marcas, F_0 = 108.31.

c) Existen diferencias entre ambos colores F_0 = 93.27; p value = 0

d) (6.4959, 8.4641)

28. a) $Ho: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: Al menos un μ_i es diferente

Poblaciones normales y muestra aleatoria independiente.

b) One-way ANOVA: dosis-25, dosis-50, dosis-75, dosis-100

Source	DF	SS	MS	F	P
Factor	3	13.281	4.427	6.83	0.004
Error	16	10.364	0.648		
Total	19	23.646			

Hay diferencia significativa, $F_0 = 6.83$; p value $= 0.004$ menor que el valor de significancia.

c) $28.73 < \mu_{dosis-75} < 30.189$

29. a)

Fuente de variación	D.F	SS	MS	F	P
Factor	3	481748.198	160582.73	28.925	0
Error	20	111034.802	5551.74		
Total	23	592783			

b) Sí hay diferencia significativa en las 4 marcas de impresora con un nivel de significancia del 3%.

c) $-41.431 < \mu_{M1} - \mu_{M3} < 138.03$.

Anexos

Anexo 1
Tabla de números aleatorios

07018	31172	12572	23968	55216	85366	56223	09300	94564	18172
52444	65625	97918	46794	62370	59344	20149	17596	51669	47429
72161	57299	87521	44351	99981	55008	93371	60620	66662	27036
17918	75071	91057	46829	47992	26797	64423	42379	91676	75127
13623	76165	43195	50205	75736	77473	07268	31330	7337	55901
27426	97534	89707	97453	90836	78967	00704	85734	21776	85764
96039	21338	88169	69530	53300	29895	71507	28517	77761	17244
68282	98888	25545	69406	29470	46476	54562	79373	72993	98998
54262	21477	33097	48125	92982	98382	11265	25366	06636	25349
66290	27544	72780	91384	47296	54892	59168	83951	91075	04724
53348	39044	04072	62210	01209	43999	54952	68699	31912	09317
34482	42758	40128	48436	30254	50029	19016	56837	05206	33851
99268	98715	07545	27317	52459	75366	43688	27460	65145	65429
95342	97178	10401	31615	95784	77026	33087	65961	10056	72834
38556	60373	77935	64608	28949	94764	45312	71171	15400	72182
39159	04795	51163	84475	60722	35268	05044	56420	39214	89822
41786	18169	96649	92406	42773	23672	37333	85734	99886	81200
95627	30768	30607	89023	60730	31519	53462	90489	81693	17849
98738	15548	42263	79489	85118	97073	01574	57310	59375	54417
75214	61575	27805	21930	94726	39454	19616	72239	93791	22610
73904	89123	19271	15792	72675	62175	48746	56084	54029	22296
33329	08896	94662	05781	59187	53284	28024	45421	37956	14252
66364	94799	62211	37539	80172	43269	91133	05562	82385	91760
68349	16984	86532	96186	53591	48268	82821	19526	63257	14288
19193	99621	66899	12351	72438	99839	24228	32079	53517	18558
49017	23489	19172	80439	76263	98918	59330	20121	89779	58862
76941	77008	27646	82072	28048	41589	70883	72035	81800	50296
55430	25875	26446	25738	32962	24266	26814	01194	48587	93319
33023	26895	65304	34978	43053	28951	22676	05303	39725	60054
87337	74487	83196	61939	05045	20405	69324	80823	20905	68727
81773	36773	21247	54735	68996	16937	18134	51873	10973	77090
74279	85087	94186	67793	18178	82224	17069	87880	54945	73489
34968	76028	54285	90845	35464	68076	15868	70063	26794	81386
99696	78454	21700	12301	88832	97796	59341	16136	01803	17537
55282	61051	97260	89829	69121	86547	62195	72492	33536	60137
31337	83886	72886	42598	5464	88071	92209	50728	67442	47529
94128	97990	58609	20002	76530	81981	30999	50147	93941	80754
06511	48241	49521	64568	69459	95079	42588	98590	12829	64366
69981	03469	56128	80405	97485	88251	76708	09558	86759	15065
23701	56612	86307	02364	88677	17192	23082	00728	78660	74196
09237	24607	12817	98120	30937	70666	76059	44446	94188	14060
11007	45461	24725	02877	74667	18427	45658	40044	59484	59966
60622	78444	39582	91930	97948	13221	99234	99629	22430	49247
79973	43668	19599	30021	68572	31816	63033	14597	28953	21162
71080	71367	23485	82364	30321	42982	74427	25625	74309	15855
9923	26729	74573	16583	37689	06703	21846	78329	98578	25447
63094	72826	65558	22616	33472	67515	75585	90005	19747	08865
19806	42212	41268	84923	21002	30588	40646	94961	31154	83133
17295	74244	43088	27056	86338	47331	09737	83735	84058	12382
59338	27190	99302	84020	15425	14748	42380	99376	30496	84523

(continúa)

Tabla de números aleatorios

(continuación)

96124	73355	00925	17210	81719	74603	30305	29383	69753	61156
31283	54371	20985	00299	71681	22496	71241	35347	37285	02028
49988	48558	20397	60384	24574	14852	26414	10767	60334	36911
82790	45529	48792	31384	55649	08779	94194	62843	11182	49766
51473	13821	75776	24401	00445	61570	80687	39454	07628	94806
07785	02854	91971	63537	84671	03517	28914	48762	76952	96837
16624	68335	46052	07442	41667	62897	40326	75187	36639	21396
28718	92405	07123	22008	83082	28526	49117	96627	38470	78905
33373	90330	67545	74667	20398	58239	22772	34500	34392	92989
36535	48606	11139	82646	18600	53898	70267	74970	35100	01291
47408	62155	47467	14813	56684	56681	31779	30441	19883	17044
56129	36513	41292	82142	13717	49966	35367	43255	06993	17418
35459	10460	33925	75946	26708	63004	89286	24880	38838	76022
61955	55992	36520	08005	48783	08773	45424	44359	25248	75881
85374	69791	18857	92948	90933	90290	97232	61348	22204	43440
15556	39555	09325	16717	74724	79343	26313	39585	56285	22525
75454	90681	73339	08810	89616	99234	36613	43440	60269	90899
27582	90856	04254	23715	00086	12164	16943	62099	32132	93031
89658	47708	01691	22284	50446	05451	68947	34932	81628	22716
57194	77203	26072	92538	85097	58178	46391	58980	12207	94901
64219	53416	03811	11439	80876	38314	77078	85171	06316	29523
53166	78592	80640	58248	68818	78915	57288	85310	43287	89223
58112	88451	22892	29765	20908	49267	18968	39165	03332	94932
14548	36314	05831	01921	97159	55540	00867	84293	54653	81281
21251	15618	40764	99303	38995	97879	98178	03701	70069	80463
30953	63369	05445	20240	35362	82072	29280	72468	94845	97004
12764	79194	36992	74905	85867	18672	28716	17995	63510	67901
72393	71563	42596	87316	80039	75647	66121	17083	07327	39209
11031	40757	10904	22385	39813	63111	33237	95008	09057	50820
91948	69586	45045	67557	86629	67943	23405	86552	17393	24221
18537	7384	13059	47389	97265	11379	24426	09528	36035	02501
66885	11985	38553	97029	88433	78988	88864	03876	48791	72613
96177	71237	08744	38483	16602	94343	18593	84747	57469	08334
37321	96867	64979	89159	33269	06367	09234	77201	92195	89547
77905	69703	77702	90176	04883	84487	88688	9360	42803	88379
53814	14560	43698	86631	87561	90731	59632	52672	24519	10966
16963	37320	40740	79330	04318	56078	23196	49668	80418	73842
87558	58885	65475	25295	59946	47877	81764	85986	61687	04373
84269	55068	10532	43324	39407	65004	35041	20714	20880	19385
94907	08019	05159	64613	26962	30688	51677	05111	51215	53285
45735	14319	78439	18033	72250	87674	67405	94163	16622	54994
11755	40589	83489	95820	70913	87328	04636	42466	68427	79135
51242	05075	80028	35144	70599	92270	62912	08859	87405	08266
00281	25893	94848	74342	45848	10404	28635	92136	42852	40812
12233	65661	10625	93343	21834	95563	15070	99901	09382	01498
88817	57827	02940	66788	76246	85094	44885	72542	31695	83843
75548	53699	90888	94921	04949	80725	72120	80838	38409	72270
42860	40656	33282	45677	05003	46597	67666	70858	41314	71100
71208	72822	17662	50330	32576	95030	87874	25965	05261	95727
44319	22313	89646	47415	21065	42846	78055	64776	64993	48051

Anexo 2
Tabla de números aleatorios

$$P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

z	0	1	2	3	4	5	6	7	8	9
-3.0	0.0013	0.0010	0.0007	0.0005	0.0003	0.0002	0.0002	0.0001	0.0001	0.0000
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0238	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0300	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0570	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.1052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Valores de la función distribución normal estándar

$$P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.516	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9648	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9978	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Anexo 3
Valores críticos para la distribución Ji cuadrado

$$P\left[\chi^2_{(n)} \leq \chi^2_0\right] = p$$

n \ p	0.005	0.01	0.025	0.05	0.1	0.25	0.75	0.9	0.95	0.975	0.99	0.995
1	-	-	0.001	0.004	0.016	0.102	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	7.841	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	4.255	9.037	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	5.071	10.219	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.899	11.389	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.737	12.549	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	7.584	13.701	17.275	19.675	21.92	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	8.438	14.845	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	9.299	15.984	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	10.165	17.117	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	11.037	18.245	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.3120	11.912	19.369	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	12.792	20.489	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	13.675	21.605	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	14.562	22.718	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	15.452	23.828	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	16.344	24.935	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	17.240	26.039	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	18.137	27.141	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	19.037	28.241	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	19.939	29.339	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	20.843	30.435	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	21.749	31.528	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	22.657	32.620	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	23.567	33.711	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	24.478	34.800	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	25.39	35.887	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	26.304	36.973	42.585	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	23.110	27.219	38.058	43.745	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	28.136	39.141	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	29.054	40.223	46.059	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	25.643	29.973	41.304	47.212	50.998	54.437	58.619	61.581
37	18.586	19.960	22.106	24.075	26.492	30.893	42.383	48.363	52.192	55.668	59.892	62.883
38	19.289	20.691	22.878	24.884	27.343	31.815	43.462	49.513	53.384	56.896	61.162	64.181
39	19.996	21.426	23.654	25.695	28.196	32.737	44.539	50.660	54.572	58.120	62.428	65.476
40	20.707	22.164	24.433	26.509	29.051	33.660	45.616	51.805	55.758	59.342	63.691	66.766
41	21.421	22.906	25.215	27.326	29.907	34.585	46.692	52.949	56.942	60.561	64.950	68.053
42	22.138	23.650	25.999	28.144	30.765	35.510	47.766	54.090	58.124	61.777	66.206	69.336
43	22.859	24.398	26.785	28.965	31.625	36.436	48.840	55.230	59.304	62.990	67.459	70.616
44	23.584	25.148	27.575	29.787	32.487	37.363	49.913	56.369	60.481	64.201	68.710	71.893
45	24.311	25.901	28.366	30.612	33.350	38.291	50.985	57.505	61.656	65.410	69.957	73.166

Anexo 4
Valores críticos para la distribución *t* de Student

$$P\left[t_{(n)} \leq t_0\right] = p$$

n \ p	0.75	0.9	0.95	0.975	0.99	0.995
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5177	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6848	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.315	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.75
31	0.6825	1.3095	1.6955	2.0395	2.4528	2.744
32	0.6822	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.6820	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.6818	1.307	1.6909	2.0322	2.4411	2.7284
35	0.6816	1.3062	1.6896	2.0301	2.4377	2.7238
36	0.6814	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.6812	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.6810	1.3042	1.686	2.0244	2.4286	2.7116
39	0.6808	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.6807	1.3031	1.6839	2.0211	2.4233	2.7045
41	0.6805	1.3025	1.6829	2.0195	2.4208	2.7012
42	0.6804	1.3020	1.6820	2.0181	2.4185	2.6981
43	0.6802	1.3016	1.6811	2.0167	2.4163	2.6951
44	0.6801	1.3011	1.6802	2.0154	2.4141	2.6923
45	0.6800	1.3006	1.6794	2.0141	2.4121	2.6896

Valores críticos para la distribución *t* de Student

$$P\left[t_{(n)} \leq t_0\right] = p$$

n \ P	0.75	0.9	0.95	0.975	0.99	0.995
46	0.6799	1.3002	1.6787	2.0129	2.4102	2.6870
47	0.6797	1.2998	1.6779	2.0117	2.4083	2.6846
48	0.6796	1.2994	1.6772	2.0106	2.4066	2.6822
49	0.6795	1.2991	1.6766	2.0096	2.4049	2.6800
50	0.6794	1.2987	1.6759	2.0086	2.4033	2.6778
51	0.6793	1.2984	1.6753	2.0076	2.4017	2.6757
52	0.6792	1.2980	1.6747	2.0066	2.4002	2.6737
53	0.6791	1.2977	1.6741	2.0057	2.3988	2.6718
54	0.6791	1.2974	1.6736	2.0049	2.3974	2.6700
55	0.6790	1.2971	1.6730	2.0040	2.3961	2.6682
56	0.6789	1.2969	1.6725	2.0032	2.3948	2.6665
57	0.6788	1.2966	1.6720	2.0025	2.3936	2.6649
58	0.6787	1.2963	1.6716	2.0017	2.3924	2.6633
59	0.6787	1.2961	1.6711	2.0010	2.3912	2.6618
60	0.6786	1.2958	1.6706	2.0003	2.3901	2.6603
61	0.6785	1.2956	1.6702	1.9996	2.3890	2.6589
62	0.6785	1.2954	1.6698	1.9990	2.3880	2.6575
63	0.6784	1.2951	1.6694	1.9983	2.3870	2.6561
64	0.6783	1.2949	1.6690	1.9977	2.3860	2.6549
65	0.6783	1.2947	1.6686	1.9971	2.3851	2.6536
66	0.6782	1.2945	1.6683	1.9966	2.3842	2.6524
67	0.6782	1.2943	1.6679	1.9960	2.3833	2.6512
68	0.6781	1.2941	1.6676	1.9955	2.3824	2.6501
69	0.6781	1.2939	1.6672	1.9949	2.3816	2.6490
70	0.6780	1.2938	1.6669	1.9944	2.3808	2.6479
71	0.6780	1.2936	1.6666	1.9939	2.3800	2.6469
72	0.6779	1.2934	1.6663	1.9935	2.3793	2.6459
73	0.6779	1.2933	1.6660	1.9930	2.3785	2.6449
74	0.6778	1.2931	1.6657	1.9925	2.3778	2.6439
75	0.6778	1.2929	1.6654	1.9921	2.3771	2.6430
76	0.6777	1.2928	1.6652	1.9917	2.3764	2.6421
77	0.6777	1.2926	1.6649	1.9913	2.3758	2.6412
78	0.6776	1.2925	1.6646	1.9908	2.3751	2.6403
79	0.6776	1.2924	1.6644	1.9905	2.3745	2.6395
80	0.6776	1.2922	1.6641	1.9901	2.3739	2.6387
81	0.6775	1.2921	1.6639	1.9897	2.3733	2.6379
82	0.6775	1.2920	1.6636	1.9893	2.3727	2.6371
83	0.6775	1.2918	1.6634	1.9890	2.3721	2.6364
84	0.6774	1.2917	1.6632	1.9886	2.3716	2.6356
85	0.6774	1.2916	1.6630	1.9883	2.3710	2.6349
86	0.6774	1.2915	1.6628	1.9879	2.3705	2.6342
87	0.6773	1.2914	1.6626	1.9876	2.3700	2.6335
88	0.6773	1.2912	1.6624	1.9873	2.3695	2.6329
89	0.6773	1.2911	1.6622	1.9870	2.3690	2.6322
90	0.6772	1.2910	1.6620	1.9867	2.3685	2.6316

Anexo 5
Resumen de fórmulas de distribuciones muestrales

Población	Parámetro	Estimador (variable aleatoria)	Distribución	Distribución muestral
Normal σ conocido	μ	$\bar{x}$	$N(\mu, \sigma^2/n)$	$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
Cualquiera μ, σ conocidos desconocidos	μ	$\bar{x}$	$N(\mu, \sigma^2/n)$ n grande (TLC)	$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
Normal μ, σ conocidos desconocidos	μ	$\bar{x}$	$t_{(n-1)\ g.l.}$	$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}$
Normal σ conocido	σ^2	s^2	χ^2 con (n-1) g.l.	$U = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$
Binomial	π	$p = \frac{x}{n}$	$N(\pi, \pi(1-\pi)/n)$ n grande	$z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$
Dos proporciones poblacionales π_1 y π_2	$\pi_1 - \pi_2$	p1 - p2	$N\left(\pi_1-\pi_2; \frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}\right)$	$z = \frac{p_1-p_2-(\pi_1-\pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1)$
$N(\mu_1, \sigma^2_1)$ $N(\mu_2, \sigma^2_2)$ σ^2_1, σ^2_2 conocidos	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$N\left(\mu_1-\mu_2, \frac{\alpha_1^2}{n_1}+\frac{\alpha_2^2}{n_2}\right)$	$z = \frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\alpha_1^2}{n_1}+\frac{\alpha_2^2}{n_2}}} \sim N(0,1)$

Población	Parámetro	Estimador (variable aleatoria)	Distribución	Distribución muestral
$N(\mu_1, \sigma^2_1)$ $N(\mu_2, \sigma^2_2)$ σ^2_1, σ^2_2 desconocidos e iguales	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$t_{(n_1 + n_2 - 2)}$.	$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$ $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$
$N(\mu_1, \sigma^2_1)$ $N(\mu_2, \sigma^2_2)$ σ^2_1, σ^2_2 desconocidos y diferentes	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$t_{(V)}$.	$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \sim t_{(v)}$ $v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1+1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2+1)}} - 2$
$N(\mu_1, \sigma^2_1)$ $N(\mu_2, \sigma^2_2)$ μ_1, μ_2, σ^2_1 y σ^2_2 desconocidos	$\frac{\sigma_1^2}{\sigma_2^2}$	$\frac{S_1^2}{S_2^2}$	$F_{(n_1 - 1,\ n_2 - 1)}$	$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F_{(n_1-1,\ n_2-1)}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

$$\bar{x} = \frac{\sum_{i=1}^{k} xi' \, fi}{\sum_{i=1}^{k} fi} = \sum_{i=1}^{k} xi' \, fri \qquad s^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{X})^2 f_i}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 f_i - n\bar{x}^2}{n-1}$$

Anexo 6
Resumen de fórmulas de intervalos de confianza

Tipo de problemas	Parámetro	Estimador puntual	Intervalo de confianza
Media poblacional μ (σ conocida)	μ	$\bar{x}$	$\left\langle \bar{x} \mp z_{\left(1-\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}} \right\rangle$
Media poblacional μ (σ desconocida)	μ	$\bar{x}$	$\left\langle \bar{x} \mp t_{\left(1-\frac{\alpha}{2};\, n-1\right)} \frac{s}{\sqrt{n}} \right\rangle$
Varianza σ^2 de una población normal	σ^2	s^2	$\left\langle \frac{(n-1)s^2}{\chi^2_{\left(1-\frac{\alpha}{2};\, n-1\right)}}; \frac{(n-1)s^2}{\chi^2_{\left(\frac{\alpha}{2};\, n-1\right)}} \right\rangle$
Proporción poblacional π (parámetro de una población binomial)	π	P	$\left\langle p \mp Z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\frac{p(1-p)}{n}} \right\rangle$
Diferencia de medias de dos poblaciones normales $\mu_1 - \mu_2$ con varianzas σ^2_1 y σ^2_2 conocidas	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$\left\langle (\bar{X}_1 - \bar{X}_2) \mp Z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\rangle$
Diferencia de medias de dos poblaciones normales $\mu_1 - \mu_2$ con varianzas σ^2_1 y σ^2_2 desconocidas pero homogéneas	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$\left\langle (\bar{X}_1 - \bar{X}_2) \mp t_{\left(1-\frac{\alpha}{2};\, n_1+n_2-2\right)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\rangle$ $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$

Tipo de problemas	Parámetro	Estimador puntual	Intervalo de confianza
Diferencia de medias de dos poblaciones normales $\mu_1 - \mu_2$ con varianzas σ^2_1 y σ^2_2 desconocidas pero homogéneas	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\left\langle (\bar{X}_1 - \bar{X}_2) \mp t_{\left(1-\frac{\alpha}{2};\, v\right)} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right\rangle$ $v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2 - 1)}}$
Razón (cociente) de las varianzas de dos poblaciones normales	σ_1^2 / σ_2^2	s_1^2 / s_2^2	$\left\langle \left(\frac{s_1^2}{s_2^2}\right) F_{\left(\frac{\alpha}{2};\, n_2-1,\, n_1-1\right)} ; \left(\frac{s_1^2}{s_2^2}\right) F_{\left(1-\frac{\alpha}{2};\, n_2-1,\, n_1-1\right)} \right\rangle$
Diferencia entre dos proporciones poblacionales $\pi_1 - \pi_2$ (o dos parámetros binomiales)	$\pi_1 - \pi_2$	$p_1 - p_2$	$\left\langle (p_1 - p_2) \mp z_{\left(1-\frac{\alpha}{2}\right)} \sqrt{\left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)} \right\rangle$
Diferencia entre medias de dos poblaciones normales para muestras pareadas $\mu_d = \mu_1 - \mu_2$	μ_d	$\bar{d}$	$\left\langle \bar{d} \mp t_{\left(1-\frac{\alpha}{2};\, n-1\right)} \frac{s_d}{\sqrt{n}} \right\rangle$

Tamaño de muestra:

$$n = \left(\frac{z}{E}\right)^2 p(1-p)$$ (Para estimar a una proporción poblacional)

$$n = \left(\frac{z\sigma}{E}\right)^2$$ (Para estimar a la media poblacional)

Anexo 7
Resumen de fórmulas de pruebas de hipótesis

Caso	H_0	Estadística de prueba	H_a	Criterio de rechazo
1	$H_0: \mu = \mu_0$ σ conocida	$z_0 = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	Ha: $\mu \neq \mu_0$ Ha: $\mu > \mu_0$ Ha: $\mu < \mu_0$	$\lvert Z_0 \rvert > Z_{(1-\alpha/2)}$ $Z_0 > Z_{(1-\alpha)}$ $Z_0 < Z_{(\alpha)}$
2	$H_0: \mu = \mu_0$ σ desconocida	$t_0 = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$	Ha: $\mu \neq \mu_0$ Ha: $\mu > \mu_0$ Ha: $\mu < \mu_0$	$\lvert t_0 \rvert > t_{(1-\alpha/2,\ n-1)}$ $t_0 > t_{(1-\alpha,\ n-1)}$ $t_0 < t_{(\alpha,\ n-1)}$
3	$H_0: \mu_1 = \mu_2$ σ_1 y σ_2 conocidas	$z_0 = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	Ha: $\mu_1 \neq \mu_2$ Ha: $\mu_1 > \mu_2$ Ha: $\mu_1 < \mu_2$	$\lvert Z_0 \rvert > Z_{(1-\alpha/2)}$ $Z_0 > Z_{(1-\alpha)}$ $Z_0 < Z_{(\alpha)}$
4	$H_0: \mu_1 = \mu_2$ $\sigma_1 = \sigma_2$ desconocidas	$t_0 = \dfrac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ $s_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$	Ha: $\mu_1 \neq \mu_2$ Ha: $\mu_1 > \mu_2$ Ha: $\mu_1 < \mu_2$	$\lvert t_0 \rvert > t_{(1-\alpha/2,\ n1+n2\ -2)}$ $t_0 > t_{(1-\alpha,\ n1+n2\ -2)}$ $t_0 < t_{(\alpha,\ n1+n2\ -2)}$
5	$H_0: \mu_1 = \mu_2$ $\sigma_1 \neq \sigma_2$ desconocidas	$t_0 = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ $\nu = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}$	Ha: $\mu_1 \neq \mu_2$ Ha: $\mu_1 > \mu_2$ Ha: $\mu_1 < \mu_2$	$\lvert t_0 \rvert > t_{(1-\alpha/2,\ \nu)}$ $t_0 > t_{(1-\alpha,\ \nu)}$ $t_0 < t_{(\alpha,\ \nu)}$
6	$H_0: \mu_d = 0$ Datos pareados	$t_0 = \dfrac{\bar{d}}{s_d/\sqrt{n}}$	Ha: $\mu_d \neq 0$ Ha: $\mu_d > 0$ Ha: $\mu_d < 0$	$\lvert t_0 \rvert > t_{(1-\alpha/2,\ n-1)}$ $t_0 > t_{(1-\alpha,\ n-1)}$ $t_0 < t_{(\alpha,\ n-1)}$
7	$H_0: \sigma_1 = \sigma_0^2$	$\chi_0^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$	Ha: $\sigma^2 \neq \sigma_0^2$ Ha: $\sigma^2 > \sigma_0^2$ Ha: $\sigma^2 < \sigma_0^2$	$\chi_0^2 > \chi^2_{(1-\alpha/2,\ n-1)}$ $\chi_0^2 < \chi^2_{(1-\alpha\alpha/2\ n-1)}$ $\chi_0^2 > \chi^2_{(1-\alpha,\ n-1)}$ $\chi_0^2 < \chi^2_{(\alpha\alpha\ n-1)}$

Caso	H_0	Estadística de prueba	H_a	Criterio de rechazo
8	$H_0: \sigma_1^2 = \sigma_2^2$	$F_0 = \frac{s_1^2}{s_2^2}$	Ha: $\sigma_1^2 \neq \sigma_2^2$ Ha: $\sigma_1^2 > \sigma_2^2$ Ha: $\sigma_1^2 < \sigma_2^2$	$F_0 > F_{(1-\alpha/2,\, n_1-1,\, n_2-1)}$ $F_0 < F_{(\alpha/2,\, n_1-1,\, n_2-1)}$ $F_0 > F_{(1-\alpha,\, n_1-1,\, n_2-1)}$ $F_0 < F_{(\alpha\alpha\, n_1-1,\, n_2-1)}$
9	$H_0: \pi = \pi_0$ σ desconocida	$z_0 = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	Ha: $\pi \neq \pi_0$ Ha: $\pi > \pi_0$ Ha: $\pi < \pi_0$	$\lvert Z_0\rvert > Z_{(1-\alpha/2)}$ $Z_0 > Z_{(1-\alpha)}$ $Z_0 < Z_{(\alpha)}$
10	$H_0: \pi_1 = \pi_2$	$z_0 = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$	Ha: $\pi_1 \neq \pi_2$ Ha: $\pi_1 > \pi_2$ Ha: $\pi_1 < \pi_2$	$\lvert Z_0\rvert > Z_{(1-\alpha/2)}$ $Z_0 > Z_{(1-\alpha)}$ $Z_0 < Z_{(\alpha)}$

α = P(rechazar Ho / Ho es verdadera)

β = P(Aceptar Ho / Ho es falsa)

Potencia = 1- β

$$n = \frac{(Z_\alpha + Z_\beta)^2 \sigma^2}{(\mu_{Ho} - \mu_{Ha})^2} \qquad n = \frac{(Z_{\alpha/2} + Z_\beta)^2 \sigma^2}{(\mu_{Ho} - \mu_{Ha})^2}$$

Prueba de bondad de ajuste

- $\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{Ei}$

  Rechazar la hipótesis de que la distribución de la población es la distribución propuesta si:

$$\chi_0^2 > \chi^2_{(1-\alpha,\, k-p-1)}$$

Prueba de independencia

- $\chi_0^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ con (I-1)(J-1) g.l.

- Se rechaza la hipótesis nula si: $\chi_0^2 > \chi^2_{(1-\alpha,(I-1)(J-1))}$

Anexo 8
Resumen de fórmulas de regresión lineal simple

Estimación puntual

- $\hat{\beta}_1 = \dfrac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$ $\qquad \hat{\beta}_0 = \dfrac{1}{n}\left(\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i\right) = \overline{Y} - \hat{\beta}_1 \overline{X}$

- $e_i = y_i - \hat{y}_i$ (Error de estimación)

Prueba de hipótesis

- $s_e = \sqrt{\dfrac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\dfrac{\sum_{i=1}^{n} Y_i^2 - \hat{a}_0 \sum_{i=1}^{n} Y_i - \hat{a}_1 \sum_{i=1}^{n} X_i Y_i}{n-2}}$

- $T_0 = \dfrac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)}$; donde: $s(\hat{\beta}_0) = s_e \sqrt{\dfrac{1}{n} + \dfrac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$; $\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$

- $T_0 = \dfrac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)}$; donde: $s(\hat{\beta}_1) = s_e \sqrt{\dfrac{1}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$

Intervalos de confianza

- $\beta_0 \varepsilon \left\langle \hat{\beta}_0 \mp t_{\left(1-\frac{\alpha}{2};\, n-2\right)} s(\hat{\beta}_0) \right\rangle$

- $\beta_1 \varepsilon \left\langle \hat{\beta}_1 \mp t_{\left(1-\frac{\alpha}{2};\, n-2\right)} s(\hat{\beta}_1) \right\rangle$

- $\mu_{Y/X_0} \varepsilon \left\langle (\hat{\beta}_0 + \hat{\beta}_1 x_0) \mp t_{\left(1-\frac{\alpha}{2};\, n-2\right)} s_e \sqrt{\dfrac{1}{n} + \dfrac{(x_0 - \overline{X})^2}{\sum_{i=1}^{n} (x_i - \overline{X})^2}} \right\rangle$

- $$Y_{/x_0}\varepsilon\left\langle(\hat{\beta}_0+\hat{\beta}_1 x_0)\mp t_{\left(1-\frac{\alpha}{2};\,n-2\right)}\,s_e\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{X})^2}{\sum_{i=1}^{n}(x_i-\bar{X})^2}}\right\rangle$$

Correlación lineal simple

- $$r=\frac{n\left(\sum_{i=1}^{n}X_i\,Y_i\right)-\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{\sqrt{\left[n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right]\left[n\sum_{i=1}^{n}Y_i^2-\left(\sum_{i=1}^{n}Y_i\right)^2\right]}}$$

Coeficiente de determinación

- $$R^2=\frac{SC\,regresión}{SC\,total}$$

Anexo 9
Resumen de fórmulas de regresión lineal múltiple

- $\hat{\beta} = (X'X)^{-1} X'Y \qquad s_e = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - k - 1}} \qquad T_0 = \frac{\hat{\beta}_i - \beta_i}{s(\hat{\beta}_i)}$

- $\beta_i \varepsilon \left\langle \hat{\beta}_i \mp t_{\left(1-\frac{\alpha}{2};\, n-k-1\right)} s(\hat{\beta}_i) \right\rangle$ (Intervalo de confianza para el coeficiente β_i)

- $\text{SC regresión} = \hat{\beta}'X'Y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \qquad \text{SC total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

- n -1 = k + (n – k - 1) (k es el número de variables independientes)

Tabla del análisis de varianza para la regresión (Anova)

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F calculado	p value
Regeresión	k	SC reg.	$\text{CM reg.} = \frac{\text{SC reg.}}{k}$	$F_0 = \frac{\text{CM reg.}}{\text{CM error}}$	P_0
Error	n – k – 1	SC error	$\text{CM error} = \frac{\text{SC error.}}{n - k - 1} = s_e^2$		
Total	n – 1	SC total	$F_{(n_1 - 1,\, n_2 - 1)}$		

- Valor P = $\mathbf{P(F_{(k,\, n-k-1)} > F_0)}$

Anexo 10
Resumen de fórmulas de diseño completamente aleatorizado

- $Y_{ij} = \mu + \iota_i + e_{ij}$ i = 1, 2, ..., a (i indica tratamiento) j = 1, 2, ...,n (j indica repetición)
  N = an (Número total de datos)

- $\sum_{j=1}^{n} Y_{ij} = Y_{i\bullet}$ Total de observaciones en el i-ésimo tratamiento.

- $\bar{Y}_{i\bullet} = \frac{Y_{i\bullet}}{n}$ Promedio de las observaciones bajo el i-ésimo tratamiento.

- $Y_{\bullet\bullet} = \sum_{i=1}^{a} \sum_{j=1}^{n} Y_{ij}$ Suma total de observaciones.

- $\bar{Y}_{\bullet\bullet} = \frac{Y_{\bullet\bullet}}{an}$ Promedio general

- $SC\ total = \sum_{i=1}^{a} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{a} \sum_{j=1}^{n} Y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{an}$

- $SC\ trat. = n \sum_{i=1}^{a} (Y_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{a} \frac{Y_{i\bullet}^2}{n} - \frac{Y_{\bullet\bullet}^2}{an}$

- SC error = SC total – SC trat.

Los grados de libertad se descomponen de la siguiente forma:

- G.L.total = G.L.trat. + G.L.error
  an –1 = (a – 1) + a (n – 1)

- $\hat{\mu} = \bar{Y}_{\bullet\bullet}$ $\hat{\mu}_i = \bar{Y}_{i\bullet}$ $\hat{\tau}_i = \bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}$ (Efecto del i-ésimo tratamiento)

- $\mu_i \varepsilon \left\langle \bar{Y}_{i\bullet} \mp t_{(1-\frac{\alpha}{2};a(n-1))} \sqrt{\frac{CMerror}{n}} \right\rangle$ (Intervalo de confianza para una media)

- $(\mu_i - \mu_j) \varepsilon \left\langle (\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}) \mp t_{(1-\frac{\alpha}{2};a(n-1))} \sqrt{\frac{2CMerror}{n}} \right\rangle$

Tabla del análisis de varianza para la igualdad de medias (Anova)

Fuente de variación	**Grados de libertad**	**Suma de cuadrados**	**Cuadrados medios**	**F calculado**	**p value**
Factor (entre grupos)	a – 1	SC trat.	$\text{CM trat.} = \frac{\text{SC trat..}}{a - 1}$	$F_0 = \frac{\text{CM trat.}}{\text{CM error}}$	P_0
Error (dentro de grupos)	a(n – 1)	SC error	$\text{CM error} = \frac{\text{SC error.}}{a(n - 1)}$		
Total	an – 1	SC total			

Bibliografía

Anderson, David R. y Thomas A. Williams. *Estadística para administración y economía*. 8.ª edición. México D.F.: Thompson, 2004.

Berenson, Mark; Levine, David y Timothy Krehbiel. *Estadística para administración.* 2.ª edición. México D.F.: Pearson Educación, 2001.

Córdova Zamora, Manuel. *Estadística descriptiva e inferencial*. 5.ª edición. Lima: Moshera, 2003.

Hanke, John y Arthur Reitsch. *Pronósticos en los negocios*. 2.ª edición. México D.F.: Prentice Hall Hispanoamericana S.A., 1996.

Hines, W. y D. Montgomery. *Probabilidad y estadística para ingeniería y administración*. 2.ª edición. México D.F.: CECSA, 1997.

Kohler, Heinz. *Estadística para negocios y economía*. México D.F.: Continental, 1996.

Mason, Robert D. *Estadística para Administración y Economía.* 11.ª edición. México D.F.: Alfaomega, 2004.

Mendenhall, William. *Estadística para administradores.* 2.ª edición. México D.F.: Iberoamérica, 1990.

Mitacc Meza, Máximo. *Tópicos de estadística descriptiva y probabilidades.* Lima: Editorial San Marcos, 1988.

Montgomery, D. y G. Runger. *Probabilidad y estadística aplicadas a la ingeniería.* 2.ª edición. México D.F.: Limusa-Wiley, 2002.

Moya Calderón, Rufino. *Estadística descriptiva: Conceptos y aplicaciones*. Lima: Editorial San Marcos, 1991.

Pérez López, César. *Estadística aplicada a través de Excel*. Madrid: Pearson Educación/Prentice Hall, 2002.

Véliz Capuñay, Carlos. *Estadística.* 3.ª edición, 1998.

Webster, Allen L. *Estadística aplicada a los negocios y la economía*. 3.ª edición. Bogotá: McGraw-Hill, 2001.

Weimer, Richard. *Estadística*. 2.ª edición. México D.F.: Continental, 1996.

Ya Lun, Chou. *Análisis estadístico.* 2.ª edición. México D.F.: Interamericana S.A., 1977.

Este libro se terminó de imprimir en enero del 2013
en Tarea Asociación Gráfica Educativa
Psje. María Auxiliadora 156-164, Breña, Lima, Perú
Teléfonos: 424-8104 / 332-3229
tareagrafica@tareagrafica.com

UNIVERSIDAD DE LIMA

Estadística aplicada

Este libro explica los principios de la estadística inferencial, las regresiones lineales simple y múltiple, y el diseño de experimentos, temas determinantes en la toma de decisiones con herramientas cuantitativas.

Dividido en cinco capítulos, en los dos primeros trata lo referente a las distribuciones de probabilidad y a los procesos de estimación puntual y por intervalos. El tercer capítulo desarrolla los aspectos relativos a la prueba de hipótesis, mientras que el cuarto considera la aplicación de las herramientas. Finalmente, en su quinto capítulo describe y analiza el diseño experimental completamente aleatorio.

Las partes mencionadas se complementan con más de trescientos ejercicios, que proporcionan el enlace necesario entre los conceptos teóricos y la práctica, lo cual constituye un aporte fundamental para los interesados en aplicaciones de la estadística en el campo de la ingeniería y las ciencias administrativas y económicas.

Emma Barreno Vereau
Licenciada en Estadística por la Universidad Nacional de Trujillo. Magíster en Estadística Matemática por el Centro Interamericano de Enseñanza de Estadística (Cienes) en convenio con la Universidad de Chile. Profesora en la Universidad de Lima.

Jorge Chue Gallardo
Ingeniero estadístico por la Universidad Nacional Agraria La Molina. Magíster en Estadística por la Universidad Nacional Mayor de San Marcos. Profesor en la Facultad de Ingeniería de Sistemas de la Universidad de Lima.

Rosa Millones Rivalles
Licenciada en Estadística por la Universidad Nacional de Trujillo. Magíster en Administración de la Educación por la Universidad de Lima, donde actualmente ejerce la docencia, así como también en la Universidad Católica Sedes Sapientiae.

Félix Vásquez Urbano
Licenciado en Estadística por la Universidad Nacional de Trujillo. Profesor en las universidades de Lima y Femenina del Sagrado Corazón.

Carlos Castillo Crespo
Licenciado en Estadística por la Universidad Nacional de Trujillo. Profesor en las universidades de Lima y Ricardo Palma.

ESCUELA DE INGENIERÍA

• FACULTAD DE INGENIERÍA INDUSTRIAL • FACULTAD DE INGENIERÍA DE SISTEMAS

ISBN 978-9972-45-237-6