



TECNOLOGICO NACIONAL DE MEXICO
INSTITUTO TECNOLÓGICO DE TIJUANA



Depto. De Metal - Mecánica

Carrera: Ingeniería Electro-Mecánica



Libro de Probabilidad y Estadística

Prof. Ing. José Alfredo Flores Pérez

Agosto / 2018

Índice

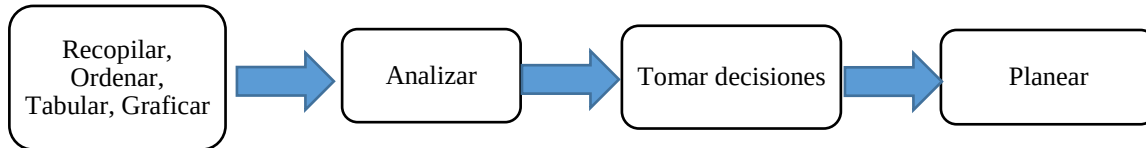
1. Estadística descriptiva	3
1.1. Población y muestra aleatoria	
1.2. Obtención de datos estadísticos	
1.3. Medidas de tendencia centra	
1.4. Medidas de dispersión	
1.5. Tabla de distribución de frecuencia	
1.6. Cuartiles	
1.7. Gráficos	
1.8. Caja y alambres	
1.9. Diagrama de Pareto	
1.10. Uso de software	
2. Probabilidad	17
2.1. Probabilidad de eventos	
2.2. Espacio muestra	
2.3. Ocurrencia de eventos	
2.4. Permutaciones y combinaciones	
2.5. Diagrama de árbol	
2.6. Axiomas de probabilidad	
2.7. Independencia y probabilidad Condicional	
2.8. Teorema de Bayes	
3. Funciones de distribución de probabilidades	30
3.1. Variables aleatorias y su clasificación	
3.2. Distribuciones de probabilidad discretas	
3.3. Distribuciones de Hipergeométrica	
3.4. Distribución Poisson	
3.5. Distribución de probabilidad continuas	
3.6. Distribución T	
3.7. Distribución chi cuadrada	
3.8. Distribución	
3.9. Esperanza matemática	

4. Estadística inferencial	35
4.1. Inferencia estadística	
4.2. Muestreo estadístico	
4.3. Estimadores	
4.4. Estimación puntual	
4.5. Estimación por intervalo	
4.6. Errores tipo I y II	
4.7. Contraste de hipótesis unilaterales y bilaterales	
 5. Regresión correlación	 57
5.1. Control de calidad	
5.2. Diagrama de dispersión	
5.3. Regresión lineal simple	
5.4. Correlación	
5.5. Determinación y análisis de los coeficientes de correlación y de determinación	
5.6. Distribución normal bidimensional	
5.7. Intervalos de confianza y pruebas para el coeficiente de correlación	
5.8. Errores de medición	

Principios Generales de Estadística

Estadística

Conjunto de técnicas que se emplean para recolectar, organizar analizar datos, los cuales pueden ser cuantitativos, valores expresados numéricamente, o cualitativos, características tales como las preferencias del consumidor organizadas en una tabla.



En estadística los datos se organizan en forma descendente o ascendente

Recopilar: Juntar o reunir datos.

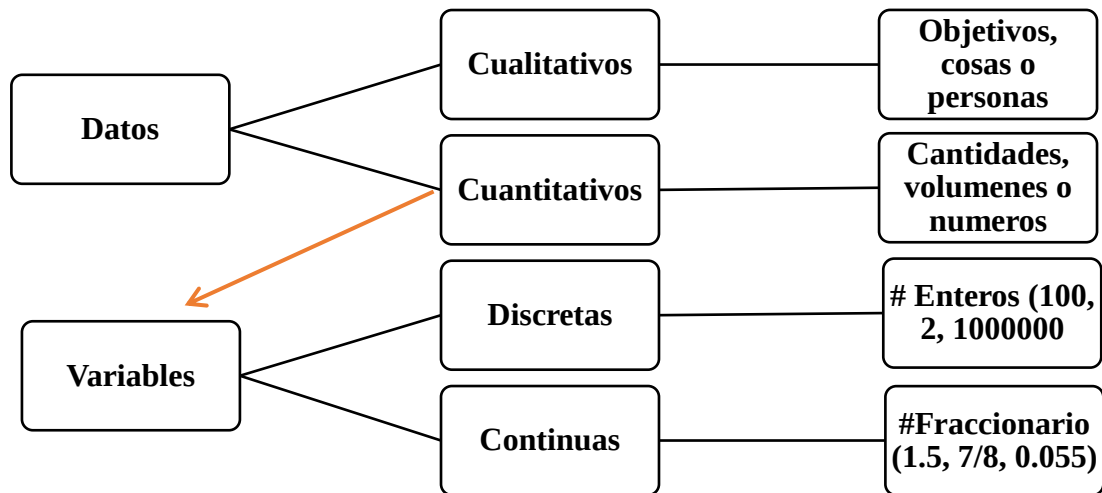
Ordenar: Ordenar los datos en forma Ascendente o Descendente.

Tabular: Acomodar los datos en una tabla (columnas y renglones).

Graficar: Es donde se plasman los datos en gráficas, existen 3 tipos de graficas: lineal, barras y de pastel.

Variables discretas y variables continuas

Una variable discreta solamente puede tomar valores en puntos aislados de una escala de valores. En la estadística de negocios datos de este tipo se generan mediante un proceso de conteo; así, por lo general estos valores se expresan mediante números enteros. Una variable continua puede tomar valores en cualquier punto a lo largo de un intervalo de valores determinado. Los datos continuos se generan mediante un proceso de medición.



Métodos de muestreo aleatorio

Muestra aleatoria simple: es aquella en la cual se escogen al azar elementos individuales de la población objeto. Esta elección al azar es similar a la extracción de números en la lotería. Sin embargo, por lo general en el muestreo estadístico se usa en una tabla de números aleatorios o un programa de computadora generador de números aleatorios para determinar los elementos numerados de la población que se elegirán para la muestra.

Muestra sistemática: es una muestra aleatoria en la cual los elementos de la población se eligen de una lista a intervalos uniformes (por ejemplo: escoger para la muestra una décima cuenta por cobrar). El primero de las 10 cuentas que se incluirán en la muestra se escoge de manera aleatoria (empleando quizás una tabla de números aleatorios). Un aspecto del muestreo sistemático es la posible existencia de algún factor periódico o cíclico en la lista de la población que pueda conducir a un error sistemático en los resultados de la muestra.

Muestreo estratificado: el investigador clasifica primero los elementos de la población en subgrupo o estratos con base a una o más características. Un plan de muestreo de este tipo se usa para asegurar que los diferentes subgrupos de la población estén presentados de manera proporcional en la muestra. Además, el tamaño requerido de muestra para alcanzar en un nivel dado de precisión es lo común menor que el que se necesita si se emplea el muestreo aleatorio simple. En lo cual se reduce el costo del muestreo.

Muestro por conglomerados: es un tipo de muestreo aleatoria en el cual los elementos de la población se presentan en forma neutral en subgrupo. Entonces se muestrean en forma aleatoria subgrupo o conglomerados completos.

Ejemplos:

Una empresa de televisores pone ofertas de descuento en sus diferentes modelos y los precios van desde 2500 pesos hasta 12500 pesos. Realiza una distribución de frecuencias en 7 clases y describe el histograma, polígono de frecuencias y la ojiva.

El número de televisores en oferta: 14, 16, 12, 20, 8, 1, 8, 10 → f

$$\text{Int. Aproximado} = \frac{12500 - 2500}{7} = \$1429.00$$

Se le suma 1429 ↓	Precio de las T.V (\$)	←Se le resta 1
	2500-3928	
	3929-5357	
	5358-6786	
	6787-8275	
	8216-9644	
	9645-110773	
	11074-12500	

Se le resta .5 ↓	Limites exactos de clase (\$)	←Se le suma .5
	2499.5-3928.5	
	3928.5-5357.5	
	5357.5-6486.5	
	6986.5-8275.5	
	8215.5-9644.5	
	9644.5-11073.5	
	11073.5- 12500.5	

Esta columna sirve para realizar el histograma y la ojiva ↑

(fa)# de T.V
14
40
52
72
80
98
108

$$\leftarrow \frac{\text{Columna 1} + \text{Columna 2}}{2}$$

Esta columna sirve para el polígono de frecuencia en el eje X ↑

Puntos Medios entre clase (\$)
3214
4643
6072
7501
8930
10354
11786

Esta columna sirve para realizar el histograma en el eje Y ↑

(f)# de T.V
14
26
12
20
8
18
10

Se suma el valor anterior
ejemplo:

$$\leftarrow 14 + 26 = 40$$

Esta columna sirve para realizar ojiva en el eje Y ↑

Distribución de frecuencia.

Distribuciones de frecuencia

Una distribución de frecuencia es una tabla en la cual los valores que puede tomar una variable se agrupan en clases, y se registra la cantidad de valores observados de cada clase. A los datos organizados en una distribución de frecuencia se les llama datos agrupados. En cambio, en el caso de datos no agrupados se hace una lista con cada uno de los valores observados de la variable aleatoria.

Ejemplo 1 En la tabla 2.1 se muestra una distribución de frecuencia de salarios semanales. Observe que las cantidades se dan redondeadas al dólar más cercano. Cuando un resto que va a redondearse es “exactamente 0.5” (exactamente \$0.50 en este caso), se ha convenido redondear al número par más cercano. Así, un salario semanal de \$259.50 se redondea a \$260 como parte del proceso de agrupación de los datos.

Tabla 2.1 Una distribución de frecuencia de los salarios semanales de 100 trabajadores por nivel de ingresos.

Salario semanal	Número de trabajadores (f)
\$240-259	7
260-279	20
280-299	33
300-319	25
320-339	11
340-359	4
	Total 100

Intervalos de clase

En cada clase de una distribución de frecuencia, los límites de clase inferior y clase superior establecidos indican los valores comprendidos en esa clase. (Véase la primera columna de la tabla 2.1.) En cambio, los límites exactos de clase, o fronteras de clase, son los puntos que sirven para separar clases contigua en una escala de medicina para variables continuas. Los límites exactos de clase pueden determinarse localizando los puntos que están a la mitad entre los límites superior e inferior, respectivamente, de clases contiguas. El intervalo de clase determina a la gama de valores comprendidos en una clase y se obtiene al sustraer del límite exacto superior de la clase el exacto inferior de esa clase. Si no se han establecido

los límites exactos, el intervalo de clase se puede determinar sustrayendo el límite inferior establecido de una clase del límite establecido de la clase superior contigua. Por último, para ciertos propósitos los valores de una clase se suelen representar por el punto medio de la clase, el cual se determina al sumar un medio del intervalo de clase al límite inferior exacto de la clase.

Ejemplo 2 La tabla 2.2 presenta los límites de clase exactos y los puntos medios de clase correspondientes a la distribución de frecuencia de la tabla 2.2.

Tabla 2.2 Salarios semanales de 100 trabajadores por nivel de ingresos

Salario semanal (límites de clase)	Límites exactos de clase*	Punto medio de la clase	Número de trabajadores
\$240-259	\$239.50-259.50	\$249.50	7
260-279	259.50-279.50	269.50	20
280-299	279.50-299.50	289.50	33
300-319	299.50-319.50	309.50	25
320-339	319.50-339.50	329.50	11
340-359	339.50-359.50	249.50	4
			Total 100

En general, en los límites de clase exactos se expresa solamente una cifra significativa adicional en comparación con los límites de clase establecidos. Sin embargo, ya que con las unidades monetarias la siguiente unidad de medición más precisa, después del “dólar más cercano por lo general se define como “el centavo más cercano”, en este caso se expresan dos cifras adicionales.

Ejemplo 3 en la tabla 2.2 el intervalo de clase de la primera clase calculado mediante los dos métodos es: \$259.50 - \$239.50= \$20 (sustrayendo el límite inferior exacto de una clase del límite superior exacto de esa clase) o \$260 – 240= \$20 (sustrayendo el límite inferior establecido de una clase, del límite inferior establecido de la clase superior siguiente).

Por lo general se prefiere que en una distribución de frecuencia todos los intervalos de clase sean iguales. Una fórmula que se usa para determinar, de manera aproximada, el intervalo de clase que debe de emplearse es

$$\text{Intervalo aproximado} = \frac{\left[\begin{array}{c} \text{valor mayor en los datos} \\ \text{no agrupados} \end{array} \right] - \left[\begin{array}{c} \text{valor menor en los datos} \\ \text{no agrupados} \end{array} \right]}{\text{numero de clases deseado}}$$

Ejemplo 4 En los datos originales que se agruparon en la tabla 2.1, suponga que el mayor salario observado sea \$358 y que el menor salario observado sea \$242. Dado que el objetivo de tener seis clases con intervalos de clases iguales,

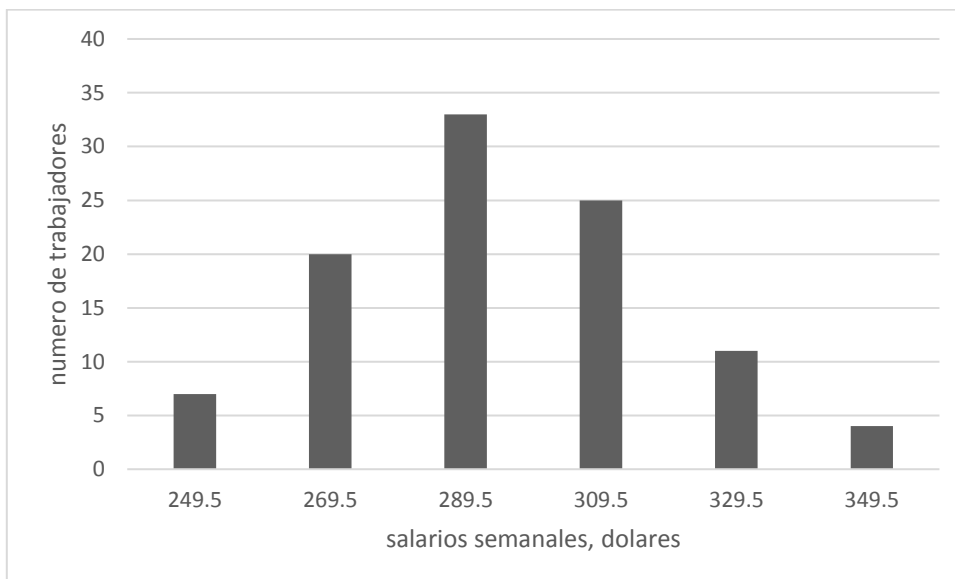
$$\text{Intervalo aproximado} = \frac{358-242}{6} = \$19.33$$

El tamaño de clase más adecuado es entonces \$20

En el caso de datos distribuidos de manera muy poco uniforme, como los salarios anuales de actividades muy distintas, puede resultar adecuado emplear intervalos de clase desiguales. En tal caso, se usan intervalos de clase mayores para rangos de valores en los cuales se encuentran relativamente pocas observaciones.

Histograma y polígonos de frecuencia

Un histograma es una gráfica de barras de una distribución de frecuencia. Como se indica en la figura 2.1, lo que se acostumbra hacer es colocar los límites exactos de clase a lo largo del eje horizontal de la gráfica mientras que el número de observadores se lista a lo largo del eje vertical. Sin embargo, también se suelen usar los puntos medios de clase en el lugar de los límites de clase para identificar las clases.



Un polígono de frecuencia es una gráfica de línea de una distribución de frecuencia. Como se indica en la figura 2.2, los dos ejes de esta grafica son similares a los del histograma salvo que el punto medio de cada clase se coloca por lo general a lo largo del eje horizontal. El número de observación de cada clase se representa por

medio de un punto sobre el punto medio de la clase y estos puntos se unen con una serie de segmento de recta para formar un polígono o “figura de varios lados”.



Curvas de frecuencia

Una curva de frecuencia es un polígono de frecuencia suavizado. En términos de sesgo: 1) negativamente sesgada: asimétrica con la “cola” a la izquierda; 2) positivamente sesgada: asimétrica con la “cola” a la derecha, o 3) simétrica.

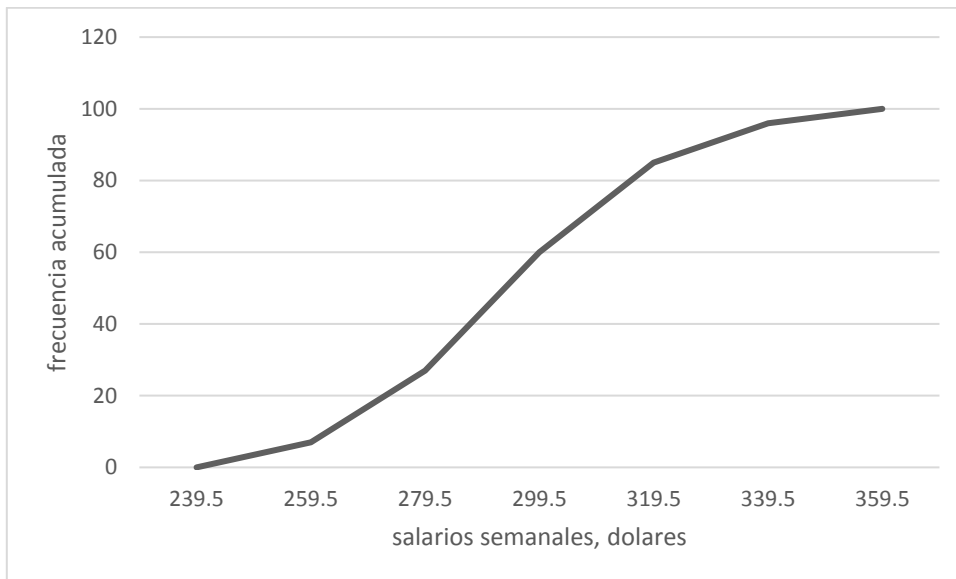
En términos de curtosis, una curva de frecuencia puede ser: 1) platicurtica: plana, la distribución de las observaciones es relativamente uniforme en todas las clases; 2) leptocurtica: picuda, las observaciones se encuentran dentro de un estrecho intervalo de valores, o 3) mesocurtica: ni plana ni picuda, en términos de la distribución de los valores observados.

Distribuciones de frecuencia acumulada

Una distribución de frecuencia acumulada indica el número de observaciones acumulado debajo del límite superior exacto de cada clase en la distribución. La frecuencia acumulada de una clase se determina sumando la frecuencia observada en esa clase a la frecuencia acumulada de la clase anterior.

salario semanal	límite de clase exacto superior	numero de trabajadores(f)	frecuencia acumulada (fa)
240-259	259.5	7	7
260-279	279.5	20	20+7=27
280-299	299.5	33	33+27=60
300-319	319.5	25	25+60=85
320-339	339.5	11	11+85=96
340-359	359.5	4	4+96=100
		total 100	

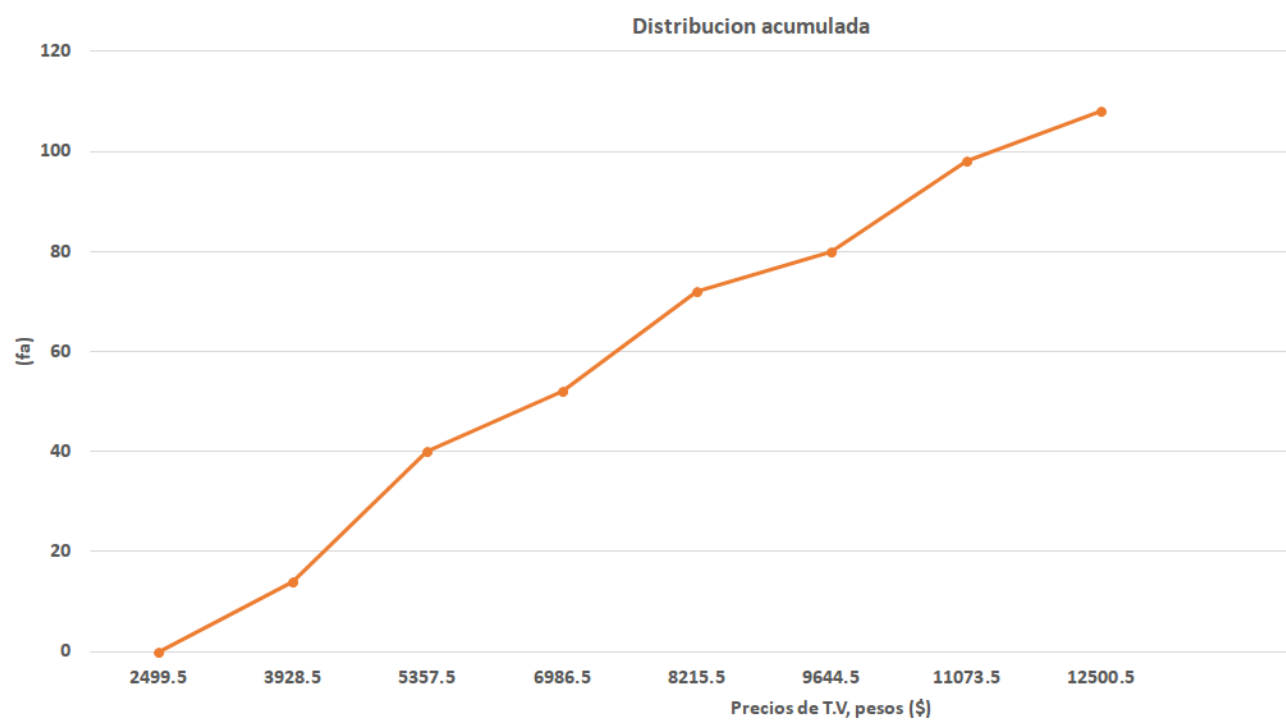
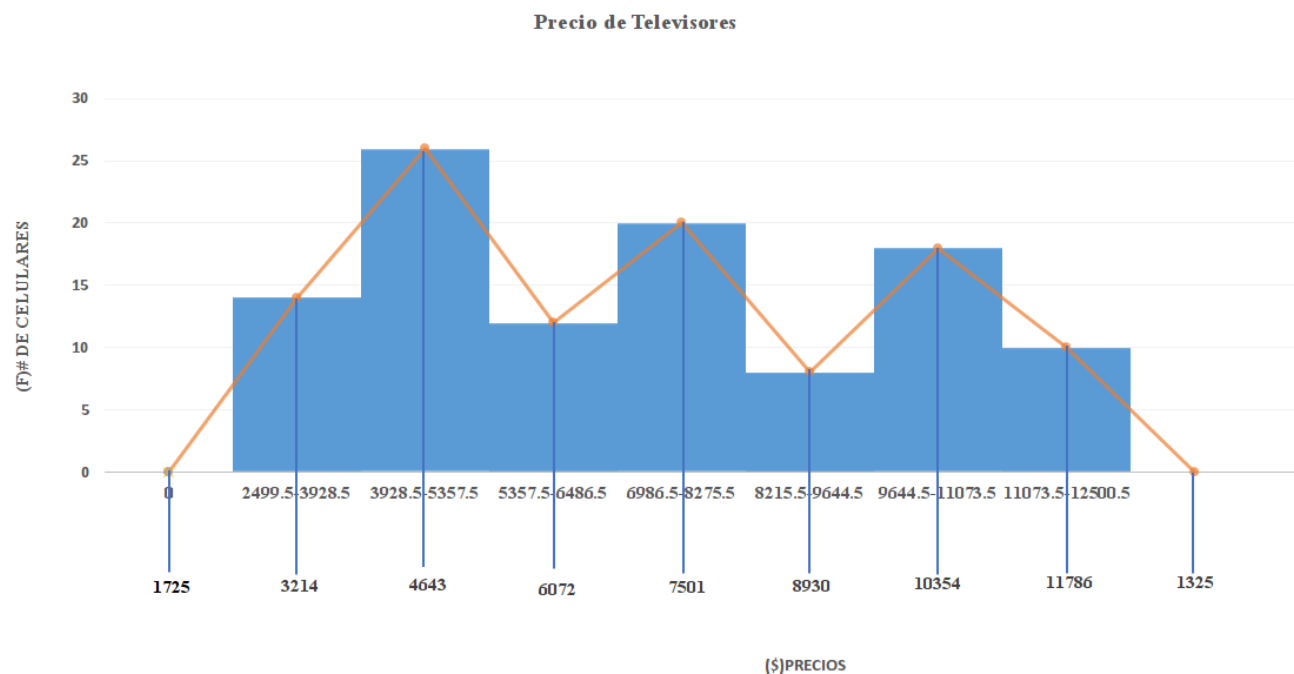
A la gráfica de una distribución de frecuencia acumulada se le llama ojiva. En el tipo de distribución acumulada “menor que”, esta grafica indica la frecuencia acumulada debajo de cada uno de los límites de clase exactos. Si una gráfica de línea de este tipo se suaviza, se le llama curva de ojiva.



Distribuciones de frecuencia relativa

Una distribución de frecuencia relativa es aquella en la cual el número de observaciones correspondientes a cada clase se ha convertido en una frecuencia relativa mediante la división entre el número total de observaciones en toda la distribución. Cada frecuencia relativa es entonces una proporción, y puede convertirse en porcentaje al multiplicarla por 100.

Una de las ventajas de una distribución de frecuencia relativa es que la distribución acumulada y la ojiva de una distribución de este tipo indican la proporción (o el porcentaje) de las observaciones acumuladas hasta los diferentes valores de la variable. Un valor de percentil es el porcentaje de observaciones acumulado hasta un valor determinado de la variable.

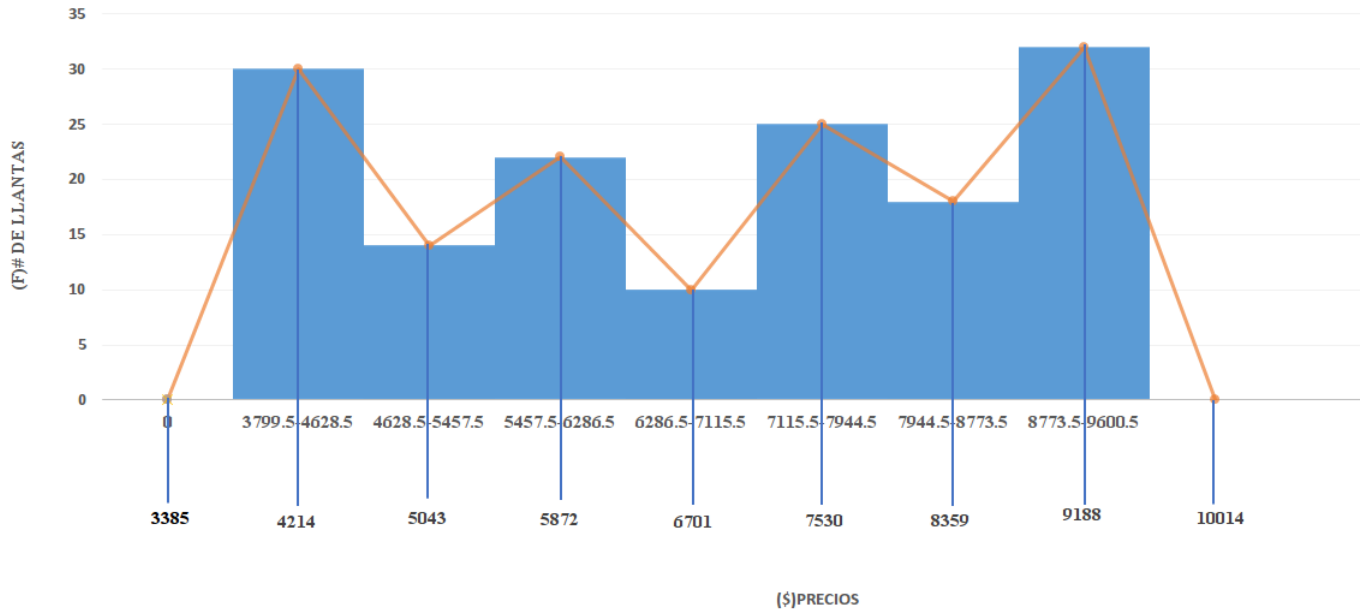


El precio de las llantas va desde 3800 hasta 9600 pesos. Realiza una distribución de frecuencia en 7 clases y describe el histograma el polígono de frecuencia y la ojiva. El # por llantas es: 30, 14, 22, 10, 25, 18, 32

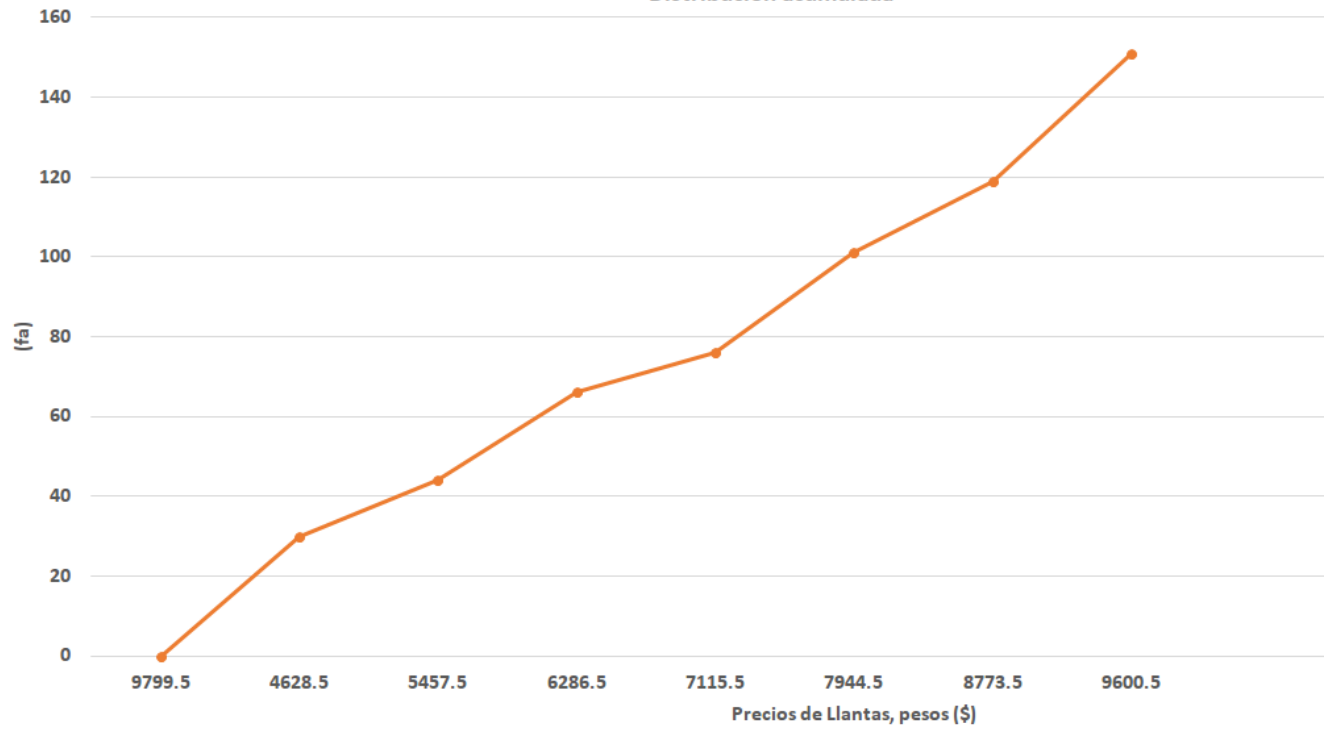
$$\text{Int. Aproximado} = \frac{9600 - 3800}{7} = \$828.57 \approx \$829.00$$

Precio de las llantas (\$)	Limites exactos de clase (\$)	Puntos medios (\$)	(f) # de llantas	f(a) # de llantas
3800-4628	3799.5-4628.5	4214	30	30
4629-5458	4628.5-5457.5	5043	14	44
5458-6287	5457.5-6286.5	5872	22	66
6287-7116	6286.5-7115.5	6701	10	76
7116-7945	7115.5-7944.5	7530	25	101
7945-8774	7944.5-8773.5	8359	18	119
8774-9600	8773.5-9600.5	9188	32	151

Precio de Llantas



Distribucion acumulada

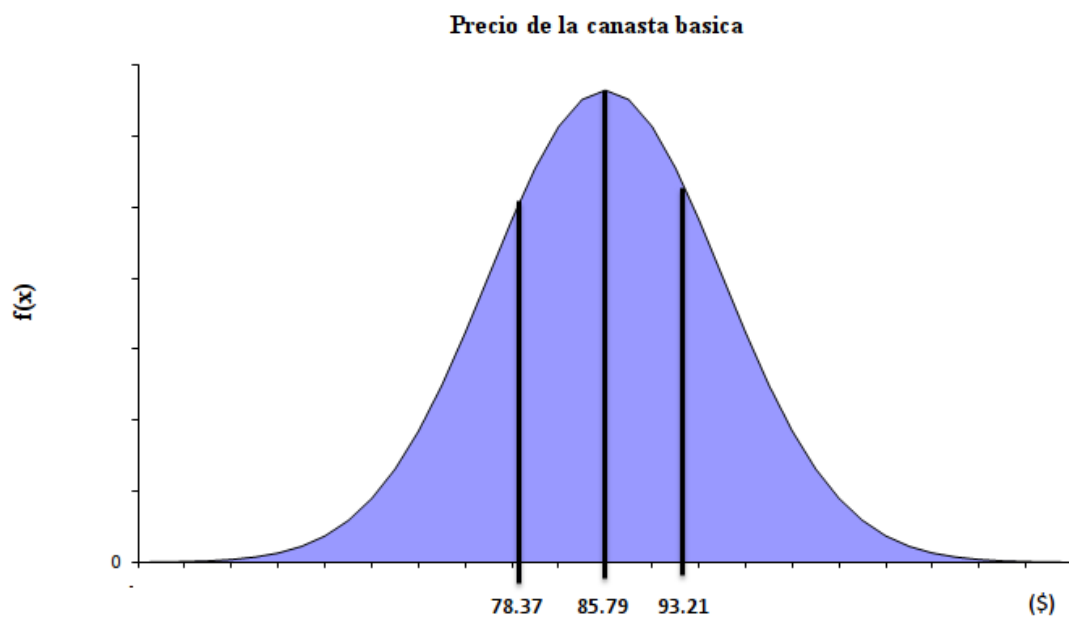


Ejemplo

El precio de 10 productos de una canasta básica es el siguiente: 89.90, 78.40, 96.70, 85.75, 90.60, 77.80, 81.90, 98.70, 81.30, 76.80. Calcular la media aritmética poblacional, la desviación estándar y graficar el resultado.

$$\begin{aligned}\mu &= \frac{\sum X}{N} \\ \mu &= \frac{857.35}{10} = 85.79 \\ \sigma &= \frac{\sqrt{\sum X^2 - N\mu^2}}{N} \\ \sigma &= \frac{\sqrt{74450.95 - 73599.24}}{10} \\ \sigma &= \frac{\sqrt{587.6}}{10} = 7.42\end{aligned}$$

x	x^2
89.90	8082.01
78.40	6246.56
96.70	9350.84
85.75	7353.06
90.60	8208.36
77.80	6052.84
81.90	6707.69
98.70	9744.69
81.30	6609.69
76.90	5898.24
$\sum = 857.85$	$\sum = 71498.24$



Medidas de tendencia central

Medidas de posición

Medidas de posición en conjuntos de datos

Una medida de posición es un valor que se calcula a partir de un grupo de datos y se usa para describir los datos. Por lo general se desea que el valor sea representativo de todos los valores del grupo, y así se busca algún promedio. Un promedio es una medida de tendencia central, este capítulo estudia los procedimientos estadísticos relacionados con las medidas de posición

La Media Aritmética

La media aritmética o promedio aritmético se define como la suma de los valores en el grupo de datos dividida entre el número de valores.

En la estadística, las medidas descriptivas de la población, se representa por lo general con una letra griega mientras que las medidas descriptivas de una muestra, se representan con una letra latina. La media aritmética de una población de valores se representa con el símbolo $\bar{X}$

En tanto que la media aritmética de valores se representa con el símbolo $\bar{x}$.

Las fórmulas para obtener la media poblacional y la media muestral son:

Operacionalmente las dos fórmulas son idénticas; en ambos casos se suman todos los valores ($\sum X$) y luego se divide el resultado entre el número de valores. La diferencia en los denominadores es que en el análisis estadístico la N por lo general indica el número de elementos de la población, mientras que la n indica el número de elementos en la muestra

La media ponderada:

La media ponderada o promedio ponderado es una media aritmética en la cual cada valor se pondera de acuerdo con su importancia en todo el grupo. Las fórmulas para las medias ponderadas poblacional y muestral son idénticas:

Cada valor en el grupo (X) se multiplica por el factor de ponderación apropiada (w), después se suman los productos y la suma se divide entre la suma de los factores de ponderación.

Sin embargo, a menos que los cuatro productos tengan ventas iguales este promedio no ponderado es incorrecto. Suponiendo que los totales de ventas sean los que se muestran en la tabla 3.1, la media ponderada describe correctamente al promedio

La Mediana

La mediana de un grupo de elementos es el valor del elemento de en medio una vez que todos los elementos del grupo se han ordenado en orden ascendente de acuerdo con su valor. En un grupo en el que el número de elementos es par, se considera que la mediana es el promedio de los dos valores de en medio. Cuando el grupo contiene un número grandes de valores, es útil emplear la siguiente fórmula para determinar la posición de la mediana en el grupo ordenado

El valor de la mediana está entre el 4to y 5to valor del grupo ordenado, ya que estos dos valores son iguales a “11” la mediana en este caso es 11.0

La Moda

Es el valor que se presenta con mayor frecuencia en un conjunto de valores. Una distribución de este tipo se dice que es uní modal. En un conjunto pequeño de datos en el cual ninguno de los valores medidos se repite, no hay moda. Cuando dos valores no contiguos tienen la misma frecuencia máxima, se dice que la distribución es bimodal. Las distribuciones de mediciones que tienen varias modas se denominan multimodales.

Relación entre Moda y Mediana

En cualquier distribución simétrica, la media, la mediana y la moda coinciden en valor. En una distribución positivamente sesgada, la media siempre es menor que la mediana. Estas dos últimas relaciones siempre son verdaderas, independientemente de si la distribución es unimodal o no. Una medida del sesgo, que se centra en la diferencia entre los valores de la media y la mediana de un grupo de valores, es el coeficiente de asimetría de Pearson, que se describe en la sección 4.12. Los conceptos de simetría y de sesgo se explican en la sección 2.4

Criterios matemáticos que satisfacen la mediana y la media

El propósito de a determinar cualquiera de las medidas de tendencia central, como la mediana o la media es usarla para representar el nivel general de los valores que conforman el grupo tanto la mediana como la media son medidas representativas, pero desde puntos de vista distintos, objetivos y criterios matemáticos. La mediana es un valor representativo que minimiza la suma de los valores absolutos de las diferencias entre cada valor del grupo y la mediana minimiza la suma de las desviaciones absolutas respecto de los valores individuales representados. La media aritmética minimiza la suma de las desviaciones al cuadrado respecto de los valores individuales del grupo. Al criterio mediante el cual el objetivo es la

minimización de la suma de las divisiones al cuadrado correspondientes a un valor representativo se le denomina criterio de los mínimos cuadrados. Este criterio es el más importante en la inferencia estadística basada en datos muestrales

Uso de la media, la mediana y la moda

El valor de la moda indica donde se localizan más de los valores observados puede ser útil como una medida descriptiva de un grupo poblacional pero solo si hay una sola moda clara la mediana siempre es una excelente medida para representar el nivel “típico” de los valores que se observan, por ejemplo, los salarios, en una población. La falta de simetría no es un problema especial porque la mediana de los salarios, por ejemplo, siempre es el salario de la ‘persona del centro’ cuando los salarios se han ordenado de acuerdo con su magnitud. En caso de datos asimétricos, los valores extremos (por ejemplo, unos cuantos salarios muy elevados de los técnicos especialistas) causan distorsión del valor de la media como valor representativo.

Ahora se considera el uso de las tres medidas de posición para datos muestrales. La moda no es buena medición de posición para datos muestrales porque su valor puede variar mucho de una muestra a otra. La mediana es mejor que la moda porque su valor es más estable y se explicara a fondo en el capítulo 8, la razón de la estabilidad relativa de la media muestral, de una muestra a otra, se debe a que es la medida de posición que satisface el criterio de los mínimos cuadrados. Así, para datos muestrales, la mejor medida de posición es por lo general la media aritmética.

Uso de la media en el control estadístico de procesos

Como ya se observó en la sección anterior, la media muestral es más estable que la mediana o la moda. Por esta razón las gráficas de corrida de datos que tienen que ver con los promedios muestrales se enfocan a graficar medias muestrales. Una gráfica de este tipo se conoce como una gráfica $\bar{X}$, y sirve de base para determinar si un proceso es estable o si existe alguna variación en el mismo que tenga una causa asignable que deba ser corregida.

Cuartiles, Deciles y Percentiles

Los cuartiles, los deciles y los percentiles son similares a la mediana en que también subdividen una distribución de mediciones de acuerdo con la proporción de frecuencias observadas. Mientras que la mediana divide a una distribución en mitades, los cuartiles la dividen en cuatro (cuatro partes iguales), los deciles la dividen en decimos (diez partes iguales) y los puntos percentiles la dividen en 100 partes.

$$Q_1(\text{primer cuartil}) = X_{[(n/4) + (1/2)]}$$

$$D_3(\text{tercer decil}) = X_{[(3n/10) + (1/2)]}$$

$$P_{70}(\text{septuagésimo percentil}) = X_{[(70n/100) + (1/2)]}$$

Distribución de probabilidad de variables aleatorias discretas: binomial, hipergeométrica y de poisson.

¿Que es una variable aleatoria?

A diferencia de los eventos categóricos, como extraer determinadas cartas de una baraja de naipes, que se vio en el Capítulo 5, una variable aleatoria se define como un evento numérico cuyo valor se determina mediante un proceso aleatorio. Si a cada uno de los posibles valores numéricos de una variable aleatoria x se le asigna un valor de probabilidad, ya sea mediante una lista o una función matemática, el resultado es una distribución de probabilidad. La suma de las probabilidades de todos los posibles resultados numéricos debe ser 1.0. cada uno de los valores de probabilidad se puede denotar mediante el símbolo $f(x)$, que indica que interviene una función matemática, mediante $p(x = X)$, que indica que la variable aleatoria puede adoptar diferentes valores, o simplemente mediante $p(X)$.

Si se trata de una variable aleatoria discreta, los valores observados solo pueden presentarse en puntos aislados a lo largo de una escala de valores. Por tanto, se pueden listar en una tabla los valores numéricos de la variable con sus probabilidades correspondientes. Hay varias distribuciones de probabilidad estándar que pueden servir como modelos para una amplia gama de variables aleatorias discretas que se utilizan en los negocios. Los modelos estándar que se describen son la probabilidad binomial, hipergeométrica y de Poisson.

En el caso de una variable aleatoria continua, no pueden listarse todos los posibles valores fraccionarios de la variable y por tanto las probabilidades se determinan a través de una función matemática y se representan gráficamente mediante una función de densidad de probabilidad o curva de probabilidad.

Demanda posible x	Número de días	Probabilidad [P(x)]
3	3	0.06
4	7	0.14
5	12	0.24
6	14	0.28
7	10	0.20
8	4	0.08
	50	1.00

Descripción de una variable aleatoria discreta

De la misma manera como ocurre con las colecciones de datos muestrales o poblacionales, con frecuencia es útil describir una variable aleatoria en términos de su media y de su varianza o desviación estándar. La media de una variable aleatoria x se llama valor esperado y se denota por $E(X)$. en el caso de una variable aleatoria discreta la media es el promedio ponderado de todos sus posibles valores numéricos empleando las probabilidades respectivas con factores de ponderación. Debido a que la suma de los factores de ponderación es 1.0, la formula se puede simplificar, y el valor esperado de una variable aleatoria discreta es

$$E(X) = \sum XP(X)$$

Demanda posible X	Probabilidad [P(X)]	Valor ponderado[XP(X)]
3	0.06	0.18
4	0.14	0.56
5	0.24	1.20
6	0.28	1.68
7	0.20	1.40
8	0.08	0.64
	1.00	E(X)=5.66

La Distribución Binomial

Es una distribución discreta de probabilidad que es aplicable como un modelo en situaciones como toma de decisiones en las que se supone que el proceso de muestreo será realizado conforme un proceso de Bernoulli. Un proceso de Bernoulli es un proceso de muestreo en el cual

- 1) En cada ensayo solo pueden presentarse dos resultados u observaciones mutuamente excluyentes. Para simplificar, a estos resultados se les llama éxitos y fracasos.
- 2) Los resultados de una serie de ensayos u observaciones constituyen eventos independientes
- 3) La probabilidad de éxitos en cada ensayo denota por P , permanece constante de un ensayo a otro. Es decir, el proceso es estacionario.

La distribución binomial se usa para determinar la probabilidad de obtener un número determinado de éxitos con un proceso de Bernoulli. Se requieren tres valores el número determinado de (X), el número de ensayos u observaciones (N) y la probabilidad de éxito para cada prueba (P). si $q=(1-P)$, la fórmula para determinar la probabilidad en un número determinado de éxitos de X en una distribución binomial es

$$P(X \text{ en } n, p) = {}^n C_x P^x Q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

Con frecuencia lo que interesa es la probabilidad acumulada de la ocurrencia “ X o más” éxitos o “ X o menos” éxitos en N ensayos. En un caso como este se determina la probabilidad de cada resultado comprendido en el intervalo determinado, y luego se suman estas probabilidades.

- 1.1. El número de automóviles vendidos por cada uno de los 10 agentes de ventas en una distribuidora de automóviles durante el mes específico. En orden ascendente, es: 2, 4, 7, 10, 10, 10, 12, 12, 14, 15. Determine a) la media, b) La mediana, c) La moda de la población del número de automóviles vendidos.

- 1.2. En una muestra de paquetes que salen de una oficina de correos, se encuentra que los pesos, dados hasta la onza más cercana, son: 21, 18, 30, 15, 14, 17, 28, 10, 16, 25, 25 onzas. Determine a) La media, b) La mediana.

- 1.3. Veinte estudiantes inscritos en un curso de análisis de decisiones obtuvieron en el examen las calificaciones siguientes, presentadas en orden ascendente: 39, 46, 57, 65, 70, 72, 72, 77, 79, 81, 81, 84, 84, 84, 87, 93, 94, 97, 97. Determine a) La media, b) La mediana y c) La moda de estas calificaciones.

- 1.4. Determine los valores de a) el primer cuartil, b) el segundo decil y c) el punto percentil 30 de los montos de venta del problema 1.1.

- 1.5. Con los datos del problema 1.2., determine los pesos correspondientes a) al tercer cuartil, b) al tercer decil y c) al punto percentil 70.

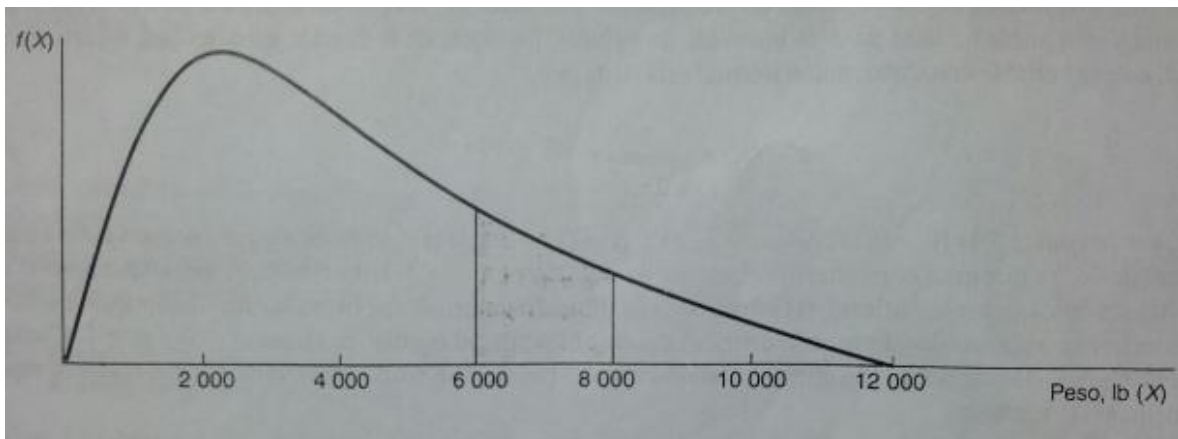
- 1.6. Las siguientes calificaciones de exámenes, colocadas en orden ascendente, fueron obtenidas por 20 estudiantes inscritos en un curso de análisis: 39, 46, 57, 65, 70, 72, 72, 75, 77, 79, 81, 81, 84, 84, 84, 87, 93, 94, 97, 97. Determine a) el rango y b) el rango del 90 por ciento central de estos datos.

- 1.7. El número de accidentes que ocurrieron durante cierto mes en los 13 departamentos de una planta industrial fue: 2, 0, 0, 3, 3, 12, 1, 0, 8, 1, 0, 5, 1. Determine a) el rango y b) el rango intercuartílico del número de accidentes.

Distribuciones de Probabilidad para variables aleatorias continuas: distribución normal y distribución exponencial.

Las variables aleatorias continuas.

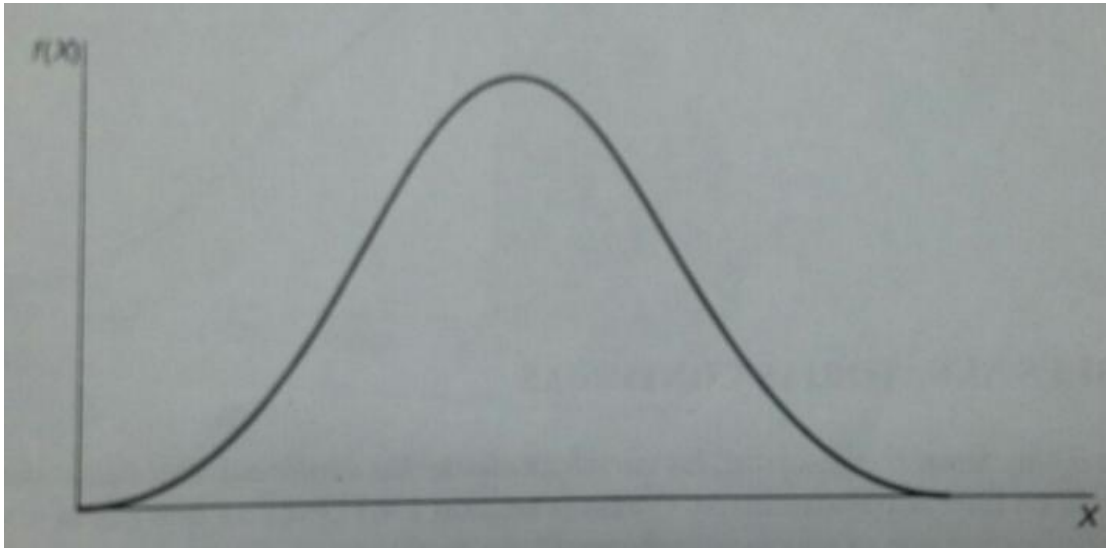
A diferencia de las variables aleatorias discretas, las variables aleatorias continuas son aquellas que pueden tomar valor dentro de un intervalo determinado. Debido a que existe una cantidad infinita de posibles valores fraccionarios que se obtengan como resultado de alguna medición, no se pueden listar todos los posibles valores con sus probabilidades correspondientes. En lugar de esto, se define una función de densidad de probabilidad. Esta expresión matemática de la $F(X)$, que se representa por el símbolo $f(X)$, corresponde a cualquier valor que se desee de la variable aleatoria X . a la gráfica de estas funciones se le llama curva de probabilidad, y el área entre dos puntos cualesquiera y bajo la curva indica la probabilidad de ocurrencia aleatoria de cualquiera de los valores entre estos dos puntos.



Hay varias distribuciones de probabilidad continuas estándar que se pueden emplear como modelo para una amplia gama de variables continuas de acuerdo de las circunstancias de las que se trate. Para estas distribuciones estándar se han elaborado tablas de probabilidad que hacen innecesario el uso de la integración para determinar las áreas bajo las curvas de probabilidad de estas distribuciones. Los modelos de probabilidad continua estándar que se describe son las distribuciones de probabilidad normal y exponencial.

La Distribución de Probabilidad Normal.

La distribución de probabilidad normal es una distribución continua que es simétrica y mesocurtica. Con frecuencia la curva de probabilidad que representa a una distribución de probabilidad normal se suele describir como una curva en forma de campana.



La distribución de probabilidad normal es importante en la estadística inferencial por tres razones:

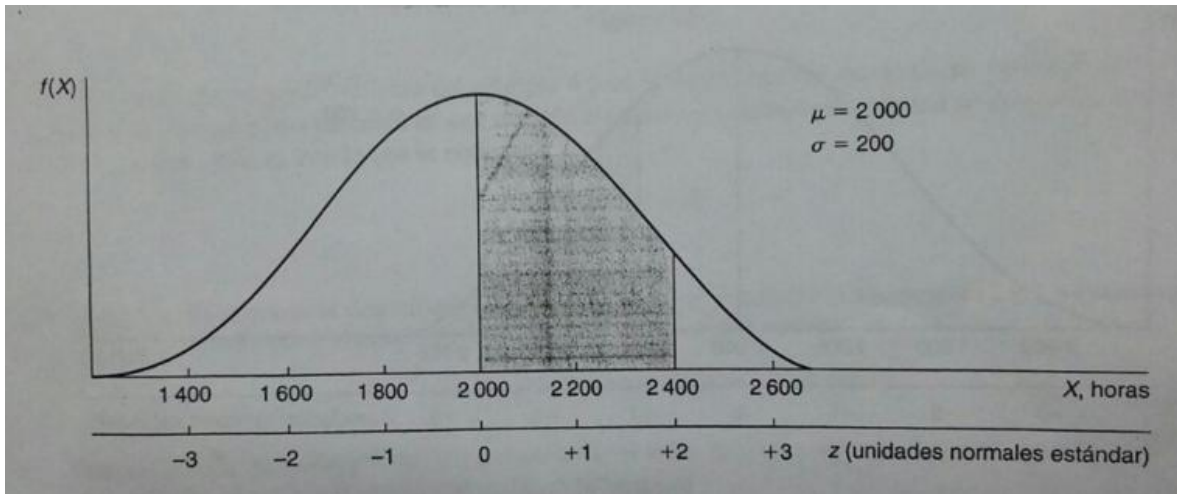
- 1) Se sabe que en muchos procesos aleatorios los datos obtenidos mediante mediciones se ajustan a este tipo de distribución.
- 2) Con frecuencia se usan las probabilidades normales para aproximar otras distribuciones de probabilidad como las distribuciones binomiales y de Poisson.
- 3) Las distribuciones de estadísticos como la media muestral y la proporción muestral cuando el tamaño de la muestra es grande, se ajustan a una distribución normal, independientemente de la distribución progenitora

Como sucede con cualquier distribución de probabilidad continua el valor de probabilidad de una variable aleatoria continua solo puede hallarse para un intervalo de valores.

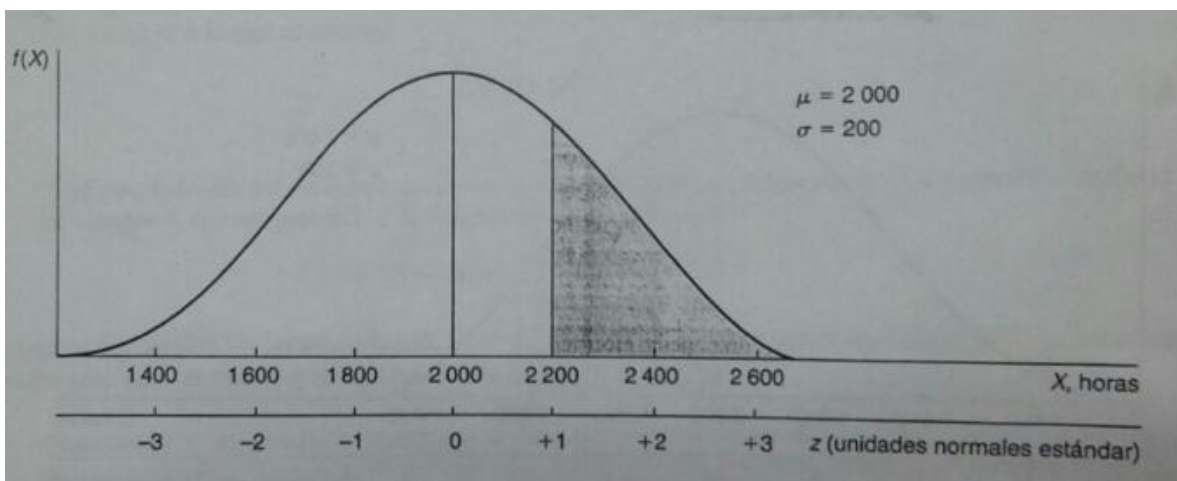
Un valor z reformula el valor original X en términos del número de unidades de desviaciones estándar en las que el valor original difiere de la medida de la

distribución. Un valor negativo de z indicara que el valor de X original es menor que el valor de la media.

El apéndice 5 de la proporción de las áreas correspondientes a diversos intervalos de valores en la distribución de probabilidad normal estándar, siendo siempre la media el límite inferior del intervalo. La conversión de los valores deseados de la variable X en valores normales permite el uso de esta tabla y hace innecesarios el uso de métodos de integración respecto de la ecuación de la función de densidad.

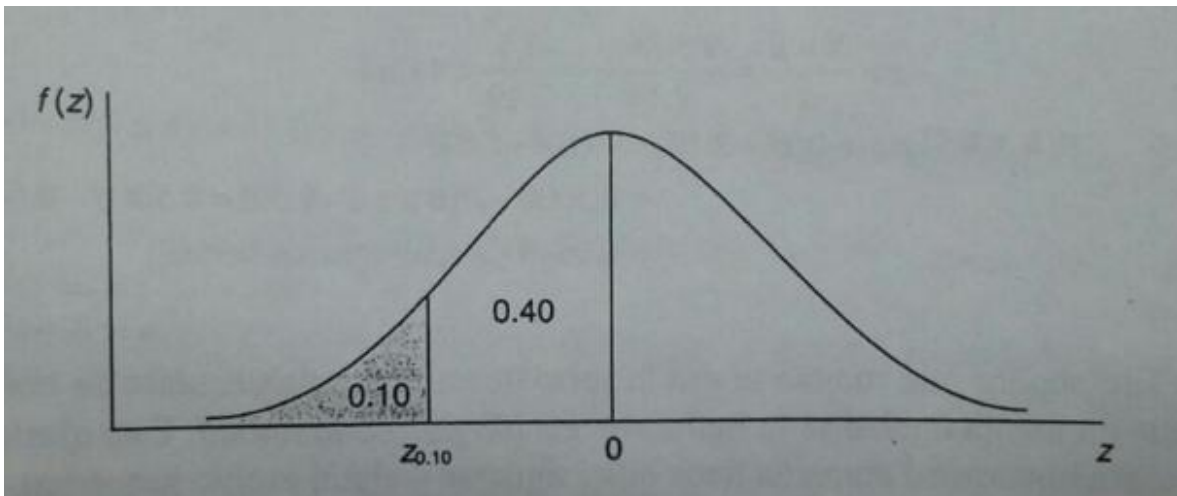
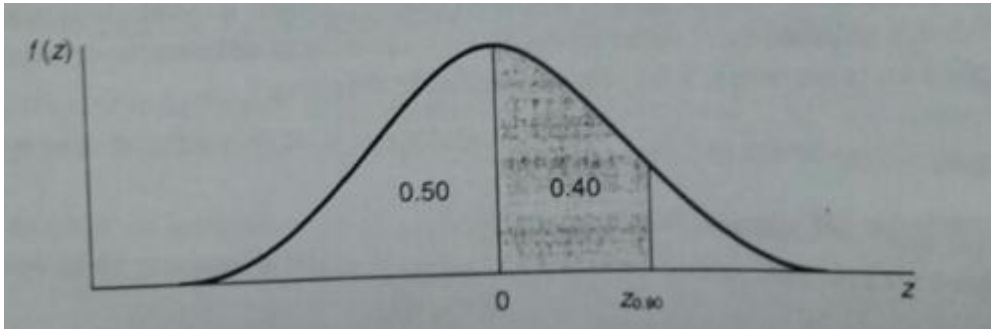


Por supuesto que no en todos los problemas el intervalo en que se trata tiene a la media como frontera interior. Sin embargo, en el apéndice 5 sirve para determinar el valor de probabilidad correspondiente a cualquier intervalo que desee, ya sea mediante una adición o sustracción de áreas o reconociendo que la curva es simétrica.



Puntos percentiles para variables con una distribución normal

Recuerde de la sección de 3.10 que, por ejemplo, el punto de percentil 90 es aquel punto en una distribución tal que 90 por ciento de los valores están debajo de este punto y 10 por ciento de los valores están arriba de este punto. En la distribución normal estándar, es el valor de z tal que la proporción total del área a la izquierda de este valor de la curva normal es 0.90.



Aproximación normal a las probabilidades binomiales

Cuando el número de observaciones o ensayos n es relativamente grande, se usa la distribución de probabilidad normal para aproximar probabilidades binomiales. Una regla útil es que esta estimación es aceptable cuando N es mayor que 30 y tanto NP mayor que 5 como NQ mayor que 5.

Cuando se usa la distribución normal de probabilidad para aproximar un valor de probabilidad binomial, la media y la desviación estándar se basan en el valor esperado y en la varianza del número de éxitos en la distribución binomial. El número de éxitos es

$$U=np$$

La desviación estándar del número de éxitos es

$$O=\sqrt{npq}$$

Aproximación normal a las probabilidades de Poisson.

Cuando la medida λ de una distribución de Poisson es relativamente grande, se puede usar la distribución de probabilidad normal para aproximar probabilidades de Poisson. Una regla útil es que esta estimación es aceptable cuando $\lambda \geq 10.0$

La media y la desviación estándar de la distribución de probabilidad normal se basan en el valor esperado y en la varianza del número de eventos en un proceso de Poisson. Esta media es

$$U=\lambda$$

La desviación estándar es

$$O= \sqrt{\lambda}$$

Pruebas de hipótesis para la media población

Introducción:

El propósito de la prueba de hipótesis es determinar si un valor propuesto (hipotético) para un parámetro poblacional, por ejemplo, para una media, debe aceptarse como plausible con base en la evidencia muestral. Recuerde las distribuciones de muestreo, en general el valor de una media muestral difiere del valor de la media poblacional. Si el valor de un estadístico muestral, como la media muestral es cercano al valor propuesto con parámetro y solo difiere en una cantidad que resulta des esperarse debido al muestreo aleatorio, entonces no se rechaza el valor hipotético. Si el estadístico muestral difiere del valor propuesto en una cantidad que no es atribuible a la casualidad, entonces se rechaza la hipótesis por no considerarse plausible. Se han desarrollado tres métodos para pruebas de hipótesis, todos ellos conducentes a la misma decisión cuando se usan los mismos estándares de probabilidad y de riesgo. Independientemente del método que se utilice en las pruebas de hipótesis, observe que cuando se rechaza el valor hipotético, y por tanto se acepta, esto no constituye una prueba de que el valor

hipotético sea correcto. La aceptación de un valor propuesto como un parámetro simplemente indica que es un valor plausible de acuerdo con el valor observado en el estadístico muestral.

PASOS BÁSICOS EN LAS PRUEBAS DE HIPÓTESIS USANDO EL MÉTODO DE VALORCRÍTICO

Paso 1.

Formular la hipótesis nula y la hipótesis alternativa. La hipótesis nula (H_0) es valor paramétrico hipotético que se compara con el resultado muestral. La hipótesis nula se rechaza solo si es poco probable que el resultado muestral siendo la hipótesis correcta. La hipótesis alternativa (H_1) se acepta solo si la hipótesis nula se rechaza. En muchos libros de texto la hipótesis alternativa también se denota como H_a .

Ejemplo 1

Un auditor toma una muestra de $n=36$ y calcula la media muestral desea probar la suposición de que el valor medio de todas las cuentas por cobrar en una determinada empresa sea \$260.00. El auditor desea rechazar este valor supuesto de \$260.00 solo si la media muestral lo contradice claramente, y así, en este procedimiento de prueba, al valor hipotético deberá otorgársele el beneficio de la duda. Las hipótesis nulas y alternativa en esta prueba son $H_0: \$260.00$ y $H_1: \mu \neq \$260.00$.

Paso 2. Especificar el nivel de significancia que habrá de usarse. El nivel de significancia es el criterio estadístico que se establece para rechazar la hipótesis nula. Si se establece 5% como nivel de significancia, entonces la hipótesis nula se rechaza solo si el resultado muestral es tan diferente del valor hipotético que la probabilidad de que una diferencia de esa magnitud o mayor se dé por casualidad es de por casualidad es de 0.05 o menos.

Observe que, si se usa como nivel de significancia 5%, existe una probabilidad de 0.05 de rechazar la hipótesis nula aun cuando sea verdadera. A esto se le conoce como error tipo I. La probabilidad de un error de tipo I es siempre igual al nivel de significancia que se utiliza como criterio para rechazar la hipótesis nula; al error tipo I se le designa mediante la letra griega minúscula α alfa y entonces α también designa el nivel de significancia. Los niveles de significancia que se usan con más frecuencia en prueba de hipótesis son los niveles de 5% y de 1%.

Un error de tipo II ocurre cuando no se rechaza la hipótesis nula, y por lo tanto se acepta, siendo falsa.

Situaciones posibles.		
	Hipótesis nula verdadera	Hipótesis nula falsa
Aceptar la hipótesis nula	Aceptación correcta	Error tipo II
Rechazar la hipótesis nula	Error tipo I	Rechazo correcto

Paso 3. Elegir el estadístico de prueba. El estadístico de prueba es el estadístico muestral o una versión estandarizada del estadístico muestral. Por ejemplo, con objeto de probar un valor hipotético de la media poblacional, como estadístico de prueba puede emplearse la media de una muestra aleatoria tomada de esa población. Sin embargo, si la distribución de muestreo para la media tiene distribución normal, entonces es común que el valor de la media muestral se convierta a un valor Z el cual sirve entonces como estadístico de prueba.

Paso 4. Establecer el valor o los valores críticos del estadístico de prueba. Una vez especificados la hipótesis nula, el nivel de significancia y el estadístico de prueba que se usaran, se establecen los valores críticos del estadístico de prueba.

Puede haber uno o dos de estos valores, dependiendo de si se trata de una prueba unilateral o bilateral. En cualquiera de los dos casos un valor crítico establece el valor del estadístico de prueba que se requiere para rechazar la hipótesis nula.

Paso 5. Determinar el valor del estadístico de prueba. Por ejemplo, al probar un valor hipotético para la media poblacional se toma una muestra aleatoria y se determina el valor de la media muestral. Si el valor crítico se fijó como un valor Z, entonces la media muestral se convierte a un valor Z.

Paso 6. Tomar la decisión. El valor del estadístico muestral obtenido se compara con los valores críticos del estadístico de prueba. A continuación, la hipótesis nula se acepta o se rechaza. Si se rechaza la hipótesis nula, se acepta la alternativa. A su vez, esta decisión tendrá relevancia en otras decisiones que tienen que tomar

los administradores de operación, por ejemplo, si se mantiene o no un estándar de rendimiento o cual de dos estrategias de mercado se debe utilizar.

La distribución de probabilidad normal se puede usar para probar un valor hipotético para la media poblacional 1) siempre que $n \geq 30$, debido al teorema del límite central, o 2) cuando $n < 30$ pero la población tiene distribución y se conoce σ .

Fórmulas para calcular valores críticos.

$$\bar{X}_{CR} = \mu_0 \pm Z \sigma_{\bar{X}}$$

$$\bar{X}_{CR} = \mu_0 \pm Z S_{\bar{X}}$$

Ejemplo 2: Dada la hipótesis nula formulada en el ejemplo 1, determine los valores críticos para la media muestral si se quiere probar la hipótesis con un nivel de significancia del 5%. Dado que se sabe que la desviación estándar de los montos de las cuentas por cobrar es $\sigma = \$43.00$ los valores críticos son:

Hipótesis: $H_0: \mu = \$260.00$ $H_1: \mu \neq \$260.00$

Nivel de significancia: $\alpha = 0.05$

Estadístico de prueba: $\bar{X}$ obtenida de una muestra de $n = 36$ en la que $\sigma = 43.00$

$$\bar{X}_{CR} = \mu_0 \pm Z \sigma_{\bar{X}} = 260 \pm 1.96 \frac{\sigma}{\sqrt{n}} = 260 \pm 1.96 \frac{43}{\sqrt{36}}$$

$$= 260 \pm 1.96 (7.17) = 260 \pm 14.05 = \$245.95 \text{ y } \$274.05$$

Por tanto, para rechazar la hipótesis nula la media muestral debe tener un valor menor que \$245.95 o mayor que \$274.05. Así, de una prueba hay dos regiones rechazo. Los ± 1.67 se usan para establecer los valores críticos, debido a



o mayor en el caso bilateral de valores Z para valores que en la

distribución normal estándar en las dos colas queda una proporción de 0.05 del área, lo que corresponde al valor se usan para establecer los valores críticos, debido a que en la distribución normal estándar en las dos colas queda una proporción de 0.05 del área, lo que corresponde al valor $\alpha=0.05$ que se fijó.

En las pruebas de hipótesis los valores críticos suelen especificarse en términos de valores de Z en lugar de establecer en términos de la media muestral. Por ejemplo, los valores críticos Z para el nivel de significancia de 5% en la prueba bilateral son -1.96 y +1.96. Cuando se determina el valor de la media muestral, este se convierte a un valor Z de modo que este valor pueda compararse con los valores críticos Z. La fórmula de conversión, de acuerdo con si se conoce o no σ , es:

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$$

$$Z = \frac{\bar{X} - \mu_0}{S_{\bar{X}}}$$

Ejemplo 3: En el problema de la prueba de hipótesis de los ejercicios 1 y 2, suponga que la media muestral es $\bar{X} = \$240$. Para determinar si se debe rechazar la hipótesis nula, esta media se convierte a un valor Z y se compara con los valores críticos ± 1.96 como sigue:

$$\sigma_{\bar{X}} = 7.17 \quad \text{Según el ejercicio 2}$$

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{240 - 260}{7.17} = \frac{-20}{7.17} = -2.79$$

En el modelo para las pruebas de hipótesis, este valor de Z se encuentra en la región de rechazo de la cola izquierda. Así la hipótesis nula se rechaza y se acepta la hipótesis alternativa $H_1: \mu \neq \$260.00$.

Observe que en el ejercicio 2 se llegara a la misma conclusión al comparar la media muestral $\bar{X} = \$240$ con los valores críticos para la media anteriormente indicados.



Ejercicio 4: De los ejercicios 1 a 3 suponga que el auditor parte de la hipótesis alternativa de que el valor medio de todas las cuentas por cobrar es menor que \$260. Dado que la media muestral es \$240, a continuación, se prueba esta hipótesis con un 5% como nivel de significancia mediante los procedimientos siguientes.

1) Determinando el valor crítico para la media muestral, cuando

($H_0: \mu = \$260.00$ y $H_1: \mu < \$260.00$.)

$$\bar{X}_{CR} = \mu_0 \pm Z\sigma_{\bar{X}} = 260 + (-1.641)(7.17) = 248.21$$

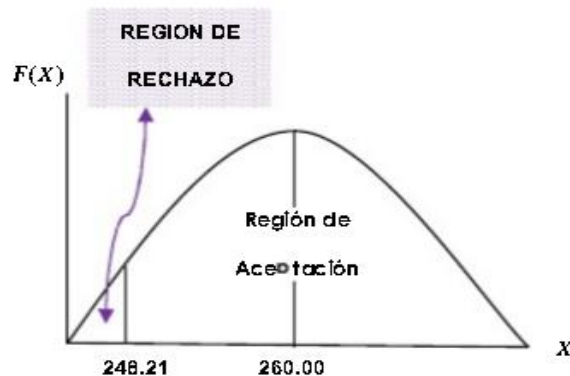
Como $\bar{X} = 240$ este valor se encuentra en la región de rechazo. Por tanto se rechaza la hipótesis nula y se acepta la hipótesis alternativa $\mu < \$260$.

1) Determinando el valor crítico en términos de Z, donde z critico

$$(\alpha = 0.05) = -1.645:$$

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{240 - 260}{7.17} = \frac{-20}{7.17} = -2.79$$

Como $Z = -2.79$, esta región de rechazo a la izquierda del valor critico -1.64, la hipótesis nula se rechaza. Y esto se representa en la gráfica siguiente.



Error tipo uno I y tipo II en pruebas de hipótesis.

Como ya se había mencionado anteriormente la probabilidad máxima de error tipo I se designa con la letra griega α alfa. Esta probabilidad es siempre igual al nivel de significancia que se usa para probar la hipótesis nula. Esto se debe a que por definición la proporción de área en la región de rechazo es igual a la proporción de resultados muestrales que se darían en esa región dado que la hipótesis nula fuera verdadera

La probabilidad del error tipo II por lo general se designa con la letra griega β beta. La única manera de determinar esta probabilidad es en relación con un valor específico en el rango de la hipótesis alternativa,

Ejemplo 5: Con el ejercicio 4, la hipótesis nula es que la media de todas las cuentas por cobrar es de \$260 y la hipótesis alternativa es que la media sea menor que esta cantidad; la prueba se realiza con 5% como nivel de significancia. El auditor indica, además, que una media verdadera de \$240 o menos, sería considerada como diferencia importante en relación con el valor hipotético \$260. Como antes $\sigma = \$43$ y el tamaño de la muestra es $n = 36$ cuentas la determinación de la probabilidad del error tipo II requiere:

1. Formular la hipótesis nula y alternativa de esta prueba.
2. Determinar el valor crítico de la media muestral necesario para probar la hipótesis nula con 5% de nivel de significancia
3. Determinar la probabilidad del error tipo I correspondiente al uso del valor crítico arriba calculado como base para la regla de decisión.
4. Determinar la probabilidad del error tipo II correspondiente a la regla de decisión dado el valor alternativo para la media \$240

La solución completa es:

$$1.- (H_0: \mu = \$260.00 \quad (H_1: \mu < \$260.00)$$

$$2.- \bar{X}_{CR} = \mu_0 \pm Z\sigma_{\bar{X}} = 260 + (-1.641)(7.17) = 248.21$$

$$\text{Donde: } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{43}{\sqrt{36}} = \frac{43}{6} = 7.17$$

3.- La probabilidad máxima de error tipo I es igual a 0.05 (el nivel de significancia que se usa para probar la hipótesis nula)

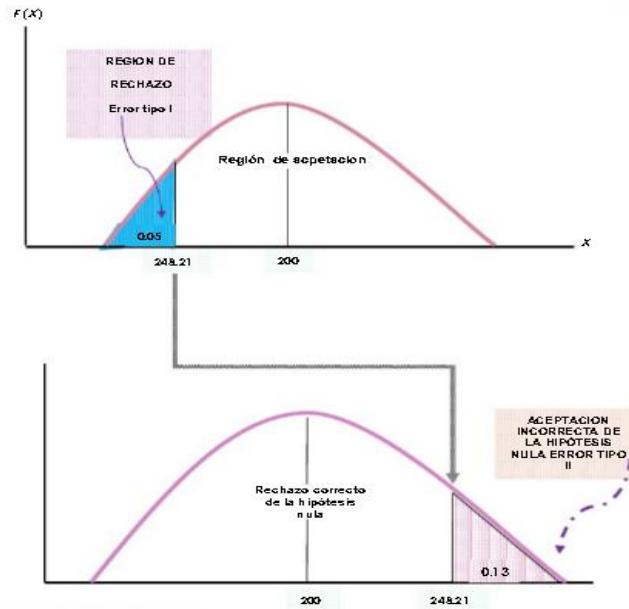
4.- La probabilidad del error tipo II es la probabilidad de que la media de la muestra aleatoria sea mayor o igual que \$284.21, dado que la media de todas las cuentas en realidad es \$240.

$$Z = \frac{\bar{X}_{CR} - \mu_1}{\sigma_{\bar{X}}} = \frac{248.21 - 240}{7.17} = \frac{8.21}{7.17} = +1.15$$

$$P(\text{error tipo II}) = P(Z \geq +1.15) = 0.5000 - 0.3749 = 0.1251 = 0.13$$

En general, el valor crítico determinado para la media respecto a la hipótesis nula se baja y se usa como valor crítico respecto de la hipótesis alternativa de que se trate.

Manteniendo constantes el nivel de significancia y el tamaño de la muestra, la probabilidad de error tipo II disminuye a medida que el valor alternativo para la media se elige más alejado de la hipótesis nula y aumenta a medida que este valor alternativo se elige más cerca del valor de la hipótesis nula.



Ejemplo 6: Suponga que un desarrollador considera una discrepancia importante en el ingreso familiar promedio o sea de \$43,500 o menor, en lugar del nivel de ingreso propuesto, que es \$45,000. Determine a) la probabilidad del error tipo I, b) la probabilidad del error tipo II.

a) $P(\text{error tipo I}) = 0.05$ (nivel α , o nivel de significancia).

b) $P(\text{error tipo II}) = P(\text{el valor crítico } \bar{X} \text{ sea sobrepasado dado que } \mu = \$43,500)$

El valor critico inferior $\bar{X} = \mu_0 + Z \sigma_{\bar{X}} = 45000 + (-1.645)(516.80) = \$43,500$

Donde $\mu_0 = \$45,000$

$Z = -1.645$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2000}{\sqrt{15}} = \frac{43}{3.87} = \$516.80$$

$$P(\text{error tipo II}) = P(\bar{X} \geq 44149.86 | \mu_1 = 43000, \sigma_{\bar{X}} = 516.80)$$

$$Z = \frac{\bar{X}_{CR} - \mu_1}{\sigma_{\bar{X}}} = \frac{44149.86 - 43500}{516.80} = \frac{649.86}{516.80} = +1.26$$

$$P(\text{error tipo II}) = P(Z \geq +1.26) = 0.500 - 0.3962 = 0.1038 = 0.10$$

Otras pruebas de hipótesis.

Prueba para la diferencia entre dos medias empleando la distribución normal.

El procedimiento para probar una hipótesis sobre la diferencia entre dos medias poblacionales es similar al que se usa para probar una hipótesis relativa al valor de una media poblacional. El procedimiento difiere solamente en que se usa el error estándar de la diferencia entre medidas para determinar el valor z (o t) correspondiente al resultado muestral. El uso de la distribución normal está basado en las mismas condiciones que en el caso de una muestra, excepto que se consideran dos muestras aleatorias independientes. La fórmula general para determinar el valor de z en una prueba de hipótesis para la diferencia entre dos medias, dependiendo de si se conocen los valores de σ de las dos poblaciones, es

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sigma_{\bar{x}_1 - \bar{x}_2}} \quad (11.1)$$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu^1 - \mu^2)_0}{s_{\bar{x}_1 - \bar{x}_2}} \quad (11.2)$$

Como implican las formulas (11.1) y (11.2), se puede partir de cualquier diferencia hipotética, $(\mu_1 - \mu_2)_0$, que desee probar. Sin embargo, la hipótesis nula acostumbrada es que las dos muestras se han obtenido de poblaciones con medias iguales. En este caso $(\mu_1 - \mu_2)_0 = 0$ y las formulas anteriores se simplifican a las siguientes:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} \quad (11.3)$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} \quad (11.4)$$

En general, el error estándar de la diferencia entre medias se calcula como se describe en la sección 9.1. Sin embargo, al probar la diferencia entre dos medidas, la hipótesis nula que suele interesar es que no solo las medias muestrales fueron obtenidas de poblaciones con medias iguales, sino que en efecto las dos muestras fueron obtenidas de la misma población de valores. Esto significa que $\sigma_1 = \sigma_2$, lo que se puede detonar simplemente con σ . Con frecuencia la supuesta varianza común se estima combinando las dos varianzas muestrales, y entonces el valor estimado de σ^2 se usa como base para el error estándar de la diferencia. El estimado conjunto de la varianza poblacional es

$$\sigma^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (11.5)$$

El error estimado de la diferencia basado en la suposición de que las desviaciones estándar poblacionales son iguales es

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \quad (11.6)$$

La suposición de que las dos varianzas muestrales fueron obtenidas de poblaciones con varianzas iguales puede probarse con hipótesis nula. Las pruebas para la diferencia entre medias pueden ser bilaterales o unilaterales, como se ilustra en el siguiente ejemplo.

Ejemplo 1

El salario semanal medio en una muestra $n_1 = 30$ empleados de una empresa grande es $\bar{x}_1 = \$280.00$ y la desviación estándar de la muestra es $s_1 = \$14.00$. En otra empresa grande, una muestra aleatoria de $n_2 = 40$ empleados tienen un salario medio de $\bar{x}_2 = \$270.00$ y la desviación estándar es $s_2 = \$10.00$. No se supone que las desviaciones estándar de las dos poblaciones de salarios sean iguales. Se prueba la hipótesis de que no hay diferencia entre los salarios semanales medios de las dos empresas empleando como nivel de significancia 5 por ciento, de la manera siguiente:

$$\begin{aligned} H_0: (\mu_1 - \mu_2) &= 0 \text{ o en forma equivalente, } \mu_1 = \mu_2 & \bar{x}_1 &= \$280.00 & \bar{x}_2 &= \$270.00 \\ H_1: (\mu_1 - \mu_2) &\neq 0 \text{ o en forma equivalente, } \mu_1 \neq \mu_2 & s_1 &= \$14.00 & s_2 &= \$10.00 \\ & & n_1 &= 30 & n_2 &= 40 \end{aligned}$$

$$Z \text{ critica } (\alpha = 0.05) = \pm 1.96$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{s} = \frac{280 - 270}{3.01} = \frac{10.0}{3.01} = +3.32$$

Donde

$$S_{\bar{x}_1} = \frac{s_1}{\sqrt{n_1}} = \frac{14.0}{\sqrt{30}} = \frac{14.0}{5.477} = 2.56$$

$$S_{\bar{x}_2} = \frac{s_2}{\sqrt{n_2}} = \frac{10}{\sqrt{40}} = \frac{10}{6.325} = 1.58$$

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{S^2_{x_1} + S^2_{x_2}} = \sqrt{(2.56)^2 + (1.58)^2} = \sqrt{6.5536 + 2.4964} = 3.01$$

La z calculada $+3.32$ está en la región de rechazo del modelo de prueba de hipótesis que se ilustra en la figura 11.1 por tanto, se rechaza la hipótesis nula y se acepta la hipótesis alternativa que dice que el salario semanal medio en las dos empresas es diferente.

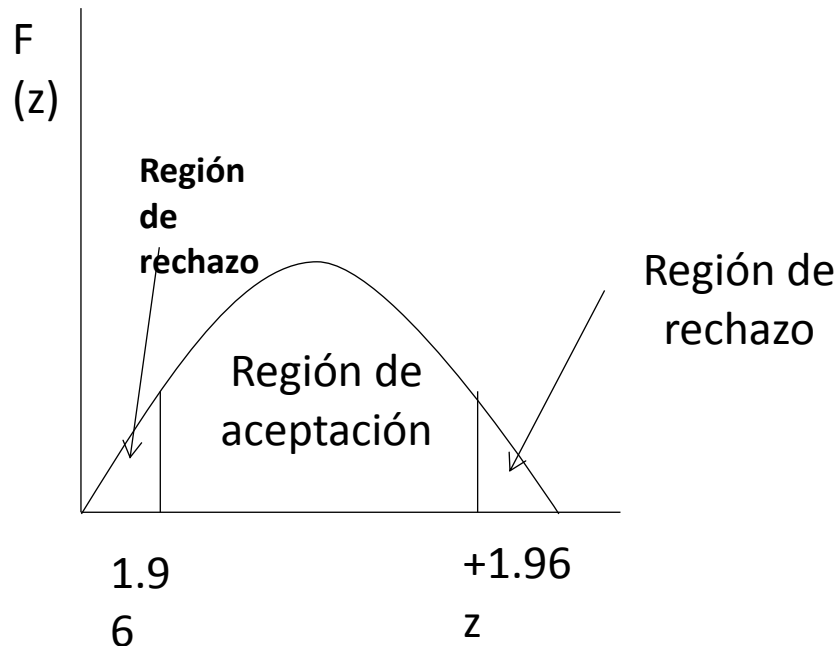


Figura 11.1

Ejemplo 2

Antes de ver los resultados muestrales del ejemplo 1, un analista de salarios cree que el salario promedio en la primera empresa es mayor que en la segunda. Con objeto de someter esta creencia a una prueba crítica, el analista concede el beneficio de la duda a la posibilidad opuesta y formula la hipótesis nula de que el salario promedio en la primera empresa es menor o igual al promedio en la segunda. A continuación, se prueba esta hipótesis con de significancia de 1 por ciento, otra vez sin suponer que las desviaciones estándar de las dos poblaciones sean iguales,

$H_0: (\mu_1 - \mu_2) \leq 0$ o en forma equivalente, $\mu_1 \leq \mu_2$

$H_1: (\mu_1 - \mu_2) > 0$ o en forma equivalente $\mu_1 > \mu_2$

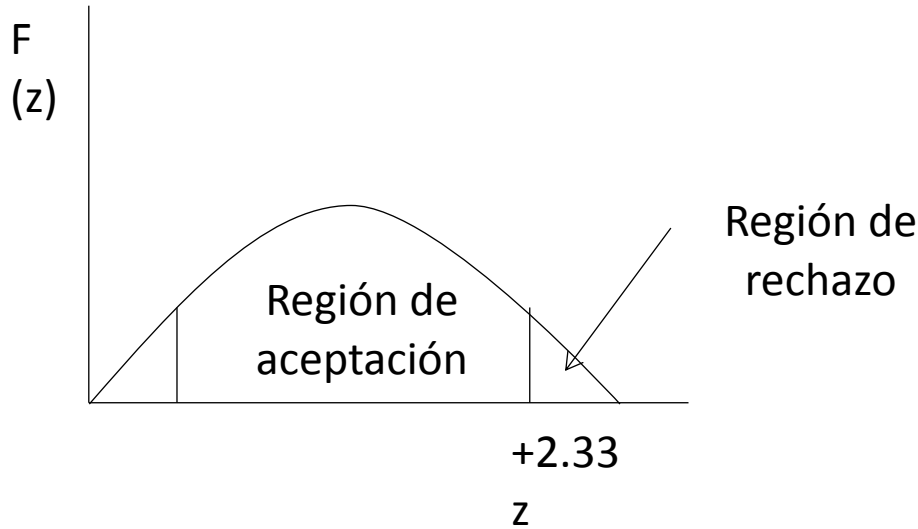
Z critica ($\alpha = 0.01$) = +2.33

Z calculada = +3.32 (de acuerdo con el ejemplo 1)

En esta prueba de cola superior, la Z calculada +3.32 es mayor que el valor critico +2.33, como se ilustra en la figura 11.2

Por tanto, se rechaza la hipótesis nula y se acepta la hipótesis alternativa que dice que el salario promedio en la primera empresa es mayor que el salario promedio en la segunda.

Figura 11.2



Prueba para la diferencia entre medias empleando la distribución t

Cuando se prueba la diferencia entre dos medias empleando la distribución t , una suposición necesaria en el proceso estándar que se emplea en la mayoría de los textos es que las varianzas de las poblaciones son iguales. Por tanto, en esta prueba el error estándar estimado para la media se calcula a partir de las formulas (11.5) y (11.6). Los requisitos para el uso apropiado de la distribución t se describen de 1 por ciento.

$$H_0: (\mu_1 - \mu_2) = 0 \quad \bar{X}_1 = 4000 \text{ h} \quad \bar{X}_2 = 4300 \text{ h}$$

$$H_1: (\mu_1 - \mu_2) \neq 0 \quad S_1 = 200 \text{ h} \quad S_2 = 250 \text{ h}$$

$$n_1 = 10 \quad n_2 = 8$$

$$gl = n_1 + n_2 - 2 = 10 + 8 - 2 = 16$$

$$t \text{ crítica } (gl = 16, \alpha = 0.01) = \pm 2.921$$

$$\hat{\sigma}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(9)(200)^2 + (7)(250)^2}{10 + 8 - 2} = \frac{360000 + 437500}{16} = 49843.75 \hat{\sigma} \bar{x}_1 - \bar{x}_2 =$$

$$\sqrt{\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}} = \sqrt{\frac{49843.75}{10} + \frac{49843.75}{8}} = \sqrt{11214.843} = 105.9$$

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma} \bar{x}_1 - \bar{x}_2} = \frac{4000 - 4300}{105.9} = \frac{-300}{105.9} = -2.833$$

El valor de t que se obtuvo (-2.833) está en la región de aceptación de la hipótesis nula. Por tanto, con un nivel de significancia de 1 por ciento no se puede rechazar la hipótesis nula.

Pruebas para la diferencia entre medias empleando pares de observaciones

Los procedimientos de las secciones 11.1 y 11.2 se basan en la suposición de que las dos muestras fueron tomadas como dos muestras aleatorias independientes. Sin embargo, en muchas situaciones las muestras se toman por pares de valores, como cuando se determina el nivel de productividad de los trabajadores antes y después de un programa de capacitación. A estas se les conoce como observaciones por pares con frecuencia se les denomina *muestras dependientes*.

En el caso de las observaciones por pares, el método apropiado para aprobar la diferencia entre las medias de dos muestras es determinar primero la diferencia d entre cada par de valores y luego probar la hipótesis nula de que la diferencia media es cero. Así, desde el punto de vista de los cálculos esta prueba se aplica a una muestra de valores d , con $H_0: \mu_d = 0$.

La media y la desviación estándar de los valores muestrales d se obtienen mediante las formulas básicas de los capítulos 3 y 4, excepto que d reemplaza a X . La diferencia media de un conjunto de diferencias entre pares de observaciones es

$$\bar{d} = \frac{\sum d}{n} \quad (11.7)$$

La fórmula de las desviaciones y la fórmula de cálculo para obtener la desviación estándar de las diferencias entre los pares de observaciones son, respectivamente,

$$S_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}} \quad (11.8)$$

$$S_d = \sqrt{\frac{\sum (d^2 - n\bar{d}^2)}{n-1}} \quad (11.9)$$

El error estándar de la diferencia media entre pares de observaciones se obtiene mediante la formula (8.4) para el error estándar de la media, excepto que de nuevo d reemplaza a X .

$$S_d = \frac{s_d}{\sqrt{n}} \quad (11.10)$$

Ya que el error estándar de la diferencia de la media se calcula con base en la desviación estándar de la muestra de diferencias (esto es, no se conoce el valor

poblacional σ_d), ya que en general se puede suponer que los valores d tienen distribución normal, la distribución apropiada para la hipótesis nula de que $\mu_d = 0$ es la distribución t .

Los grados de libertad están dados por el número de diferencias menos uno, o $n - 1$. Como se estudió en la sección 8.8, la distribución normal estándar z se puede usar como aproximación de las distribuciones t cuando $n \geq 30$.

El ejemplo 4 ilustra una distribución bilateral, mientras que el problema 11.5 ilustra una prueba unilateral. El estadístico de prueba que se usa para aprobar la hipótesis de que no hay diferencia entre las medias de un conjunto de observaciones por pares es

$$t = \frac{\bar{d}}{s_{\bar{d}}} \quad (11.11)$$

Ejemplo 4

Un fabricante de automóviles colecta datos de rendimiento de gasolina en una muestra de $n = 10$ automóviles con diferente peso usando un tipo estándar de gasolina con y sin un aditivo. Por supuesto que los motores se ajustan a las mismas especificaciones antes de cada corrida y se usan los mismos conductores con los dos tipos de gasolina (sin que el conductor sepa que gasolina se está usando en cada corrida). Dados los datos de la tabla 11.1 a continuación se prueba la hipótesis de que no hay diferencia entre el rendimiento medio con y sin el aditivo, usando el nivel de significancia de 5%.

$$\text{Promedio con aditivo} = \frac{276.8}{10} = 27.68 \text{ mpg}$$

$$\text{Promedio sin aditivo} = \frac{275.1}{10} = 27.51 \text{ mpg}$$

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

$$T \text{ critico } (gl = 9, \alpha = 0.05) = \pm 2.262$$

$$\bar{d} = \frac{\sum d}{n} = \frac{1.7}{10} = 0.17$$

$$S_d = \sqrt{\frac{\sum d^2 - n\bar{d}^2}{n-1}} = \sqrt{\frac{1.31 - 10(0.17)^2}{10-1}} = \sqrt{\frac{1.31 - 10(0.0289)}{9}} = \sqrt{0.1134} = 0.337$$

$$S_{\bar{d}} = \frac{S_d}{\sqrt{n}} = \frac{0.337}{\sqrt{10}} = \frac{0.337}{3.16} = 0.107$$

$$T = \frac{\bar{d}}{S_{\bar{d}}} = \frac{0.17}{0.107} = +1.59$$

Tabla 11.1 Datos del rendimiento y hoja de cálculo para la diferencia media y la desviación estándar de la diferencia.

Automóvil	Recorrido con aditivo	Recorrido sin aditivo	D	d ²
1	36.7	36.2	0.5	0.25
2	35.8	35.7	0.1	0.01
3	31.9	32.3	-0.4	0.16
4	29.3	29.6	-0.3	0.09
5	28.4	28.1	0.3	0.09
6	25.7	25.8	-0.1	0.01
7	24.2	23.9	0.3	0.09
8	22.6	22	0.6	0.36
9	21.9	21.5	0.4	0.16
10	20.3	20	0.3	0.09
Total	276.8	275.1	+1.7	1.31

El valor de t que se calculó (+ 1.59) no está en la región de rechazo de la hipótesis nula. Por lo tanto, la hipótesis nula de que no hay diferencia en el rendimiento por galón con el aditivo o sin el aditivo se acepta como plausible.

Pruebas de hipótesis para la proporción poblacional

Como se explicó en la sección 7.4, la distribución normal se puede usar como aproximación a una distribución binomial cuando $n \geq 30$ y tanto $np \geq 5$ como $n(q) \geq 5$, donde $q = 1 - p$. esta es la base sobre la cual se determinan los intervalos de confianza para una proporción en la sección 9.3, donde también se estudia el error estándar para la proporción. Sin embargo, en el caso de los intervalos de confianza por lo general se requiere que el tamaño de la muestra sea cuando menos $n = 100$, como se explicó en la sección 9.3.

En la determinación de los intervalos de confianza en la sección 9.3 la proporción muestral $\hat{p}$ sirve como basa para el error estándar. En las pruebas de hipótesis el

valor del error estándar para la proporción se basa por lo general en el valor hipotético π_0 :

$$\sigma_{\hat{p}} = \sqrt{\frac{\pi_0(1-\pi_0)}{n}} \quad (11.12)$$

El procedimiento para la prueba de un valor hipotético de la proporción poblacional es idéntico al que se describe en la sección 10.3, excepto que la hipótesis nula considera el valor de la proporción poblacional en lugar de la media poblacional. Así, la fórmula para el estadístico z en pruebas de hipótesis para la proporción poblacional es

$$Z = \frac{\hat{p} - \pi_0}{\sigma_{\hat{p}}} \quad (11.13)$$

Ejemplo 5

El director de una oficina universitaria de colocaciones asegura que por lo menos 50 por ciento de los estudiantes graduados han concretado arreglos para un empleo hacia el 1 de marzo. Suponga que se toma una muestra aleatoria de $n = 30$ estudiantes graduados y solo 10 de ellos indicaron haber concretado arreglos para tener un empleo hacia el primer día de marzo. ¿Puede rechazarse la afirmación del director de colocaciones con un nivel de significancia de 5 por ciento? Se utiliza z como el estadístico de prueba, como sigue:

$$H_0: \pi = 0.50 \quad H_1: \pi < 0.50$$

$$Z \text{ critico } (\alpha = 0.05) = -1.645$$

[En el uso de la distribución normal se justifica porque $n \geq 30$, $n\pi_0 \geq 5$, y $n(1-\pi_0) \geq 5$.]

$$\sigma$$

En esta prueba de cola inferior el valor calculado para z, -1.88, es menor que el valor crítico, -1.645; por tanto, se rechaza la afirmación del director de colocaciones con el nivel de significancia de 5 por ciento.

Determinación del tamaño de muestra requerido para pruebas de proporciones

Antes de coleccionar una muestra se puede determinar el tamaño de muestra requerido para probar una hipótesis acerca de la proporción poblacional especificando 1) el valor hipotético de la proporción, 2) un valor alternativo para la proporción de modo que su diferencia con el valor de la hipótesis nula se considere importante, 3) el nivel de significancia que debe usarse en la prueba, y 4) la probabilidad del error tipo II

que ha de permitirse. La fórmula para determinar el tamaño mínimo de muestra requerido para probar un valor hipotético para la proporción es

$$n = \left\lceil \frac{z_0\sqrt{\pi_0(1-\pi_0)} - z_1\sqrt{\pi_1(1-\pi_1)}}{\pi_1 - \pi_0} \right\rceil \quad (11.14)$$

En la fórmula (11.14), z_0 es el valor crítico de z que se usa en conjunción con el nivel de significancia especificado (nivel α), mientras que z_1 es el valor de z respecto de la probabilidad elegida para el error tipo II (nivel β). Como en la sección 10.5 para la determinación del tamaño de la muestra para una prueba para la media, z_0 y z_1 siempre tienen signos algebraicos opuestos. Esto da como resultado que los dos productos del numerador siempre se acumulen. La fórmula (11.14) se puede usar para 1 o 2 colas, y todo tamaño fraccionario de muestra se redondea hacia arriba. Por último, el tamaño de la muestra debe ser lo suficientemente grande para justificar el uso de la distribución de probabilidad normal en conjunción con π_0 y π_1 , como se vio en la sección 11.4.

Ejemplo 6

Un miembro del congreso desea probar, usando el nivel de significancia del 5%, la hipótesis de que cuando menos 60% de los votantes están a favor de una nueva legislación laboral. La discrepancia de esta hipótesis se considera importante si solamente 50% (o menos) está a favor de la legislación, y el riesgo de error tipo II de $\beta = 0.05$ es aceptable. El tamaño mínimo de la muestra que debe tomarse para satisfacer estas especificaciones para la toma de una decisión es

$$n = \left\lceil \frac{z_0\sqrt{\pi(1-\pi)} - z_1\sqrt{\pi(1-\pi)}}{\pi - \pi} \right\rceil^2 = \left\lceil \frac{-1.645\sqrt{(0.060)(0.40)} - (+1.645)\sqrt{(0.50)(0.50)}}{0.50 - 0.60} \right\rceil^2 = \left\lceil \frac{-1.645(0.49) - 1.645(0.50)}{-0.10} \right\rceil^2 = (16.28)^2 = 265.04 = 266$$

Pruebas para una proporción en un proceso en el control estadístico de procesos

Como se estudió en la sección 10.9, el uso y la interpretación de las gráficas de control en el control estadístico de procesos es una aplicación directa de los métodos y conceptos de las pruebas de hipótesis. Como en el caso de la media del proceso en la sección 10.9, los límites de control para una proporción de procesos se definen a ± 3 unidades de error estándar del valor hipotético (aceptable). La metodología para realizar e interpretar las gráficas de control se estudia en el capítulo 19. La sección 19.9 se ocupa principalmente de proporciones de procesos.

Prueba para la diferencia entre dos proporciones poblacionales

Cuando se desea probar la hipótesis de que las dos proporciones en las dos poblaciones no son diferentes, las dos proporciones muestrales se emplean para determinar el error estándar de la diferencia entre proporciones. Observe que esto difiere del procedimiento que se usó en la sección 9.5 para la estimación estadística, en donde no se hizo la suposición de no diferencia. Además, el presente procedimiento es conceptualmente similar al que se presentó en la sección 11.1, en el que se empleaban las dos varianzas muestrales para calcular el error estándar de la diferencia entre medias. La estimación conjunta de la proporción poblacional, basada en las proporciones obtenidas en dos muestras independientes, es

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \quad (11.15)$$

El error estándar de la diferencia entre proporciones que se usa para probar la suposición de no diferencia es

$$\sigma_{p_1 - p_2} = \sqrt{\frac{\pi(1-\pi)}{n_1} + \frac{\pi(1-\pi)}{n_2}} \quad (11.16)$$

La fórmula para obtener el estadístico z para probar la hipótesis nula de que no hay diferencia entre dos proporciones poblacionales es

$$z = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}} \quad (11.17)$$

Una prueba para la diferencia entre proporciones puede realizarse como una prueba unilateral (véase el problema 11.8) o como una prueba bilateral (véase el ejemplo 8)

Ejemplo 8

Una muestra de 50 familias de una comunidad muestra que 10 de ellas están viendo un programa especial de televisión sobre la economía nacional. En una segunda comunidad, 15 familias de una muestra aleatoria de 50 están viendo el programa especial de TV. A continuación, se prueba la hipótesis de que la proporción general de televidentes en las dos comunidades no difiere, usando el nivel de significancia de 1%, como sigue:

$H_0: (\pi_1 - \pi_2) = 0$ o lo que es equivalente, $\pi_1 = \pi_2$

$H_1: (\pi_1 - \pi_2) \neq 0$ o lo que es equivalente, $\pi_1 \neq \pi_2$

Z crítico ($\alpha = 0.01$) = ± 2.58

$$\hat{\pi} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{50(0.20) + 50(0.30)}{50 + 50} = \frac{10 + 15}{100} = 0.25$$

$$\hat{\sigma}_{p1-p2} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}} = \sqrt{\frac{(0.25)(0.75)}{50} + \frac{(0.25)(0.75)}{50}} = \sqrt{0.00375 + 0.00375} = 0.087$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\hat{\sigma}_{p1-p2}} = \frac{0.20 - 0.30}{0.087} = \frac{-0.10}{0.087} = -1.15$$

El valor calculado para Z, - 1.15, está en la región de aceptación de la hipótesis nula. Por tanto, no se puede rechazar la hipótesis de que no hay diferencia en la proporción de televidentes en las dos comunidades.

Prueba para un valor hipotético de la varianza empleando la distribución chi cuadrada

Como se explicó en la sección 9.6, para una población con distribución normal (n-1) s^2 / σ^2 sigue una distribución de probabilidad χ^2 , habiendo una distribución chi cuadrada diferente de acuerdo con los grados de libertad (n-1). Por lo tanto, el estadístico que se usa para probar una hipótesis acerca del valor de la varianza poblacional es

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \quad (11.18)$$

Esta prueba basada en la formula (11.18) puede ser unilateral o bilateral, aunque lo más frecuente es que la hipótesis acerca de la varianza poblacional sean pruebas unilaterales. El apéndice 7 puede usarse para determinar el (los) valor (es) crítico (s) del estadístico chi cuadrado para varios niveles de significancia.

Ejemplo 9 la vida media de operación de una muestra aleatoria de n = 10 focos es $\bar{X} = 4000$ h y la desviación estándar es s = 200 h. Por lo general se supone que la vida de operación de los focos tiene distribución normal. Suponga que antes de tomar la muestra se formulo la hipótesis de que la desviación estándar de la población no es mayor que $\sigma = 150$. Basándose en los resultados muestrales, se prueba esta hipótesis con un nivel de significancia de 1%, como sigue

$$H_0: \sigma^2 = 22500 \text{ [ya que } \sigma^2 = (150)^2 = 22500] \quad H_1: \sigma^2 > 22500$$

$$\chi^2 \text{ critica (gl = 9, } \alpha = 0.01) = 21.67$$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(9)(40000)}{22500} = \frac{360000}{22500} = 16.0$$

Ya que en esta prueba de cola superior el estadístico de prueba calculado 16.0 no es mayor que el valor crítico 21.67, no se puede rechazar la hipótesis nula de que $\sigma \leq 150$ con un nivel de significancia de 1%.

Prueba para la variabilidad de un proceso en el control estadístico de procesos

Como se describió en la sección 10.9, acerca de las pruebas para la media de un proceso y en la sección 11.6 acerca de las pruebas para la proporción de un proceso, el uso e interpretación de graficas en el control de procesos es una aplicación directa de los métodos y conceptos de las pruebas de hipótesis. La variabilidad del proceso se vigila y controla ya sea en relación con la desviación estándar del proceso o con el rango del proceso. Como en las gráficas de control para la media de un proceso y para la proporción de un proceso, los límites de control se establecen a ± 3 de error estándar respecto al valor central esperado en la grafica dado que la hipótesis nula de que no hay variación asignable a alguna causa es verdadera. La metodología para trazar e interpretar las gráficas de control se estudia en el capítulo 19. La sección 19.7 se ocupa de la desviación estándar del proceso y la sección 19.8 considera el rango del proceso.

Distribución f y prueba para la igualdad de dos varianzas poblacionales

Puede mostrarse que la distribución F es el modelo de la probabilidad apropiado para el cociente de las varianzas de dos muestras extraídas independientemente de una misma población con distribución normal, y hay una distribución F diferente para cada una de las combinaciones de grados de libertad gl correspondiente a cada muestra. En cada muestra, $gl = n - 1$. El estadístico que se usa para probar la hipótesis nula de que las dos varianzas poblacionales son iguales es

$$F_{gl1, gl2} = \frac{S_1^2}{S_2^2} \quad (11.19)$$

Ya que cada varianza muestral es un estimado insesgado de una varianza poblacional, el valor esperado a largo plazo del cociente anterior es aproximadamente 1.0. [Nota: el valor esperado no es exactamente 1.0, sino mas bien $gl_2 / (gl_2 - 2)$, por razones matemáticas que están fuera del alcance de esta exposición.] Sin embargo, en cualquier par dado de muestras es poco probable que las varianzas muestrales tengan valores idénticos, aun cuando la hipótesis nula sea verdadera. Como se sabe que este cociente sigue una distribución F, esta distribución de probabilidad se puede usar en la prueba de la diferencia entre dos varianzas. Aunque una suposición matemática necesaria es que las dos poblaciones tienen distribución normal, la prueba F ha demostrado ser relativamente robusta e insensible a desviaciones de la normalidad cuando cada población es unimodal y los tamaños de las muestras son aproximadamente iguales.

El apéndice 8 da los valores de F que dejan a la derecha probabilidades de 0.05 y 0.01. Los grados de libertad gl en el numerador del cociente que se emplea son los rótulos de las columnas de esta tabla y los grados de libertad en el denominador son los rótulos de los renglones. La tabla no proporciona valores críticos de F en la cola inferior de la distribución, en parte debido a que la distribución F se usa, por lo general, para pruebas que requieren solo probabilidades para cola superior. Esto ocurre especialmente en el uso de la distribución F en el análisis de varianza. Otra razón por la que solo se proporcionan los valores de F en la cola superior es que los valores de la cola inferior pueden calcularse como se indica a continuación mediante la llamada propiedad recíproca de la distribución F:

$$F_{gl1,gl2,inferior} = \frac{1}{F_{gl1,gl2,superior}} \quad (11.20)$$

Mediante la fórmula (11.20), introduciendo en el denominador el valor F correspondiente al punto que deja en la cola superior 5%, se determina el valor de F correspondiente al punto que deja en la cola inferior 5%. Hay que observar que los valores de los dos gl en el denominador están en orden inverso al que se requiere para el valor de F que se busca.

Ejemplo 11

En los datos del ejemplo 3 se supone que la vida de los focos tiene distribución normal. A continuación, se prueba la hipótesis nula de que las muestras se obtuvieron de poblaciones con varianzas iguales, usando el nivel de significancia de 10% y empleando la distribución F:

$$H_0: \sigma^2 = \sigma^2 \quad \text{o lo que es equivalente} \quad \frac{\sigma_1^2}{\sigma_2^2} = 1.0 \quad s_1 = 200 \text{ h} \quad s_2 = 250 \text{ h}$$

$$H_1: \sigma_1^2 \neq \sigma_2^2 \quad \text{o lo que es equivalente} \quad \frac{\sigma_1^2}{\sigma_2^2} \neq 1.0 \quad s_2 = 40000 \text{ h} \quad s_2^2 = 62500$$

$$n_1 = 10 \quad n_2 = 8$$

Esta prueba con un nivel de significancia de 10%, los puntos de F que dejan en las colas superior e inferior 5% son los valores críticos:

$$F_{9,7} \text{ critico (5\% superior)} = 3.68$$

$$F_{9,7} \text{ critico (5\% inferior)} = \frac{1}{F_{7,9}(5\% \text{ superior})} = \frac{1}{3.29} = 0.304$$

$$F_{gl1,gl2} = \frac{s_1^2}{s_2^2} \quad F_{9,7} = \frac{40000}{62500} = 0.64$$

Como el cociente que se calcula da un valor F que no es menor que 0.304 ni mayor que 3.68, este valor de F se encuentra en la región de aceptación de la hipótesis nula. Así, la suposición de que las varianzas de las dos poblaciones son iguales no pueden rechazarse con un nivel de significancia de 10%.

Otros métodos para probar hipótesis nulas

Como se estudió en la sección 10.7 y 10.8, los métodos del valor P y del intervalo de confianza son alternativas para el método del valor crítico para pruebas de hipótesis usado en las secciones anteriores de este capítulo.

En el método del valor P, en lugar de comparar el valor observado de un estadístico de prueba con un valor crítico, se determina la probabilidad de la ocurrencia del valor del estadístico de la prueba, dado que la hipótesis nula es verdadera, y esta probabilidad se compara con el nivel de significancia α . La hipótesis nula se rechaza si el valor P es menor que la α elegida. Los problemas 11.13 y 11.14 ilustran la aplicación del método del valor P a pruebas bilaterales y unilaterales, respectivamente, para la diferencia entre medias.

En el método del intervalo de confianza se construye el intervalo de confianza $1 - \alpha$ para el valor del parámetro de interés. Si el valor hipotético del parámetro no queda comprendido en el intervalo, entonces se rechaza la hipótesis nula. Los problemas 11.15 y 11.16 ilustran la aplicación del método del intervalo de confianza en pruebas bilaterales y unilaterales, respectivamente, para la diferencia entre medias.,

Uso de Excel y de minitab

Para la mayor parte de las aplicaciones de pruebas de hipótesis que se describen en este capítulo se cuenta con software. Para pruebas de hipótesis por lo general se usa el método del valor P, como se describió en la sección 11.11. En los problemas 11.17 y 11.18, respectivamente, se ilustra el uso de Excel y de minitab para probar la diferencia de medias entre dos muestras independientes, mientras que en los problemas 11.19 y 11.20 se ilustran el uso de Excel y minitab con pares de observaciones para probar la diferencia entre medias.

Regresión lineal y análisis de correlación

Objetivos y suposiciones en el análisis de regresión

El objetivo principal del análisis de regresión es estimar el valor de una variable aleatoria (la variable dependiente) dado que se conoce el valor de una variable asociada (la variable independiente). La variable dependiente también se denomina variable de respuesta, mientras que la variable independiente también se denomina variable productora. La ecuación de regresión es la fórmula algebraica mediante la cual se determina el valor estimado de la variable dependiente o variable de respuesta.

El término análisis de regresión simple significa que se estima el valor de una variable dependiente sobre la base de una variable independiente o predictora. El análisis de regresión múltiple considera la estimación del valor de una variable dependiente sobre la base de dos o más variables independientes.

Las suposiciones generales que subyacen al modelo de análisis de regresión que se presenta en este capítulo son que 1) la variable dependiente es una variable aleatoria y 2) la variable independiente y dependiente están relacionadas en forma lineal. La suposición 1) establece que, aunque se puedan controlar los valores de la variable independiente, los valores de la variable dependiente deben obtenerse a través del proceso de muestreo aleatorio.

Si en el análisis de regresión se realizan una estimación de intervalo o una prueba de hipótesis, se requieren tres hipótesis adicionales que son 3) que las varianzas de las distribuciones condicionales de la variable dependientes, dados diferentes valores de la variable independiente, sean todas iguales, 4) que las distribuciones condicionales de la variable dependiente, dados diferentes valores de la variable independiente, estén distribuidas todas de manera normal en la población de valores, y 5) que los valores observados de la variable dependiente sean independientes entre sí.

Ejemplo 1: un analista desea estimar el tiempo de entrega como variable dependiente, basándose en la distancia como variable independiente en el caso de partes industriales transportadas por camino. Suponga que se escogen 10 fletes recientes de los registros de la empresa de modo que las distancias en la autopista que intervienen tengan una dispersión aproximadamente igual entre 100 y 1 000 millas, y se registra el tiempo de entrega para flete. Ya que la distancia en la autopista va a ser la variable independiente, es aceptable esta selección de viajes con distancias específicas. Por otro lado, la variable dependiente, que es el tiempo de entrega, en este estudio es una variable aleatoria y se ajusta a la suposición que subyace al análisis de regresión. Si las dos variables guardan o no una relación lineal se investiga, por lo general, mediante la construcción de un diagrama de dispersión o una gráfica de residuos. Estos diagramas también se usan para observar si la dispersión vertical (varianza) es aproximadamente igual a lo largo de la recta de regresión.

El diagrama de dispersión

Un diagrama de dispersión es una gráfica en la cual cada punto representa un par de valores observados de las variables independiente y dependiente. El valor de la variable independiente X se grafica en el eje horizontal, y el valor de la variable dependiente Y se grafica en él es vertical. La forma de la relación que se representa en el diagrama de dispersión puede ser curvilínea en lugar de lineal. Aun cuando el análisis de regresión para relaciones curvilíneas esta fuera del alcance de este libro, en la sección se presenta un estudio limitado del análisis de la tendencia curvilínea. En la relación que no es lineales, lo que se suele hacer es determinar un método para transformar los valores de una o de ambas variables de modo que la regresión entre los valores transformados sea lineal. Entonces se puede aplicar el análisis de regresión lineal a los valores transformados y los valores estimados de la variable dependiente pueden transformarse nuevamente a la escala original de mediciones.

Ejemplo 2: Un ejemplo de una relación curvilínea seria la relación entre los años de antigüedad en una empresa y el nivel de ventas, dado que cada año el nivel de ventas ha aumentado el mismo el mismo porcentaje respecto al año anterior. La curva resultante con pendiente creciente será inductiva de una relación exponencial.

Si el diagrama de dispersión indica una relación que generalmente es lineal, entonces se ajusta una línea recta a los datos. La posición precisa de esa recta se determina mediante el método de los mínimos cuadrados. Una recta de regresión con pendiente positiva indica una relación directa entre las variables, una pendiente negativa indica una relación inversa entre las variables y una pendiente de cero indica que las variables no están relacionadas. Además, el alcance de la dispersión vertical de los puntos graficados respecto a la recta regresión indica el grado de relación entre las dos variables.

Método de los mínimos cuadrados para ajustar una recta de regreso

La ecuación lineal que representa al modelo de regresión lineal simple es:

(14.1)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

y_i = Valor de la variable dependiente en el ensayo u observación i- esima

β_0 = Primer parámetro de la ecuación de regresión, indica el valor de Y cuando X=0

β_1 = Segundo parámetro de la ecuación de regresión, llamado coeficiente de regresión, que indica la pendiente de la recta de regresión

x_i = Valor especificado de la variable independiente en el ensayo u observación i-esima

ε_i = error aleatorio de muestreo en el ensayo u observación i-esima (ε es la epsilon giego)

Los parámetros de β_0 y β_1 , en el modelo represión lineal se estiman mediante los valores β_0 y β_1 , que están basados en los datos muestrales. Así, la ecuación de regresión lineal que está basada en los datos muestrales y que se usa para estimar un solo valor (condicional de la variable dependiente, donde acento circunflejo sobre la Y indica que es un valor estimado es

(14.2)

$$\hat{Y} = b_0 + b_1 X$$

Dependiendo del criterio matemático que se emplee, se puede dar distintas ecuaciones lineales diferentes para un mismo diagrama de dispersión dado. Con el criterio de mínimos cuadrados, la recta de regresión mejor ajuste (y la mejor ecuación) es aquella en la cual se minimiza la suma de las desviaciones al cuadrado entre los valores estimado y real de la variable dependiente para el dato muestra correspondiente. Las fórmulas de calcular mediante las cuales se puede determinar los valores de β_0 y β_1 para la ecuación de regresión que satisface el criterio de mínimos cuadrados son:

(14.3)

$$\frac{\sum XY - n\bar{Y}\bar{X}}{\sum X^2 - n\bar{X}^2}$$

(14.4)

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Una vez formulada la ecuación de regresión, se puede usar para estimar el valor de la variable dependiente. Sin embargo, esta estimación solo debe hacerse dentro del intervalo de los valores de la variable independiente tomados en la muestra, ya que no hay una base estadística para suponer que la recta de regresión sea apropiada fuera de estos límites. Además, debe determinarse si la relación expresada por la ecuación de regresión es real o pudo haber ocurrido en los datos muestrales simplemente al azar (véase la sección 14.6). El problema 14.2 ilustra la determinación de una ecuación de regresión basada en datos muestrales.

Residuos y graficas de residuales

Para un valor dado en X de la variable independiente, al valor Y de la recta de regresión con frecuencia se le denomina valor ajustado de la variable dependiente. La diferencia entre el valor observado " Y " y el valor ajustado $\hat{Y}$ se determina el residuo para esa observación y se denota por e :

(14.5)

$$e = Y - \hat{Y}$$

El problema 14.3 reporta el conjunto completo de residuos para datos muestrales específicos. Una gráfica de residual se obtiene al graficar los residuos " e " respecto la variable independiente " X " o en relación con los valores Y de la recta de regresión ajustable. Esta grafica se usa como alternativa al diagrama de dispersión para investigar si se satisfacen las suposiciones de linealidad y de igualdad de las varianzas condicionales. El problema 14.3 presenta una múltiple, como se describe en el capítulo 15.

El conjunto de residuos para los datos muestrales también sirve como base para calcular el error estándar de estimación, como se estudia en la siguiente sección.

14.5 El error estándar de estimación

El error estándar de estimación es la desviación estándar condicional de la variable dependiente Y dado un valor de la variable independiente X . para datos poblacionales, el error estándar de estimación se representa con el símbolo $\sigma_{Y.X}$. La fórmula de desviación mediante la cual se estima ese valor con base en los datos muestrales es:

(14.6)

$$S_{y.x} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}} = \sqrt{\frac{\sum e^2}{n - 2}}$$

Observe que el numerador de la formula (14.6) es la suma de los cuadrados de los residuos descritos en la sección anterior. Aunque la formula (14.6) refleja claramente la idea de que el error estándar de estimación es la desviación estándar respecto de la recta de regresión (es decir, es la desviación estándar de la "dispersión" vertical respecto a la recta), desde el punto de vista de los cálculos la formula requiere que se calcule cada uno de los valores ajustados $\hat{Y}$ correspondientes a los datos muestrales. La fórmula de cálculo alternativa, que no requiere la determinación de cada uno de los valores ajustados y que por tanto generalmente es de uso más sencillo, es:

(14.7)

$$S_{y.x} = \sqrt{\frac{\sum \hat{Y}^2 - b_0 \sum Y - b_1 \sum XY}{n - 2}}$$

Como se verá en las secciones siguientes, el error estándar de estimación sirve como piedra angular los diferentes errores estándar que se usan de los procedimientos de prueba de hipótesis y de estimación de intervalo en el análisis de regresión. Los problemas 14.5 y 14.6 ilustran el cálculo de error estándar de estimación mediante la fórmula de cálculo y la formula de las desviaciones, respectivamente

Inferencias relativas a la pendiente

Antes de usar una ecuación de regresión con el propósito de estimar o predecir se debe determinar primero si existe una relación entre dos variables de la población, o relaciones observada en la muestra pudo haber ocurrido por casualidad. En ausencia en cualquier relación en la población, la pendiente de la recta de regresión poblacional será, por definición $\beta_1 = 0$. Por tanto, la usual hipótesis nula es $H_0: \beta_1 = 0$. La hipótesis nula también se puede formular como una prueba de una cola, en cuyo caso la hipótesis alternativa no es simplemente que las dos variables estén relacionadas, si no que la relación es de un tipo específico (directa o inversa). Se prueba un valor hipotético de la pendiente mediante el cálculo de un estadístico t y el uso de n-2 grados de libertad. Se pierden dos grados de libertad en el proceso de la inferencia porque se estiman dos parámetros de la ecuación de regresión, b_0 y b_1 . La fórmula estándar es:

(14.8)

$$t = \frac{b_1 - (\beta_1)_0}{s_{b_1}}$$

Donde

(14.9)

$$s_{b_1} = \frac{s_{Y.X}}{\sqrt{\sum X^2 - n\bar{X}^2}}$$

Sin embargo, cuando la hipótesis nula es que la pendiente es 0, que es lo más frecuente entonces la formula (14.8) se simplifica y se tiene:

(14.10)

$$t = \frac{b_1}{s_{b_1}}$$

Los problemas 14.7 y 14.8 ilustran una prueba de dos colas y una prueba de 1 cola para la pendiente, respectivamente. El intervalo de confianza para la pendiente poblacional β_1 , en donde los grados de libertad para t son una vez más n-2, se construye como sigue:

(14.11)

$$b_1 \pm ts_{b_1}$$

El problema 14.9 ilustra la construcción de un intervalo de confianza para pendiente poblacional

Intervalo de confianza para la media condicional

La estimación puntual de la media condicional de la variable pendiente, dado un valor específico de X, es el valor Y de la recta de regresión. Cuando se usa la ecuación de regresión para estimar la media condicional, el símbolo apropiado para la media condicional de Y que se estima es $\hat{u}_y$:

(14.12)

$$\hat{u}_y = b_0 + b_1X$$

Con base en estos datos muestrales, el valor del error estándar para la media condicional varía de acuerdo con el valor designado de X y es

(14.13)

$$s_{\bar{y}.x} = \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - [(\sum X)^2/n]}}$$

Dados la estimación puntual de la media condicional y el error estándar de la media condicional, el intervalo de confianza para la media condicional, usando n-2 grados la libertad es:

(14.14)

$$\hat{u}_y = \pm ts_{\bar{y}.x}$$

De nuevo ahí n-2 grados de libertad porque se requieren las estimaciones de los dos parámetros β_0 y β_1 para la ecuación de regresión. El problema 14.10 ilustra la construcción de un intervalo de confianza para la media condicional.

Intervalos de predicción para los valores individuales de la variable pendiente

A diferencia de un intervalo de confianza, que sirve para estimar un parámetro poblacional un intervalo de predicción sirve para estimar un valor individual y por tanto es un intervalo de probabilidad. Puede parecer que un intervalo de predicción podría construirse usando el error estándar de estimación que se define en las formulas (14.6) (14.7). sin embargo, un intervalo así estaría incompleto, porque el error estándar de estimación no considera la incertidumbre relacionada con el hecho de que la posición de la recta de regresión basada en datos muestrales incluye al error de muestreo y por lo general no es idéntica a la recta de regresión poblacional. Al error estándar completo para un intervalo de predicción se le denomina error estándar de la predicción, e incluye la incertidumbre relacionada con la “dispersión” vertical respecto de la recta de regresión más la incertidumbre relacionada con la posición del valor de recta de regresión. La fórmula básica para el error estándar de predicción es:

(14.15)

$$S_{y(siguiente)} = \sqrt{s_{y \cdot x}^2 + s_{\bar{y}}^2 \cdot x^2}$$

La versión de cálculo de la fórmula para el error estándar de pendiente es

(14.16)

$$S_{y(siguiente)} = s_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - [(\sum X)^2/n]}}$$

Por último, el intervalo de predicción para un valor individual de la variable pendiente, usando n-2 grados de libertad es

(14.17)

$$\hat{Y} \pm t_{S_{y(siguiente)}}$$

La determinación del error estándar de la pendiente y la construcción de un intervalo de predicción se ilustran en el problema 14.11.

Objetivos y suposiciones en el análisis de correlación

A diferencia del análisis de regresión, el análisis de correlación mide el grado de relación entre dos variables. Tal como el estudio del análisis de regresión. La suposiciones poblacionales que subyacen en el análisis de correlación simple son que 1) la relación entre las dos variables es lineal, 2) ambas variables son aleatorias, 3) para cada variable las variancias condicionales, dados valores diferentes de la otra variable, son iguales, 4) para cada variable, los valores observados son independientes de los otros valores observados de esta variable, 5) para cada variable las distribuciones condicionales, dados valores diferentes de la otra variable, están todos normalmente distribuidas.

Ejemplo 4: refiérase al procedimiento de colección de datos del ejemplo 1, que considera las variables distancias de recorridos y tiempo en una muestra de 10 fletes recientes de partes industriales que se transportan en camión. En vez del análisis en el cual se escogen 10 embarques en modo de que las distancias tengan una dispersión aproximadamente igual desde 100 millas hasta 1,000 millas, los 10 fletes se eligen en forma completamente aleatoria, sin considerar ni la distancia de recorrido ni el tiempo de entrega de cada observación. A diferencia del ejemplo 1, en el cual solo el tiempo de entrega es una variable aleatoria, es un plan revisado de muestreo ambas variables son variables aleatorias y por tanto son claramente adecuadas para el análisis de correlación.

El coeficiente de determinación

Considere que un valor individual de la variable pendiente Y se estima sin el conocimiento de valor de cualquier otra variable, entonces la varianza relacionada con esta estimación, y la base para la construcción de un intervalo de predicción, sería la varianza σ_y^2 . Sin embargo, dado un valor de X , la varianza relacionada con esta estimación se reduce y se representa con $\sigma_{y.x}^2$ (o $s_{y.x}^2$ para datos muestrales), como se describe en la sección 14.5. Si existe relación entre las dos variables, entonces $\sigma_{y.x}^2$ será siempre menor que σ_y^2 . En una relación perfecta, en la cual todos los valores de la variable dependiente son iguales a los valores respectivos de la recta de regresión ajustada, $\sigma_{y.x}^2=0$. Por tanto, en ausencia de una relación perfecta, el valor $\sigma_{y.x}^2$ indica la incertidumbre que resta después de la consideración del valor de variable independiente. O se puede decir que el cociente de $\sigma_{y.x}^2$ entre σ_y^2 indica la proporción de la varianza (incertidumbre) en la variable dependiente que resta no explicada después de haber dado un valor específico de la variable independiente:

(14.18)

$$\left(\frac{\text{proporcion de la varianza}}{\text{que resta no explicada}} \right) = \frac{\sigma_{y.x}^2}{\sigma_y^2} = \frac{\text{varianza no explicada restante en Y}}{\text{varianza total de Y}}$$

En la formula (14.18) $\sigma_{y.x}^2$ Se determina mediante el procedimiento que se describe en la redacción 14.5 (excepto que los datos poblacionales son supuestos) y σ_y^2 se calcula con las formulas generales que se presentan en las secciones 4.6 y 4.7.

Dada la proporción de la varianza no explicada, una medida útil de la relación es el coeficiente de determinación-el complemento del cociente anterior-, que indica la proporción de varianza en la variable dependiente que es estadísticamente explicada por la ecuación de regresión (es decir, por el conocimiento de la variable X correspondiente). Para datos poblacionales, el cociente de determinación se representa con la letra griega ρ^2 (ro a cuadrado) y se determina mediante

(14.19)

$$\rho^2 = 1 - \frac{\sigma_{y.x}^2}{\sigma_y^2}$$

Proporción de varianza explicada=1-(proporción de varianza no explicada)

Para los datos muestrales, el valor estimado del coeficiente de determinación se puede obtener con la formula correspondiente:

(14.20)

$$r^2 = 1 - \frac{s_{y.x}^2}{s_y^2}$$

Para propósitos del cálculo, la siguiente formula es conveniente para el coeficiente de determinación de la muestra:

(14.21)

$$r^2 = \frac{b_0 \sum Y + b_1 \sum XY - n \bar{Y}^2}{\sum Y^2 - n \bar{Y}^2}$$

La población de la formula (14.21) se ilustra en el problema 12.12. Aunque esta fórmula de uso frecuente para calcular el coeficiente de determinación para los datos muestrales, no importa ninguna corrección por sesgo e incluye un sesgo ligeramente positivo.

El coeficiente de correlación

Aun cuando el coeficiente de determinación r^2 es relativamente fácil de interpretar, no se presenta para pruebas estadísticas. Sin embargo, la raíz cuadrada del coeficiente de determinación, que se denomina coeficiente de correlación r , ciertamente se presta para pruebas estadísticas porque se puede usar para definir un estadístico de prueba que tiene la misma distribución que la distribución t cuando la correlación poblacional ρ es igual a cero. El valor del coeficiente de correlación va desde -1.00 hasta +1.00. El signo aritmético del coeficiente de correlación, que es siempre el mismo de β_1 en la ecuación de regresión, indica la relación entre X y Y (positivo=directa; negativo=inversa). El coeficiente de correlación para los datos poblacionales, con el signo aritmético siendo el mismo que el de la pendiente β_1 en la ecuación de regresión, es

(14.22)

$$\rho = \sqrt{\rho^2}$$

El coeficiente de correlación para los datos muestrales, con el signo aritmético siendo el mismo que el de la pendiente muestral b_1 , es

(14.23)

$$r = \sqrt{r^2}$$

El signo del coeficiente de correlación indica la dirección de la relación entre las variables X y Y , mientras que el valor absoluto del coeficiente indica la magnitud de la relación. El cuadrado del coeficiente de correlación es el coeficiente de determinación r^2 e indica la proporción de varianza de Y explicada mediante el conocimiento de X (y viceversa).

Como alternativa a la fórmula (14.21), la fórmula siguiente depende de una determinación previa del valor de regresión b_0 y b_1 . Esta fórmula se usa cuando el propósito del análisis es determinar la magnitud y el tipo de la relación entre dos variables, pero sin interés en estimar Y dado X . Cuando se usa esta fórmula, el signo del coeficiente de correlación se determina de manera automática, sin necesidad de observar o calcular la pendiente de la recta de regresión. La fórmula es

(14.24)

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2} \sqrt{n\sum Y^2 - (\sum Y)^2}}$$

La aplicación de la formula (14.24) se ilustra en el problema 14.13.

El coeficiente de correlación muestral r es un poco sesgado como estimador de ρ , con un valor absoluto que es demasiado grande. Este factor no se menciona en muchos libros de texto ya que la cantidad de sesgo es mínima, excepto en muestras muy pequeñas. Un estimador insesgado de coeficiente de determinación para la población se obtiene como sigue:

(14.25)

$$\hat{\rho}^2 = 1 - (1 - r^2) \left(\frac{n-1}{n-2} \right)$$

Método de la varianza para entender el coeficiente de correlación

Otra medida que se usa para expresar la relación entre dos variables aleatorias es la covarianza muestral. La covarianza mide la magnitud en la que dos variables “varían de manera conjunta” y la usan los análisis financieros para determinar el riesgo total relacionado con inversiones interrelacionadas. Como en el caso del coeficiente de correlación, signo positivo indica una relación directa mientras que signo negativo indica una relación inversa. La fórmula para la covarianza muestral es:

(14.26)

$$cov(X, Y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n - 1}$$

Al estudiar la formula (14.26) se concluye, por ejemplo, que cuando el valor observado de Y tiende a variar en la misma dirección, respecto a la media, que el valor observado de X , también con respecto a la media, entonces los productos de estas desviaciones tienden a ser valores positivos. Por tanto, la suma de estos productos será positiva, lo que indica una relación directa y no inversa. Mientras que el coeficiente de correlación es una medida estandarizada de una relación cuyo valor varía entre -1.00 y + 1.00, el valor de la varianza se convierte en el coeficiente de correlación es:

(14.27)

$$r = \frac{cov(X, Y)}{s_x s_y}$$

El problema 14.14 ilustra la determinación de una covarianza muestral y su conversión en el coeficiente de correlación correspondiente.

Pruebas de significancia con respecto al coeficiente de correlación

Es común que la hipótesis nula de interés sea que la correlación poblacional $\rho = 0$, ya que si esta hipótesis se rechaza para un nivel α determinados, se incluye que existe relación en las dos variables en la población. También se puede formular la hipótesis como una prueba de una cola. Dado que se satisfacen las suposiciones de la sección 14.9, el siguiente estadístico muestral en el que se usa r está distribuido como la distribución t con $gl=n-2$ cuando $\rho=0$.

(14.28)

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

La prueba de la hipótesis nula $\rho=0$ equivale a la prueba de la hipótesis nula de que $\beta = 0$ en la ecuación de regresión. (Problema 14.15)

Fallas y limitaciones relacionadas con el análisis de regresión y de correlación

- 1) En el análisis de regresión, un valor de Y no se debe estimar el valor de X esta fuera del rango de valores que sirvieron como base para obtener la ecuación de regresión.
- 2) Si la estimación de Y involucra la predicción de resultado que todavía no ha ocurrido, los datos históricos que sirvieron como base para la ecuación de regresión pueden no ser relevantes para eventos futuros.
- 3) El uso de un intervalo de predicción o un intervalo de confianza se basa en la suposición de que las distribuciones condicionales de Y , y por tanto de los residuales, son normales y tienen varianzas iguales.
- 4) Un coeficiente de correlación significativo no necesariamente indica una relación de causa y efecto, sino que puede indicar un nexo común con otros eventos.
- 5) Una correlación significativa no es necesariamente una correlación importante. Dada una muestra grande, una correlación de, digamos, $r=+0.10$ puede ser significativamente de 0 para $\alpha = 0.05$. Sin embargo, el coeficiente de determinación

$r^2 = 0.01$ en este ejemplo indica que solo 1 por ciento de la varianza en Y es estadísticamente explicable conociendo X.

6) La interpretación de los coeficientes de correlación y de determinación está basada en la suposición de una distribución normal derivada para la población y, para cada variable, varianzas condicionales iguales.

7) Tanto en el análisis de regresión como en el de correlación se supone un modelo lineal. Para una relación que sea curvilínea, se puede utilizar alguna transformación para obtener la linealidad. Otra posibilidad es restringir el análisis al rango de valores dentro del cual la relación sea en esencia lineal.

Uso de Excel y de Minitab

Para realizar los diferentes procedimientos del análisis de regresión y de correlación simples que se estudian en este capítulo se puede emplear el software. El problema resuelto 14.17 ilustra el uso de Excel y compara el resultado con los resultados que se obtienen mediante el cálculo manual en los problemas que lo preceden al final del capítulo. El problema 14.18 ilustra el uso de Minitab y compara el resultado con los resultados que se obtienen mediante calculo manual en los problemas que lo preceden al final del capítulo.

Determine la probabilidad de cada uno de los eventos siguientes.

- a) La probabilidad de tomar de manera aleatoria una cuenta por cobrar en moratoria, dado que 5 por ciento de las cuentas están en esa condición.
- b) La probabilidad de que una inversión en un terreno tenga éxito. En el área en cuestión, por lo general solo la mitad de estas inversiones tienen éxito, pero los métodos de decisión del inversionista de que se trata lo han conducido a tener un record de 30 por ciento mejor que el inversionista promedio en el área.
- c) La probabilidad de que la suma de los puntos de dos dados después de lanzados sea siete.

De 500 empleado, 200 participan en el plan de reparto de utilidades de una empresa (P). 400 tienen seguro de gastos médicos mayores (M), y 200 participan en ambos programas. Construya un diagrama de Venn para ilustrar los eventos denotados por P y M.

Utilizando el ejercicio anterior a) determine la probabilidad de que un empleado participe en el plan de reparto de utilidades (P) siendo que el empleado tiene cobertura de gastos médicos mayores (M) y b) determine si los dos eventos son independientes o dependientes haciendo referencia al valor de la probabilidad condicional.

Durante un periodo específico, 80 por ciento de las emisiones de acciones comunes en una industria que solo 10 empresas han incrementado su valor de mercado. Si un inversionista escoge dos de estas emisiones en forma aleatoria, ¿Cuál es la probabilidad de que ambas hayan incrementado su valor de mercado durante este periodo?

El número de automóviles vendidos por cada uno de los 10 agentes de ventas en una distribuidora de automóviles durante el mes específico. En orden ascendente, es: 2, 4, 7, 10, 10, 10, 12, 12, 14, 15. Determine a) el rango, b) el rango intercuartilico y c) el rango del 80 por ciento central de estos datos.

En una muestra de paquetes que salen de una oficina de correos, se encuentra que los pesos, dados hasta la onza más cercana, son: 21, 18, 30, 15, 14, 17, 28, 10, 16, 25, 25 onzas. Determine a) el rango y b) el rango intercuartilico de estos pesos.

Suponga que hay dos urnas U_1 y U_2 . La primera contiene dos bolas rojas y una bola verde, mientras que la segunda contiene una roja y dos bolas verdes.

- a) Se elige en forma aleatoria y luego se saca también en forma aleatoria una bola de la urna. La bola es roja. ¿Cuál es la probabilidad de que la urna elija sea la U_1 ?
- b) Se elige una urna en forma aleatoria y luego se toman de la urna, en la misma forma, dos bolas (sin reemplazo). La primera bola es roja y la segunda es verde. ¿Cuál es la probabilidad de que la urna elegida sea la U_1 ?

Ochenta por ciento del material de vinilo que se recibió del proveedor de A es la calidad excepcional, mientras que solamente 50 por ciento del material de vinilo que se recibió del proveedor B es de calidad excepcional. Sin embargo, la capacidad de fabricación del proveedor A es limitada, y por esta razón solo 40 por ciento del material vinilo que se compra de la empresa proviene del proveedor A. EL otro 60 por ciento proviene del proveedor B. Se inspecciona un flete de material de vinilo que va entrando y se encuentra que es de calidad excepcional. ¿Cuál es la probabilidad de que provenga del proveedor A?

En los varios de miles de artículos que tienen en existencia una empresa de ventas por correo, existe una probabilidad general de 0.08 de que un artículo determinado (considerado características tales como tamaño y color específicos, etc.) no se encuentren en existencia?

Durante el periodo pico de 4 p.m. en una estación de servicio automotriz, en promedio cada 3 minutos llega un automóvil. ¿Cuál es la probabilidad de que entre 4 p.m. y 5 p.m. lleguen por lo menos 25 automóviles a la estación de servicio?

Se ha encontrado que la vida útil de cierta marca de llantas de alto rendimiento se ajusta a una distribución normal $\mu = 38\,000$ y $\sigma = 3\,000$ millas. A) ¿Cuál es la probabilidad de que una llanta elegida en forma aleatoria tenga una vida útil de por lo menos 35 000 millas? B) Cual es la probabilidad de que dure más de 45 000 millas?

Un distribuidor compra para reventa 500 de las llantas del problema 7.29. ¿Cuántas llantas duraran aproximadamente a) entre 40 000 y 45 000 millas? B) 40 000 millas o más?

Se ha encontrado que el tiempo requiere una persona en la ventanilla de un cajero de banco tiene una distribución aproximadamente normal con $\mu = 130$ segundos y $\sigma = 45$ segundos. ¿Cuál es la probabilidad de que una persona elegida en forma aleatoria a) requiera menos de 100 segundos para realizar una transacción? b) demore entre 2.0 y 3.0 minutos en la ventanilla del cajero?

Existe 90 por ciento de las posibilidades de que un determinado tipo de componente funcione de manera adecuada de bajo temperatura elevada. Si un dispositivo tiene cuatro de estos componentes, determine la probabilidad de cada uno de los siguientes eventos mediante el uso de la fórmula de probabilidades binomiales.

- a) Que todos los componentes funcionen de manera adecuada, y por tanto el dispositivo también
- b) Que el dispositivo no funcione porque falle exactamente uno de los cuatro componentes.
- c) Que el dispositivo no funcione porque fallen uno o más de los componentes.

En una clase de 20 estudiantes, 15 están descontentos con el libro de texto que se usa. Si a una muestra aleatoria de cuatro estudiantes se les pregunta acerca del texto, determine la probabilidad de que a) exactamente tres y b) cuando menos tres estén descontentos con el libro de texto.

En una tienda de departamentos, seis personas por hora en promedio usan un cajero automático durante horas pico. ¿Cuál es la probabilidad de que

- a) Exactamente seis personas usen el cajero durante una hora tomada al azar?
- b) Menos de cinco personas usen el cajero durante una hora tomada al azar?
- c) Nadie use la instalación durante un intervalo de 10 minutos?
- d) Nadie usa la instalación durante un intervalo de 5 minutos?