

Unidad I Estadística Descriptiva

PRESENTACIÓN DEL CURSO

La ESTADISTICA es la parte de las matemáticas encargada de la presentación y análisis de los datos de un experimento.

Normalmente la estadística se divide en:

- Estadística Descriptiva
- Estadística Inferencial

ESTADÍSTICA DESCRIPTIVA: se encarga de la presentación adecuada de la información (tablas, gráficas, histogramas, etc.)

ESTADÍSTICA INFERENCIAL: se especializa en la estimación e inferencia de parámetros (promedio, desviación estándar, etc.).

Experimentos probabilísticos y determinísticos

Un EXPERIMENTO es un procedimiento mediante el cual se puede obtener información acerca de un sistema físico ó Matemático.

El objetivo principal de realizar experimentos es obtener información acerca de sistema bajo estudio, y a partir de ella obtener conclusiones.

Los DATOS son en generalmente la forma en que se presenta la información obtenida de un experimento.

Los datos pueden clasificarse primeramente como:

DATOS NUMERICOS.- son aquellos que como su nombre indica pueden representarse mediante un número real el cual representa su magnitud y sus respectivas unidades de medición, por ejemplo los obtenidos de la medición de una cantidad física como longitud, masa, tiempo, energía, etc.

DATOS DE ATRIBUTO. Son aquellos datos que no se pueden expresar como datos numéricos, por ejemplo, sabor, color, sexo, nombre, país, nacionalidad, etc.

Se dice que un EXPERIMENTO ES DETERMINÍSTICO si al realizarse bajo las mismas condiciones se obtiene invariablemente el mismo resultado o dato, en el caso de que se obtenga resultados o datos diferentes se dirá que es un EXPERIMENTO PROBABILISTICO ó ALEATORIO.

Población muestra, eventos

La POBLACION es el conjunto total de datos que se obtienen al realizar un experimento.

La MUESTRA es una parte ó subconjunto de la población.

Los EVENTOS están formados generalmente por muestras a las cuales se les pide que cumplan con alguna condición o condiciones.

ORGANIZACIÓN DE DATOS

Una vez que se ha realizado un experimento el resultado generalmente es un conjunto de datos u observaciones, sin embargo, tal como aparecen pueden no resultar adecuados para obtener información de ellos, por lo que es necesario realizar en la mayoría de los caso un trabajo mínimo que consiste en la organización y presentación de los datos de manera adecuada. Esto es precisamente el objetivo de la estadística descriptiva.

Como primer paso los datos pueden ser acomodados en un ARREGLO, el cual tiene el objetivo de presentar los datos con un mínimo de orden. Es deseable que este orden sea descendente o ascendente, como se muestra a continuación.

NUMERO DE PERSONAS VIVIENDO EN UN GRANJAS

2	4	5	6	6	7	8	8	9	10
2	4	5	6	7	7	8	9	9	11
3	4	5	6	7	7	8	9	10	11
3	5	5	6	7	7	8	9	10	12
4	5	6	6	7	8	8	9	10	12

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

A partir de los datos ordenados en un arreglo se puede presentar los datos en una DISTRIBUCION DE FRECUENCIAS. Para realizar la distribución de frecuencias se puede seguir el siguiente procedimiento:

a) Localice el valor máximo (X_{\max}) y mínimo (X_{\min}) del conjunto de datos, y a partir de ellos Obtégase el RANGO como:

$$R = X_{\max} - X_{\min}$$

b) Ahora proceda a dividir el rango en INTERVALOS DE CLASE, se sugiere que el número de intervalos de clase no sea menor a 6 ni mayor a 20.

c) La LONGITUD DE EL INTERVALO de cada clase debe ser la misma en todas las clases y deberá ser de tal que el punto medio de cada intervalo tenga en mismo número de dígitos y precisión que los datos originales.

d) Una vez definidos adecuadamente los intervalos proceda a contar los datos que se encuentren dentro de su límite inferior y su límite superior, el número de datos que caen dentro de dicho intervalo, constituye la FRECUENCIA DE CLASE.

e) Tome en cuenta que cada dato solo pertenece solamente a una clase, por lo que no debe haber ambigüedad en su pertenencia a alguna clase.

f) El punto medio de cada intervalo es llamado LA MARCA DE CLASE y representará a todos los puntos que caigan dentro del intervalo.

g) LA TABLA DE DISTRIBUCIÓN DE FRECUENCIA se construye colocando en la primera columna (ó fila) los intervalos de clase y/o las marcas de clase y en la siguiente columna (ó fila) las frecuencias correspondientes.

EJEMPLOS

1. Obtenga la tabla de la distribución de frecuencias para los datos siguientes.

NÚMERO DE PERSONAS VIVIENDO EN UN GRANJAS

2 4 5 6 6 7 8 8 9 10
 2 4 5 6 7 7 8 9 9 11
 3 4 5 6 7 7 8 9 10 11
 3 5 5 6 7 7 8 9 10 12
 4 5 6 6 7 8 8 9 10 12

Por la naturaleza de los datos presentados en la tabla se puede optar por que cada uno de los valores: 2, 3, 4, 5, 6, 7, 8, 9, 10 11 y 12 sean los "intervalos", entonces

X	2	3	4	5	6	8	9	10	11	12
FR(X)	2	2	4	6	7	7	6	4	2	2

(2) Obtenga la tabla de la distribución de frecuencias para los datos siguientes. Divida en 7 clases.

2.3 3.7 4.3 4.7 5.4
 2.3 3.8 4.4 4.8 5.5
 2.4 3.8 4.4 4.8 5.6
 2.6 3.9 4.4 4.9 5.7
 2.8 3.9 4.5 4.9 5.8
 3.0 4.0 4.5 5.0 5.9
 3.4 4.0 4.6 5.0 6.0
 3.5 4.1 4.6 5.1 6.4
 3.5 4.1 4.6 5.1 6.5
 3.6 4.3 4.6 5.3 7.1

El rango es $R = 7.1 - 2.3 = 4.8$.

Dividiendo el rango en $N = 7$ intervalos ancho $= 4.8 / 7 = 0.6857$

Como el ancho tiene muchos dígitos, el ancho se puede redefinir como ancho $= 0.7$

Pero en este caso la longitud total de los intervalos es Longitud $= (7) (0.7) = 4.9$

Esta longitud excede en $4.9 - 4.8 = 0.1$ al rango, este excedente se puede repartir entre las clase extremas, por ejemplo, el límite inferior de la primera clase es 2.25 y el superior $2.25 + 0.7 = 2.95$. Para la segunda clase se considera como límite inferior el límite superior de la primera clase, su correspondiente límite superior es $2.95 + 0.7 = 3.65$, el proceso anterior se repite para cada una de las clases posteriores.

Los resultados son colocados en la siguiente tabla

Clases	Marca de Clase	Frecuencia FR(X)
2.25 -2.95	2.6	5
2.95 -3.65	3.3	5
3.65 - 4. 35	4.0	11
4.35 -5.05	4.7	16
5.05 -5.75	5.4	6
5.75 -6.45	6.1	5
6.45 -7.15	6.8	2

Tabla 1. Distribución de frecuencias problema 2

PRESENTACIÓN GRÁFICA DE DATOS.

HISTOGRAMA Y POLÍGONO DE FRECUENCIAS

La tabla de distribución de frecuencias puede ser utilizada para obtener una gráfica en la cual se coloca en el eje X los puntos medios de las clases y en el eje Y las correspondientes frecuencias de la clase. La gráfica descrita se conoce como HISTOGRAMA.

Un histograma se puede convertir en un POLÍGONO DE FRECUENCIAS simplemente conectando los puntos medios o marcas de clase con líneas rectas, pero es necesario agregar dos puntos medios extras, uno correspondiente a una previa a la primera clase y con frecuencia cero y otro posterior a la última clase con frecuencia cero.

OJIVA

Para algunas aplicaciones es requerido obtener la tabla de las FRECUENCIAS ACUMULADAS la cual se obtiene sumando las frecuencias precedentes a cada una de las clases. La gráfica de las clases vs las frecuencias acumuladas es conocida como OJIVA

EJEMPLOS

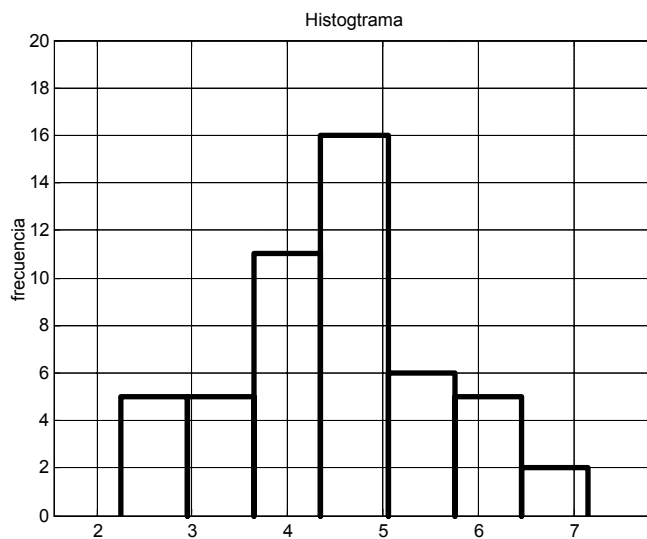
3. Utilice el resultado de problema (2) anterior para obtener el histograma, polígono de frecuencias y ojiva.

SOLUCION: Primero se obtiene la frecuencia acumulada de los datos.

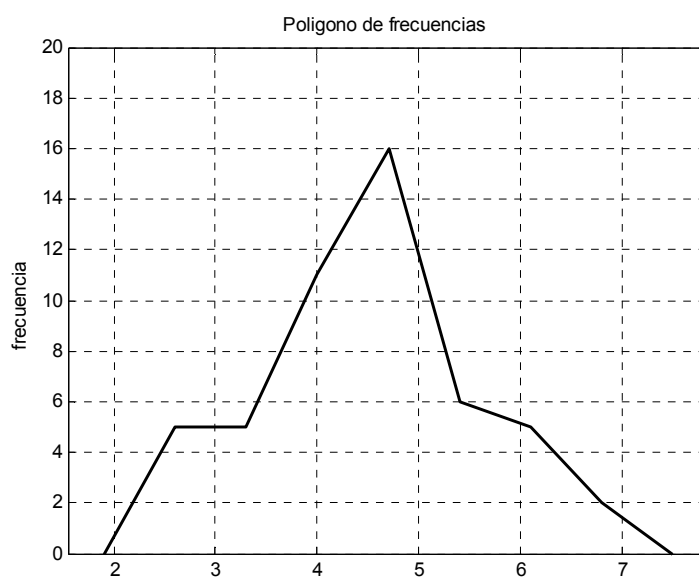
Clases	Marca de Clase	Frecuencia FR(X)	Frecuencia acumulada
2.25 -2.95	2.6	5	5
2.95 -3.65	3.3	5	10
3.65 - 4. 35	4.0	11	21
4.35 -5.05	4.7	16	37
5.05 -5.75	5.4	6	43
5.75 -6.45	6.1	5	48
6.45 -7.15	6.8	2	50

Tabla 1. Distribución de frecuencias y frecuencias acumuladas ejemplo1

A continuación se presentan cada una de las gráficas solicitadas a partir de los datos de la tabla anterior

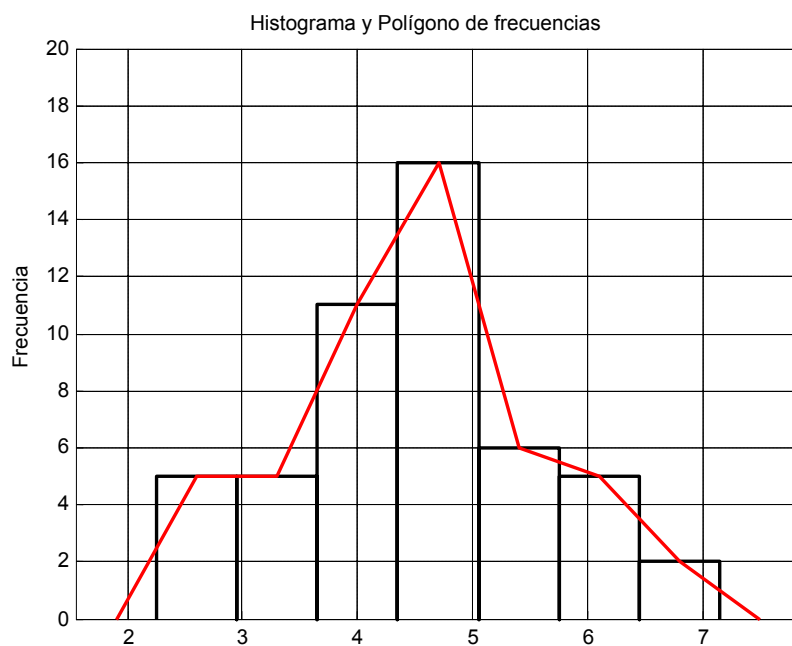


Histograma del ejemplo 1

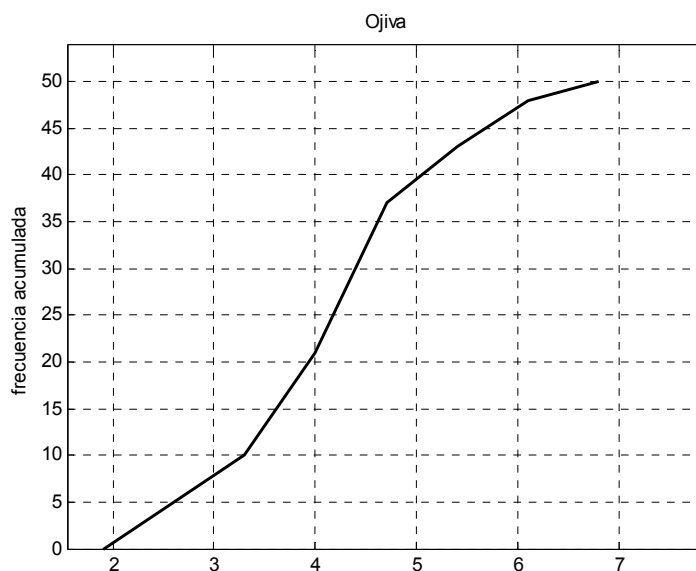


Gráfica del polígono de frecuencias del ejemplo 1

Las gráficas anteriores representan a la distribución de frecuencias, por lo que pueden ser representadas juntas como se observa a continuación.



Histograma y polígono de frecuencias del ejemplo 1

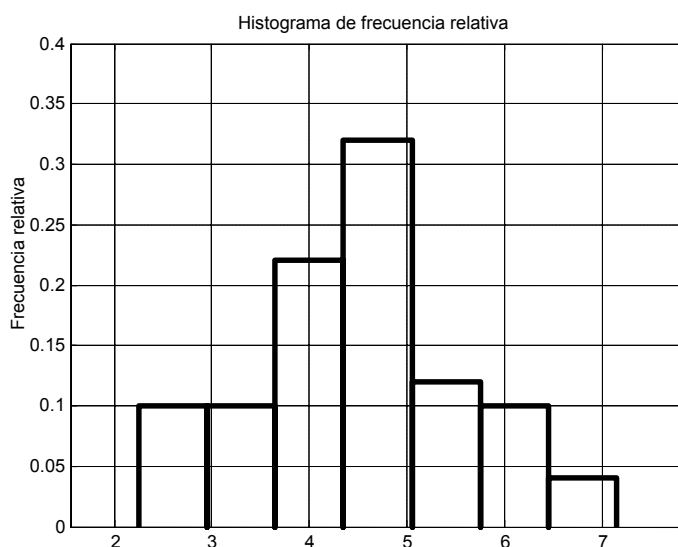


Ojiva o gráfica de las frecuencias acumuladas del problema 1

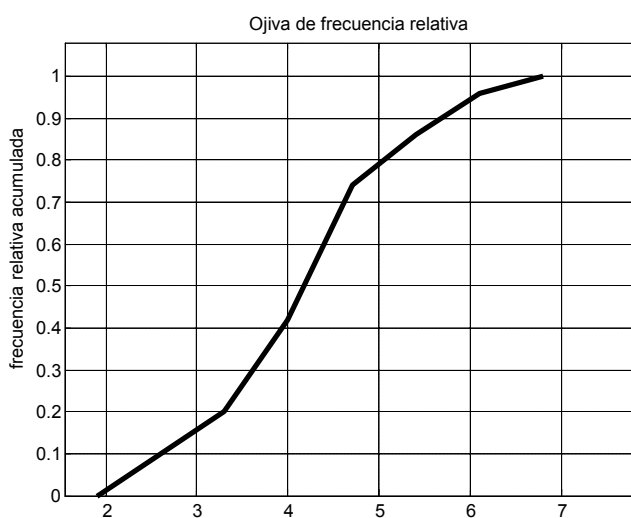
Histograma de frecuencias relativas

Si se dividen las frecuencias obtenidas en la tabla de distribución de frecuencias entre el total de datos se obtiene la llamada LA TABLA DE DISTRIBUCIÓN DE FRECUENCIA RELATIVA, y su respectiva gráfica se llama HISTOGRAMA DE FRECUENCIAS RELATIVAS. Lo anterior se puede aplicar también a la tabla de frecuencias acumuladas obteniéndose LA TABLA DE FRECUENCIAS ACUMULADAS RELATIVAS y su respectiva gráfica se llama OJIVA DE FRECUENCIAS RELATIVAS. La ventaja del uso de las frecuencias relativas es su inmediata relación con la probabilidad, es decir, la frecuencia relativa de una clase es la probabilidad de que los datos considerados se encuentren en dicho intervalo.

(2) A continuación se muestran algunas de las gráficas del problema 2 para el caso de frecuencias relativas.



Histograma de frecuencias relativas del ejemplo 1



Ojiva de frecuencias relativas acumuladas del ejemplo 1

4. Se realiza una investigación a los vendedores de una cadena nacional de tiendas de departamentos para determinar el patrón de sus ingresos diarios. Se seleccionan una muestra aleatoria de 50 vendedores y se obtienen sus ingresos durante cierto día.

53	57	58	61	61
63	64	66	67	68
69	70	71	72	73
74	74	74	74	77
77	77	78	81	79
79	79	81	78	81
82	82	83	83	84
85	85	86	87	87
88	90	90	90	90
92	93	94	96	97

a) Organice los datos en una tabla. Las clases son 52.5 - 57.5, 57.5 - 62.5, 62.5 - 67.5,..., 92.5 - 97.5

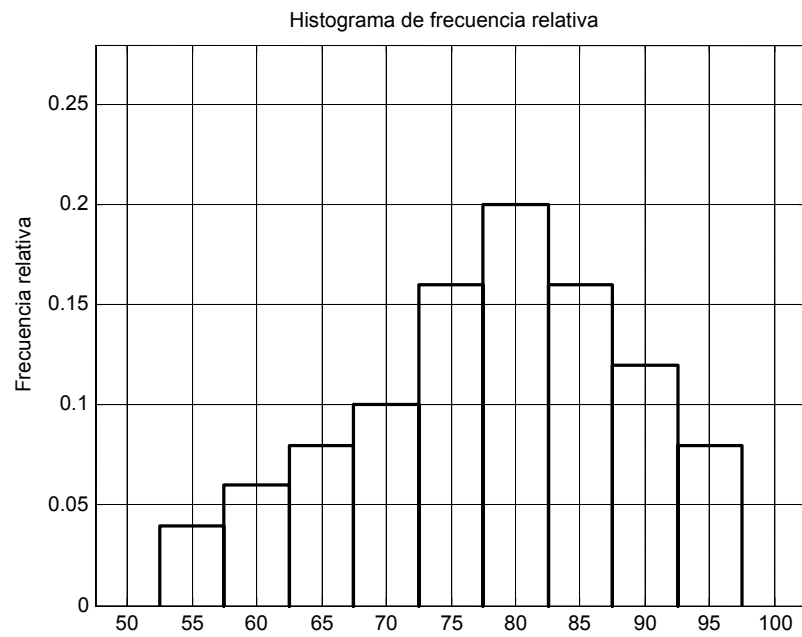
b) Conviértase en frecuencias relativas y relativas acumuladas. Obténgase el Histograma de frecuencias relativas y la ojiva de frecuencias relativas.

SOLUCION

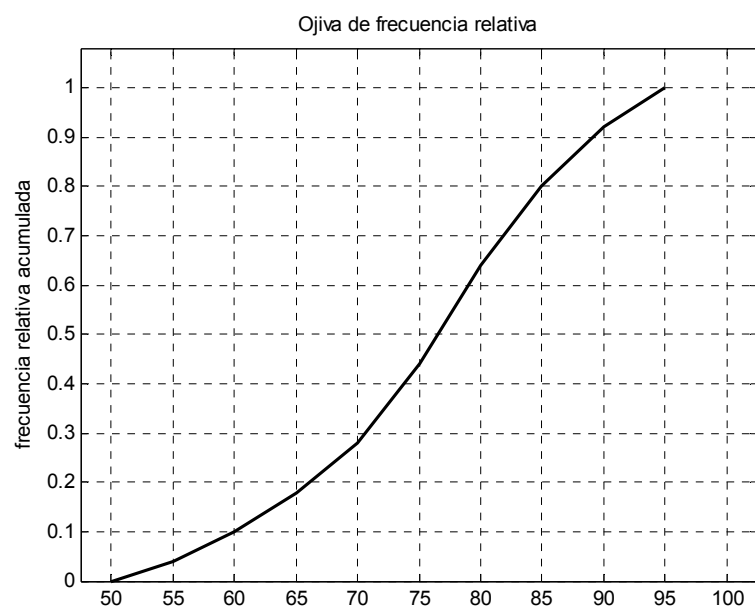
A partir de los datos y las clases propuestas se determina la siguiente tabla.

Clases	Marca de Clase	Frecuencia FR(X)	Frecuencia acumulada	Frecuencia relativa FR(X)	Frecuencia relativa acumulada
52.5 -57.5	55	2	2	0.0400	0.0400
57.5 - 62.5	60	3	5	0.0600	0.1000
62.5- 67.5	65	4	9	0.0800	0.1800
67.5 -72.5	70	5	14	0.1000	0.2800
72.5 - 77.5	75	8	22	0.1600	0.4400
77.5 - 82.5	80	10	32	0.2000	0.6400
82.5 - 87.5	85	8	40	0.1600	0.8000
87.5 - 92.5	90	6	46	0.1200	0.9200
92.5 - 97.5	95	4	50	0.0800	1.0000

Tabla 2. Distribución de frecuencias, frecuencias acumuladas y relativas de ejemplo 2



Histograma de frecuencias relativas del ejemplo 2



Ojiva de frecuencias relativas acumuladas del ejemplo 1

MEDIDAS DE TENDENCIA CENTRAL

Las MEDIDAS TENDENCIA CENTRAL ó DE CENTRALIZACION de tienen como objetivo es tratar de localizar (ó encontrar) el centro de la distribución. Las más conocidas son la MEDIA ARITMETICA MEDIANA y MODA.

Es costumbre representar algunas propiedades y definiciones mediante la notación sigma:

$$\sum_{i=1}^N a_i = a_1 + a_2 + a_3 + \dots + a_N$$

Como se puede observar es utilizada para representar la suma de de elementos también conocida como serie. A continuación se presentan algunas de las propiedades más importantes, las cuales se utilizarán posteriormente.

Propiedades de la notación sigma

Sean $\sum_{i=1}^N a_i$ y $\sum_{i=1}^N b_i$ dos sumatorias y c una constante, entonces:

$$a) \sum_{i=1}^N (a_i + b_i) = \sum_{i=1}^N a_i + \sum_{i=1}^N b_i$$

$$b) \sum_{i=1}^N ca_i = c \sum_{i=1}^N a_i$$

MEDIA ARITMÉTICA, PROMEDIO \bar{X}

La media aritmética, promedio o simplemente media es denotada por: \bar{X} , es simplemente la suma de todas las observaciones $X_1, X_2, X_3, \dots, X_N$, dividida entre el número N total de datos, esto es:

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N} \quad (1.1)$$

Es posible dar una justificación matemática a la definición anterior. Para tal fin, supongamos que se define la función $D(X)$ como a continuación se indica

$$S(a) = \sum_{i=1}^N (X_i - a)$$

Donde X_i son los datos y a es una constante, el menor valor de la función es $S(a) = 0$, entonces

$$S(a) = \sum_{i=1}^N (X_i - a) = 0$$

Aplicando las propiedades de la notación sigma

$$\sum_{i=1}^N X_i - \sum_{i=1}^N a = 0$$

$$\sum_{i=1}^N X_i - Na = 0$$

Despejando a a

$$a = \frac{\sum_{i=1}^N X_i}{N}$$

La cual corresponde a la definición del promedio.

Para datos agrupados se calcula la media mediante la ecuación.

$$\bar{X} = \frac{\sum_{i=1}^N f(x_i)x_i}{\sum_{i=1}^n f(x_i)} \quad (1.2)$$

La suma de las frecuencias individuales es igual al número total de datos, esto es

$$N = \sum_{i=1}^n f_i(x_i)$$

Entonces

$$\bar{X} = \frac{\sum_{i=1}^n f(x_i)x_i}{N} \quad (1.3)$$

MEDIANA \tilde{X}

Para el caso de datos no agrupados, la mediana \tilde{X} , es el número que divide el conjunto de datos en dos partes iguales $\frac{N}{2}$.

En el caso de datos agrupados, la mediana se define como el valor \tilde{X} que divide al histograma correspondiente en dos partes con áreas iguales. Para datos agrupados la mediana se puede obtener mediante

$$\tilde{X} = L_i(x_m) + \frac{N/2 - CF(x_{m-1})}{F(x_m)} w \quad (1.4)$$

Donde

$L_i(x_m)$ Límite inferior de la clase que contiene a la mediana-

$N/2$ Mitad de los datos.

$CF(x_{m-1})$ Frecuencia acumulada hasta la clase anterior a la que contiene a la mediana.

$F(x_m)$ Frecuencia de la clase que contiene a la mediana.

w Ancho de la clase.

MODA \hat{X}

La moda \hat{X} es el valor que más veces aparece en un conjunto de datos.

EJEMPLO

5. Determine media, mediana y moda para la distribución de frecuencias siguiente y localice sobre el histograma cada una de ellas sobre el histograma correspondiente.

Clases	X	F(x)
52.5 -57.5	55	2
57.5 - 62.5	60	3
62.5- 67.5	65	4
67.5 -72.5	70	5
72.5 - 77.5	75	8
77.5 - 82.5	80	10
82.5 - 87.5	85	8
87.5 - 92.5	90	6
92.5 - 97.5	95	4
TOTAL		50

SOLUCION

Es recomendable construir la tabla siguiente a partir de los datos dados:

Clases	X	F(x)	X F(X)
52.5 -57.5	55	2	110
57.5 - 62.5	60	3	180
62.5- 67.5	65	4	260
67.5 -72.5	70	5	350
72.5 - 77.5	75	8	600
77.5 - 82.5	80	10	800
82.5 - 87.5	85	8	680
87.5 - 92.5	90	6	540
92.5 - 97.5	95	4	380
TOTAL		50	3900

La media se obtiene a partir de la definición de datos agrupados

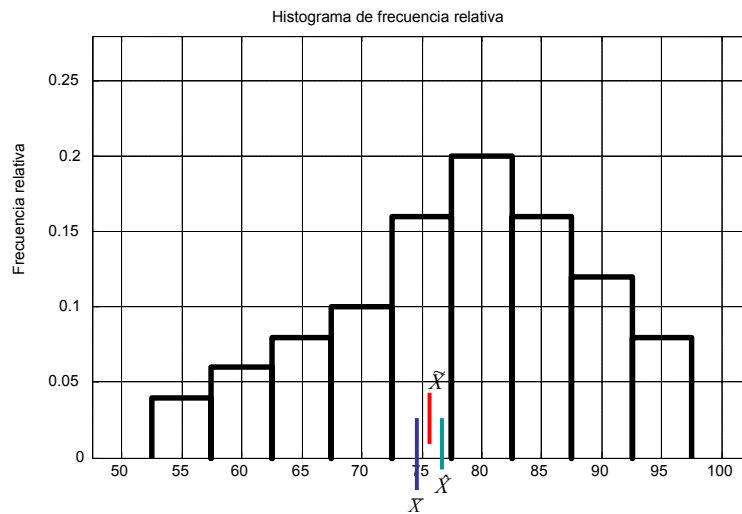
$$\bar{X} = \frac{\sum_{i=1}^n f(x_i)x_i}{N} = \frac{3900}{50} = 78$$

La clase que contiene a la mediana se ha sombreado en la tabla anterior. La mediana se obtiene aplicando la ecuación para datos agrupados

$$\tilde{X} = L_i(x_m) + \frac{\frac{N}{2} - CF(x_{m-1})}{F(x_m)} w = 77.5 + \left(\frac{50/2 - 22}{10} \right) 5 = 79$$

La moda es simplemente $\hat{X} = 80$

La gráfica siguiente muestra que las tres medidas de centralización, las cuales son muy cercanas entre si y se localizan como debe ser en el centro del histograma.



MEDIDA DE DISPERSIÓN

DESVIACIÓN TÍPICA Ó ESTÁNDAR

La desviación típica ó estándar: es la medida de dispersión más representativa de un conjunto de datos. Se define utilizando como

$$S_N = \left[\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N} \right]^{\frac{1}{2}} \quad (1.5)$$

La fórmula anterior es conocida como **desviación típica ó estándar sesgada**

Para datos agrupados la fórmula anterior se escribe como

$$S_N = \left[\frac{\sum_{i=1}^N f(x_i)(x_i - \bar{x})^2}{N} \right]^{\frac{1}{2}} \quad (1.6)$$

VARIANZA

El valor de la desviación estándar al cuadrado es conocido como la Varianza, esto es

$$\text{Varianza} = S^2$$

Una forma alternativa para el cálculo de la varianza y/o de la desviación estándar sesgada se obtiene desarrollando la definición dada, esto es

$$\begin{aligned} S_N^2 &= \frac{\sum (x_i - \bar{x})^2}{N} = \frac{1}{N} \sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{N} \left(\sum x_i^2 - \sum 2x_i\bar{x} + \sum \bar{x}^2 \right) \\ &= \frac{1}{N} \left(\sum x_i^2 - 2\bar{x} \sum x_i + \bar{x}^2 \sum 1 \right) \\ &= \frac{1}{N} \left(\sum x_i^2 - 2\bar{x} N\bar{x} + N\bar{x}^2 \right) \\ &= \frac{1}{N} \sum x_i^2 - \bar{x}^2 \end{aligned}$$

Entonces

$$S_N^2 = \frac{1}{N} \sum x_i^2 - \bar{x}^2 \quad (1.7)$$

Notación

Normalmente las letras latinas \bar{x}, S, S^2 , etc., representan los *estadísticos de una muestra* y las letras griegas $\bar{\mu}, \sigma, \sigma^2$, etc., representan los *estadísticos de una población*.

Existe una forma para la varianza muestral S^2 que proporciona una estimación más precisa de la varianza de la población, en particular, cuando la muestra es pequeña ($N \leq 36$); es conocida como **varianza insesgada de la población** y se calcula mediante

$$S_{N-1}^2 = \frac{\sum (x_i - \bar{x})^2}{N-1} \quad (1.8)$$

De aquí se calcula mediante la raíz cuadrada la **desviación estándar insesgada**

$$S_{N-1} = \left[\frac{\sum (x_i - \bar{x})^2}{N-1} \right]^{\frac{1}{2}} \quad (1.9)$$

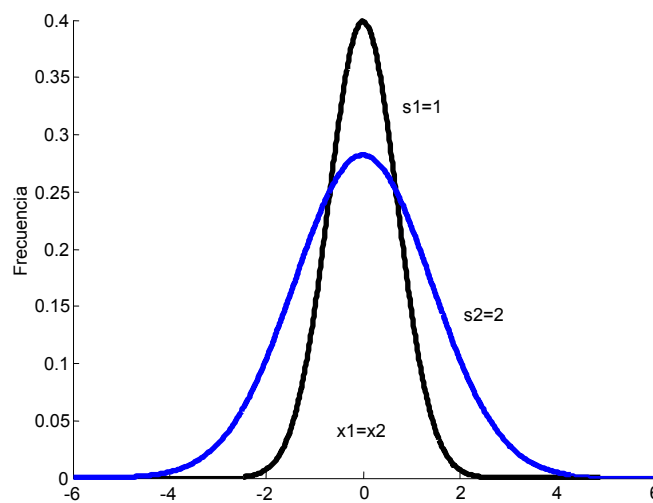
Procediendo de manera similar al caso sesgado se puede obtener una fórmula directa para calcular la varianza y/o desviación estándar insesgada

$$\begin{aligned} S_{N-1}^2 &= \frac{\sum (x_i - \bar{x})^2}{N-1} = \left(\frac{1}{N-1} \right) \sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \left(\frac{1}{N-1} \right) \left(\sum x_i^2 - \sum 2x_i\bar{x} + \sum \bar{x}^2 \right) \\ &= \left(\frac{1}{N-1} \right) \left(\sum x_i^2 - 2\bar{x} \sum x_i + \bar{x}^2 \sum 1 \right) \\ &= \left(\frac{1}{N-1} \right) \left(\sum x_i^2 - 2 \frac{\sum x_i}{N} \sum x_i - N \left(\frac{\sum x_i}{N} \right)^2 \right) \\ &= \left(\frac{1}{N-1} \right) \left(\sum x_i^2 - \frac{(\sum x_i)^2}{N} \right) \end{aligned}$$

Por lo tanto

$$S_{N-1}^2 = \left(\frac{1}{N-1} \right) \left(\sum x_i^2 - \frac{(\sum x_i)^2}{N} \right) \quad (1.10)$$

La desviación estándar como se ha indicado anteriormente es una medida de la dispersión de los datos, esta dispersión se mide a partir de la media de la distribución de datos; por ejemplo, supóngase que se comparan dos conjuntos de datos obtenidos a partir de la misma población, los cuales tienen el mismo número de datos ($N_1 = N_2$), el mismo promedio ($\bar{x}_1 = \bar{x}_2$), entonces, si la desviación del primer conjunto es menor que la del segundo conjunto, ($s_1 < s_2$), es posible afirmar que los datos del primer conjunto se encuentran más concentrados que los de la segundo y la altura del primer conjunto de datos es mayor que la del segundo. La figura siguiente compara dos distribuciones continuas con las características descritas anteriormente.



Comparación de dos distribuciones de frecuencia con diferentes desviaciones estándar $s_1 < s_2$

La desviación estándar se puede emplear también para medir las variaciones con respecto a la media de los valores con respecto a la media. Un valor pequeño de la desviación típica ó estándar indica una mayor probabilidad de obtener un valor más cercano a la media. Esta idea se expresa en un teorema enunciado por el matemático ruso Tchebycheff.

Teorema de Tchebycheff

La proporción de cualquier conjunto de valores que caerá dentro k desviaciones típicas a partir de la media es al menos $1-1/k^2$, donde k es cualquier número mayor que 1.

Por ejemplo, para el caso de $k = 2$, el teorema anterior garantiza que sin importar como es la distribución de frecuencias, existe $1-1/2^2=0.75$ de los datos se encuentran dentro del intervalo comprendido por $[\bar{x} - 2s, \bar{x} + 2s]$. En la figura 1, se muestra la idea del teorema de Tchebycheff para $k = 2$.

Regla de la normal

En muchas ocasiones el histograma que representa la distribución de frecuencia tiene una forma de campana simétrica, este tipo de distribución puede ser comparada con una distribución teórica continua llamada curva normal. Es posible aplicar las características de la curva normal a este tipo de distribuciones muestrales para determinar la proporción de datos contenidos dentro de una, dos y tres desviaciones estándar. A continuación se enuncia la regla de la normal.

Para distribuciones de frecuencia simétricas en forma de campana, aproximadamente el 68 % de los datos caerán en el intervalo $[\bar{X} - S, \bar{X} + S]$, el 95 % de los datos caerán en el intervalo $[\bar{X} - 2S, \bar{X} + 2S]$, y casi el 100 % de los datos caerán en el intervalo $[\bar{X} - 3S, \bar{X} + 3S]$.

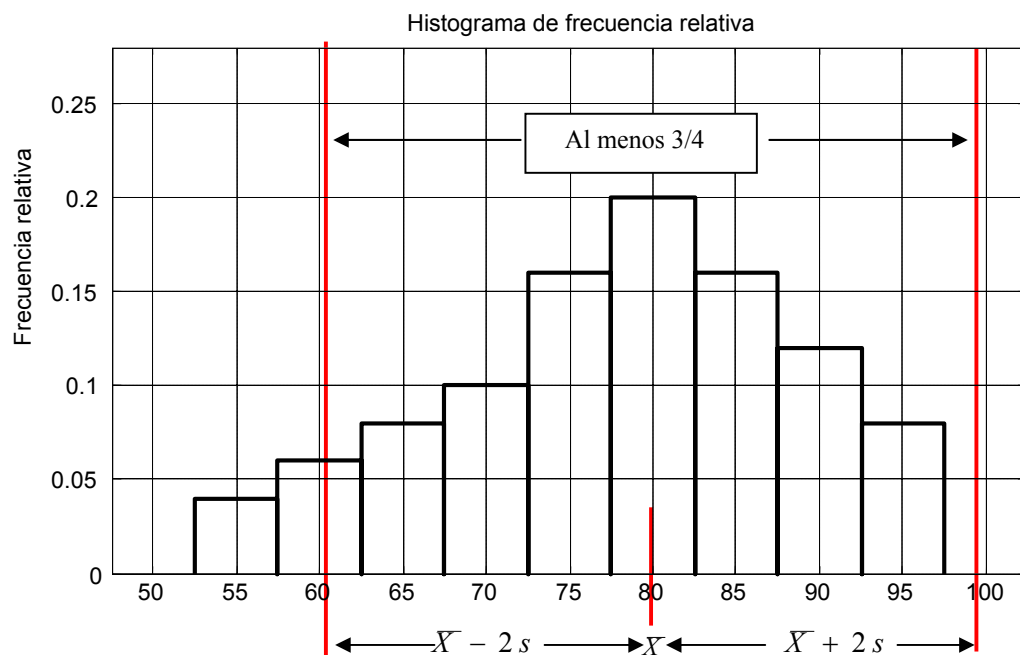


Figura 1, Teorema de Tchebycheff proporción de datos $1 - 1/k^2$ para el caso $k = 2$.

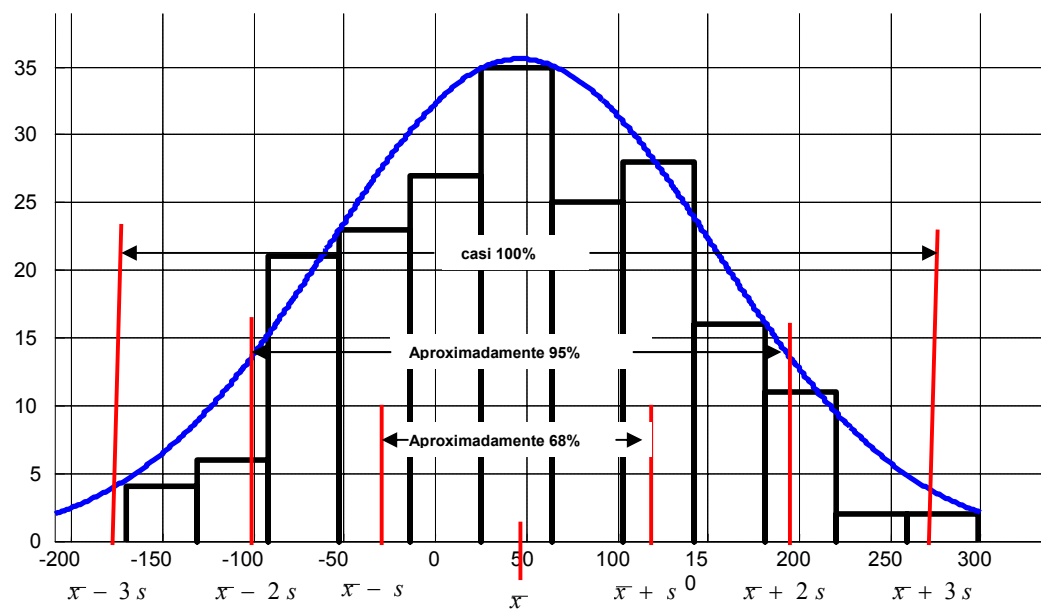


Figura 2, Regla de Normal. 68 % de los datos en el intervalo $[\bar{X} - S, \bar{X} + S]$, el 95 % en $[\bar{X} - 2S, \bar{X} + 2S]$, y casi el 100 % en $[\bar{X} - 3S, \bar{X} + 3S]$.

EJEMPLOS

6. Determine la desviación estándar sesgada e insesgada para el conjunto de datos siguientes.

X	F(x)
55	2
60	3
65	4
70	5
75	8
80	10
85	8
90	6
95	4
	50

SOLUCION

Es recomendable construir la tabla siguiente a partir de los datos dados:

X	F(x)	X F(X)	X ² F(X)
55	2	110	6050
60	3	180	10800
65	4	260	16900
70	5	350	39200
75	8	600	45000
80	10	800	64000
85	8	680	57800
90	6	540	48600
95	4	380	36100
	50	3900	309750

Utilizando los resultados de la tabla en las ecuaciones respectivas

$$S_N^2 = \frac{1}{N} \sum f(x_i) x_i^2 - \bar{x}^2 = \frac{1}{50} (309750) - \left(\frac{3900}{50} \right)^2 = 111$$

$$S_N = \sqrt{111} = 10.54$$

$$S_{N-1}^2 = \left(\frac{1}{N-1} \right) \left(\sum f(x_i) x_i^2 - \frac{(\sum f(x_i) x_i)^2}{N} \right) = \left(\frac{1}{50-1} \right) \left(309750 - \frac{(3900)^2}{50} \right) = 113.27$$

$$S_N = \sqrt{113.27} = 10.64$$

7. Obtenga la mediana para el conjunto de datos siguiente

53	57	58	61	61
63	64	66	67	68
69	70	71	72	73
74	74	74	74	77
77	77	78	81	79
79	79	81	78	81
82	82	83	83	84
85	85	86	87	87
88	90	90	90	90
92	93	94	96	97

SOLUCION

La mediana debe dividir los datos en la mitad, esto es en 25 datos a la izquierda y 25 a la derecha. Puesto que los datos se encuentran acomodados en orden ascendente, se puede observar el dato $X_{25} = 79$ y el dato $X_{26} = 79$, por lo tanto

$$\bar{X} = \frac{X_{25} + X_{26}}{2} = \frac{79 + 79}{2} = 79$$

8. Cierta tarde del sábado 30 estudiantes universitarios de primer semestre trabajaron. A continuación se muestra la distribución de frecuencias de sus ganancias.

- Obtenga la media, mediana y moda
- Obtenga la desviación estándar S_n , S_{n-1}

Ganancia x	Frecuencia f(x)
10	2
15	5
20	9
25	6
30	3
35	5
	30

SOLUCION

Primero se realiza la siguiente tabla a partir de la anterior

x	f(x)	$x_i f(x_i)$	$f(x_i) x_i^2$
10	2	20	200
15	5	75	1125
20	9	180	3600
25	6	150	3750
30	3	90	2700
35	5	175	6125
Σ	30	690	17500

Promedio

$$\bar{X} = \frac{\sum f(x_i)x_i}{N} = \frac{690}{30} = 23$$

Mediana

De los datos de la tabla

Límite inferior de la clase $L_i(x_m) = 17.5$

Frecuencia acumulada hasta antes de la clase m $CF(x_{m-1})$ $m=7$

Frecuencia de la clase donde está la mediana = 9 $F(x_m)$

Ancho de la clase $w = 5$

$$\tilde{X} = L_i(x_m) + \frac{N/2 - CF(x_{m-1})}{F(x_m)} w = 17.5 + \left(\frac{\frac{30}{2} - 7}{9} \right) (5) = 22.22$$

Moda

El valor con mayor frecuencia es $\hat{x} = 20$

Desviación estándar sesgada

$$S_N^2 = \frac{1}{N} \sum f(x_i)x_i^2 - \bar{x}^2 = \frac{1}{30} (17500) - (23)^2 = 54.33$$

Entonces $S = \sqrt{54.33} = 7.37$

Desviación estándar insesgada

$$S_{N-1}^2 = \frac{1}{N-1} \left(\sum f(x_i)x_i^2 - \frac{(\sum f(x)x_i)^2}{N} \right) = \frac{17500 - \frac{(690)^2}{30}}{30-1} = 56.21$$

Por lo tanto $S_{N-1} = \sqrt{56.21} = 7.50$

9. Las mediciones en la escala de Richter correspondientes a los 50 terremotos más recientes en el mundo son dadas en la tabla.

- Constrúyanse una distribución de frecuencias con límites de clase de 2.25 a 2.75, 2.75 a 3.25, etc.
- Trácese el histograma y polígono de frecuencias
- Obtenga la media, mediana y moda
- Obtenga la desviación estándar S_n , S_{n-1}

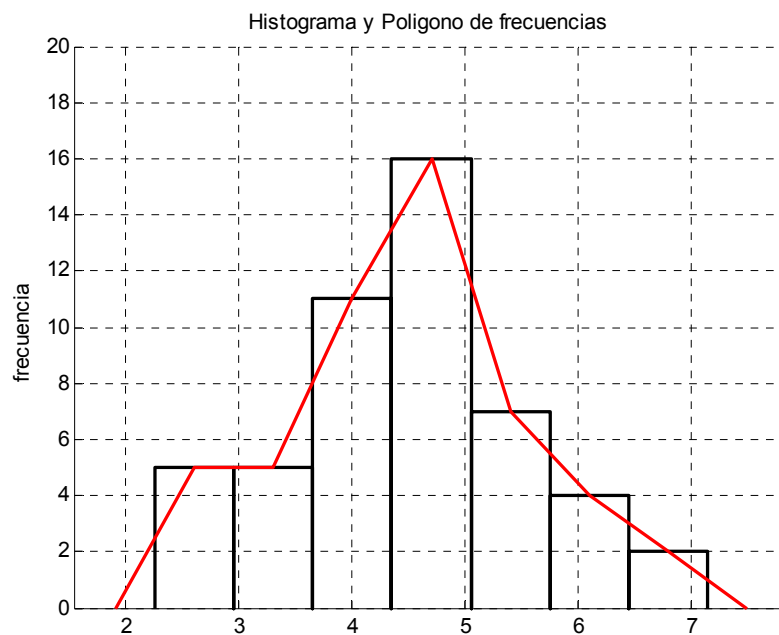
2.3	3.7	4.3	4.7	5.4
2.3	3.8	4.4	4.8	5.5
2.4	3.8	4.4	4.8	5.6
2.6	3.9	4.4	4.9	5.7
2.8	3.9	4.5	4.9	5.8
3.0	4.0	4.5	5.0	5.9
3.4	4.0	4.6	5.0	6.0
3.5	4.1	4.6	5.1	6.4
3.5	4.1	4.6	5.1	6.5
3.6	4.3	4.6	5.3	7.1

SOLUCION

(a) Utilizando las clases sugeridas se determinan las respectivas marcas de clase, frecuencias y se evalúan de $xf(x)$ y $x^2 f(x)$, acomodando los resultados en la siguiente tabla

clase	x	f(x)	$x(f(x))$	$x^2 f(x)$
2.25-2.95	2.6	5	13	33.8
2.95-3.65	3.3	5	16.5	54.45
3.65-4.35	4.0	11	44	176
4.35-5.05	4.7	16	75.2	353.44
5.05-5.75	5.4	7	37.8	204.12
5.75-6.45	6.1	4	24.4	148.84
6.45-7.15	6.8	2	13.6	92.48
Σ		50	224.5	1106.313

(b) Histograma y polígono de frecuencias.



(b) A partir de los datos de la tabla de frecuencia se puede determinar los estadísticos solicitados

Media

$$\bar{x} = \frac{\sum (f_i)(x_i)}{N} = \frac{2245}{50} = \frac{44.9}{16} = 4.49$$

Moda $\hat{x} = 4.7$

Mediana

Para los datos no agrupados

$$\tilde{x} = \frac{\text{dato}\left(\frac{N}{2}\right) + \text{dato}\left(\frac{N}{2} + 1\right)}{2} = \frac{4.5 + 4.5}{2} = 4.5$$

Para los datos agrupados

$$\tilde{X} = L_i(x_m) + \frac{\frac{N}{2} - CF(x_{m-1})}{F(x_m)} w = 4.35 + \left(\frac{\frac{50}{2} - 21}{16}\right)(0.7) = 4.54$$

Desviación estándar sesgada

$$S_N^2 = \frac{1}{N} \sum_{i=1}^n f_i(x_i) x_i^2 - \bar{x}^2 = \frac{1}{50} (1063.13) - (4.49)^2 = 1.1025$$

Entonces

$$S = \sqrt{1.1025} = 1.05$$

Desviación estándar insesgada

$$S_{N-1}^2 = \left(\frac{1}{N-1}\right) \left[\sum f(x_i) x_i^2 - \frac{(\sum f_i(x) (x_i))^2}{N} \right] = \left(\frac{1}{50-1}\right) \left[1063.13 - \left(\frac{(224.5)^2}{50}\right) \right] = 1.125$$

Por lo tanto

$$S_{N-1} = \sqrt{1.125} = 1.0606$$

10. Supóngase que cierto conjunto de observaciones tiene una $\bar{x} = 100$ y una $S^2 = 225$

Conteste las siguientes preguntas, de acuerdo al teorema de Tchebycheff.

- ¿Al menos qué porcentaje de todas las observaciones caerá entre 70 y 130?
- ¿A menos que porcentaje de las observaciones caerá entre 25 y 175?

SOLUCION

- De los datos se obtiene $\bar{x} = 100$ $S = 15$

En general el valor de k correspondiente a un valor X cualquiera se puede determinar a partir de la

ecuación
$$k = \frac{X - \bar{x}}{S}$$

Los valores de k correspondientes a 70 y a 130 son $k_1 = \frac{70-100}{15} = -2$ y $k_2 = \frac{130-100}{15} = 2$

Es un intervalo simétrico a partir de la media con k =2. De acuerdo al teorema de Tchebycheff

$$\text{Proporción al menos} = \left(1 - \frac{1}{k^2}\right)100 = \left(1 - \frac{1}{2^2}\right)100 = 75 \%$$

(b) Procediendo de manera similar al inciso anterior, los valores de k correspondientes a 25 y a 175 son

$$k_1 = \frac{25-100}{15} = -5 \text{ y } k_2 = \frac{175-100}{15} = 5$$

Es un intervalo simétrico a partir de la media con k =5. De acuerdo al teorema de Tchebycheff

$$\text{Proporción al menos} = \left(1 - \frac{1}{k^2}\right)100 = \left(1 - \frac{1}{5^2}\right)100 = 96 \%$$

11. De acuerdo con la regla normal ¿Cuál es la proporción aproximada de un conjunto de observaciones que caerá por debajo de $\bar{x} - 2S$

SOLUCION

De acuerdo a la regla de la Normal dentro del intervalo $[\bar{x} - 2S, \bar{x} + 2S]$ hay aproximadamente el 95 % de los datos, quedando fuera el 5 %, pero como solo se consideran los que están por debajo de $\bar{x} - 2S$ esto corresponde a la mitad, o sea al 2.5% ó equivalentemente a 0.0250 de los datos.

12. Una muestra de 100 trabajadores tiene una producción promedio por hora de 60 unidades y una desviación típica de 10 unidades. De acuerdo con la regla de la normal, ¿aproximadamente cuántos trabajadores tienen una producción entre 40 y 80 unidades?

SOLUCION

El número de desviaciones estándar a partir de la media se puede determinar con $k = \frac{X - \bar{x}}{S}$

Del problema $\bar{x} = 60$ y $S = 10$ entonces, para los valores de 40 y 80 se tiene que

$$k_1 = \frac{40-60}{10} = -2 \text{ y } k_2 = \frac{80-60}{10} = 2$$

Lo cual corresponde a dos desviaciones a la izquierda y a la derecha del promedio, que de acuerdo a la regla de la normal corresponde al 95 % de los datos ó al 0.95 del total de datos, por lo tanto

Número de trabajadores = Total x Fracción

$$N = 100 \times 0.95 = 95$$

Unidad II Probabilidad

CONJUNTOS Y ÁLGEBRA DE CONJUNTOS

DEFINICIÓN DE CONJUNTO.

Conceptos básicos de la teoría de conjuntos:

CONJUNTO: es una colección de objetos, datos, que pueden cumplir una o varias condiciones.

Notación de conjunto: comúnmente se representa a los conjuntos mediante letras mayúsculas A, B, C, U, Z, W, Φ , Ω

ELEMENTO: en un único objeto o dato que es parte de un conjunto

Notación de elemento: los elementos se denotan con letras minúsculas a, b, c, α , ϕ , v, w, θ

Los conjuntos pueden describirse de dos maneras, de forma explícita y/o implícita.

La forma explícita corresponde cuando los elementos del conjunto son mostrados directamente

EJEMPLO

$$A = \{a, e, i, o, u\}$$

$$B = \{1, 2, 3, 4, 5, 6, \dots\}$$

$$C = \{\dots -4, -2, 0, 2, 4, 6, \dots\}$$

La forma implícita corresponde cuando los elementos del conjunto no son mostrados directamente y son definidos mediante una condición o condiciones.

$$A = \{x. | x \text{ es una vocal del abecedario}\}$$

$$B = \{x. | x \text{ es un número natural}\}$$

$$C = \{x. | x \text{ es un número par}\}$$

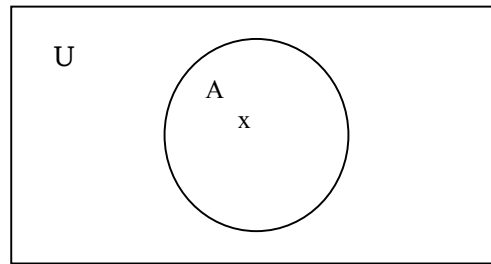
El CONJUNTO UNIVERSO denotado generalmente por U es el conjunto más grande que es utilizado en un problema particular y contiene a todos los elementos.

En el ámbito de la Estadística se relaciona directamente el conjunto universo con la población y el caso de la Probabilidad con el llamado espacio muestral.

Se dice que un elemento **x pertenece a un conjunto A** si x es parte del conjunto A.

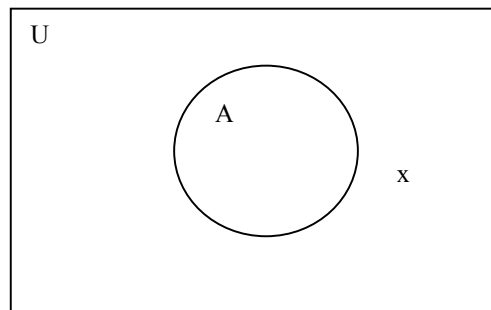
Notación: $x \in A$.

En forma gráfica la condición se representa mediante el diagrama siguiente



Si x **no pertenece** a un conjunto A , entonces x no es parte del conjunto A .

Notación: $x \notin A$.

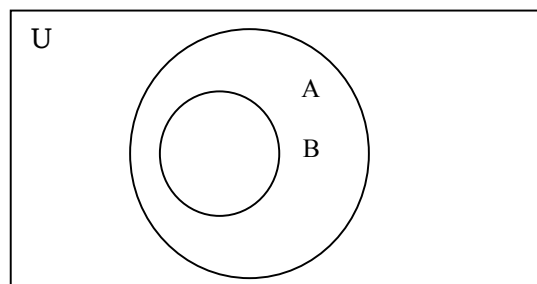


Un **conjunto es finito** si se pueden contar sus elementos, esto es, existe un número total de elementos.
 $\# A = n$

Si el $\# A = \infty$ entonces el **conjunto es infinito**.

Se dice que un conjunto B está **CONTENIDO** en un conjunto A ó es **SUBCONJUNTO** de A si y solo si todo elemento $x \in B$, x también $x \in A$.

Notación: $B \subset A$.



Para facilitar la escritura de algunas expresiones matemáticas a continuación se presentan algunos símbolos y su significado

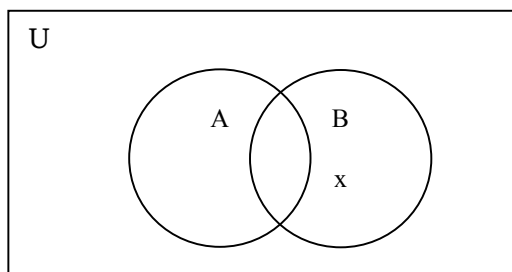
\forall	Para todo.
\leftrightarrow	Si y solo si.
\rightarrow	Entonces.
\exists	Existe.
\therefore	Por lo tanto.

La definición de CONTENIDO o CONTENCION anterior se puede escribir como:

$$B \subset A \leftrightarrow \forall x \in B, x \in A$$

Si algún $x \in B$ pero $x \notin A$ entonces se dirá que B NO ESTA CONTENIDO A ó que B no es SUBCONJUNTO de A. En forma compacta: $\exists x \in B \quad x \notin A \rightarrow B \not\subset A$.

Notación: $B \not\subset A$.



ÁLGEBRA DE CONJUNTOS (OPERACIONES BÁSICAS)

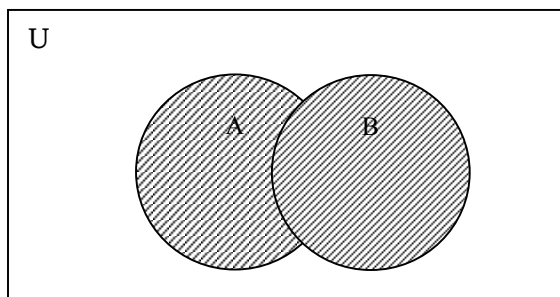
Las operaciones entre conjuntos permiten obtener nuevos conjuntos a partir de conjuntos más simples ó representar conjuntos complejos mediante conjuntos más simples.

Todas las operaciones que se define a continuación son de gran importancia para el desarrollo de la probabilidad, por lo que se recomienda aprenderlas y aplicarlas correctamente cada una de ellas. Cabe mencionar que estas operaciones no se deben comparar con las operaciones algebraicas entre números como son la suma, resta y multiplicación-

UNIÓN DE CONJUNTOS

$$A \cup B = \{x \mid x \in A \text{ ó } x \in B\}$$

Notación: $A \cup B$



EJEMPLO

$$A = \{a, b, c, d\}$$

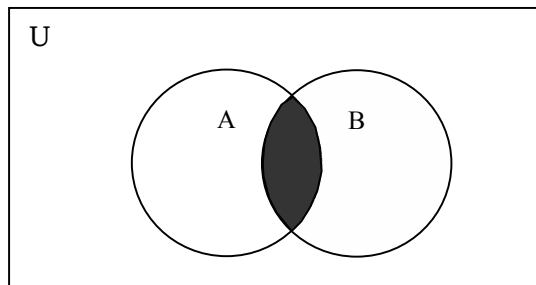
$$B = \{a, b, c, d, f, g, h\}$$

$$C = A \cup B = \{a, b, c, d, f, g, h\}$$

INTERSECCIÓN DE CONJUNTOS

$$A \cap B = \{x \mid x \in A \text{ y } x \in B\}$$

Notación: $A \cap B$

**EJEMPLO**

$$A = \{a, b, c, d\}$$

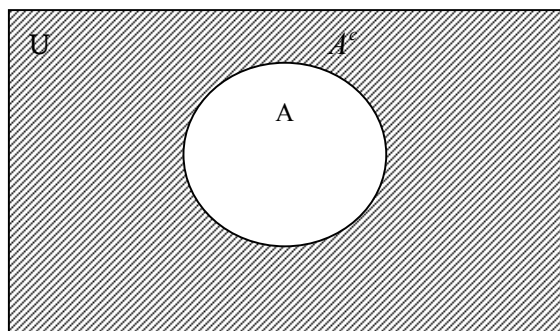
$$B = \{a, b, c, d, f, g, h\}$$

$$A \cap B = \{c, d\}$$

COMPLEMENTO

$$A^c = \{x \mid x \notin A \text{ y } x \in U\}$$

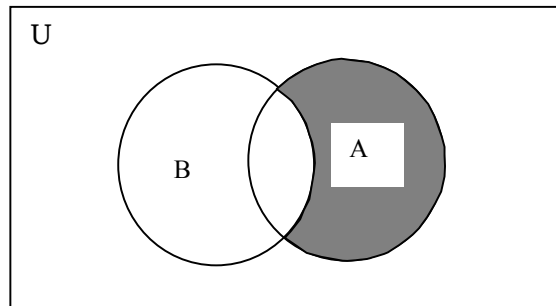
Notación: A^c



Complemento relativo:

$$B / A = \{x \mid x \notin B \text{ y } x \in A\}$$

Notación: A^c

**EJEMPLO**

Utilizando los conjuntos anteriores

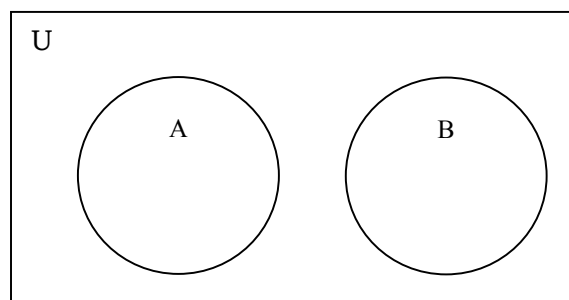
$$B / A = \{ \} = \Phi$$

$$A / B = \{g, f, h\}$$

Siendo $\Phi = \{ \}$ conjunto vacío

A partir de las operaciones anteriores entre conjuntos se pueden definir y obtener nuevas propiedades entre conjuntos, las cuales serán utilizadas en secciones posteriores y en particular en el tema de probabilidad.

Se dice que dos conjuntos A y B son AJENOS si solo si $A \cap B = \Phi$,



PROPIEDADES BÁSICAS DE LOS CONJUNTOS

Sean A, B dos conjuntos generales dentro de un conjunto universo U entonces se cumplen las siguientes condiciones

- a) $A \cup A = A$
- b) $A \cap A = A$
- c) $A \cup A^c = U$
- d) $A \cap A^c = \Phi$
- e) $U^c = \Phi$
- f) $\Phi^c = U$
- g) $A \cup \Phi = A$
- h) $A \cap \Phi = \Phi$
- i) $A = (A \cap B) \cup (A \cap B^c)$

Si $B \subset A$, entonces:

- j) $A \cup B = A$
- k) $A \cap B = B$

Leyes conmutativas

- l) $A \cup B = B \cup A$
- m) $A \cap B = B \cap A$

Leyes distributivas

- n) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- o) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Leyes de Morgan

- p) $(A \cup B)^c = A^c \cap B^c$
- q) $(A \cap B)^c = A^c \cup B^c$

EXPERIMENTOS PROBABILÍSTICOS Y DETERMINÍSTICOS

Como ya se ha mencionado en la unidad anterior:

Un EXPERIMENTO ES DETERMINÍSTICO si al realizarse bajo las mismas condiciones se obtiene invariablemente en mismo resultado o dato, en el caso de que se obtenga resultados o datos diferentes se dirá que el es un EXPERIMENTO PROBABILISTICO ó ALEATORIO.

POBLACIÓN MUESTRA, EVENTOS

A continuación se dan nuevamente las definiciones de población, muestra y eventos.

La POBLACION es el conjunto total de datos que se obtienen al realizar un experimento.

La MUESTRA es una parte ó subconjunto de la población.

Los EVENTOS están formados generalmente por muestras a las cuales se les pide que cumplan con alguna condición o condiciones.

Teoría elemental del muestreo

La toma de datos ó muestras de un experimento aleatorio en general se debe realizar de tal manera que todos los posibles resultados del experimento tenga la misma oportunidad ó probabilidad de se elegidos, lo anterior constituye el PRINCIPIO FUNDAMENTAL DEL MUESTREO.

El principio anterior es conocido también como MUESTREO AL AZAR y tiene la finalidad de obtener una muestra lo más representativa del experimento.

El muestreo al azar se puede realizar de dos maneras CON REEMPLAZO y SIN REEMPLAZO.

En el caso de **reemplazo** una vez elegido un objeto este es regresado de nuevo al conjunto y por lo tanto puede ser nuevamente seleccionado, por otra parte si el muestreo se lleva a cabo **sin reemplazo** el objeto que es seleccionado no se regresa al conjunto y por lo tanto nunca más podrá se seleccionado. En aplicaciones prácticas aparecen ambos tipos de muestreo.

Para efectuar un muestreo adecuado se debe evitar posibles tendencias al realizar un experimento, por ejemplo, para la elección de muestras de un lote se puede recurrir a tablas ó programas que generan números aleatorios para evitar tendencias y realizar una correcta selección de las muestras

El muestreo de datos se puede realizar al azar con o sin reemplazo

El estudio de la Probabilidad permite dar una respuesta a problema de la elección adecuada de cuando una muestra es representativa de un experimento aleatorio o población.

ESPACIO MUESTRAL

El ESPACIO MUESTRAL es el conjunto de todos los resultados posibles de un evento aleatorio ó probabilístico.

Normalmente el espacio muestral se representa por la letra S y en términos de conjuntos es el equivalente al conjunto universo.

Un EVENTO O SUCESO: es un subconjunto del espacio muestral.

DEFINICIÓN DE PROBABILIDAD

La PROBABILIDAD DE UN EVENTO se puede definir en el caso de conjuntos finitos como:

$$P(E) = \frac{N.(E)}{N.(S)}$$

$N(E)$:= número de elementos independientes de E.

$N(S)$ = número total de elementos independientes.

En algunos casos sencillos es posible conocer fácilmente el número total de elementos que conforman cada uno de los conjuntos, sin embargo, esto no es posible para la mayoría de los demás caso, por lo que es conveniente recurrir en principio a las técnicas de conteo para determinar las probabilidad.

TÉCNICAS DE CONTEO

PRINCIPIO FUNDAMENTAL DEL CONTEO.

Si un evento n_1 se puede realizar de N_1 formas y otro evento se puede realizar de N_2 formas, entonces el evento conjunto se puede realizar de $N_1 \cdot N_2$ formas.

$$N = N_1 \cdot N_2 \quad (2.1)$$

El principio fundamental del conteo se puede representar gráficamente mediante el llamado **diagrama de árbol**. Cada trayectoria en el diagrama de árbol representa un posible resultado o forma de realizarse el experimento.

En la figura 1 se muestra el diagrama de árbol para el caso de $N_1=4$ y $N_2 = 2$, con lo que se obtienen $N_1 \cdot N_2 = 4 \cdot 2 = 8$ trayectorias ó formas

Por otra parte el principio fundamental del conteo se puede generalizar a k eventos, esto es, si el evento i puede ocurrir de N_i formas, entonces el evento total conjunto de los k eventos, se puede realizar de $N_1 \cdot N_2 \dots N_i \dots N_k$ formas.

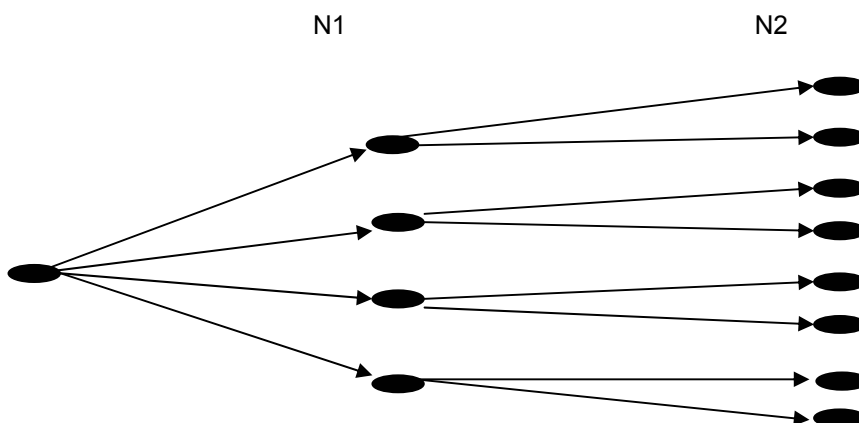


Figura 1. Diagrama de árbol que representa el principio fundamental del conteo $N_1 \cdot N_2 = 4 \cdot 2 = 8$

EJEMPLOS

1. Determine el número total de combinaciones de un candado formado por 3 discos giratorios y cada uno de los cuales puede ser colocado en los números 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Combinación de un candado.

SOLUCION

De acuerdo a lo indicado en el problema cada uno de los discos puede ser colocado en 10 formas, esto es $N_1=10$; $N_2=10$, y $N_3=10$. Aplicando el principio fundamental del conteo se obtiene:

10	10	10
----	----	----

$$= 10^3 = 1000 \text{ combinaciones}$$

2. Una moneda es arrojada 2 veces consecutivas. Obtenga el espacio muestral del conjunto.

SOLUCION

Una moneda tiene dos resultados posibles, Águila (A) ó Sol (S), si la moneda es arrojada dos veces entonces

$$N = N_1 \cdot N_2 = 2 \cdot 2 = 4 \text{ eventos independientes}$$

Cada uno de los eventos individuales se muestran a continuación:

$$S = \{ (A,A), (A,S), (S,A), (S,S) \}$$

3. Un experimento consiste en arrojar una moneda 4 veces, lístense todas las posibilidades:

SOLUCION

El número total de posibles eventos independientes es $N = (2, 2, 2, 2) = 2^4 = 16$

Puede utilizarse un diagrama de árbol para listar correctamente todas las posibilidades, estas son:

A, A, A, A	S, A, A, A
A, A, A, S	S, A, A, S
A, A, S, A	S, A, S, A
A, A, S, S	S, A, S, S
A, S, A, A	S, S, A, A
A, S, A, S	S, S, A, S
A, S, S, A	S, S, S, A
A, S, S, S	S, S, S, S

4. Obtenga el número total de eventos independientes que se obtiene al arrojar una moneda 5 veces consecutivas.

SOLUCION

En cada uno de los 5 casos de arrojar una moneda está puede tener solamente dos resultados posibles, Águila (A). ó Sol (S), entonces:

2	2	2	2	2
---	---	---	---	---

 $= 2^5 = 32 \text{ posibles}$

5. Obtenga el espacio muestral que se genera al arrojar un dado 2 veces

SOLUCION

El dado tiene 6 caras y por lo tanto existen 6 posibilidades para cada vez que es arrojado, entonces como es arrojado 2 veces:

6	6
---	---

 $= 6^2 = 36 \text{ resultados}$

Los eventos independientes pueden obtenerse fácilmente mediante un diagrama de árbol.

$S = \{ (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6) \}$

6. Determine el número posible de combinación de placas válidas si la placa esta formada por 3 números consecutivos y 3 letras del abecedario.

SOLUCION

Existen 10 posibilidades para cada uno de los números y 26 posibilidades para cada una de las letras (no se incluyen letras dobles RR, CH, LL y la letra Ñ), entonces:

METODO I

Números

10	10	10
----	----	----

Letras

26	26	26
----	----	----

Placas

$= (10^3) (26^3)$

En el cálculo anterior se han incluido placas que no existen para fines prácticos, por ejemplo:

La placa

0	0	0	A	A	A
---	---	---	---	---	---

No existe

En general las placas no pueden tener un cero o ceros antes que un número diferente de cero.

Por otra parte no existen las placas

0	num	num	letra	letra	letra	No existen
1	9	10	26	26	26	$= (90)(26^3)$

0	0	num	letra	letra	letra	No existen
1	1	9	26	26	26	$= (9)(26^3)$

0	0	0	letra	letra	letra	No existen
1	1	1	26	26	26	$= 26^3$

Número de placas no validas $= (90)(26^3) + (9)(26^3) + 26^3 = (10^2)(26^3) = (100) (26^3)$

Entonces

Número de placas validas $= \text{Número total} - \text{Número de placas no validas.}$
 $= (10^3) (26^3) - (100) (26^3) = (900) (26^3) = 15\,818\,400 \text{ placas.}$

METODO II

La primer casilla de número no puede ser cero, por lo tanto se reduce sus posibles valores a $N1=9$ Manteniéndose los demás valores iguales al método I

Números			Letras			Placas
9	10	10	26	26	26	$= (900) (26^3)$

Número de placas no validas $= (900) (26^3) = 15\,818\,400 \text{ placas.}$

El principio fundamental del conteo permite obtener fórmulas matemáticas para algunos casos generales que ocurren comúnmente en aplicaciones prácticas, como son, las permutaciones y las combinaciones

PERMUTACIONES

La permutación aparece cuando se tienen N objetos DISTINGUIBLES SIN REEMPLAZO y estos pueden ocupar r lugares o posiciones. Lo anterior se representa gráficamente como

Lugar 1	Lugar 2	Lugar 1	Lugar 1	...	Lugar r
---------	---------	---------	---------	-----	-----------

Aplicando el principio fundamental del conteo y recordando que en el primer lugar puede ser ocupado por los n objetos, el segundo lugar por los $n-1$ restantes y así sucesivamente hasta el lugar r donde solamente puede ser ocupado por $n-r+1$ objetos

n	$n-1$	$n-2$	$n-3$...	$n-r+1$
-----	-------	-------	-------	-----	---------

$$\text{Permutaciones} = n(n-1)(n-2)(n-3)\dots(n-r+1)$$

Existe un caso particular en el cual el número de objetos n es igual al número de posiciones que pueden ocupar, esto es, $r = n$. por lo tanto el producto anterior se convierte en el producto de los enteros consecutivos del 1 al n .

n	$n-1$	$n-2$	$n-3$...	1
-----	-------	-------	-------	-----	---

$$\text{Permutaciones} = n(n-1)(n-2)(n-3)\dots 1$$

Este producto particular es conocido como el FACTORIAL

$$n! = n(n-1)(n-2)(n-3)\dots 1 \quad (2.2)$$

Propiedades elementales del factorial

$$(a) \ n! (n+1) = (n+1)!$$

$$(b) \ 0! = 1$$

Las permutaciones para n objetos ocupando r lugares ó casillas pueden definirse en términos del factorial y sus propiedades anteriores como;

$${}_n P_r = \frac{n!}{(n-r)!} \quad (2.3)$$

EJEMPLOS

7. Mostrar que la definición de las permutaciones en términos de factoriales es correcta

SOLUCION

Partiendo de la definición dada

$${}_n P_r = \frac{n!}{(n-r)!} = \frac{n(n-1)(n-2)\dots(n-r+1)(n-r)\dots 3 \cdot 2 \cdot 1}{(n-r)(n-r-1)\dots 3 \cdot 2 \cdot 1}$$

Simplificando términos

$${}_n P_r = \frac{n!}{(n-r)!} = n(n-1)(n-2)\dots(n-r+1)$$

para el caso particular de $n = r$

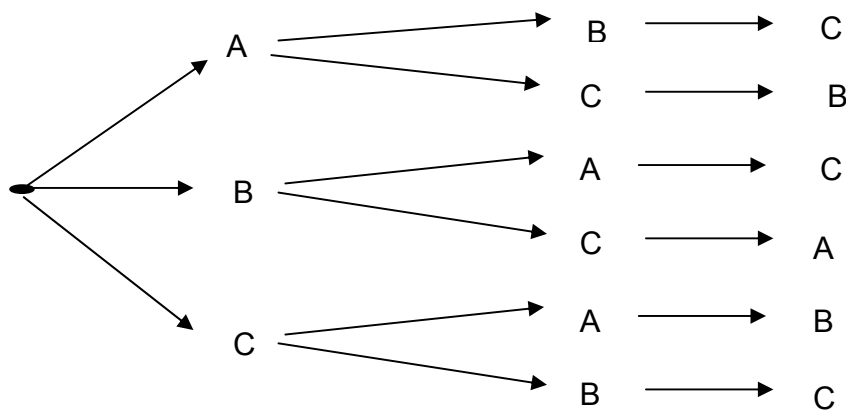
$${}_nP_n = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n!$$

8. Determinar cuantas formas hay de acomodar las letra A,B,C sin reemplazo en tres lugares consecutivos. Muestre explícitamente cuales son estas posibilidades.

Para el problema $n=3$ y $r=3$,

$${}_3P_3 = 3! = 1 \cdot 2 \cdot 3 = 6$$

Explícitamente las permutaciones se pueden obtener a partir del diagrama de árbol siguiente



(A,B,C), (A,C,B), (B,A,C), (B,C,A), (C,A,B) y (C,B,A)

9. Utilizando el problema anterior determine ¿en cuántos casos las letra A y B permanecen juntas en todo momento?

SOLUCION

MÉTODO I

Directamente del problema anterior se pueden observar directamente que los casos que cumplen que A y B estén siempre juntas son:

(A,B,C), (B,A,C), (C,A,B) y (C,B,A), esto es, solo hay 4 casos

MÉTODO II (formación de bloques)

Si las letras A y B deben permanecer juntas, entonces ambas forman un bloque, con lo cual el bloque en conjunto se puede considerar como un "elemento", en términos de permutaciones $n=2$ $r=2$

Bloque		letra
A	B	C

2	1	=2!
---	---	-----

Pero en el bloque formado por las letras A, B estas puede permutarse y mantenerse juntas entre si, por los que hay que tomar en cuenta está posibilidad donde también $n = 2$ $r = 2$

B	A	C
---	---	---

2	1
---	---

 $= 2!$

Sumando las posibilidades anteriores se tiene $TOTAL = 2! + 2! = 2 + 2 = 4$ permutaciones

En términos de notación de permutaciones: $TOTAL = 2P2 * 2P2 = 2! + 2! = 2 + 2 = 4$ permutaciones

10. ¿De cuántas formas se pueden acomodar 10 libros distintos en un estante

SOLUCION

Aplicando el principio fundamental del conteo

10	9	8	7	6	5	4	3	2	1
----	---	---	---	---	---	---	---	---	---

 $= 10! = 3\,628\,800$

Mediante permutaciones $n = 10$ y $r = 10$, entonces

$$10P10 = 10! = 3\,628\,800 \text{ Formas}$$

11. Se tienen 8 libros 3 de matemáticas, 3 de física y 2 de biología.

¿De cuántas maneras se pueden acomodar de tal manera que los libros de cada materia queden siempre juntos?

SOLUCION

Los tipos de libros para mantenerse juntos forman bloques de cada tipo, por lo que hay tres bloques, los cuales se pueden acomodar de las siguientes $N1 = 3P3 = 3!$

3	2	1
---	---	---

 $= 3!$

Bloque 1 bloque 2 bloque 3

Supóngase ahora que se tiene por ejemplo el siguiente acomodo particular de los bloques

3	2	1	3	2	1	2	1
---	---	---	---	---	---	---	---

 $= 3! \cdot 3! \cdot 2!$

Matemáticas

Física

Biología

Dentro de cada bloque se pueden permutar los libros de cada sección y tal como se observa se tendrían $N2 = (3P3)(3P3)(2P2) = 3! \cdot 3! \cdot 2!$ Permutaciones

Aplicando el principio fundamental de conteo en número total es

$$N2 = 3P3 \cdot 3P3 \cdot 2P2 = 3!$$

$$N = N1 \cdot N2 = 3! \cdot (3! \cdot 3! \cdot 2!) = 432$$

12. Diez personas se encuentran esperando ser atendidas en una oficina de gobierno, pero la secretaria les informa que solo se atenderán a seis personas, ¿cuál es la cantidad de posibles opciones para atender a las personas?

SOLUCION

Para este problema se tienen $n = 10$ personas y solo se cuenta con $r = 6$ lugares, entonces

$$N = nPr = 10P6 = \frac{10!}{(10-6)!} = \frac{10!}{4!} = 151\,200 \text{ opciones}$$

COMBINACIONES

Para entender las como se obtienen las combinaciones primero hay que observar lo que sucede cuando los objetos que son considerados distinguibles se transforman en indistinguibles.

Como ejemplo considere las permutaciones de las letras A, B, C y posteriormente hagamos que $A = B$

A, B, C diferentes	A = B, C diferente	reducción
A, B, C	A, A, C	$\left. \begin{array}{l} A, A, C \\ A, C, A \\ A, C, A \\ A, A, C \\ C, A, A \\ C, A, A \end{array} \right\} \begin{array}{l} A, A, C \\ A, C, A \\ C, A, A \end{array}$
A, C, B	A, C, A	
B, C, A	A, C, A	
B, A, C	A, A, C	
C, A, B	C, A, A	
C, B, A	C, A, A	

Las permutaciones se reducen a 3 casos únicamente.

Si ahora se las tres letras son indistinguibles entre si ó equivalentemente $A=B=C$

A, B, C diferentes	A = B = C	reducción
A, B, C	A, A, A	$\left. \begin{array}{l} A, A, A \\ A, A, A \\ A, A, A \\ A, A, A \\ A, A, A \\ a, A, A \end{array} \right\} \begin{array}{l} , \\ A, A, A \\ , , \end{array}$
A, C, B	A, A, A	
B, C, A	A, A, A	
B, A, C	A, A, A	
C, A, B	A, A, A	
C, B, A	a, A, A	

Las permutaciones se reducen a 1 caso únicamente.

Utilizando los ejemplos anteriores es posible deducir una fórmula simple. Si se tienen n objetos que pueden ocupar r lugares y entre ellos hay l_1 objetos indistinguibles, l_2 objetos indistinguibles,..., l_k objetos indistinguibles, que cumplen $l_1 + l_2 + \dots + l_k = n$, entonces en numero total de permutaciones se reduce a:

$$N = \frac{nPr}{l_1!l_2!\dots l_k!} \quad (2.4)$$

Para el primer caso $n = r$, $l_1 = 2$

$$N = \frac{3!}{2!} = \frac{1.2.3}{1.2} = 3$$

Para el segundo caso $n = r$, $l_1 = 3$

$$N = \frac{3!}{3!} = 1$$

EJEMPLO

13. Se tienen 8 libros, 3 de matemáticas, 3 de física y 2 de biología. Si los 3 libros de matemáticas son iguales y los 2 de biología son iguales ¿Cuántas formas posibles existen de acomodarlos en un librero?

SOLUCION

De acuerdo a los datos del problema, $n=8$ libros, $l_1 = 3$ libros de matemáticas iguales, $l_2 = 2$ libros de biología iguales, entonces

$$N = \frac{8!}{3!2!} = \frac{1.2.3.4.5.6.7.8}{1.2.3.1.2} = 3360$$

Las COMBINACIONES de n objetos en r lugares se obtiene cuando en una permutación de estos objetos la posición relativa no importa a pesar de ser diferentes entre ellos, por ejemplo todas las permutaciones (A,B,C), (A,C,B), (B,A,C), (B,C,A), (C,A,B) y (C,B,A) son equivalentes a (A,B,C), en este caso se puede considerar que existe un conjunto con $l = r$ objetos iguales por lo tanto utilizando la fórmula (2.4)

$$nCr = \frac{nPr}{r!} = \frac{n!}{(n-r)!r!} \quad (2.5)$$

Las combinaciones pueden escribirse también como

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

EJEMPLOS

14. Un contratista de construcción ofrece casas con cinco distintos tipos de distribución, tres tipos de techo y dos tipos de alfombrado. ¿De cuántas formas diferentes puede un comprador elegir una casa?

SOLUCION

Hay $N_1 = 5$ distribuciones $N_2 = 3$ tipos de techos y $N_3 = 2$ tipos de alfombra, entonces, aplicando el principio fundamental del conteo

N=	N1	N2	N3	=	5	3	2
----	----	----	----	---	---	---	---

$$= 30 \text{ elecciones de casa diferentes}$$

15. Se tiran seis dados. ¿De cuántas formas diferentes pueden quedar las caras hacia arriba?

SOLUCION

Hay 6 posibles resultados de cara para cada uno de los 6 dados, entonces, aplicando el principio fundamental del conteo

N=	6	6	6	6	6	6
----	---	---	---	---	---	---

 $= 6^6 = 46656$ formas diferentes

16. Las placas de matrícula de automóviles emitidas por cierto estado tienen dos letras seguidas por tres dígitos. ¿Cuántas placas diferentes pueden emitirse si no hay restricciones?

SOLUCION

Para las letras hay 26 posibles resultados y para los números hay 10 posibles valores, por lo tanto mediante el principio fundamental del conteo

	Letra	letra	Num	Num	Num
N=	26	26	10	10	10

 $= 26^2 \cdot 10^3 = 676000$

17. Una clase consiste en diez estudiantes. ¿De cuántas formas puede seleccionarse un comité de tres estudiantes

SOLUCION

Este problema corresponde a un caso clásico de combinaciones donde $n = 10$ estudiantes, $r = 3$ estudiantes, entonces

$$N = \frac{10!}{(10-3)!3!} = 120 \text{ comités.}$$

18. Un club consta de 30 miembros. 15 blancos, 10 negros y 5 de otras razas. Debe formarse un comité de 6 miembros. Si los 3 grupos deben estar representados, con proporciones iguales, ¿de cuántas formas puede hacerse esto?

SOLUCION

Los 30 miembros son divididos en 3 clases: 15 blancos, 10 negros, 5 de otros

Como las proporciones deben de ser iguales y el comité está formado por 6 miembros a cada clase le corresponden 2 miembros para el comité

Se pueden elegir

$$\binom{15}{2} = \frac{15!}{(15-2)!2!} = 105 \text{ comités de blancos}$$

$$\binom{10}{2} = \frac{10!}{(10-2)!2!} = 45 \text{ comités de blancos}$$

$$\binom{5}{2} = \frac{5!}{(5-2)!2!} = 10 \text{ comités de otros}$$

Un posible caso de de comité es

2 blancos 2 negros 2 de otros

N=	105	45	10
----	-----	----	----

 $= 47\,250$ comités

19. En una clase de 30 estudiantes, hay 20 hombres y 10 mujeres.

a. ¿De cuántas formas puede seleccionarse un comité de tres hombres y dos mujeres?

b. ¿De cuántas formas puede seleccionarse un comité de cinco estudiantes?

c. ¿De cuántas formas puede seleccionarse un comité de cinco estudiantes si los cinco deben de ser del mismo sexo?

SOLUCION

a. Procediendo como en el problema anterior

	3 hombres de 20	2 mujeres de 10	
N=	$\binom{20}{3}$	$\binom{10}{2}$	$= (1140)(45) = 51\,300$ comités

b. Hay $n = 30$ estudiantes para ocupar $r = 5$ lugares

$$\binom{n}{r} = \binom{30}{5} = \frac{30!}{(30-5)!5!} = 142\,506 \text{ comités.}$$

c. Puede haber un comité formado por 5 hombres ó un comité formado por 5 mujeres, entonces el resultado es la suma de cada uno de los casos

	5 hombres de 20		5 mujeres de 10	
N=	$\binom{20}{5}$	+	$\binom{10}{5}$	$= 15\,504 + 45 = 15\,549$ comités

20. Una "mano de póker" consiste en 5 naipes sacados de una baraja ordinaria 52 naipes. ¿Cuántas manos diferentes pueden formarse a partir de la baraja completa?

SOLUCION

Se tiene $n = 52$ naipes para seleccionar una combinación $r = 5$, entonces

$$nC_r = \frac{52!}{(52-5)!5!} = 2\,598.960 \text{ manos}$$

La probabilidad de un evento se definió en párrafos anteriores como:

$$P(E) = \frac{N.(E)}{N.(S)}$$

$N.(E)$:= número de elementos independientes de E.

$N.(S)$ = número total de elementos independientes.

Es de mencionar que la definición anterior está dada particularmente para conjuntos finitos y existen otras definiciones para conjuntos infinitos, por ejemplo par el caso de conjuntos representados mediante áreas, la probabilidad se puede definir como el cociente de el área que representa al evento E entre el área total que representa al espacio muestral.

La probabilidad se puede interpretar como la medida de la ocurrencia de un evento que es parte de un evento E que es parte de un espacio muestral ó experimento aleatorio.

EJEMPLOS

21. En una votación preliminar simulada para determinar la probabilidad de cierto candidato para la presidencia de los E.U.A., se encontró que 495 de 1000 votantes seleccionados aleatóriamente están a favor de dicho candidato. ¿Cuál es la probabilidad de que cualquiera de los votantes favorezca a este candidato?

SOLUCION

$N(S) = 1000$ y $N(E) = 495$ entonces aplicando la definición directa de la probabilidad

$$P = \frac{495}{1000} = 0.495$$

22. Supóngase que estadísticas recopiladas por la oficina meteorológica de Los Ángeles muestran que ha llovido durante el desfile de las Rosas en Pasadena 14 veces durante los últimos 80 años.

- ¿Cuál es la probabilidad de que llueva durante el desfile de las Rosas el próximo día de año nuevo?
- ¿Cuál es la probabilidad de que no llueva?

SOLUCION

Si $E = \{x \mid x \text{ es un año lluvioso el día del desfile de las Rosas}\}$, entonces
 $E^c = \{x \mid x \text{ es un año no lluvioso el día del desfile de las Rosas}\}$,

Como $N(E) = 14$, entonces $N(E^c) = 80 - 14 = 66$

$$\text{a) } P(E) = \frac{N(E)}{N(S)} = \frac{14}{80} = \frac{7}{40}$$

$$\text{b) } P(E^c) = \frac{N(E^c)}{N(S)} = \frac{66}{80} = \frac{33}{40}$$

23. Un club tiene 30 miembros: 25 hombres y 5 mujeres. Va a constituirse un comité de 5 miembros. ¿Cuál es la probabilidad de que las 5 mujeres se incluyan en el comité, si los miembros de éste se seleccionan aleatóriamente?

SOLUCION

El número total de comités con $r = 5$ miembros que se pueden formar con $n = 30$ miembros es

$$N(S) = {}^{30}C_5 = 142\,506$$

El número de comités con $r = 5$ mujeres que se pueden formar con $n = 5$ mujeres es

$$N(E) = {}^5C_5 = 1$$

Por lo tanto

$$P(E) = \frac{N(E)}{N(S)} = \frac{1}{142506}$$

24. Sea el espacio muestral $S = \{\text{arrojan una moneda legal 8 veces}\}$ y sea el evento $E = \{\text{Salen 5 águilas exactamente}\}$. Determine la probabilidad $P(E)$.

SOLUCION

El número de elementos que forman el espacio muestral es:

$$N(S) = \boxed{2} \boxed{2} \boxed{2} \boxed{2} \boxed{2} \boxed{2} \boxed{2} \boxed{2} = 2^8 = 256$$

Un esquema de un elemento del evento E es mostrado a continuación

A	A	A	A	A	S	S	S
---	---	---	---	---	---	---	---

Para determinar el número total de elementos que forman el evento E se puede aplicar la ecuación 4, en la cual se considera que $n = 8$, $r = 8$, $l_1 = 5$ y $l_2 = 3$.

$$N(E) = \frac{nPr}{l_1!l_2!} = \frac{8!}{5!3!} = 56$$

Entonces

$$P(E) = \frac{N(E)}{N(S)} = \frac{56}{256} = \frac{7}{32}$$

25. Una tienda de aparatos de sonido acaba de recibir un embarque de diez nuevos aparatos, siete de modelo X y tres de modelo Y. Si se venden aleatoriamente cuatro aparatos, ¿cuál es la probabilidad de que se vendan dos de cada modelo?

SOLUCION

Hay $n_x = 7$ aparatos tipo X, $n_y = 3$ aparatos tipo Y, se seleccionan $r = 4$ aparatos, $n = n_x + n_y = 10$.

Sea E el evento de que se vendan dos de cada modelo ó equivalentemente dos aparatos del modelo X y dos aparatos del modelo Y, el evento puede representarse como: [X, X, Y, Y]

Se deben de elegir $r_x = 2$ aparatos tipo x de 7 existentes y $r_y = 2$ aparatos tipo Y de 3 existentes, entonces,

$$N(E) = \binom{n_x}{r_x} \binom{n_y}{r_y} = \binom{7}{2} \binom{3}{2} = \frac{7!}{(7-2)!2!} \frac{3!}{(3-2)!2!} = (21)(3) = 63$$

y

$$N(S) = \binom{n}{r} = \binom{10}{4} = \frac{10!}{(10-4)!4!} = 210$$

por lo tanto

$$P(E) = \frac{N(E)}{N(S)} = \frac{63}{210} = \frac{3}{10}$$

26. Debe seleccionarse un comité de tres personas del consejo directivo de una compañía. El consejo consta de quince miembros, un tercio de los cuales son mujeres y dos tercios hombres. ¿Cuál es la probabilidad de que las tres personas del comité sean todas del mismo sexo?

SOLUCION

De acuerdo a los datos $n = 15$ personas, $n_H = 10$ hombres y $n_M = 5$ mujeres, se debe seleccionar un comité $r = 3$ personas

Sean los conjuntos $A = \{\text{comité de 3 mujeres}\}$ y $B = \{\text{comité de 3 hombres}\}$ entonces

$C = \{\text{en comité de personas del mismo sexo}\} = \{\text{las tres personas sean mujeres o sean hombres}\}$

$$C = A \cup B$$

Puesto que $A \cap B = \emptyset$ se tiene que $N(C) = N(A) + N(B)$

$$N(C) = \binom{n_H}{r} + \binom{n_M}{r} = \frac{10!}{(10-3)! 3!} + \frac{5!}{(5-3)! 3!} = 120 + 10 = 130 \text{ comités}$$

y

$$N(S) = \binom{n}{r} = \binom{15}{3} = \frac{15!}{(15-3)! 3!} = 455 \text{ comités}$$

finalmente

$$P(E) = \frac{N(E)}{N(S)} = \frac{130}{455} = \frac{2}{7}$$

27. Una "mano de póker" consta de cinco naipes. ¿Cuál es la probabilidad de que los cinco naipes sean del mismo palo?

SOLUCION

En un problema previo se sabe que $n = 52$ cartas, $r = 5$ cartas y

$$N(S) = \binom{n}{r} = \binom{52}{5} = \frac{52!}{(52-5)! 5!} = 2\,598\,960 \text{ manos}$$

El mazo de cartas es esta formado por 4 figuras: diamantes♦, corazones♥, picas♣ y tréboles♠ por lo que cada tipo de figuras está conformado por $n_p = 13$ cartas.

Sea el conjunto $B = \{5 \text{ cartas del mismo palo}\}$ y $A_i = \{5 \text{ cartas del mismo palo tipo } i\}$, para $i = 1, 2, 3$ y 4 .

Entonces resulta que $B = A_1 \cup A_2 \cup A_3 \cup A_4$, y además $A_1 \cap A_2 \cap A_3 \cap A_4 = \Phi$, por lo tanto se cumple que

$$N(B) = N(A_1) + N(A_2) + N(A_3) + N(A_4)$$

Utilizando los datos se puede determinar el número de elementos para cada uno de los conjuntos A_i , $i = 1, 2, 3$ y 4 como las combinaciones de $n_p = 13$ cartas tomadas de $r = 5$ cartas.

$$N(A_i) = \binom{n_p}{r} = \binom{13}{5} = \frac{13!}{(13-5)! 5!} = 1287$$

por lo tanto

$$N(B) = 4 \binom{13}{5} = 4(1287) = 5148$$

$$P(B) = \frac{5148}{2598960} = \frac{33}{16660}$$

28. Se están formando grupos de cuatro letras empleando las letras A E I O U X Y.

a. ¿Cuántos grupos pueden formarse si no deben repetirse las letras?

b. ¿Cuántos grupos pueden formarse si cualquier letra puede repetirse tan veces como se desee?

A E I O U X Y

SOLUCION

a) Este caso corresponde a una permutación puesto que todas las letras son diferentes con $n = 7$, $r = 4$,

$$N = {}_7P_4 = \frac{7!}{(7-4)! 4!} = 840$$

b) El caso corresponde a un caso de elección con reemplazo donde en cada elección se puede seleccionar cualquiera de las 7 letras para ocupar los 4 lugares, entonces

$$N = (7)(7)(7)(7) = 7^4 = 2401$$

29. Un vendedor de automóviles acaba de recibir un embarque de ocho automóvil nuevos, cinco de los cuales son compactos y tres modelos de lujo. Si se venden aleatoriamente cuatro automóviles, obténgase la probabilidad de que se hayan vendido dos de cada modelo

SOLUCION

$n = 8$ automóviles 5 compactos, 3 de lujo, se venden $r = 4$

$S = \{\text{vender 4 modelos de 8 disponibles}\}$

$E = \{2 \text{ de cada modelo}\} = \{2 \text{ modelos compactos y 2 modelos de lujo}\}$

$$N(S) = \binom{8}{4} = \frac{8!}{4!4!} = 70 \quad \text{Total de posibles ventas}$$

$$N(E) = \binom{5}{2} \binom{3}{2} = \frac{5!}{3!2!} \cdot \frac{3!}{1!2!}$$

$$P(A) = \frac{N(E)}{N(S)} = \frac{30}{70} = \frac{3}{7} = 0.128$$

30. Si en una estación televisora se debe seleccionar cuatro de entre diez programas de media hora para emitirlos cada mañana de 8:30 a 10:30, ¿de cuántas formas posibles puede arreglarse la programación?

SOLUCION

De 8:30 a 10:30 solo se pueden acomodar $r = 4$ programas de media hora, de $n = 10$ disponibles, como en la programación hay orden, entonces el número de formas posibles de acomodar la programación es:

$$N = {}_{10}P_4 = \frac{10!}{(10-4)! 4!} = 5040$$

31. Supóngase que una compañía que fabrica relojes y una compañía que fabrica máquinas de escribir deben elegir para embarcar sus productos entre tren (T), camión (C) y avión (A). Ninguno de los fabricantes tiene preferencia en cuanto a la forma de envío, de manera que cada resultado es equiprobable.

a. Muéstrase el espacio muestral en un plano bidimensional, señalando las selecciones del fabricante de relojes en el eje horizontal y las del fabricante de máquinas de escribir en el eje vertical.

b. ¿Cuál es la probabilidad de que solamente uno de los fabricantes seleccione avión para el embarque de sus productos?

SOLUCION

(a) $R = \text{FABRICANTE DE RELOJES} = \{T, C, A\}$
 $M = \text{FABRICANTE DE MAQUINAS} = \{T, C, A\}$

$S = M \times R = \{ (x, y) \mid x \in M \text{ y } y \in R \}$
 $= \{(T, T), (T, C), (T, A), (C, T), (C, C), (C, A), (A, T), (A, C), (A, A)\}$

(b) $E = \{\text{solamente uno de los fabricantes seleccione avión}\} = \{(T, A), (C, A), (A, T), (A, C)\}$

32. Un comprador de un automóvil nuevo puede elegir entre cinco estilos de carrocería, con o sin transmisión automática, con o sin aire acondicionado, con o sin asientos individuales y entre diez colores. ¿De cuántas formas puede realizar su elección el comprador?

SOLUCION

Aplicando directamente el principio fundamental del conteo

$N_1=5$ carrozas (carrocerías)
 $N_2=2$ transmisión automática
 $N_3=2$ aire acondicionado
 $N_4=2$ asientos individuales
 $N_5=10$ colores

$$N = N_1 N_2 N_3 N_4 N_5 = (5).(2).(2).(2).(10)=400$$

33. ¿De cuántas formas puede elegirse un cuarteto (grupo de cuatro jugadores) de entre doce miembros de un club de golf?

SOLUCION

El problema corresponde directamente a el caso típico de combinaciones donde $n=12$ y $r=4$, entonces

$${}_{12}C_4 = \frac{12!}{(12-4)!4!} = 495$$

34. Si 20 estaciones de servicio constituyen una población, ¿cuál es la probabilidad de que se seleccione como muestra aleatoria una combinación de cuatro estaciones en particular?

SOLUCION

Para el problema $n=20$ y $r=4$, entonces

$${}_{20}C_4 = \frac{20!}{(20-4)!4!} = \frac{2.432902008 \times 10^{18}}{2.092278989 \times 10^3 (24)} = 4845$$

y por lo tanto la probabilidad de que se seleccione una estación de servicio es:

$$P = \frac{1}{4845} = 2.06 \times 10^{-4} \quad P(A) = \frac{\#A}{\#S}$$

AXIOMAS BÁSICOS DE LA PROBABILIDAD

Aunque la definición dada anteriormente de la PROBABILIDAD permite calcularla a partir del conteo de los conjuntos, es necesario definir nuevas propiedades que permitan calcularla para los casos en que no sea posible aplicar dicha definición.

Sean S el espacio muestral y E un evento cualquiera, entonces

- a) $P(S)=1$ evento seguro
- b) $P(\phi)=0$ evento imposible
- c) $0 \leq P(E) \leq 1$

Es importante resaltar la propiedad c) ya que señala que ningún evento puede de ninguna manera tener una probabilidad negativa ni nunca puede ser mayor que la unidad. Por lo tanto, si al resolver algún problema se obtiene una probabilidad que no cumpla la propiedad c) se puede afirmar que el problema está mal resuelto.

REGLA DE LA ADICIÓN DE PROBABILIDAD PARA EVENTOS AJENOS

(c) Si $A \cap B = \phi$ es decir A y B son conjuntos ajenos, entonces

$$P(A \cup B) = P(A) + P(B) \quad (2.6)$$

(d) Si $E_i \cap E_j = \phi$ para $i \neq j$ $i, j = 1, 2, 3, \dots, n$, entonces

$$P(E_1 \cup E_2 \cup \dots \cup E_n) = P(E_1) + P(E_2) + \dots + P(E_n) \quad (2.7)$$

(e) como $S = A \cup A^c$ y $A \cap A^c = \emptyset$ entonces $P(S) = P(A \cup A^c) = P(A) + P(A^c)$

Por otra parte $P(S) = 1$ por lo tanto $1 = P(A) + P(A^c)$

Despejando a $P(A)$

$$P(A) = 1 - P(A^c) \quad (2.8)$$

REGLA GENERAL DE LA ADICIÓN DE PROBABILIDAD.

(f) Si $A \cap B \neq \emptyset$ entonces

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (2.9)$$

Nota: La regla (f) se reduce a la regla (c) en el caso de conjuntos ajenos.

La regla es difícil de generalizar para un número grande de conjuntos. Por ejemplo, a continuación se muestra la regla de adición para el caso de tres conjuntos A, B, C cualquiera, no necesariamente ajenos

$$\begin{aligned} P(A \cup B \cup C) &= P(A \cup (B \cup C)) = P(A) + P(B \cup C) - P(A \cap (B \cup C)) \\ &= P(A) + P(B) + P(C) - P(B \cap C) - P((A \cap B) \cup (A \cap C)) \\ &= P(A) + P(B) + P(C) - P(B \cap C) - (P(A \cap B) - P(A \cap C) + P(A \cap B \cap C)) \\ &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \quad (2.10)$$

CALCULO DE PROBABILIDADES APLICANDO LAS REGLAS BÁSICAS.

EJEMPLOS

35. En el experimento de arrojar tres monedas, se considera que los ocho posibles resultados son equiprobables. Si E_1 denota al evento de que ocurran dos soles y E_2 al evento de que ocurran tres soles, ¿cuál es la probabilidad de que ocurra ya sea E_1 ó E_2 ? Esto es, ¿cuál es $P(E_1 \cup E_2)$?

SOLUCION

El espacio muestral del problema y cada uno de los eventos E_1 y E_2 son mostrados a continuación

$S = \{\text{arrojar 3 monedas}\} = \{SSS, SSA, SAS, SAA, ASS, ASA, AAS, AAA\}$

$E_1 = \{\text{dos soles}\} = \{SSA, SAS, ASS\}$

$E_2 = \{\text{3 soles}\} = \{SSS\}$

$P(E_1) = 3/8, \quad P(E_2) = 1/8,$

$E_1 \cup E_2 = \{\text{dos soles ó tres soles}\} = \{SSA, SAS, ASS, SSS\}$

$E_1 \cap E_2 = \emptyset$

$P(E_1 \cup E_2) = P(E_1) + P(E_2) = 3/8 + 1/8 = 4/8 = 1/2$

36. En el problema anterior, si A denota al evento de que ocurran dos o más soles y B denota al evento de que ocurran dos o menos soles, ¿cuál es la probabilidad de que ocurra ya sea A o B? Esto es ¿cuánto es, vale $P(A \cup B)$?

SOLUCION

Del espacio muestral del problema anterior se tiene que

$$A = \{2 \text{ ó más soles} \} = \{ASS, SAS, SSA, SSS\}$$

$$B = \{2 \text{ ó menos soles} \} = \{ASS, SAS, SSA, AAS, ASA, SAA, AAA\}$$

$$A \cap B = \{ASS, SAS, SSA\}$$

Debido a que los conjuntos no son ajenos, se debe aplicar la ecuación (8)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 4/8 + 7/8 - 3/8 = 1$$

37. Supóngase que una bolsa contiene 10 esferas marcadas 1, 2, 3, . . . , 10. Sea E el evento de extraer una esfera marcada con un número par y F el evento de extraer una esfera marcada con un número 5 o mayor. ¿Son E y F mutuamente excluyentes? Obténgase $P(E \cup F)$.

SOLUCION

El espacio muestral y cada uno de los eventos se describen a continuación

$$S = \{\text{extraer una esfera marcada del 1 al 10}\} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

$$E = \{\text{par}\} = \{2, 4, 6, 8, 10\}$$

$$F = \{5 \text{ ó mayor}\} = \{5, 6, 7, 8, 9, 10\}$$

Para que los eventos sena excluyente se debe tener que $P(E \cap F) = P(E) P(F)$

Como $E \cap F = \{6, 8, 10\}$ se tiene que $P(E \cap F) = 3/10$

Y puesto que $P(E) P(F) = (5/10)(6/10) = 3/10$, entonces los conjuntos E y F son excluyentes.

Entonces No son excluyentes

Aplicando la regla general de la adición

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) = 5/10 + 6/10 - 3/10 = 8/10 = 4/5$$

38. Si se extrae aleatoriamente un naipe de una baraja ordinaria de 52 naipes bien barajados, (a) ¿cuál es la probabilidad de extraer un trébol o un corazón o un diamante? (b) ¿Cuál es la probabilidad de extraer un diamante o un as?

SOLUCION

Hay que recordar que la baraja está formada por 4 conjuntos de 13 cartas, y que cada uno de los conjuntos está corresponde a las figuras de tréboles, corazones, diamantes y picas.

El conveniente definir los siguientes conjuntos:

$A = \{\text{la carta elegida es un trébol}\}$

$B = \{\text{la carta elegida es un corazón}\}$

$C = \{\text{la carta elegida es un diamante}\}$

$D = \{\text{la carta elegida es una pica}\}$

$E = \{\text{la carta elegida es un as}\}$

Los eventos A, B, C y D son mutuamente ajenos. Por lo tanto:

$$(a) P(A \cup B \cup C) = P(A) + P(B) + P(C) = 13/52 + 13/52 + 13/52 = 3/4.$$

(b) En este $C \cap E = \{\text{as de diamantes}\}$, o sea los eventos no son ajenos, por lo que:

$$P(C \cup E) = P(C) + P(E) - P(C \cap E) = 13/52 + 4/52 - 1/52 = 4/13$$

39. Supóngase que el 80% de todos los estadounidenses que vacacionan en el lejano oriente visitan Tokio, 80% visitan Hong Kong y 70% visitan tanto Tokio como Hong Kong. ¿Cuál es la probabilidad de que un turista estadounidense vacacionando en el Lejano Oriente visite o Tokio o Hong Kong? ¿Cuál es la probabilidad de que el turista no visite ninguna de estas ciudades?

SOLUCION

Sean

$A = \{\text{visitan Tokio}\}$

$$P(A) = 0.8$$

$B = \{\text{visitan Hong Kong}\}$

$$P(B) = 0.8$$

$A \cap B = \{\text{visitan Tokio y Hong Kong}\}$

$$P(A \cap B) = 0.7$$

La probabilidad de la unión se obtiene utilizando

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = 0.8 + 0.8 - 0.7 = 0.9$$

$$P(A \cup B) = 0.9$$

$C = (A \cup B)^c$ representa a el conjunto de los turistas que no visitan a Tokio ó Hong Kong

La probabilidad $P(C)$ puede ser calculada mediante

$$P(C) = 1 - P(C)^c$$

$$P(C) = 1 - P(A \cup B)$$

$$P(C) = 1 - 0.9 = 0.10$$

40. Las probabilidades de que un vendedor de automóviles venda en una semana cero, uno, dos, tres, cuatro o cinco o más automóviles son 0.05, 0.10, 0.18, 0.25, 0.20 y 0.22, respectivamente.

a. ¿Cuál es la probabilidad de que venda tres o más automóviles en una semana?

b. ¿Cuál es la probabilidad de que venda tres o menos automóviles en una semana?

SOLUCION

Los datos para la probabilidad de venta en una semana son:

Venda	0	1	2	3	4	5
Prob.	0.05	0.10	0.18	0.25	0.20	0.22

(a) Sean lo eventos

$E_1 = \{\text{venta 3 automóviles}\} \quad 0.25$

$E_2 = \{\text{venta 4 automóviles}\} \quad 0.20$

$E_3 = \{\text{venta 5 automóviles}\} \quad 0.22$

Los cuales cumplen $E_i \cap E_j = \Phi$ para $i, j = 1, 2, 3$, entonces

$A = \{\text{venta 3 ó mas automóviles}\} = E_1 \cup E_2 \cup E_3$, así se tiene que

$$P(A) = P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) = 0.25 + 0.20 + 0.22 = 0.67$$

(b) Sean lo eventos

$F_1 = \{\text{no venta}\} \quad 0.05$

$F_2 = \{\text{venta 1 auto}\} \quad 0.10$

$F_3 = \{\text{venta 2 autos}\} \quad 0.18$

$F_4 = \{\text{venta 3 autos}\} \quad 0.25$

Los cuales cumplen $F_i \cap F_j = \Phi$ para $i, j = 1, 2, 3, 4$, entonces

$B = \{\text{venta 3 ó menos automóviles}\} = F_1 \cup F_2 \cup F_3 \cup F_4$ así se tiene que

$$P(B) = P(F_1 \cup F_2 \cup F_3 \cup F_4) = P(F_1) + P(F_2) + P(F_3) + P(F_4) = 0.05 + 0.10 + 0.18 + 0.25 = 0.58$$

Unidad III Probabilidad condicional y variables aleatorias

PROBABILIDAD CONDICIONAL

Eventos independientes y dependientes

Se dice que dos eventos A y B son EVENTOS INDEPENDIENTES si y solo si la ocurrencia de uno de ellos no afecta la ocurrencia del otro.

Si A y B son EVENTOS INDEPENDIENTES entonces, la probabilidad de que ocurran tanto A como B es igual al producto de sus probabilidades respectivas, esto es:

$$P(A \cap B) = P(A) \cdot P(B) \quad (3.1)$$

En el caso de que la ocurrencia de un evento A afecte la ocurrencia del evento B entonces se tiene el caso de EVENTOS DEPENDIENTES ó de la PROBABILIDAD CONDICIONAL, la cual se denota por:

$$P(B | A) \quad \text{"La probabilidad de B dado que ha ocurrido A"}$$

En general la probabilidad de la intersección de los eventos $A \cap B$, cuando son dependientes se obtiene mediante la expresión:

$$P(A \cap B) = P(A)P(B | A). \quad (3.2)$$

Despejando a $P(B | A)$.

$$P(B | A) = \frac{P(A \cap B)}{P(A)} \quad (3.3)$$

EJEMPLOS

1. Determine si los eventos $A = \{\text{sol en la primera tirada}\}$ $B = \{\text{sol en la segunda tirada}\}$ son independientes en el experimento de arrojar una moneda dos veces.

SOLUCION

El espacio muestral del problema es $S = \{(S,S), (S,A), (A,S), (A,A)\}$

Para la parte izquierda de la ecuación (10)

$$E = \{\text{dos soles al arrojar una moneda dos veces}\} = A \cap B = \{(S,S)\}$$

$$P(A \cap B) = N(E)/N(S) = 1/4$$

$$\text{Para la parte derecha de la ecuación (10)} \quad P\{A\} = 1/2 \quad P\{B\} = 1/2$$

$$P(A) \cdot P(B) = (1/2)(1/2) = 1/4$$

Entonces se cumple que $P(A \cap B) = P(A) \cdot P(B)$, por lo que los eventos son independientes.

2. Una caja contiene diez esferas. Cinco de ellas son blancas, tres rojas y dos negras. Se selecciona aleatoriamente una esfera .sin *reemplazo*.

- ¿Cuál es la probabilidad de extraer dos esferas blancas una después de otra?
- ¿Cuál es la probabilidad de extraer una esfera roja y después una negra?
- ¿Cuál es la probabilidad de extraer tres esferas rojas, una después de otra?
- ¿Cuál es la probabilidad de extraer una esfera negra, después un roja y finalmente un blanca?

SOLUCION

Los datos del problema son: total de esferas $n = 10$ repartidas en 5 blancas, 3 rojas y 2 negras. El experimento se realiza sin *reemplazo*, por lo que los eventos son dependientes. Definiendo los siguientes conjuntos

$B_1 = \{\text{Sacar bola blanca en la 1}^{\text{a}} \text{ extracción}\}$
 $B_2 = \{\text{Sacar bola blanca en la 2}^{\text{a}} \text{ extracción}\}$
 $B_3 = \{\text{Sacar bola blanca en la 3}^{\text{a}} \text{ extracción}\}$
 $R_1 = \{\text{Sacar bola roja en la 1}^{\text{a}} \text{ extracción}\}$
 $R_2 = \{\text{Sacar bola roja en la 2}^{\text{a}} \text{ extracción}\}$
 $R_3 = \{\text{Sacar bola roja en la 3}^{\text{a}} \text{ extracción}\}$
 $N_1 = \{\text{Sacar bola negra en la 1}^{\text{a}} \text{ extracción}\}$
 $N_2 = \{\text{Sacar bola negra en la 2}^{\text{a}} \text{ extracción}\}$

- $P(\{2 \text{ blancas una después de la otra}\}) = P(B_1 \cap B_2) = P(B_1) P(B_2 | B_1) = (5/10)(4/9) = 2/9$
- $P(\{Una roja y una negra\}) = P(R_1 \cap N_2) = P(R_1) P(N_2 | R_1) = (3/10)(2/9) = 1/15$
- $P(\{Tres rojas después de otra\}) = P(R_1) \cdot P(R_2 | R_1) \cdot P(R_3 | R_2 \cap R_1) = (3/10)(2/9)(1/8) = 1/120$
- $P(\{Negra, después roja, y finalmente blanca\}) = P(R_1) \cdot P(R_2 | R_1) \cdot P(R_3 | R_2 \cap R_1) = (3/10)(2/9)(5/8) = 1/24$

3. El Sr. Huerta y su esposa tienen 55 y 50 años de edad, respectivamente. Si la probabilidad de que un hombre de 55 años de edad viva al menos otros 15 años es de 0.70, y la probabilidad de que una mujer de 50 años de edad viva al menos otros 15 años es de 0.85, ¿cuál es la probabilidad de que tanto el Sr. Huerta como su esposa continúen vivos dentro de 15 años? (Considérese que las longevidades del esposo y esposa son independientes.)

SOLUCION

Se definen los eventos:

$A = \{\text{el señor viva más de 15 años}\}$, entonces, $P(A) = 0.70$
 $B = \{\text{la señora viva más de 15 años}\}$, entonces $P(B) = 0.85$

Entonces $C = A \cup B = \{\text{El señor y la señora vivan más 15 de años}\}$

Considerando los eventos independientes se tiene que $P(A \cap B) = P(A) \cdot P(B) = (0.70)(0.85) = 0.595$

$P(C) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.70 + 0.85 - 0.595 = 0.955$

4. Se dispone de dos máquinas contra incendios para casos de emergencia. La probabilidad de que cualesquier de las dos máquinas esté lista cuando se necesite es de 90%. Se considera que la disponibilidad de una máquina es independiente de la otra. a. En el caso de una alarma por incendio, ¿cuál es la probabilidad de que ambas máquinas estén listas? b. ¿Cuáles la probabilidad de que ambas máquinas no estén listas? c. ¿Cuál es la probabilidad de que solamente una máquina esté lista?

SOLUCION

Es conveniente definir los eventos

$A = \{\text{la máquina 1 esté lista}\}$ $P(A) = 0.9$

$B = \{\text{la máquina 2 esté lista}\}$ $P(B) = 0.9$

Entonces, cada uno de los incisos se puede resolver como se indica a continuación

a) $P(A \cap B) = P(A) \cdot P(B) = (0.9)(0.9) = 0.81$

b) $P(A^c \cap B^c) = P(A^c) \cdot P(B^c) = (1 - P(A))(1 - P(B)) = (0.1)(0.1) = 0.01$

c) El evento de que al menos una de las máquinas esté disponible es $C = (A \cap B^c) \cup (A^c \cap B)$

$$P(C) = P((A \cap B^c) \cup (A^c \cap B)) = P(A \cap B^c) + P(A^c \cap B) - P(A \cap B) \cap (A^c \cap B) \\ = P(A) \cdot P(B^c) + P(A^c) \cdot P(B) = (0.9)(1 - 0.9) + (1 - 0.9)(0.9) = 0.09 + 0.09 = 0.18$$

5. A continuación se encuentra una tabla probabilística acerca del sexo y el estado civil de los empleados de una gran institución.

Estado civil	Mujeres F	Hombres F'	Total
Casados (M)	0.42	0.18	0.60
Solteros (M')	0.28	0.12	0.40
Total	0.70	0.30	1.0

a. ¿Son independientes el sexo y estado civil? ¿Por qué si o por qué no?

b. Obténgase $P(M | F)$, $P(M | F')$ y $P(M)$. (La barra vertical "I" significa "dado que".)

c. Obténgase $P(F | M)$, $P(F | M')$ y $P(F)$.

d. Obténgase $P(M' | F)$, $P(M' | F')$, y $P(M')$.

e. Obténgase $P(F' | M)$, $P(F' | M')$, y $P(F')$.

SOLUCION

(a) Para contestar esta pregunta hay que aplicar la ecuación (12) para determinar la probabilidad condicional en cada una de las combinaciones señaladas en los incisos siguientes

(b) $P(M | F) = P(M \cap F) / P(F) = 0.42 / 0.70 = 0.6$

$P(M | F') = P(M \cap F') / P(F') = 0.18 / 0.30 = 0.6$

$P(M) = 0.6$

Entonces $P(M | F) = P(M | F') = P(M)$

(c) $P(F | M) = P(F \cap M) / P(M) = 0.42 / 0.60 = 0.7$

$P(F | M') = P(F \cap M') / P(M') = 0.28 / 0.4 = 0.7$

$P(F) = 0.7$

Entonces $P(F | M) = P(F | M') = P(F)$

(d) $P(M' | F) = P(M' \cap F) / P(F) = 0.28 / 0.70 = 0.4$

$P(M' | F') = P(M' \cap F') / P(F') = 0.12 / 0.30 = 0.4$

$P(M') = 0.4$

Entonces $P(M' | F) = P(M' | F') = P(M') = 0.4$

(e) $P(F' | M) = P(F' \cap M) / P(M) = 0.18 / 0.60 = 0.3$

$P(F' | M') = P(F' \cap M') / P(M') = 0.12 / 0.4 = 0.3$

$P(F') = 0.3$

Entonces $P(F' | M) = P(F' | M') = P(F') = 0.3$

Como se observa de cada uno de los incisos anteriores, los eventos de sexo y estado civil son independientes uno del otro.

6. Se extraen naipes de una baraja ordinaria. Si los naipes que se han extraído no se reemplazan antes de extraer el siguiente, ¿cuál es la probabilidad de extraer

- Cuatro ases y después cualesquier de los otros naipes;
- Tres ases y después dos reyes;
- Cinco naipes del mismo palo?

SOLUCION

- a) Un caso posible se muestra a continuación

A	A	A	A	B
---	---	---	---	---

Definiendo los eventos:

A_1 ={As en la primera elección}

A_2 ={As en la segunda elección }

A_3 ={As en la tercera elección}

A_4 ={As en la cuarta elección}

B ={cualquiera en la quinta elección }

Entonces:

$$P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap B) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2) \cdot P(A_4 | A_1 \cap A_2 \cap A_3) \cdot P(B | A_1 \cap A_2 \cap A_3 \cap A_4) \\ = (4/52)(3/51)(2/50)(1/49)(48/48) = 1152/3118752000 = 1/270725$$

- b) El caso es mostrado

A	A	A	K	K
---	---	---	---	---

Utilizando lo eventos anteriores y

K_4 ={Rey en la cuarta elección}

K_5 ={Rey en la quinta elección}

$$P(A_1 \cap A_2 \cap A_3 \cap K_4 \cap K_5) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2) \cdot P(K_4 | A_1 \cap A_2 \cap A_3) \cdot P(K_5 | A_1 \cap A_2 \cap A_3 \cap K_4) \\ = 4/52(3/51)(2/50)(4/49)(3/48) = 288/31879.220 = 1/10820900$$

- c) Hay 4 palos y 13 figuras por palo, para cada uno de los palos, por ejemplo, corazones sean los eventos:

C_1 ={Corazón en la primera elección}

C_2 ={ Corazón en la segunda elección }

C_3 ={ Corazón en la tercera elección}

C_4 ={ Corazón en la cuarta elección}

C_5 ={ Corazón en la quinta elección }

$$P(C_1 \cap C_2 \cap C_3 \cap C_4 \cap C_5) = P(C_1) \cdot P(C_2 | C_1) \cdot P(C_3 | C_1 \cap C_2) \cdot P(C_4 | C_1 \cap C_2 \cap C_3) \cdot P(C_5 | C_1 \cap C_2 \cap C_3 \cap C_4) \\ = (13/52)(12/51)(11/50)(10/49)(9/48) = 15440/311873200 = 1/209.39$$

Finalmente multiplicando por 4

$$P(\{5 \text{ naipes del mismo palo}\}) = (4)(1/209.39) = 4/209.39$$

7. Un cartón contiene 20 huevos, 5 de los cuales están descompuestos. Si se seleccionan aleatoriamente tres huevos sin reemplazo, ¿cuál es la probabilidad de que los tres estén descompuestos?

SOLUCION

De acuerdo a la información de $n = 20$ hay 5 descompuestos y hay que elegir 3 sin reemplazo, entonces, definiendo los eventos

$D_i = \{\text{Huevo defectuoso en la elección } i\}$ para $i = 1, 2, 3$.

$$P(\{3 \text{ huevos descompuestos}\}) = P(D_1 \cap D_2 \cap D_3) = P(D_1)P(D_2 | D_1)P(D_3 | D_1 \cap D_2) = (5/20)(4/19)(3/18) = 1/114$$

8. Supóngase que la política de cierta compañía de seguros es que sus vendedores realicen visitas de casa en casa. De acuerdo a la experiencia anterior, el 20 % de las visitas dan como resultado una venta (S), o $P(S) = 0.20$, y 80% de las visitas no (S') o $P(S') = 0.80$. De las familias que han adquirido pólizas de seguros el 30% viven en casas unifamiliares de dos pisos (T) o $P(T | S) = 0.30$. Los restantes compradores (70%) viven en otros tipos de edificios (T') o $P(T' | S) = 0.70$. De aquellas familias que no adquirieron una póliza, el 60% vivían en casas unifamiliares de dos pisos o $P(T | S') = 0.60$ y el 40% vivían en otros tipos de casas o $P(T' | S') = 0.40$.

a. ¿Cuál es la probabilidad de que la siguiente visita dé como resultado una venta si los posibles clientes viven en una casa unifamiliar de dos pisos? Es decir, ¿cuánto vale $P(S|T)$?

b. ¿Cuál es la probabilidad de que la siguiente visita no dé como resultado una venta si la familia vive en cualquier otro tipo de edificio? Es decir, ¿cuánto vale $P(S' | T')$?

(Sugerencia: calcúlense las probabilidades conjuntas)

SOLUCION

La información se puede resumir como:

$$\begin{array}{lll} P(S) = 0.20 & P(T | S) = 0.30 & P(T' | S) = 0.70 \\ P(S') = 0.80 & P(T | S') = 0.60 & P(T' | S') = 0.40 \end{array}$$

La cual puede ser utilizada para calcular las probabilidades conjuntas

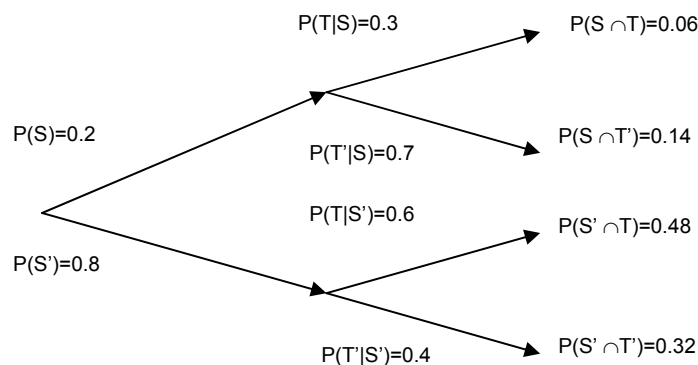
$$P(S \cap T) = P(S) P(T | S) = (0.20)(0.30) = 0.06$$

$$P(S \cap T') = P(S) P(T' | S) = (0.20)(0.70) = 0.14$$

$$P(S' \cap T) = P(S') P(T | S') = (0.80)(0.60) = 0.48$$

$$P(S' \cap T') = P(S') P(T' | S') = (0.80)(0.40) = 0.32$$

El resultado anterior puede ser representado gráficamente con un diagrama de árbol



Por otra parte

$$T = (S \cap T) \cup (S' \cap T)$$

$$T' = (S \cap T') \cup (S' \cap T')$$

Entonces

$$P(T) = P(S \cap T) + P(S' \cap T) = 0.06 + 0.48 = 0.54$$

$$P(T') = P(S \cap T') + P(S' \cap T') = 0.14 + 0.32 = 0.46$$

Con la información anterior

$$(a) P(S | T) = \frac{P(S \cap T)}{P(T)} = \frac{0.06}{0.54} = \frac{1}{9}$$

$$(b) P(S' | T') = \frac{P(S' \cap T')}{P(T')} = \frac{0.32}{0.46} = \frac{16}{23}$$

9. En una encuesta aplicada a los estudiantes que se gradúan en el colegio de cierta comunidad, se determinó que el 40% de los estudiantes continuarán estudiando alguna especialización en otra universidad (T) y el 60% no lo harán (T'). Dadas estas dos categorías de estudiantes, la proporción de estudiantes que han obtenido calificaciones promedio de A, B y C o menos se muestran a continuación,

Estudiantes	Calificaciones promedio			TOTAL
	A	B	C o menos	
T	0.10	0.30	0.60	1
T'	0.05	0.40	0.55	1

- Se selecciona aleatoriamente un estudiante y su calificación promedio es A. ¿Cuál es la probabilidad de que continúe estudiando?
- ¿Cuál es la probabilidad de que no continúe su educación si la calificación promedio es de B?

SOLUCION

Utilizando $P(T)=0.4$, $P(T')=0.6$ y la tabla se puede calcular la probabilidad conjunta

$$P(T \cap A) = P(T) P(A | T) = (0.40)(0.10) = 0.04$$

$$P(T' \cap A) = P(T') P(T' | A) = (0.60)(0.05) = 0.03$$

$$P(T \cap B) = P(T) P(B | T) = (0.40)(0.30) = 0.12$$

$$P(T' \cap B) = P(T') P(T' | B) = (0.60)(0.40) = 0.24$$

$$P(T \cap C) = P(T) P(C | T) = (0.40)(0.60) = 0.24$$

$$P(T' \cap C) = P(T') P(T' | C) = (0.60)(0.55) = 0.33$$

Además

$$P(A) = P(A \cap T) + P(A \cap T') = 0.04 + 0.03 = 0.07$$

$$P(B) = P(B \cap T) + P(B \cap T') = 0.12 + 0.24 = 0.36$$

Por lo tanto

$$(a) P(T | A) = \frac{P(T \cap A)}{P(A)} = \frac{0.04}{0.07} = \frac{4}{7}$$

$$(b) P(T' | B) = \frac{P(T' \cap B)}{P(B)} = \frac{0.24}{0.38} = \frac{12}{19}$$

Regla de Bayes o teorema de Bayes

Algunos de los problemas resueltos en la sección anterior son problemas que pueden ser resueltos mediante el Teorema de Bayes, el cual se detalla a continuación.

Sean los conjuntos $A_1, A_2, A_3, \dots, A_n$, conjuntos mutuamente excluyentes, esto es, $A_i \cap A_j = \emptyset$

Y que además $\bigcup_{i=1}^n A_i = S$

Por lo tanto cualquier conjunto B puede ser representado por los $A_1, A_2, A_3, \dots, A_n$ de la forma:

$$B = \bigcup_{i=1}^n (A_i \cap B) = (A_1 \cap B) \cup (A_2 \cap B) \cup \dots \cup (A_n \cap B)$$

Entonces

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \dots + P(A_n \cap B)$$

Además como

$$P(A_i \cap B) = P(B) \cdot P(B | A_i), \text{ para } i = 1, 2, 3, \dots, n$$

Así se tiene que

$$P(B) = P(B) \cdot P(B | A_1) + P(B) \cdot P(B | A_2) + \dots + P(B) \cdot P(B | A_n)$$

Por otra parte adecuando la ecuación (12) al problema

$$\begin{aligned} P(A_i | B) &= \frac{P(A_i \cap B)}{P(B)} = \frac{P(B \cap A_i)}{P(B)} \\ &= \frac{P(A_i)P(B | A_i)}{P(A_1)P(B | A_1) + P(A_2)P(B | A_2) + \dots + P(A_n)P(B | A_n)} \end{aligned} \quad (3.4)$$

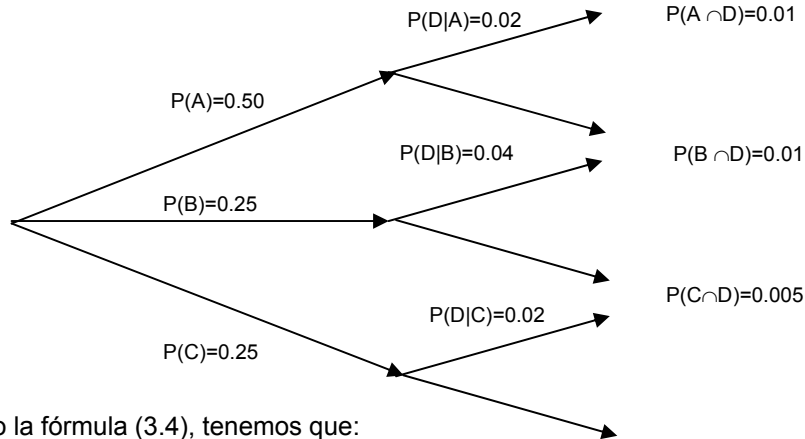
La ecuación anterior establece una forma para invertir la probabilidad condicional, esto es se puede pasar de $P(B | A_i)$ a $P(A_i | B)$.

EJEMPLOS

10. Una gran caja contiene transistores fabricados en tres máquinas. La máquina A es el doble de rápida que la máquina B o C. La tasa de defectos para la máquina A es 0.02 para B es 0.04 y para C es 0.02. Se selecciona al azar un transistor de la caja y resulta defectuoso.
¿Cuál es la probabilidad de que la haya producido la máquina C?

SOLUCION

El uso de un diagrama de árbol es útil para representar los datos y calcular la probabilidad conjunta



Utilizando la fórmula (3.4), tenemos que:

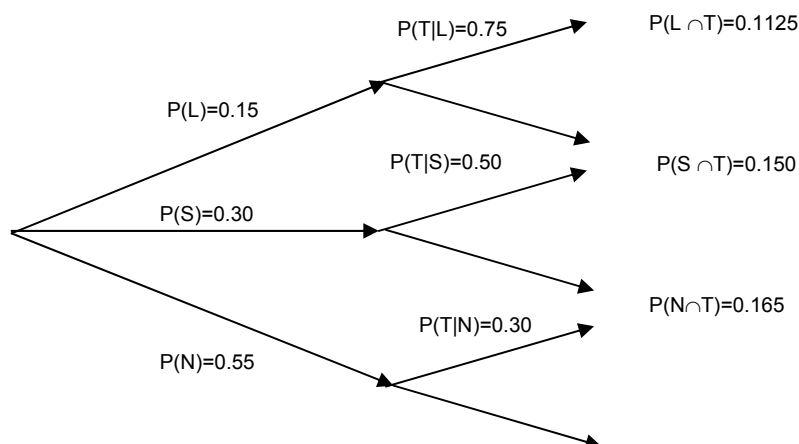
$$P(C | D) = \frac{P(C)P(D | C)}{P(A)P(D | A) + P(B)P(D | B) + P(C)P(D | C)}$$

$$P(C | D) = \frac{(0.25)(0.02)}{(0.50)(0.02) + (0.25)(0.04) + (0.25)(0.02)} = \frac{1}{5} = 0.20$$

11. Una vendedora realiza su trabajo haciendo visitas domiciliarias. Durante los años de experiencia ha acumulado los siguientes datos: de todas las visitas realizadas el 15% dieron como resultado lo que ella considera como grandes ventas (L), 30% ventas pequeñas (S) y 55% no fueron ventas (N). Además, de aquellos que hicieron grandes compras, el 75% viven en casas unifamiliares de dos pisos (T); de los que realizaron pequeñas compras, el 50% viven en casas de este tipo; entre quienes no realizaron compras el 30% viven en casas de este tipo. Si la siguiente casa que visita es una casa unifamiliar de dos pisos, ¿cuál es la probabilidad de que dé como resultado una gran venta? ¿Una venta pequeña? ¿Ninguna venta?

SOLUCION

Representando los resultados en un diagrama de árbol



Utilizando la fórmula (3.4)

$$P(L | T) = \frac{P(L)P(T | L)}{P(L)P(T | L) + P(S)P(T | S) + P(N)P(T | N)} = \frac{0.1125}{0.1125 + 0.150 + 0.165} = \frac{0.1125}{0.4275} = \frac{5}{19}$$

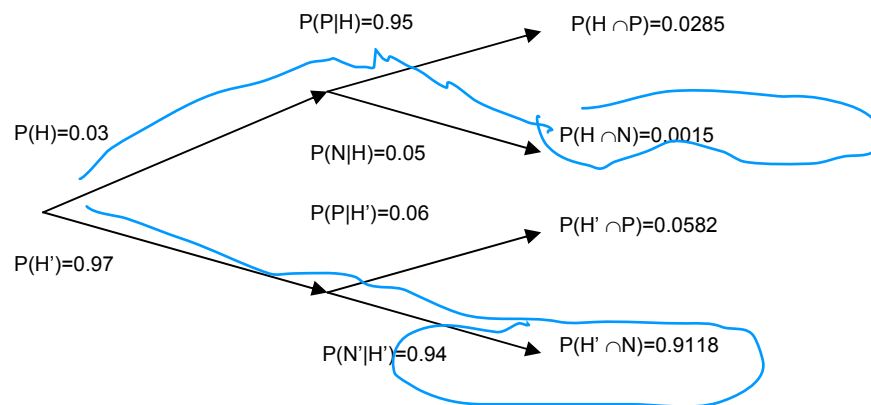
$$P(S | T) = \frac{P(S)P(T | S)}{P(L)P(T | L) + P(S)P(T | S) + P(N)P(T | N)} = \frac{0.150}{0.1125 + 0.150 + 0.165} = \frac{0.150}{0.4275} = \frac{20}{57}$$

$$P(N | T) = \frac{P(N)P(T | N)}{P(L)P(T | L) + P(S)P(T | S) + P(N)P(T | N)} = \frac{0.165}{0.1125 + 0.150 + 0.165} = \frac{0.165}{0.4275} = \frac{22}{57}$$

12. Como muchos saben la hepatitis se detecta comúnmente realizando pruebas sanguíneas. Supóngase que en un cierto grupo de personas, el 3% realmente tiene hepatitis (H) y el 97% no (H'). Supóngase además que si una persona tiene la enfermedad, el 95% de las pruebas sanguíneas la detectan (P), pero el 5% no la detectan (N). Para las personas que no tienen la enfermedad, el 6% de las pruebas muestran resultados positivos y el 94% muestran resultados negativos. Si la prueba sanguínea de una persona es negativa, ¿cuál es la probabilidad de que en realidad tenga la enfermedad?

SOLUCION

Representando los resultados en un diagrama de árbol



Entonces

$$P(H | N) = \frac{P(H)P(N | H)}{P(H)P(N | H) + P(H')P(N | H')} = \frac{0.0015}{0.0015 + 0.9118} = \frac{0.0015}{0.9133} = 1.6451 \times 10^{-3}$$

VARIABLES ALEATORIAS

Una **función** es una asociación tal que a cada elemento X de un conjunto llamado **dominio** le asocia un único elemento Y de otro conjunto llamado **rango**.

La variable X se les conoce como **variable independiente** y la variable Y como **variable dependiente**.

La **variable aleatoria** es una función que asigna valores numéricos a los resultados de un experimento aleatorio. La variable aleatoria se denota normalmente con letras mayúsculas X, Y, Z, \dots , etc.

TIPOS VARIABLES ALEATORIAS

Una variable aleatoria que toma que toma un número finito o infinito contable de valores se denomina **variable aleatoria discreta**, mientras que la que toma un número infinito ó continuo de valores se llama **variable aleatoria continua**

DISTRIBUCIONES DE PROBABILIDAD DE LAS VARIABLES DISCRETAS Y CONTINUAS

Si X es una variable aleatoria discreta ó continua la cual tiene un conjunto de valores x_1, x_2, x_3, \dots , ordenados de forma creciente y además la probabilidad de la variable aleatoria tome cada uno de los valores x_k es

$$P(X = x_k) \quad k = 1, 2, 3, \dots,$$

Es posible entonces definir una **función de probabilidad** para la variable aleatoria discreta como:

$$f(x_k) = P(X = x_k) \quad k = 1, 2, 3, \dots, \quad (3.5)$$

y para el caso continuo en una variable

$$f(x) = P(X = x) \quad x \in [a, b] \quad (3.6)$$

En general se dice que una función $f(x)$ es una distribución de probabilidad si satisface las siguientes propiedades

Para el caso discreto

- (a) $0 \leq f(x_k) \leq 1$ para $k = 1, 2, 3, \dots$,
- (b) $\sum_k f(x_k) = 1$ para $k = 1, 2, 3, \dots$,

Para el caso continuo

- (a) $0 \leq f(x) \leq 1$ para $x \in [a, b]$
- (b) $\int_a^b f(x) dx = 1$ para $x \in [a, b]$

La **función de distribución acumulada** para una variable aleatoria X se define como

$$F(x) = P(X \leq x)$$

Lo cual se traduce para el caso discreto en

$$F(x) = \sum_{j \leq k} f(x_j)$$

Y para el caso continuo

$$F(x) = \int_a^x f(x) dx$$

Las ideas anteriores pueden generalizarse para el caso de más variables aleatorias, por ejemplo, para el caso de dos variables aleatorias X y Y , se define la **función de probabilidad conjunta** como

$$f(x, y) = P(X = x, Y = y) \quad (3.7)$$

Donde la función $f(x, y)$ satisface para el caso discreto

- (a) $0 \leq f(x_j, y_k) \leq 1$ para $j = 1, 2, 3, \dots$, y $k = 1, 2, 3, \dots$,
- (b) $\sum_j \sum_k f(x_k) = 1$ para $j = 1, 2, 3, \dots$, y $k = 1, 2, 3, \dots$,

Para el caso continuo

- (a) $0 \leq f(x, y) \leq 1$ para $x \in [a, b]$ y $y \in [c, d]$
- (b) $\int_c^d \int_a^b f(x, y) dx dy = 1$ para $x \in [a, b]$ y $y \in [c, d]$

Se dice que dos variables aleatorias X y Y discretas son **variables aleatorias independientes** si y solo si los eventos $X=x$ y $Y=y$ son independientes para todo x, y . Para este caso se dice que la distribución conjunta de probabilidad satisface

$$P(X = x, Y = y) = P(X = x) P(Y = y)$$

o de igual forma

$$f(x, y) = f(x)f(y)$$

VALOR ESPERADO DE LA DISTRIBUCIÓN DE PROBABILIDAD

Un concepto importante para las distribuciones de probabilidad es el **valor esperado** ó **esperanza matemática** la cual se define como:

$$E(X) = \sum_{i=1}^n f(x_i)x_i \quad (3.8)$$

Y para el caso continuo

$$E(X) = \int_a^b x f(x) dx \quad (3.9)$$

La esperanza matemática $E(X)$ se puede considerar como el promedio de la distribución de probabilidad, la cual se denota por la letra griega μ .

Propiedades de la esperanza matemática

- | | | |
|-----|-----------------------------------------------------------------|--------------------------|
| (a) | Si c es una constante, entonces | $E(cX) = cE(X)$ |
| (b) | Si X, Y son variables aleatorias, entonces | $E(X + Y) = E(X) + E(Y)$ |
| (c) | Si X, Y son dos variables aleatorias independientes, entonces | $E(XY) = E(X)E(Y)$ |

VARIANZA Y DESVIACIÓN TÍPICA DE UNA DISTRIBUCIÓN DE PROBABILIDAD

Otra estadística importante en la probabilidad y la estadística es la Varianza la cual se denota por σ^2 y se define para el caso de distribuciones de probabilidad como

$$Var(X) = E((X - \mu)^2) \quad (3.10)$$

La varianza $Var(X)$ se relaciona con la desviación típica de una variable aleatoria σ_X mediante $\sigma_X^2 = Var(X)$. Por lo que la varianza puede ser representada mediante cualquiera de las notaciones anteriores.

Por su definición la varianza nunca puede tomar valores negativos, y su interpretación es idéntica a la que se dio para la distribuciones de frecuencia en la sección de la estadística descriptiva.

Desarrollando la definición anterior y aplicando las propiedades de la esperanza matemática

$$\begin{aligned} \sigma_X^2 &= E[(X - \mu)^2] = E[X^2 - 2X\mu + \mu^2] = E(X^2) - 2\mu E(X) + \mu^2 E(1) \\ &= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2 \end{aligned}$$

esto es

$$\sigma_X^2 = E(X^2) - \mu^2 \quad (3.11)$$

Para una distribución discreta la varianza se calcula mediante

$$\sigma_X^2 = \sum_{i=k}^n x_k^2 f(x_k) - \mu^2 \quad (3.12)$$

y para el continuo

$$\sigma_X^2 = \int_a^b x^2 f(x) dx - \mu^2 \quad (3.13)$$

Propiedades de la Varianza

- (a) Si c es una constante, entonces $Var(cX) = cVar(X)$
- (b) La cantidad $E[(X + a)^2]$ es mínima cuando $a = \mu$
- (c) Si X, Y son dos variables aleatorias independientes, entonces
 $Var(X \pm Y) = Var(X) + Var(Y)$ ó $\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y$

EJEMPLOS

13. Se dice que un juego es "legal" si al jugar el juego el valor esperado de ganar ó perder es cero. Diga usted si el juego de los "volados" con una moneda balanceada es un juego "legal".

SOLUCION

El juego consiste en lo siguiente:

- Se tira la moneda, la persona pide sol y cae sol, gana 1 peso.
- Se tira la moneda, la persona pide águila y cae águila, gana 1 peso.
- Se tira la moneda, la persona pide sol y cae águila, pierde 1 peso.
- Se tira la moneda, la persona pide águila y cae sol, pierde 1 peso.

La variable aleatoria del experimento se puede definir como $X = \{-1, 1\}$

Definiendo los eventos $S_1 = \{\text{la persona pide sol}\}$, $S_2 = \{\text{cae sol}\}$
 $A_1 = \{\text{la persona pide águila}\}$, $A_2 = \{\text{cae águila}\}$

Entonces las respectivas probabilidades de cada valor de la variable aleatoria son:

$$f(1) = P(X=1) = P(S_1 \cap S_2) + P(A_1 \cap A_2) = P(S_1) \cdot P(S_2) + P(A_1) \cdot P(A_2) = (1/2)(1/2) + (1/2)(1/2) = (1/2)$$

$$f(-1) = P(X=-1) = P(S_1 \cap A_2) + P(A_1 \cap S_2) = P(S_1) \cdot P(A_2) + P(A_1) \cdot P(S_2) = (1/2)(1/2) + (1/2)(1/2) = (1/2)$$

Los resultados generalmente se pueden acomodar para las variables discretas en una tabla

x	-1	1
f(x)	1/2	1/2

De la tabla anterior se puede calcular la esperanza matemática del experimento

$$E(X) = \sum_{i=1}^n x_i f(x_i) = (-1)(1/2) + (1)(1/2) = -1/2 + 1/2 = 0$$

El resultado indica que el juego es legal.

14. Denótese mediante X al número de caras obtenidas en la tirada de dos monedas ¿Cuál es la media y la varianza de X ?

SOLUCION

La tabla de la distribución de probabilidad se da a continuación

x	0	1	2
f(x)	1/4	1/2	1/4

Entonces

$$\mu = E(x) = \sum_{i=1}^n x_i f(x_i) = 0(1/4) + 1(1/2) + 2(1/4) = 1$$

$$E(x^2) = \sum_{i=1}^n x_i^2 f(x_i) = 0^2(1/4) + 1^2(1/2) + 2^2(1/4) = 1/2 + 1 = 3/2$$

$$\sigma_x^2 = E(x^2) - [E(x)]^2 = (3/2)^2 - 1^2 = 3/4$$

15. En un estudio acerca de las actitudes de los consumidores hacia cierto producto nuevo, se pregunta lo siguiente: "¿Le agrada el nuevo producto?" Para esta pregunta hay solamente dos posibles respuestas, "sí" y "no", a las cuales se les asignan los valores de 1 y 0, respectivamente. Sea p la probabilidad de que ocurra el evento de una respuesta "sí". (a) ¿Cuál es la distribución probabilística de W, variable aleatoria de este experimento?, (b) su Valor esperado y (c) su desviación típica.

SOLUCION

(a) De acuerdo a los datos del problema, la variable aleatoria W toma los valores $W = \{0, 1\}$ y $f(1) = P(X = 1) = p$

Como la distribución de probabilidad de la variable aleatoria W debe cumplir la propiedad $\sum_k f(x_k) = 1$, entonces

$$f(0) + f(1) = 1 \quad f(0) = 1 - f(1) = 1 - p$$

Entonces la tabla de distribución de probabilidad de W es

W	0	1
f(W)	1-p	p

$$(b) E(X) = \sum_{i=1}^n x_i f(x_i) = (0)(1-p) + (1)(p) = p$$

$$(c) \sigma_x^2 = \sum_{i=k}^n x_k^2 f(x_k) - \mu^2 = (0)^2(1-p) + (1)^2(p) - p^2 = p - p^2 = p(1-p)$$

entonces $\sigma_x = \sqrt{p(1-p)}$

16. Sea X la variable aleatoria correspondiente al número de soles obtenidas en la tirada de cuatro monedas balanceadas. Obténgase la distribución probabilística de X. y su valor esperado.

SOLUCION

De la definición de la variable aleatoria se tiene que $X = \{0, 1, 2, 3, 4\}$

En general para un evento cualquiera de arrojar una moneda balanceada n veces la probabilidad de cada evento simple es: $P(E) = \frac{1}{2^n}$

Por otra parte, si en el evento se lanzan n monedas y aparecen r soles, entonces aparecerán $n-r$ águilas y el número de eventos simples que contienen r soles se determina utilizando las técnicas de conteo:

$$\frac{n!}{r!(n-r)!}$$

Entonces la probabilidad de que ocurran en n tiradas r soles es

$$P(r \text{ soles}) = \frac{n!}{r!(n-r)!} \frac{1}{2^n}$$

Aplicando el resultado anterior para cada uno de los valores de la variable aleatoria

$$f(0) = P(X=0) = \frac{4!}{0!(4-0)!} \frac{1}{2^4} = \frac{1}{16}$$

$$f(1) = P(X=1) = \frac{4!}{1!(4-1)!} \frac{1}{2^4} = \frac{4}{16} = \frac{1}{4}$$

$$f(2) = P(X=2) = \frac{4!}{2!(4-2)!} \frac{1}{2^4} = \frac{6}{16} = \frac{3}{8}$$

$$f(3) = P(X=3) = \frac{4!}{3!(4-3)!} \frac{1}{2^4} = \frac{4}{16} = \frac{1}{4}$$

$$f(4) = P(X=4) = \frac{4!}{4!(4-4)!} \frac{1}{2^4} = \frac{1}{16}$$

Acomodando los resultados en la tabla siguiente

x	0	1	2	3	4
f(x)	1/16	1/4	3/8	1/4	1/16

Utilizando la tabla anterior

$$E(X) = \sum_{i=1}^n x_i f(x_i) = (0) \left(\frac{1}{16}\right) + (1) \left(\frac{1}{4}\right) + (2) \left(\frac{3}{8}\right) + (3) \left(\frac{1}{4}\right) + (4) \left(\frac{1}{16}\right) = 2.$$

17. Sea X la variable aleatoria correspondiente al número de caras obtenidas en la tirada de cuatro monedas balanceadas. a. Obténgase la distribución probabilística de X . b. La media de la distribución. c. La desviación típica.

SOLUCION

(a) El espacio muestral del experimento es

$S = \{ (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6) \}$

Entonces los valores posibles de la variable aleatoria son $X = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ y sus respectivas probabilidades se pueden calcular directamente del espacio muestral

$$f(2) = P(X = 2) = \frac{1}{36} \quad f(3) = P(X = 3) = \frac{2}{36} = \frac{1}{18} \quad f(4) = P(X = 4) = \frac{3}{36} = \frac{1}{12}$$

$$f(5) = P(X = 5) = \frac{4}{36} = \frac{1}{9} \quad f(6) = P(X = 6) = \frac{5}{36} \quad f(7) = P(X = 7) = \frac{6}{36} = \frac{1}{6}$$

$$f(8) = P(X = 6) = \frac{5}{36} \quad f(9) = P(X = 9) = \frac{4}{36} = \frac{1}{9} \quad f(10) = P(X = 10) = \frac{3}{36} = \frac{1}{12}$$

$$f(11) = P(X = 11) = \frac{2}{36} = \frac{1}{18} \quad f(12) = P(X = 12) = \frac{1}{36}$$

Colocando los resultados en una tabla.

x	2	3	4	5	6	7	8	9	10	11	12
f(x)	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

$$(b) \mu = \sum_{i=1}^n x_i f(x_i) = (2) (1/36) + (3) (1/18) + (4) (1/12) + (5) (1/9) + (6) (5/36) + (7) (1/6) + (8) (5/36) + (9) (1/9) + (10) (1/12) + (11) (1/18) + (12) (1/36) = 7.$$

$$(c) \sigma_X^2 = \sum_{i=k}^n x_k^2 f(x_k) - \mu^2 = (2)^2 (1/36) + (3)^2 (1/18) + (4)^2 (1/12) + (5)^2 (1/9) + (6)^2 (5/36) + (7)^2 (1/6) + (8)^2 (5/36) + (9)^2 (1/9) + (10)^2 (1/12) + (11)^2 (1/18) + (12)^2 (1/36) - 7^2 = .35/6 = 5.83333$$

entonces $\sigma_X = 2.4152$

18. Un juego llamado CHICOS Y GRANDES consiste primero en arrojar dos dados y se suman los puntos de sus caras. Los resultados de la suma son divididos en CHICOS si su valor es menor que siete, CASA si cae siete y GRANDES si valor es mayor que siete, tal como se muestra en la siguiente figura

2, 3, 4, 5, 6 chicos	7 Casa	8, 9, 10, 11, 12 grande
-------------------------	-----------	----------------------------

Las condiciones de juego son las siguientes:

- Si apuesta 1 peso a chicos y sale chicos, gana 1 peso.
- Si apuesta 1 peso a grandes y sale grandes, gana 1 peso.
- Si apuesta 1 peso a chicos y sale grandes ó casa, pierde 1 peso
- Si apuesta 1 peso a grandes y sale chico ó casa, pierde 1 peso
- Si apuesta 1 peso a la casa y sale casa gana 2 pesos.

f. Si apuesta 1 peso a la casa y sale chicos ó grandes, pierde 1 peso.

Diga usted si el juego es legal o no.

SOLUCION

La variable aleatoria adecuada al juego es $X = \{-1, 1, 2\}$

La distribución de probabilidad para la suma de los puntos de las caras de un dado son

y	2	3	4	5	6	7	8	9	10	11	12
f(y)	1/36	1/18	1/12	1/9	5/56	1/6	5/56	1/9	1/12	1/18	1/36

Definiendo los siguientes eventos $C_H = \{\text{CHICOS}\}$ $C_A = \{\text{CASA}\}$ y $G = \{\text{GRANDES}\}$, utilizando las condiciones de juego y tabla anterior

$$f(-1) = P(X = -1) = P(C_H \cap C_H') + P(G \cap G') + P(C_A \cap C_A') = P(C_H)P(C_H') + P(G)P(G') + P(C_A)P(C_A') = \\ = (15/36) + (15/36) + (6/36) = 36/36 = 1$$

$$f(1) = P(X = 1) = P(C_H \cap C_H) + P(G \cap G) = P(C_H)P(C_H) + P(G)P(G) = \\ = (15/36)(15/36) + (15/36)(15/36) = 25/72$$

$$f(2) = P(X = 2) = P(C_A \cap C_A) = P(C_A)P(C_A) = (6/36)(6/36) = 1/36$$

Por lo tanto se tiene la tabla

x	-1	1	2
f(x)	5/8	25/72	1/36

Entonces

$$\mu = \sum_{i=1}^n x_i f(x_i) = -1(5/8) + 1(25/72) + 2(1/36) = -5/8 + 1/8 = -4/8 = -0.5$$

Como el resultado es negativo el juego no solamente no es legal sino que es desfavorable al jugador.

19. Un vendedor ofrece dos modelos distintos de receptores de estéreo, H y T. Considérese que los dos modelos son igualmente populares: el 50% de todos los posibles compradores prefieren el Modelo H y el 50% prefieren el Modelo T. Además, considérese que el vendedor tiene en existencia tres receptores de cada modelo y que en un solo día se venden tres receptores.

a. Defínase la variable aleatoria de este experimento.

b. ¿Cuál es la distribución probabilística de la variable aleatoria?

SOLUCION

En total hay $n = 6$ receptores, 3 modelo H y 3 modelo T y la venta o selección consiste en $r = 3$ aparatos

(a) La variable aleatoria X del experimento es el número de aparatos tipo H vendidos, entonces si la venta consiste solamente de 3 aparatos X puede tomar los siguientes valores: $X = \{0, 1, 2, 3\}$,

(b) Las probabilidades de la variable aleatoria X se determinan mediante las técnicas de conteo

$$f(0) = \frac{{}_3C_3}{{}_6C_3} = \frac{1}{20}$$

$$f(1) = \frac{{}_3C_2 {}_3C_1}{{}_6C_3} = \frac{(3)(3)}{20} = \frac{9}{20}$$

$$f(2) = \frac{{}_3C_1 {}_3C_2}{{}_6C_3} = \frac{(3)(3)}{20} = \frac{9}{20}$$

$$f(3) = \frac{{}_3C_3}{{}_6C_3} = \frac{1}{20}$$

La respectiva distribución de probabilidad se resume en la tabla siguiente

x	0	1	2	3
f(x)	1/20	9/20	9/20	1/20

20. La inversión realizada por el Sr. Aranda podrían dar como resulta siguientes beneficios, con las probabilidades indicadas:

Beneficio	Probabilidad
\$1 millón	0.2
2 millones	0.3
3 millones	0.2
4 millones	0.2
5 millones	0.1
Total	1.0

Sea X el beneficio de su inversión. Obténganse la varianza y desviación típica de X.

SOLUCION

$$\mu = \sum_{i=1}^n x_i f(x_i) = (1)(0.2) + (2)(0.3) + (3)(0.2) + (4)(0.2) + (5)(0.1) = 2.7 \text{ millones}$$

$$E(x^2) = \sum_{i=1}^n X_i^2 f(x_i) = 1^2(0.2) + 2^2(0.3) + 3^2(0.2) + 4^2(0.2) + 5^2(0.1) = 8.9 \text{ millones}$$

$$\sigma_x^2 = E(x^2) - \mu^2 = 8.9 - 2.7^2 = 1.61.$$

$$\sigma_x = \sqrt{1.61} = 1.27 \text{ millones}$$

21. Supóngase que un aparato de televisión tiene ocho bulbos, dos de los cuales dos son defectuosos. Se seleccionan sucesivamente dos bulbos y se quitan del aparato para inspeccionarlos. Sea X el número de bulbos defectuosos en la muestra de dos bulbos. ¿Cuál es el valor esperado de X y su respectiva desviación típica?

SOLUCION

El número total de bulbos es n = 8 tubos, 2 defectuosos 6 sin defecto. La muestra a considerar es r = 2.

La variable aleatoria es X = {No. de defectuosos en la muestra} = {0, 1, 2}

$$f(0) = P(X = 0) = \frac{{}_6C_2}{{}_8C_2} = \frac{15}{28}$$

$$f(1)=P(X=1)=\frac{{}_2C_1({}_6C_1)}{{}_8C_2}=\frac{12}{28}$$

$$f(2)=P(X=2)=\frac{{}_2C_2}{{}_8C_2}=\frac{1}{28}$$

Entonces la tabla de la distribución de frecuencia es

x	0	1	2
f(x)	15/28	12/28	1/28

por lo tanto

$$\mu = E(x) = \sum x_i f(x_i) = 0(15/28) + 1(12/28) + 2(1/28) = 1/2$$

$$E(x^2) = \sum x_i^2 f(x_i) = 0^2(15/28) + 1^2(12/28) + 2^2(1/28) = 4/7$$

$$\sigma_x^2 = E(x^2) - \mu^2 = 4/7 - (1/2)^2 = 9/28$$

$$\sigma_x = \sqrt{9/28} = \frac{3}{\sqrt{28}} = 0.5666$$

22. Un jugador arroja tres monedas ideales. Gana \$3 si ocurren tres caras, \$2~ ocurren dos caras y \$1 si ocurre una cara. Si el juego es justo, ¿cuánto debería pagar si no aparece ninguna cara?

SOLUCION

La distribución de probabilidad del experimento de arrojar tres monedas legales es

x	0	1	2	3
f(x)	1/8	3/8	3/8	1/8

La variable aleatoria del experimento es $Y = \{y_1, 1, 2, 3\}$, donde y_1 representa el valor que debe pagar el jugador si en el resultado de arrojar las monedas no sale ninguna cara y los demás valores representan la ganancia igual al número de caras que aparecen. La distribución de probabilidad de la variable aleatoria Y es la siguiente

Ganancia

y	y_1	1	2	3
f(y)	1/8	3/8	3/8	1/8

Para que un juego sea legal se requiere que $E(y)=0$, entonces

$$(1/8)(y_1) + 1(3/8) + 2(3/8) + 3(1/8) = 0$$

despejando $y_1 = -12$

23. Supóngase que se van a vender 10 000 boletos a \$1 cada uno en una lotería realizada para ayudar en las investigaciones contra el cáncer. El premio es un automóvil con valor de \$ 4000. Si usted compró cinco boletos, ¿cuál es su contribución esperada a la investigación en contra del cáncer?

SOLUCION

Debido a que solamente se compran 5 de los 1000 boletos la probabilidad

de ganar es $P(\text{ganar})=5/10000$

y la de perder $P(\text{perder})=9995/10000$

El premio es 4000 pesos pero, se resta 5 porque se ha pagado por el boleto $4000-5=3995$ y la pérdida es 5.

La variable aleatoria del experimento Y es la ganancia y/o pérdida, $Y = \{-5, 3995\}$, entonces la correspondiente distribución de probabilidad de Y es

y	-5	3995
f(y)	9995/10000	5/10000

El valor esperado de la variable aleatoria es

$$E(Y)=3995(5/10000)+(-5)(9995/10000)=1.9975-4.9475=-31$$

Distribución de la media muestral \bar{X}

Considérese una población compuesta por los siguientes elementos $P = \{1, 3, 5, 7\}$, los cuales tienen una distribución de probabilidad uniforme, esto es, todos los elementos tienen la misma probabilidad de ser seleccionados, lo anterior es mostrado en la siguiente tabla de distribución de probabilidad

x	1	3	5	7
p(x)	1/4	1/4	1/4	1/4

Su respectiva media y su varianza son

$$\mu_X = E(X) = \sum x_i f(x_i) = 1(1/4) + 3(1/4) + 5(1/4) + 7(1/4) = 16/4 = 4$$

$$\begin{aligned} \sigma_X^2 &= E(x_i)^2 - E(x)^2 \\ &= \sum x_i^2 f(x) - \mu_X^2 = 1^2 \left(\frac{1}{4} \right) + 3^2 \left(\frac{1}{4} \right) + 5^2 \left(\frac{1}{4} \right) + 7^2 \left(\frac{1}{4} \right) - 4^2 = 5 \end{aligned}$$

Supóngase ahora que se realiza el experimento de seleccionar una muestra de dos números (X_1, X_2) de la población anterior **con reemplazo** y además se define la variable aleatoria $\bar{X} = \frac{(X_1 + X_2)}{2}$ (el promedio de los valores resultantes). Se pueden obtener un número infinito de muestras, pero muchas

de la muestra obtenidas serán idénticas, es decir tendrán el mismo resultado, aplicando las técnicas de conteo se sabe que hay solamente

4	4	=16	Diferentes muestras.
---	---	-----	----------------------

Explícitamente las muestras son:

$$S = \{ (1,1), (1,3), (1,5), (1,7), (3,1), (3,3), (3,5), (3,7), (5,1), (5,3), (5,5), (5,7), (7,1), (7,3), (7,5), (7,7) \}$$

Aplicando la definición de la variable aleatoria \bar{X} se obtienen siguientes valores

$$\bar{X} = \{1, 2, 3, 4, 5, 6, 7\}$$

Con los resultados anteriores es posible construir una distribución de probabilidad para la variable aleatoria \bar{X} a partir de todas las muestras posibles del mismo tamaño de una población dada, lo anterior se denomina **distribución muestral de la media**.

La distribución muestral de la media se puede obtener a partir de la siguiente tabla:

muestra	X_1	X_2	Total	Promedio
1	1	1	2	1
2	1	3	4	2
3	1	5	6	3
4	1	7	8	4
5	3	1	4	2
6	3	3	6	3
7	3	5	8	4
8	3	7	10	5
9	5	1	6	3
10	5	3	8	4
11	5	5	10	5
12	5	7	12	6
13	7	1	8	4
14	7	3	10	5
15	7	5	12	6
16	7	7	14	7

\bar{x}	$f(\bar{x})$
1	1/36
2	2/36
3	3/36
4	4/36
5	3/36
6	2/36
7	1/36

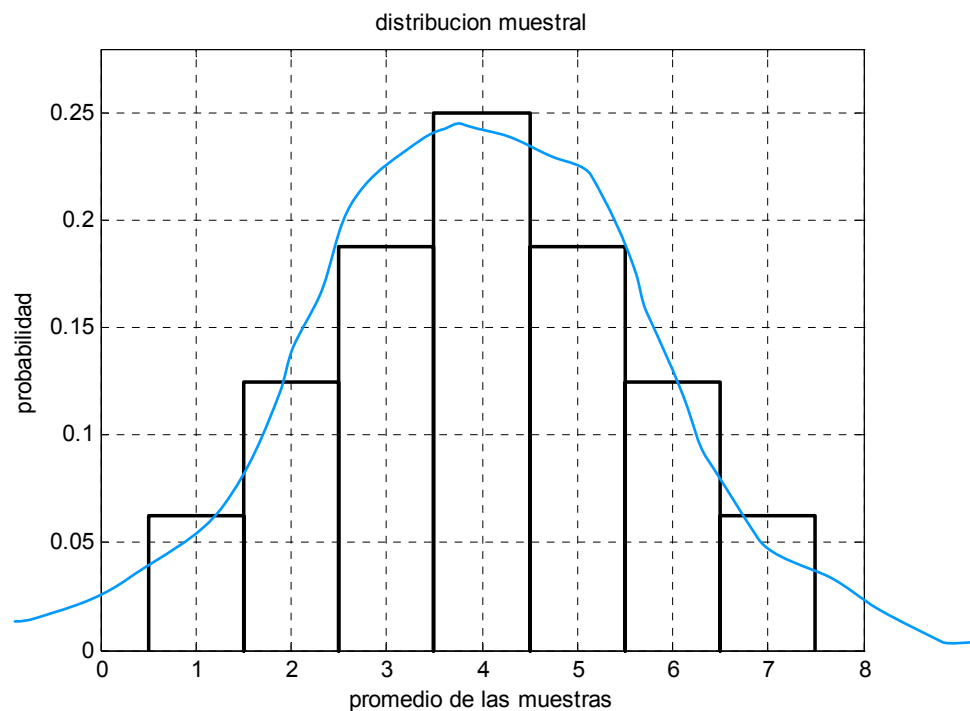
Las distribuciones probabilísticas de todos los diferentes valores de un estadístico muestral

El valor esperado de la media muestral y su varianza son.

$$\mu_{\bar{X}} = E(\bar{X}) = \sum \bar{x}_i f(\bar{x}_i) = 1(1/36) + 2(2/36) + 3(3/36) + 4(4/36) + 3(5/36) + 2(6/36) + 1(7/36) = 4$$

$$\sigma_{\bar{X}}^2 = E(\bar{X}^2) - E(\bar{X})^2$$

$$\begin{aligned}
 \sigma_{\bar{X}}^2 &= E(\bar{X})^2 - E(\bar{X})^2 \\
 &= \sum \bar{x}_i^2 f(\bar{x}) - \mu_{\bar{X}} = 1^2 \left(\frac{1}{16} \right) + 2^2 \left(\frac{2}{16} \right) + 3^2 \left(\frac{3}{16} \right) + 4^2 \left(\frac{4}{16} \right) + 5^2 \left(\frac{3}{16} \right) + 6^2 \left(\frac{2}{16} \right) + 7^2 \left(\frac{1}{16} \right) - 4^2 \\
 &= \frac{5}{2}
 \end{aligned}$$



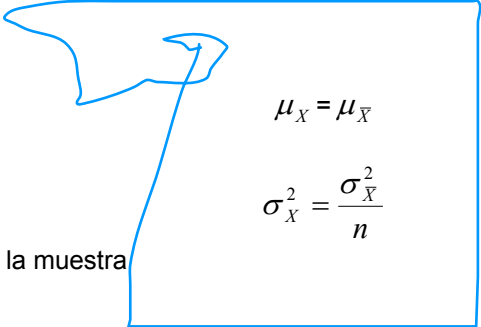
Distribución muestral de \bar{X} con $n = 2$ muestras

Como se puede apreciar en la gráfica anterior, la distribución muestral de la media \bar{X} tiene una forma totalmente simétrica. Si el experimento se realiza con una población y muestras más grandes se observaría el mismo comportamiento, es más, en el caso límite de una población y muestras infinitas la distribución se transformaría en una **distribución normal** con media $\mu_{\bar{X}}$ y varianza $\sigma_{\bar{X}}^2$, para más detalles de esta distribución ver la siguiente sección.

Unas preguntas interesantes son ¿Cuál es la relación entre la media muestral μ_X y $\mu_{\bar{X}}$?, y ¿Cuál es la relación entre la media muestral σ_X^2 y $\sigma_{\bar{X}}^2$?

De el problema anterior se observa que $\mu_X = \mu_{\bar{X}}$ y $\sigma_X^2 = \frac{\sigma_{\bar{X}}^2}{2}$

Aunque el problema anterior es un ejemplo de muchos posibles, las relaciones anteriores se cumplen en todos los casos de muestreo con reemplazo, esto es,



$$\mu_X = \mu_{\bar{X}} \quad (27)$$

$$\sigma_X^2 = \frac{\sigma_{\bar{X}}^2}{n} \quad (28)$$

Donde n = tamaño de la muestra

EJEMPLOS

24. Supóngase que una variable aleatoria X tiene la siguiente distribución probabilística

x	1	2	3
f(x)	1/3	1/3	1/3

- Obtégase la media y varianza de la población de X.
- Sea \bar{X} la media de una muestra aleatoria de dos observaciones tomadas con reemplazo a partir de esta población. Obtégase la distribución muestral de \bar{X} y preséntese gráficamente.
- Obtégase la media y la varianza de X con base a la distribución muestral y verifíquese las ecuaciones (27) y (28).

SOLUCION

Los valores de la media y varianza de la población son

a)
$$\mu_X = E(x) = \sum_{i=1}^n x_i f(x_i) = 1(1/3) + 2(1/3) + 3(1/3) = 2$$

$$\sigma_X^2 = E(X^2) - E(X)^2 = 1^2(1/3) + 2^2(1/3) + 3^2(1/3) - 2^2 = 1/3 + 4/3 + 9/3 - 4 = 14/3 - 4 = 2/3$$

b) los valores posibles del promedio $\bar{x} = \frac{x_1 + x_2}{2}$ de dos observaciones (n=2) son $\bar{X} = \{1 \frac{1}{2}, 1, 2, 2 \frac{1}{2}, 3\}$

Explícitamente las muestras son S = {(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3)}

Entonces

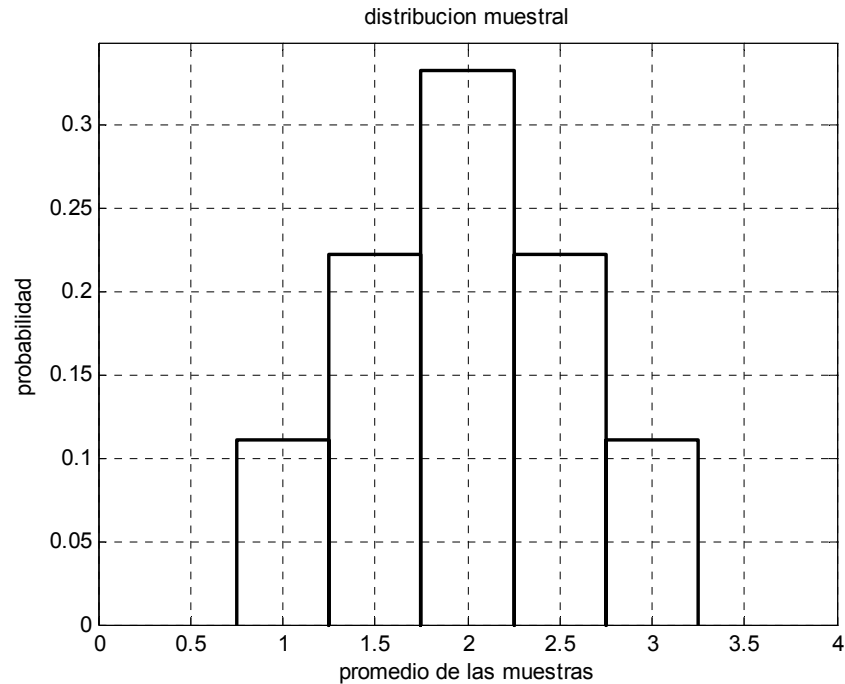
$$f(1) = P(\bar{X} = 1) = 1/9 \quad f(3/2) = P(\bar{X} = 3/2) = 2/9 \quad f(2) = P(\bar{X} = 2) = 3/9$$

$$f(5/2) = P(\bar{X} = 5/2) = 2/9 \quad f(3) = P(\bar{X} = 3) = 1/9$$

Por lo tanto la distribución de probabilidad para la media muestral \bar{X} es

\bar{x}	1	3/2	2	5/2	3
$f(\bar{x})$	1/9	2/9	3/9	2/9	1/9

Su gráfica respectiva se muestra a continuación



c)

$$\mu_X = E(\bar{X}) = \sum_{i=1}^n \bar{x}_i f(\bar{x}_i) = 1(1/9) + (3/2)(2/9) + 2(3/9) + (5/2)(2/9) + 3(1/9) = 2$$

$$\sigma_{\bar{X}}^2 = E(\bar{X}^2) - E(\bar{X})^2 = 1^2(1/9) + (3/2)^2(2/9) + 2^2(3/9) + (5/2)^2(2/9) + 3^2(1/9) - 2^2 = 13/3 - 4 = 1/3$$

Comparando los resultados $\mu_X = \mu_{\bar{X}} = 2$ y $\sigma_X^2 = \frac{\sigma_{\bar{X}}^2}{n} = (2/3)/2 = 1/3$

Lo cual verifica las ecuaciones (27) y (28)

25. Se sabe que la varianza de una variable aleatoria Y es 225. Si \bar{Y} es la media de una muestra aleatoria de 36 observaciones para , obténgase el error típico de \bar{Y} .

SOLUCION

Se sabe que $\sigma_Y^2 = 225$ y $n = 36$ observaciones, entonces utilizando la ecuación 28

$$\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n} \quad \text{ó}$$

$$\sigma_{\bar{Y}} = \sqrt{\frac{\sigma_Y^2}{n}} = \frac{\sqrt{\sigma_Y^2}}{\sqrt{n}} = \frac{\sqrt{225}}{\sqrt{36}} = 15/6$$

26. Sea X la duración en millas de cierta marca de neumáticos para automóvil. Supónganse que la media y desviación típica de X son, respectivamente, 30 000 y 200 mi. Si se selecciona una muestra aleatoria de 16 neumáticos, ¿cuáles serán el valor esperado y error típico de la media muestral?

SOLUCION

Tenemos una variable X, tiene media $\mu_X=30,000$, desviación típica $\sigma_X=200$ y el tamaño de la muestra es $n=16$

Entonces de las ecuaciones (27) y (28)

$$\mu_{\bar{X}} = \mu_X = 30,000 \text{ mi}$$

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{200}{\sqrt{16}} = \frac{200}{4} = 50 \text{ mi}$$

26. Cierta población tiene una media de 36 y una desviación típica de 5. Se extrae de esta población una muestra de 1000 y se calcula la media de la muestra.

- Obtégase el valor esperado de la media muestral.
- Obtégase el error típico de la media muestral.

SOLUCION

Tenemos una variable X, tiene media $\mu_X=36$, desviación típica $\sigma_X=5$ y el tamaño de la muestra es $n=1000$

Entonces de las ecuaciones (27) y (28)

$$\mu_{\bar{X}} = \mu_X = 36$$

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{5}{\sqrt{1000}} = 0.158$$

Unidad IV Distribuciones paramétricas

DISTRIBUCIONES DISCRETAS DE PROBABILIDAD

ENSAYO DE BERNOULLI

Un **Ensayo de Bernoulli**: es un experimento con dos resultados posibles uno llamado ÉXITO y el otro FRACASO. La variable aleatoria es X es tal que $X(\text{ÉXITO})=1$ y $X(\text{FRACASO})=0$, por otra parte, la probabilidad $P(X=1)=p$ y por lo tanto $P(X=0)=q=1-p$

La distribución de probabilidad del ensayo de Bernoulli se representa en la siguiente tabla

x	0	1
f(x)	q	p

MEDIA Y VARIANZA DEL MODELO DE BERNOULLI

A partir de la distribución de probabilidad se puede obtener su respectiva media y desviación típica

$$\mu = \sum x_i f(x_i) = (0)(q) + (1)(p) = p$$

entonces $\mu = p$

$$E(X^2) = \sum x_i^2 f(x_i) = (0)^2(q) + (1)^2(p) = p$$

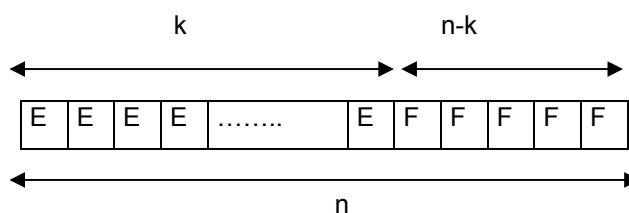
$$\sigma^2 = E(X^2) - \mu^2 = p - p^2 = p(1 - p) = pq$$

por lo tanto $\sigma = \sqrt{pq}$

DISTRIBUCION BINOMIAL

El experimento binomial consiste en n ensayos independientes de Bernoulli. Para cada ensayo la probabilidad de éxitos $P(E)=p$ y por lo tanto de fracaso es $P(F)=q=1-p$. La variable aleatoria del experimento es $X = \{\text{el número de éxitos en } n \text{ ensayos}\}$

Para el cálculo de la probabilidad en un caso general de el experimento binomial obsérvese el caso mostrado en la figura siguiente, donde se muestran k EXITOS y por lo tanto $n - k$ FRACASOS.



Se muestra solamente un resultado posible de el total de eventos que tienen k éxitos,

En número de eventos que contienen k éxitos se puede determinar utilizando las técnicas de conteo, esto es

$$N(k \text{ ÉXITOS}) = \frac{n!}{(n-k)! k!}$$

La probabilidad del evento individual mostrado se obtiene aplicando la condición de que cada ensayo de Bernoulli es independiente y por lo tanto su probabilidad es el producto de las probabilidades individuales

$$\begin{aligned} P(E E E \dots E F F F \dots F) &= P(E)P(E)P(E) \dots P(E)P(F)P(F)P(F) \dots P(F) \\ &= (p)(p)(p) \dots (p)(q)(q)(q) \dots (q) = p^k q^{n-k} \end{aligned}$$

Así pues la probabilidad de obtener $X = k$ éxitos en n ensayos es

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

Escribiendo el resultado anterior de otra forma

$$f(k) = \binom{n}{k} p^k q^{n-k}. \quad (4.1)$$

Por otra parte es conocido que el BINOMIO DE NEWTON tiene la forma:

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$$

de donde se observa inmediatamente que si se realiza el cambio de variable $a = p$ y $b = q$ se tiene que el término dado en la sumatoria es igual al obtenido en la ecuación (29), de ahí el nombre de la distribución binomial.

Por otra parte se puede verificar inmediatamente que (4.1) cumple con la propiedad

$$(p + q)^n = \sum_{k=0}^n \binom{n}{k} p^k q^{n-k}$$

$$1^n = \sum_{k=0}^n \binom{n}{k} p^k q^{n-k}$$

$$\sum_{k=0}^n \binom{n}{k} p^k q^{n-k} = 1$$

MEDIA Y VARIANZA DE LA DISTRIBUCIÓN BINOMIAL

No es fácil determinar la media y desviación típica de la distribución binomial directamente, pero se puede obtener aplicando las propiedades del valor esperado y la varianza para la suma de eventos independientes.

La variable aleatoria se puede representar mediante la suma de las variables aleatorias individuales de cada uno de los ensayos de Bernoulli

$$X = X_1 + X_2 + X_3 + \dots + X_n$$

Entonces para la media μ

$$\begin{aligned}\mu &= E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n) \\ &= p + p + \dots + p = np\end{aligned}$$

Por lo que

$$\mu = np \quad (4.2)$$

Y para la desviación típica

$$\begin{aligned}Var(X_1 + X_2 + \dots + X_n) &= Var(X_1) + Var(X_2) + \dots + Var(X_n) \\ &= pq + pq + \dots + pq = n pq\end{aligned}$$

Entonces

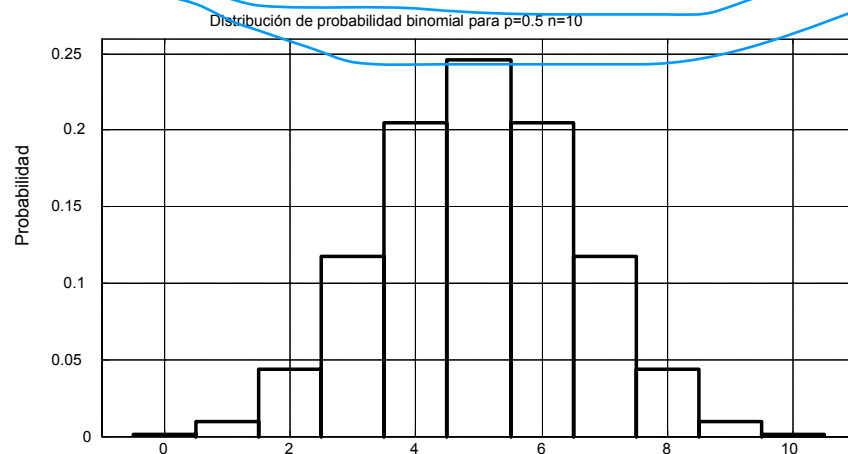
$$\sigma = \sqrt{n pq} \quad (4.3)$$

Los coeficientes binomiales dados por la ecuación (4.1) se pueden calcular mediante el uso de una calculadora o recurrir a las tablas donde se encuentran previamente evaluados.

Para el caso particular de $n = 10$ y $p = 0.5$ se tienen la siguiente distribución de probabilidad

x	0	1	2	3	4	5	6	7	8	9	10
f(x)	0.00098	0.00977	0.04395	0.11719	0.20508	0.24609	0.20508	0.11719	0.04395	0.00977	0.00098

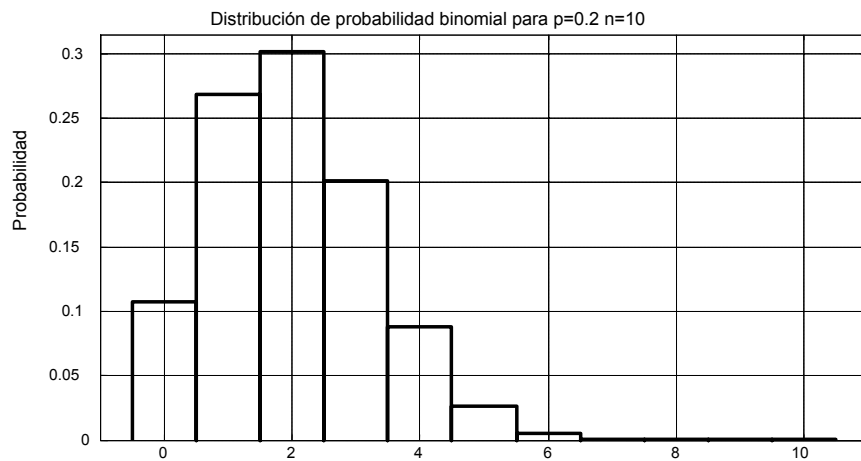
El histograma correspondiente muestra una distribución simétrica



Distribución binomial para $n = 10$ y $p = 0.5$

Para $n = 10$ y $p = 0.2$ se obtiene la siguiente distribución de probabilidad

x	0	1	2	3	4	5	6	7	8	9	10
f(x)	0.10737	0.26843	0.30198	0.20133	0.0880	0.02642	0.00550	0.00079	0.00007	0.0000	0.0000



Distribución binomial para $n = 10$ y $p = 0.2$

APLICACIONES DE LA DISTRIBUCIÓN BINOMIAL

EJEMPLOS

1. Obténganse los valores de las siguientes expresiones.

a. $C_1^3 (0.4)^1 (0.6)^2$ b. $C_2^4 (0.7)^2 (0.3)^2$

SOLUCION

a) $C_1^3 (0.4)^1 (0.6)^2 = \frac{3!}{1!(3-1)!} = (0.4)^1 (0.6)^2 = 0.2492$

b) $C_2^5 (0.6)^2 (0.4)^3 = \frac{5!}{2!(5-2)!} = (0.6)^2 (0.4)^3 = 0.2304$

2. Obténganse los valores de las siguientes expresiones.

a. $\sum_{x=0}^1 C_x^3 (0.5)^x (0.5)^{3-x}$

b. $\sum_{x=0}^2 C_x^5 (0.5)^x (0.5)^{5-x}$

c. $P(X \leq 2 | n = 5 \text{ y } p = 0.5)$

SOLUCION

$$b) \sum_{x=0}^2 C_x^5 (0.5)^x (0.5)^{5-x} = C_0^5 (0.5)^0 (0.5)^5 + C_1^5 (0.5)(0.5)^4 + C_2^5 (0.5)^2 (0.5)^3$$

$$= 0.03125 + 0.15625 + 0.3125 = 0.5000$$

$$c) P(X \leq 2, n = 5 \text{ y } P = (0.5)) = \sum_{x=0}^2 C_x^5 (0.5)^x (0.5)^{5-x} = 0.5000$$

3. Supóngase que en una prueba se incluyen diez preguntas de opción múltiple, con cinco respuestas para cada pregunta, de las cuales una es correcta. Si una estudiante responde las preguntas simplemente adivinando, ¿cuál es la probabilidad de que

- conteste correctamente cinco preguntas;
- conteste correctamente tres o menos preguntas;
- conteste correctamente cinco o más preguntas?

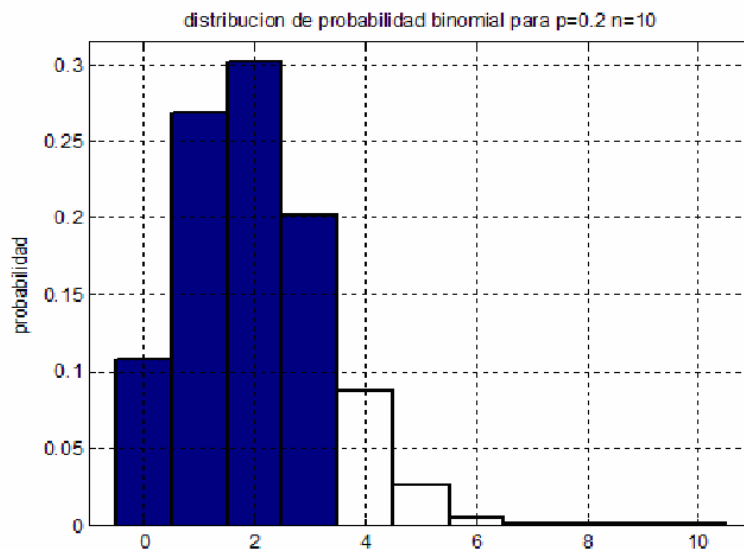
SOLUCION

Puesto que son diez preguntas $n = 10$ y debido a que se contesta al azar y cada pregunta contiene cinco posibles respuestas de las cuales solo una es correcta la probabilidad de ÉXITO es $p = 1/5 = 0.2$ y por lo tanto la de FRACASO $q = 1 - 1/5 = 4/5 = 0.8$

Para obtener la evaluación de cada una de las preguntas se puede recurrir a las tablas correspondientes de la distribución binomial

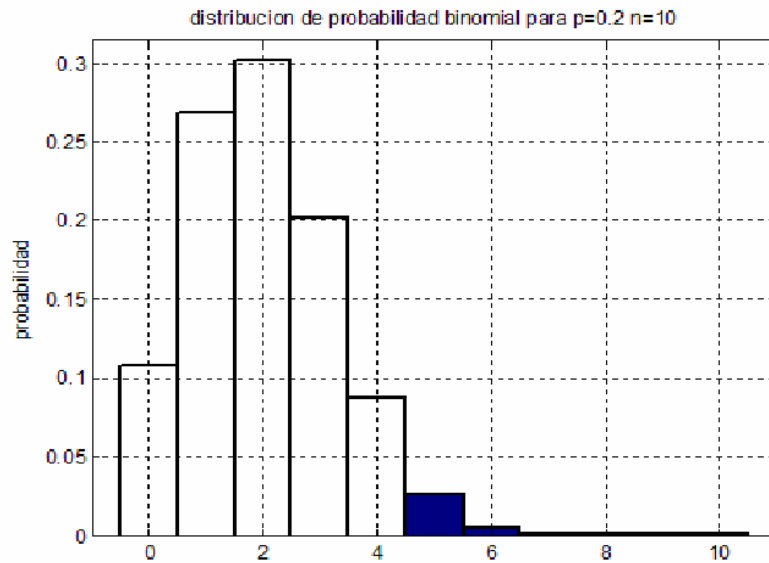
$$a) P(X = 5, n = 10, p = 0.2) = C_5^{10} (0.2)^5 (0.8)^5 = 0.02642$$

$$b) P(X \leq 3, n = 10, p = 0.2) = \sum_{x=0}^3 C_x^{10} (0.2)^x (0.8)^{10-x} = 0.87913$$



$$c) P(5 \leq X, n = 10, p = 0.2) = 1 - P(X < 5, n = 10, p = 0.2) = 1 - \sum_{x=0}^4 C_x^{10} (0.2)^x (0.8)^{10-x} =$$

$$= 1 - 0.96721 = 0.03279$$



4. Supóngase que diez aparatos de radar están operando independientemente uno del otro, y que la probabilidad de que uno solo de los aparatos detecte un cohete enemigo es de 0.80. ¿Cuál es la probabilidad de que nueve aparatos de radar detecten el cohete?

SOLUCION

De los datos proporcionados por el problema $n = 10$ y la probabilidad de ÉXITO es $p = 0.8$ y la de FRACASO $q = 1 - p = 1 - 0.80 = 0.20$

La pregunta se refiere a que nueve de los aparatos exactamente tengan éxito en detectar el cohete enemigo esto es $k = 9$, entonces

$$P(k = 9, n = 10, p = 0.8) = C_{10}^9 (0.8)^9 (0.20)^1 = 0.26844$$

5. Si se sabe que el 90% de los estudiantes que toman un curso elemental de economía aprueban, ¿cuál es la probabilidad de que al menos 3 estudiantes en una clase de 15 no aprueben el curso?

SOLUCION

Para este problema $n = 15$ la probabilidad de éxito es $p = 0.9$ y de fracaso $q = 1 - p = 1 - 0.9 = 0.1$

La pregunta se puede traducir al lenguaje simbólico como

$$P(3 \leq k, n = 15, p = 0.8) = \sum_{k=3}^{15} C_{15}^k (0.9)^k (0.2)^{15-k}$$

Puesto que las tablas de distribución binomial acumulada dan la sumatoria empiezan en cero, se puede transformar la expresión anterior al complemento

$$P(3 \leq k, n = 15, p = 0.8) = 1 - P(0 \leq k < 3, n = 15, p = 0.8) = 1 - \sum_{k=0}^2 C_{15}^k (0.9)^k (0.2)^{15-k}$$

$$= 1 - 0.81594 = 0.18406$$

6. De la clase del último semestre, 60% son muchachas. ¿Cuál es la probabilidad de que en un grupo de 10 estudiantes seleccionados aleatoriamente de esta clase haya

- a. cinco muchachas;
- b. al menos 5 muchachas;
- c. cuando más 5 muchachas;
- d. entre 4 y 6 muchachas, inclusive?

SOLUCION

La clase corresponde a $n = 10$ estudiantes con probabilidad de ser muchachas $p = 0.6$ y la de muchachos $q = 1 - p = 1 - 0.60 = 0.40$

Traduciendo correctamente cada una de las preguntas al lenguaje matemático

- a) $P(X = 5, n = 10, p = 0.4) = 0.20066$
- b) $P(5 \leq X, n = 10, p = 0.4) = 1 - P(X \leq 4, n = 10, p = 0.4) = 1 - 0.16624 = 0.83376$
- c) $P(X \leq 5, n = 10, p = 0.4) = 0.36640$
- d) $P(4 \leq X \leq 6, n = 10, p = 0.4) = P(X \leq 6, n = 10, p = 0.4) - P(X \leq 3, n = 10, p = 0.4)$
 $= 0.61772 - 0.05476 = 0.56296$

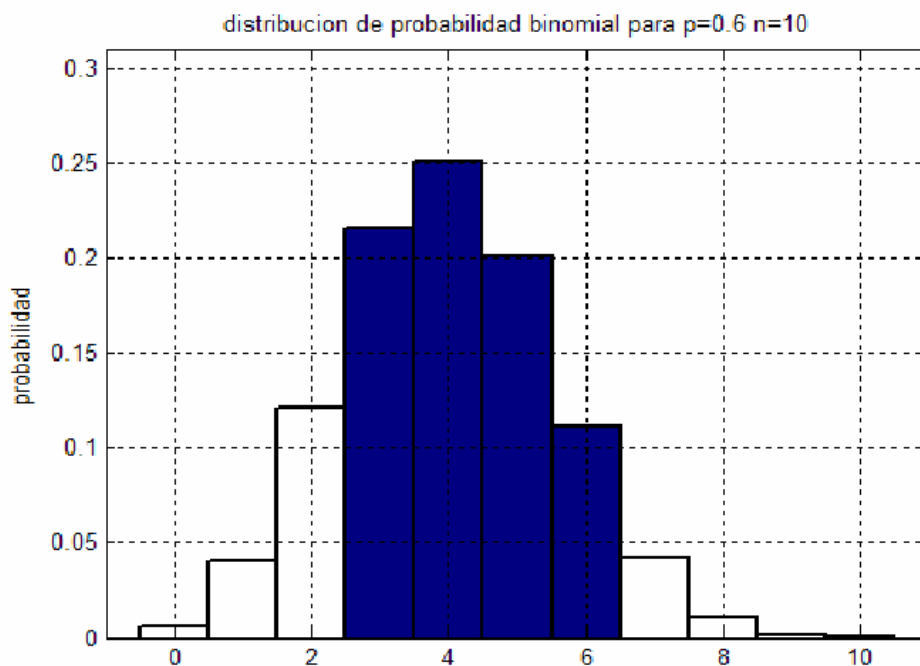


Figura. La figura muestra la interpretación gráfica del inciso d)

7. Supóngase que la probabilidad de que al tirar un dado quede hacia arriba un número no de puntos es 0.4: ¿Cuál es la probabilidad de que en cinco tiradas del dado el número de veces que aparezca un número no de puntos sea

- menos de dos;
- más de dos;
- entre dos y cuatro, inclusive?

SOLUCION

El número de tiradas es $n = 5$ y la probabilidad de que quede un número no es $p = 0.4$, entonces la probabilidad de que quede un número par es $q = 1 - p = 1 - 0.4 = 0.6$

- $p(X < 2, n = 5, p = 0.4) = p(X \leq 1, n = 5, p = 0.4) = 0.33696$
- $p(X > 2, n = 5, p = 0.4) = 1 - p(X \leq 2, n = 5, p = 0.4) = 1 - 0.68256 = 0.31744$
- $p(2 \leq X \leq 4, n = 5, p = 0.4) = p(0 \leq X \leq 4, n = 5, p = 0.4) - p(X \leq 1, n = 5, p = 0.4)$
 $= 0.98976 - 0.33696 = 0.6528$

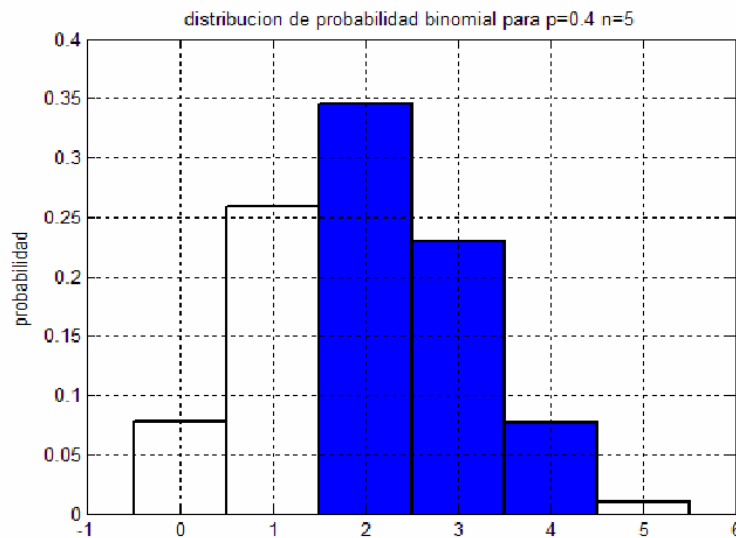


Figura. La figura muestra la interpretación gráfica del inciso c)

8. Considérese que el 50% de todos los empleados de una gran compañía están casados. Sea X el número de empleados casados en una muestra aleatoria de empleados. Obténganse la media y desviación típica de X.

SOLUCION

La probabilidad de estar casado es $p = 0.5$ y el número de empleados es $n=100$
 Aplicando directamente las ecuaciones (30) y (31)

$$\mu = np = 100(0.5) = 50$$

$$\sigma^2 = npq = 100(0.5)(1-0.5) = 25$$

$$\sigma = \sqrt{25} = 5$$

9. De acuerdo con los registros de producción de cierta compañía, el 10% de tornillos producidos por cierta máquina son defectuosos. Obténganse la media y la desviación típica para X si ésta es el número de tornillos defectuosos en cualquier muestra aleatoria de tamaño 100.

SOLUCION

Como la variable aleatoria es el número de tornillos defectuosos en la muestra $n = 100$, la probabilidad "éxito" en este caso es $p = 0.1$

Aplicando directamente las ecuaciones (30) y (31)

$$\mu = np = 100(0.1) = 10$$

$$\sigma^2 = npq = 100(0.1)(1-0.1) = 9$$

$$\sigma = \sqrt{9} = 3$$

DISTRIBUCIÓN CONTINÚA DE PROBABILIDAD

DISTRIBUCIÓN NORMAL

Es una distribución continua descrita por la siguiente función de probabilidad

$$p(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- Se aplica a MEDICIONES de cantidades físicas continuas como longitud, masa, tiempo, voltaje corriente, energía, temperatura, etc.
- Es la aproximación de TEOREMA DE LIMITE CENTRAL
- Es una aproximación de la distribución binomial para $n \geq 35$ y $p \approx 0.5$

La distribución Normal depende de dos parámetros el valor esperado o media μ y la desviación típica σ , Por lo que para cada uno de los valores de estos parámetros se tiene una gráfica diferente, pero todas estas

$$N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (4.4)$$

La variación del parámetro μ ocasiona un desplazamiento de la gráfica a la izquierda para valores negativos y a la derecha para valores positivos. La Figura siguiente muestra el efecto descrito para las graficas de la distribución normal con desviación típica $\sigma = 1$, y tres diferentes medias $\mu = -2$ $\mu = 0$ y $\mu = 2$.

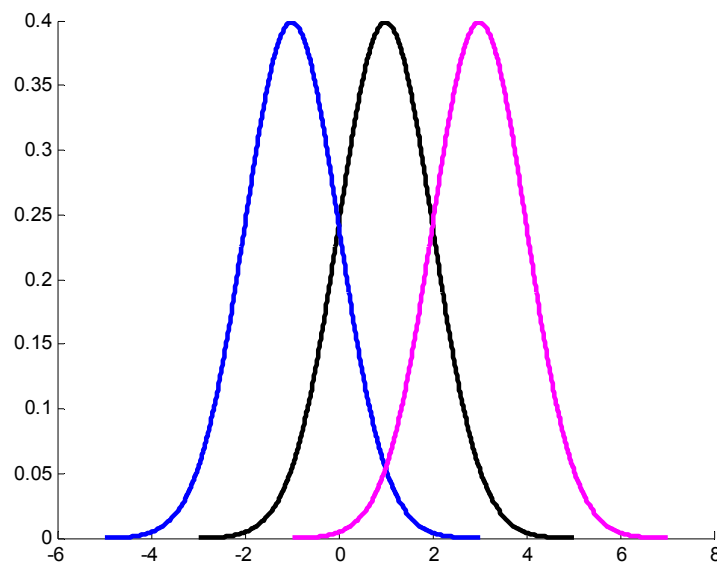


Figura. Efecto de desplazamiento para $\sigma = 1$ $\mu = -2$, $\mu = 0$ y $\mu = 2$

Por otra parte la variación del parámetro σ hace que la altura y la anchura de la distribución de probabilidad cambien, esto es, si σ es grande la distribución será más ancha (más dispersa) y su altura disminuirá, pero si σ es pequeña su anchura disminuirá (más concentrada) y su altura será más grande.

La siguiente figura muestra el efecto de modificar la desviación típica para una media dada $\mu = 0$, y tres diferentes desviaciones $\sigma = 1$, $\sigma = 4$ y $\sigma = \frac{1}{2}$.

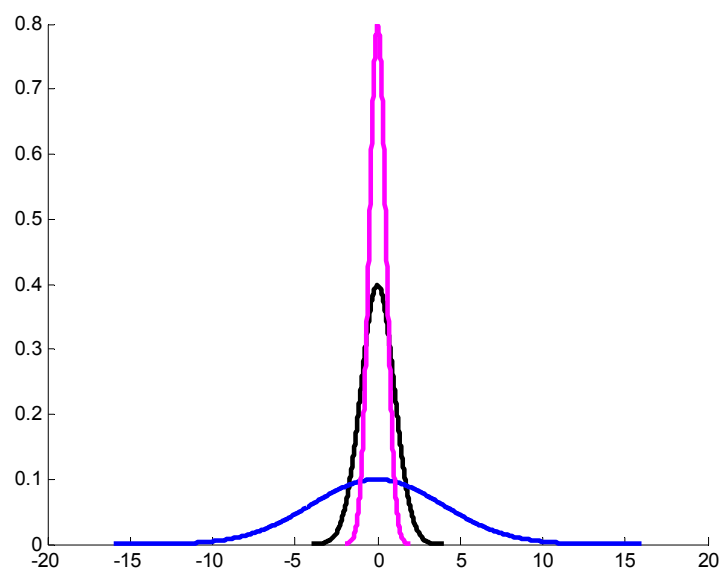


Figura. Efecto de estiramiento o estrechamiento para $\mu = 0$, $\sigma = 1$, $\sigma = 4$ y $\mu = \frac{1}{2}$

La probabilidad de que la variable aleatoria X tome un conjunto de valores en un intervalo (a, b) se obtiene a partir de la siguiente integral

$$p(a < X < b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dy \quad (4.5)$$

La figura siguiente muestra la gráfica del área bajo la distribución normal en un intervalo (a, b)

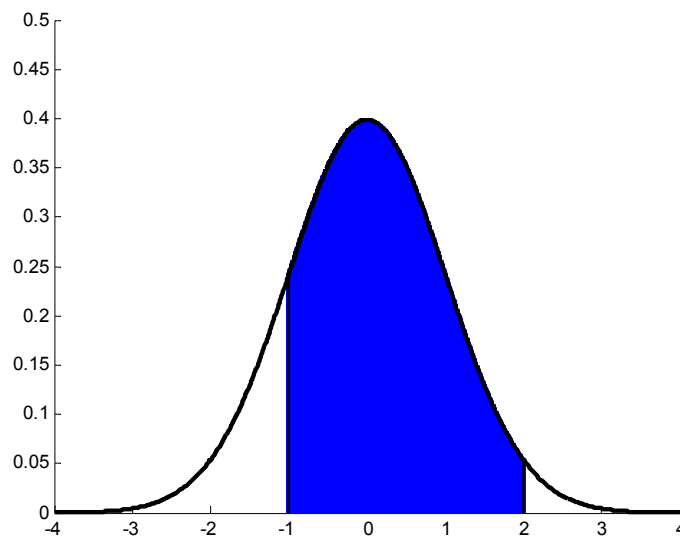


Figura. Área bajo la curva normal en un intervalo (a, b)

Resulta que la integral anterior no tiene primitiva, esto es, no existe una función cuya derivada de cómo resultado la función de distribución normal dada por la ecuación (32). Por lo que la integral anterior se obtiene mediante integración numérica ó series. El problema anterior de determinar la probabilidad en un intervalo conduce a la elección de una distribución normal representativa la cual es conocida como distribución normal estándar.

Distribución normal estándar

La distribución normal estándar es aquella en la cual se tiene que $\mu = 0, \sigma = 1$, por lo que la ecuación (4.4) y (4.5) se transforman en

$$N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad (4.6)$$

$$\int_a^b N(0,1) dx = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}x^2} dx \quad (4.7)$$

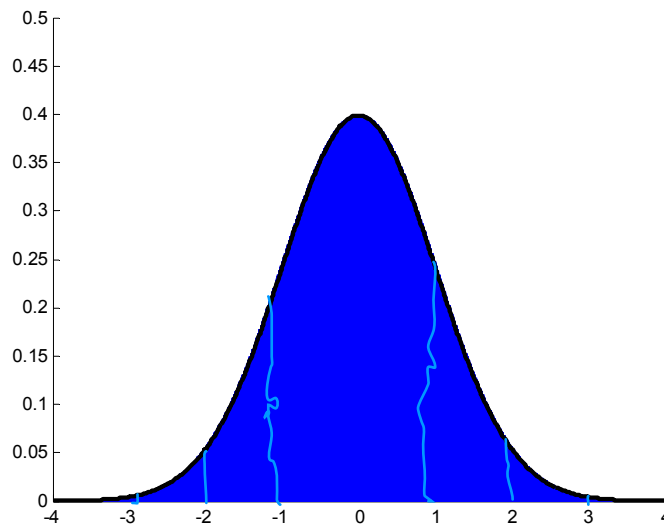
Cualquier distribución normal con media μ y desviación típica σ puede ser relacionada con la distribución normal mediante el cambio de variable

$$Z = \frac{x - \mu}{\sigma}$$

(4.8)

La variable Z es conocida con *variable tipificada*

El área bajo la curva normal estándar se puede consultar en tablas respectivas para los valores más comúnmente utilizados. Las tablas disponibles en general solo abarcan un rango para la variable tipificada de $-3.4 \leq Z \leq 3.4$, esto es debido a que la probabilidad de valores de Z mayores que 3.4 y menores que -3.4 tienen una probabilidad muy baja, y la probabilidad el área o bajo la curva normal estándar es prácticamente 1.



El área bajo la distribución normal estándar en el intervalo $-3.4 \leq Z \leq 3.4$ es prácticamente 1.

APLICACIONES DE LA DISTRIBUCIÓN BINOMIAL

EJEMPLOS

10. Obténganse las siguientes probabilidades.

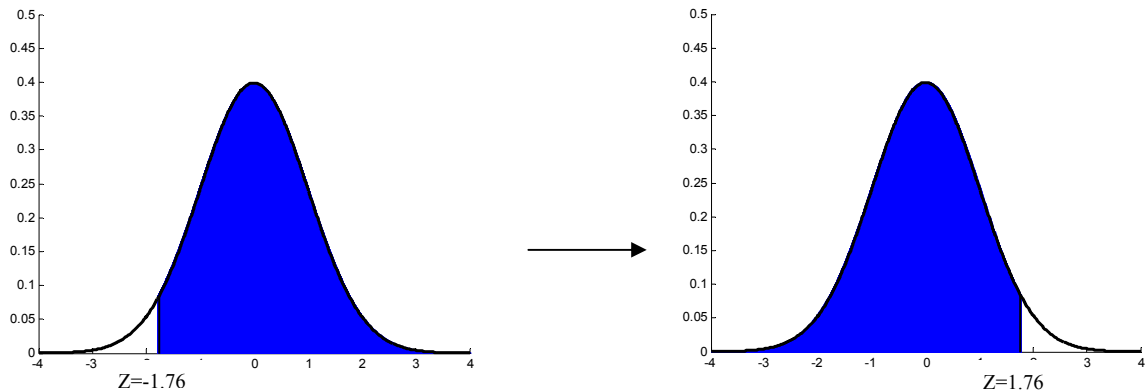
- | | |
|---------------------------|-------------------------|
| a. $P(Z < 2.0)$ | b. $P(Z < 1.45)$ |
| c. $P(Z > -1.76)$ | d. $P(Z > -1.65)$ |
| e. $P(1.0 < Z < 1.89)$ | f. $P(-1.4 < Z < 1.75)$ |
| g. $P(-2.15 < Z < -0.55)$ | |

SOLUCION

Lo valores de los incisos a) y b) se obtiene directamente de la tabla del área bajo la curva de la distribución normal.

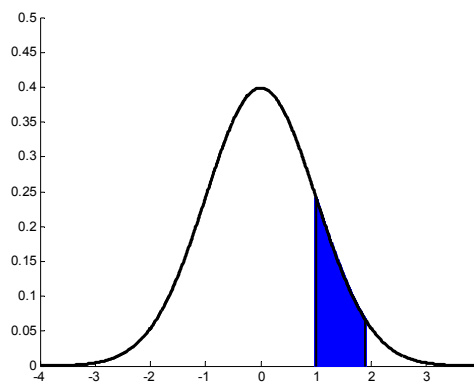
- a) $p(Z < 2.00) = 0.9772$
 b) $p(Z < 1.45) = 0.9265$

Para los incisos c) y d) se procede como se indica a continuación. El área para valores de Z mayores que un número negativo es equivalente al área por debajo del valor absoluto de Z , en la cual se utiliza la simetría de la distribución normal. Lo anterior es mostrado en la figura siguiente.

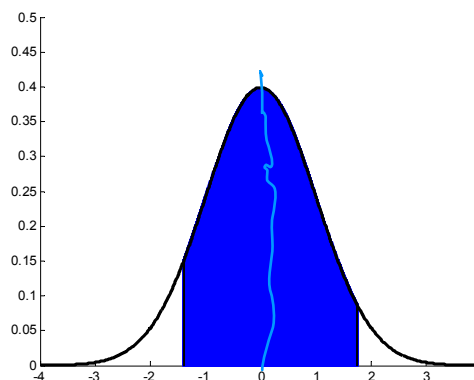


- c) $p(Z > -1.76) = p(Z < 1.76) = 0.9608$
 d) $p(Z > -1.65) = p(Z < 1.65) = 0.9505$

En el inciso e) la probabilidad solicitada es igual al área entre los valores $Z_1 = 1.00$ y $Z_2 = 1.89$, que de acuerdo a la figura y a la tabla se puede obtener mediante la diferencia de áreas



- e) $p(1.0 < Z < 1.89) = p(Z < 1.89) - p(Z < 1) = 0.9706 - 0.8413 = 0.1293$
 f) El área buscada es mostrada en la figura siguiente:



Se puede descomponer en la suma de dos áreas, el área comprendida de -1.40 a 0 mas el área de 0 a 1.75. Para calcular la primera área se utiliza la simetría de la distribución normal esto es

$$P(-1.40 < Z \leq 0) = P(0 \leq Z < 1.40) = P(Z < 1.40) - 0.50$$

Para la segunda área se procede de manera semejante

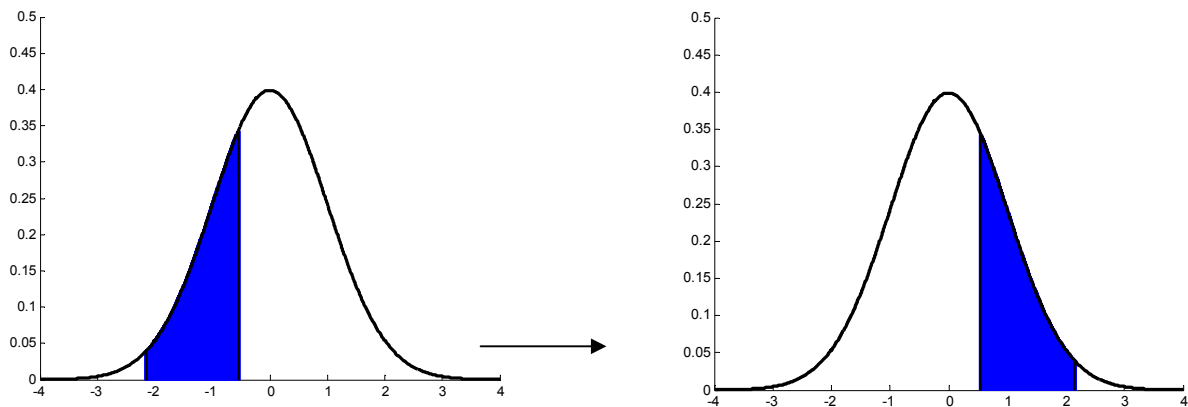
$$P(0 \leq Z < 1.75) = P(Z < 1.75) - 0.50$$

Entonces sumando las áreas

$$\begin{aligned} P(-1.40 < Z < 1.75) &= P(Z < 1.40) - 0.50 + P(Z < 1.75) - 0.50 = P(Z < 1.40) + P(Z < 1.75) - 1 \\ &= 0.9192 + 0.9599 - 1.0000 = 0.8792 \end{aligned}$$

g) Utilizando la simetría de la normal el problema es equivalente a

$$\begin{aligned} P(-2.15 < Z < -0.55) &= P(0.55 < Z < 2.15) = P(Z < 2.15) - P(Z < 0.55) \\ &= 0.9842 - 0.7088 = 0.2754 \end{aligned}$$



11. Obténgase el valor de Z para cada una de las siguientes áreas bajo la curva normal estándar.

- A la izquierda de Z el área es 0.9949
- A la izquierda de Z el área es de 0.9951
- A la derecha de Z el área es de 0.005.
- A la izquierda de Z el área es de 0.9412.
- A la izquierda de Z el área es de 0.0582.
- A la derecha de Z el área es de 0.2810.
- A la derecha de z el área es de 0.0228.

SOLUCION

- Se busca en la tabla el valor del área respectiva $a = 0.9949$ que corresponde a $Z = 2.57$.
- procediendo de igual que el inciso anterior para $a = 0.9951$ $Z = 2.58$.

- Se requiere el valor de área a la izquierda, por complemento este valor es $a = 1 - 0.005 = 0.9950$

En la tabla no existen el valor exacto de Z que conduzca al área $= 0.9950$, los valores más aproximados de Z son $Z_1 = 2.57$ que conduce a $a_1 = 0.9949$ y $Z_2 = 2.58$ que $a_2 = 0.9951$, entonces el valor de Z buscado se encuentra entre estos dos valores de Z ya que el área solicitada se encuentra entre las dos áreas $a = 0.9950$.

Como los valores son muy cercanos se puede aproximar el resultado pensando que la relación es lineal, esto es

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1} (x - x_1)$$

donde $x_1 = a_1$ = área 1 correspondiente a $y_1 = Z_1$ y $x_2 = a_2$ = área 2 correspondiente a $y_2 = Z_2$, entonces

$$Z - Z_1 = \left(\frac{Z_2 - Z_1}{a_2 - a_1} \right) (a - a_1)$$

Despejando a y y sustituyendo a $x = a$

$$Z = \left(\frac{Z_2 - Z_1}{a_2 - a_1} \right) (a - a_1) + Z_1 = \left(\frac{2.58 - 2.57}{0.9951 - 0.9949} \right) (0.9950 - 0.9949) + 2.57 = 2.575$$

d) Buscando en la tabla los valores más cercanos a el área $a = 0.9412$ son $Z_1 = 1.56$ con $a_1 = 0.9406$ y $Z_2 = 1.57$ con $a_2 = 0.9418$. Utilizando el resultado anterior

$$Z = \left(\frac{Z_2 - Z_1}{a_2 - a_1} \right) (a - a_1) + Z_1 = \left(\frac{1.57 - 1.56}{0.9418 - 0.9406} \right) (0.9412 - 0.9406) + 1.56 = 1.565$$

e) Los valores de áreas menores que 0.5 en la tabla corresponden a valores negativos de Z , el problema se puede cambiar por el valor positivo pero para el área $= 1 - 0.0582 = 0.9418$ que buscando en la tabla corresponde a $Z = 1.57$, por lo tanto el resultado es $Z = -1.57$.

f) Aplicando el complemento $a = 1 - 0.2810 = 0.7190$, buscando en las tablas el valor correspondiente es $Z = 0.58$

g) Aplicando el complemento $a = 1 - 0.0228 = 0.9772$, buscando en las tablas el valor correspondiente es $Z = 2.00$

12. Una variable aleatoria (X) se distribuye normalmente, con una media de 100 y una desviación típica de 15. Obténgase la probabilidad de que

- a. X sea menor de 80.5; b. X sea mayor de 116.5;
- c. X sea menor de 112; d. X esté entre 91 y 109;
- e. X esté entre 85 y 97.

SOLUCION

Para el problema $\mu = 100$ y $\sigma = 15$

$$a) \quad p(X < 80.5) = P\left(Z < \frac{80.5 - 100}{15}\right) = P(Z < -1.30) = 1 - P(Z < 1.30) = 1 - 0.9032 = 0.0968$$

$$b) \quad p(X > 116.5) = P\left(Z > \frac{116.5 - 100}{15}\right) = P(Z > 1.1) = 1 - P(Z < 1.1) = 1 - 0.8643 = 0.1357$$

$$c) \quad p(X < 112) = P\left(Z < \frac{112 - 100}{15}\right) = P(Z < 0.8) = 0.7881$$

$$\begin{aligned} \text{d) } P(91 < X < 109) &= P\left(\frac{91-100}{15} < Z < \frac{109-100}{15}\right) = P(-0.6 < Z < 0.6) \\ &= 2 * (0.7257) - 1 = 0.4515 \end{aligned}$$

$$\begin{aligned} \text{e) } P(85 < X < 97) &= P\left(\frac{85-100}{15} < Z < \frac{97-100}{15}\right) = P(-1 < Z < -0.2) \\ &= P(Z < -1) - P(Z < -0.2) = 0.8413 - 0.5793 = 0.2620 \end{aligned}$$

13. Una variable aleatoria (X) se distribuye normalmente con media 70 y desviación típica de 5. Obténgase la probabilidad de que

- a. X sea mayor de 66;
- b. X sea mayor de 63;
- c. X sea mayor de 71 y menor de 75;
- d. X sea mayor de 79 o menor de 61.

SOLUCION

Para todos los incisos $\mu=70$, $\sigma=5$ y el cambio de variable a la variable tipificada se realiza mediante

$$Z = \frac{X - \mu}{\sigma}$$

$$\text{a) } P(X > 66) = P\left(Z > \frac{66-70}{5}\right) = P(Z > -0.8) = P(Z < 0.8) = 0.7881$$

$$\text{b) } P(X > 63) = P\left(Z > \frac{63-70}{5}\right) = P(Z > -1.4) = P(Z < 1.4) = 0.9192$$

$$\begin{aligned} \text{c) } P(71 < X < 75) &= P\left(\frac{71-70}{5} < Z < \frac{75-70}{5}\right) = P(0.2 < Z < 1) = P(Z < 1) - P(Z < 0.2) \\ &= 0.8413 - 0.5793 = 0.2620 \end{aligned}$$

$$\begin{aligned} \text{d) } P(X > 79) + P(X < 61) &= P\left(Z > \frac{79-70}{5}\right) + P\left(Z < \frac{61-70}{5}\right) = P(Z > 1.8) + P(Z < -1.8) \\ &= 2(1 - P(Z < 1.8)) = 2(1 - 0.9641) = 0.0718 \end{aligned}$$

14. Un profesor de inglés ha determinado que el tiempo necesario para que los estudiantes concluyan un examen final se distribuye normalmente con media de 110 min y desviación típica de 10 min.

- a. ¿Cuál es la probabilidad de que un estudiante de inglés elegido aleatoriamente concluya el examen en menos de dos horas?
- b. ¿Cuál es la probabilidad de que un estudiante de inglés seleccionado aleatoriamente concluya el examen en 125 min o más?
- e. Si hay 50 estudiantes en la clase, ¿cuántos de ellos concluirán el examen antes de una hora ~~50~~ minutos?

SOLUCION

La media y la desviación típica son $\mu=110$ y $\sigma=10$

- a) Dos horas representan 120 minutos, entonces

$$P(X < 120) = P(Z < (120 - 110)/10) = P(Z < 1) = 0.8413$$

b) Si el estudiante debe resolver el examen en 125 o más

$$P(125 \leq X) = P(Z \leq (125 - 110)/10) = P(1.5 \leq Z) = 1 - P(Z < 1.5) = 1 - 0.9332 = 0.0668$$

c) Primero se debe determinar la probabilidad de que los alumnos terminen antes de 110 min.

$$P(X \leq 110) = P(Z \leq (110 - 110)/10) = P(Z \leq 0) = 0.5$$

Entonces el número de alumnos que terminen antes de 110 min es $n = N \cdot P(X \leq 110) = (50)(0.5) = 25$

15. Supóngase que la longitud promedio de la estancia de los pacientes en cierto hospital es de diez días y la desviación típica es de dos días. Considérese que tales duraciones se distribuyen normalmente.

a. ¿Cuál es la probabilidad de que el siguiente paciente que se reciba permanezca más de nueve días?

b. Si el día de hoy se admitieron 200 pacientes, ¿cuántos continuarán en el hospital dentro de dos semanas?

SOLUCION

La media y la desviación típica son $\mu = 10$, $\sigma = 2$

$$a) \quad P(X \geq 9) = P(Z \geq (9 - 10)/2) = P(Z \geq -0.5) = P(Z \leq 0.5) = 0.6915$$

$$b) \quad N = 200, X = 2 \text{ semanas} = 14 \text{ días}$$

$$P(X \geq 14) = P(Z \geq (14 - 10)/2) = P(Z \geq 2) = 1 - P(Z < 2) = 1 - 0.9772 = 0.0228$$

Entonces el número de pacientes después de dos semanas es $n = N \cdot P(X \geq 14) = (200)(0.0228) = 4.56$

16. Supóngase que las calificaciones de prueba de un examen estándar se distribuyan normalmente, ¿Cuál es el valor aproximado correspondiente al percentil 75 -ésimo?

SOLUCION

El percentil corresponde a el porcentaje del área total, entonces $P(Z \leq Z_0) = 0.75$

Buscando en la tabla los valores más cercanos a el área $a = 0.75$ son $Z_1 = 0.67$ con $a_1 = 0.7486$ y $Z_2 = 0.68$ con $a_2 = 0.7517$. la aproximación lineal

$$Z = \left(\frac{Z_2 - Z_1}{a_2 - a_1} \right) (a - a_1) + Z_1 = \left(\frac{0.68 - 0.67}{0.7517 - 0.7486} \right) (0.7500 - 0.7486) + 0.67 = 0.6745$$

TEOREMA DEL LÍMITE CENTRAL

El **teorema del límite central** establece que si X es cualquier variable aleatoria con media μ y desviación típica σ la distribución de la media muestral \bar{X} será aproximadamente normal con media

$\mu_{\bar{X}} = \mu_X = \mu$ y desviación típica $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$ sin importar la forma de la distribución de probabilidad de X siempre y cuando el tamaño de la muestra sea grande $n > 30$

Por lo anterior la variable tipificada para determinar la probabilidad de la variable aleatoria \bar{X} es

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (4.9)$$

EJEMPLOS

17. Supóngase que la distribución de las edades de los empleados de una gran compañía tiene una media de 35 años y una desviación típica de 6 años. Se considera que la distribución no es normal. Si se selecciona una muestra aleatoria de 36 empleados, y se calcula su edad promedio, ¿cuál es la probabilidad de que la edad promedio de la muestra sea

- a. de más de 37.5 años; b. de menos de 33 años;
c. de entre 34.25 y 34.75 años; d. de entre 36 y 37.75 años?

SOLUCION

La media y desviación típica de la población es $\mu=35$, $\sigma=6$ y el tamaño de la muestra $n = 36$

$$a) P(37.5 < \bar{x}) = P\left(\frac{37.5 - 35}{6/\sqrt{36}} < Z\right) = P(Z > 2.5) = 1 - P(Z < 2.5) = 1 - 0.9958 = 0.0042$$

$$b) P(\bar{x} < 33) = P\left(\frac{33 - 35}{6/\sqrt{36}} < Z\right) = P(Z < -2) = 1 - P(Z < 2) = 1 - 0.9772 = 0.0228$$

$$c) P(34.25 < \bar{x} < 34.75) = P\left(\frac{34.25 - 35}{6/\sqrt{36}} < Z < \frac{34.75 - 35}{6/\sqrt{36}}\right) = P(-0.75 < Z < -0.25) \\ = P(0.25 < Z < 0.75) = P(Z < 0.75) - P(Z < 0.25) = 0.7734 - 0.5987 = 0.1747$$

$$d) P(36 < \bar{x} < 37.75) = P\left(\frac{36 - 35}{6/\sqrt{36}} < Z < \frac{37.75 - 35}{6/\sqrt{36}}\right) = P(1 < Z < 2.75) \\ = P(2 < Z < 2.75) = P(Z < 2.75) - P(Z < 1) = 0.9970 - 0.8413 = 0.1557$$

18. La distribución de los 10 dígitos aleatorios 0, 1, 2, ..., y 9 se considera como uniforme, ya que la probabilidad de que aparezca cada dígito es de 0.1. Supóngase que se selecciona una muestra aleatoria de 100 dígitos, ya sea utilizando la tabla de dígitos aleatorios o mediante el método de la urna con reemplazo, y se calcula una media muestral. Obténganse las siguientes probabilidades.

- a. $P(\bar{x} < 4.84)$ b. $P(\bar{x} > 4.79)$
c. $P(4.18 < \bar{x} < 4.87)$ d. $P(4.00 < \bar{x} < 4.90)$

SOLUCION

Para la distribución uniforme

x	0	1	2	3	4	5	6	7	8	9
f(x)	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10

Por lo tanto la media μ y la desviación típica σ poblacionales son

$$\mu = E(x) = \sum x_i f(x_i) = 0(1/10) + 1(1/10) + 2(1/10) + 3(1/10) + 4(1/10) + 5(1/10) + 6(1/10) + 7(1/10) + 8(1/10) + 9(1/10) = 4.5$$

$$E(x^2) = \sum x_i^2 f(x_i) = 0^2 (1/10) + 1^2 (1/10) + 2^2 (1/10) + 3^2 (1/10) + 4^2 (1/10) + 5^2 (1/10) + 6^2 (1/10) + 7^2 (1/10) + 8^2 (1/10) + 9^2 (1/10) = 28.5$$

$$\sigma^2 = E(x^2) - \mu^2 = 28.5 - (4.5)^2 = 8.25$$

$$\sigma = \sqrt{8.25} = 2.87$$

Entonces para una muestra $n=100$

$$a) P(\bar{x} < 4.84) = P\left(Z < \frac{4.84 - 4.5}{2.87/\sqrt{100}}\right) = P(Z < 1.19) = 0.8830$$

$$b) P(\bar{x} > 4.79) = P\left(\frac{4.79 - 4.5}{2.87/\sqrt{100}} > Z\right) = P(1.01 > Z) = 1 - P(Z \leq 1.01) = 1 - 0.8438 = 0.1562$$

$$c) P(4.18 < \bar{x} < 4.57) = P\left(\frac{4.18 - 4.5}{2.87/\sqrt{100}} < Z < \frac{4.57 - 4.5}{2.87/\sqrt{100}}\right) = P(-1.11 < Z < 1.29) \\ = P(Z < 1.29) - P(Z < -1.11) = 0.9015 - 0.1357 = 0.7658$$

$$d) P(4.00 < \bar{x} < 4.90) = P\left(\frac{4.00 - 4.5}{2.87/\sqrt{100}} < Z < \frac{4.90 - 4.5}{2.87/\sqrt{100}}\right) = P(-1.74 < Z < 1.39) \\ = P(Z < 1.39) - P(Z < -1.74) = 0.9177 - 0.0411 = 0.8766$$

19. Supóngase que a fin de mes los saldos de las cuentas de cheques en bancos se distribuyen normalmente con media \$250 y desviación típica \$15.

a. ¿Cuál es la probabilidad de que una cuenta seleccionada aleatoriamente tenga un saldo de más de \$272.50?

b. ¿Cuál es la probabilidad de que el promedio de una muestra aleatoria de 25 cuentas sea de más de \$257.50?

SOLUCION

De el problema se obtiene que $\mu=250$, $\sigma= 15$ y $n = 25$

$$a) P(272.5 < x) = P\left(\frac{272.5 - 250}{15} < Z\right) = P(1.5 < Z) = 1 - P(Z \leq 1.5) = 1 - 0.9332 = 0.0668$$

$$b) P(257.5 < \bar{x}) = P\left(Z < \frac{257.5 - 250}{15/\sqrt{25}}\right) = P(2.5 < Z) = 1 - P(Z \leq 2.5) = 1 - 0.9938 = 0.0062$$

Aproximación de la distribución binomial mediante la distribución normal.

La distribución binomial con variable aleatoria X que representa el número de éxitos con probabilidad p puede ser aproximada mediante una distribución normal si cumple que el número de muestras es grande, esto es, $n > 30$ y con probabilidad $p \approx 0.5$. Si la probabilidad p está alejada de 0.5, entonces es posible que se requiera un mayor número de datos para obtener una mejor aproximación.

La media a utilizar por parte de la normal

$$\mu = np \quad (4.10)$$

y la desviación típica o estándar

$$\sigma = \sqrt{npq} \quad (4.11)$$

La aproximación se puede llevar a cabo para un número n menor siempre y cuando el producto de np y $n(1-p)$ sea mayores a 5, por ejemplo para el caso $n=15$ y $p=0.4$ se tiene que $np=6$ y $n(1-p)=9.6$, entonces es posible aproximar la distribución binomial mediante la distribución normal para este caso. La figura siguiente muestra la distribución binomial y la normal para $n=15$ y $p=0.4$.

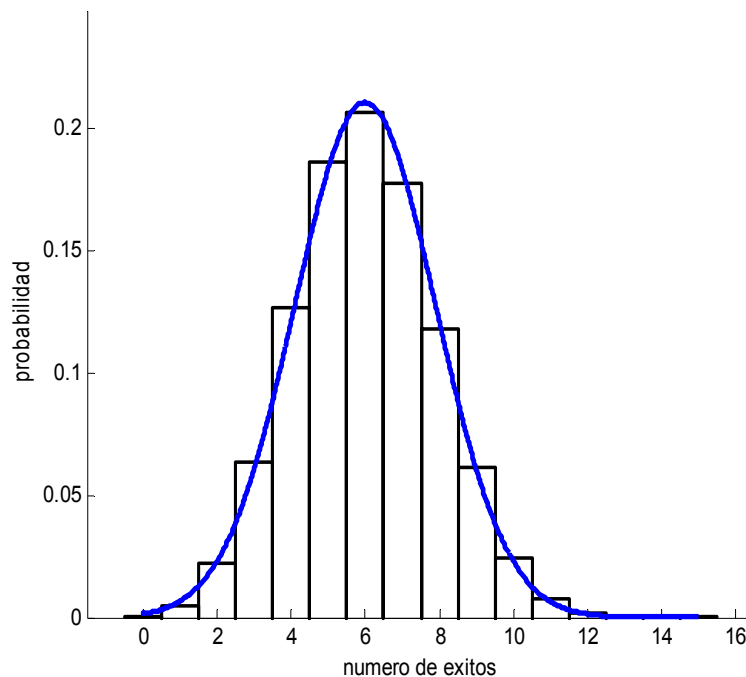


Figura. Aproximación de la binomial mediante la distribución normal, $n=15$ y $p=0.4$

Si ahora $n = 15$ y $p = 0.3$ se tiene que $np = 4.5$ y $n(1-p) = 10.5$, entonces, para este caso no es adecuado aproximar la distribución binomial mediante la distribución normal. La figura siguiente muestra la distribución binomial y la normal para $n = 15$ y $p = 0.3$.

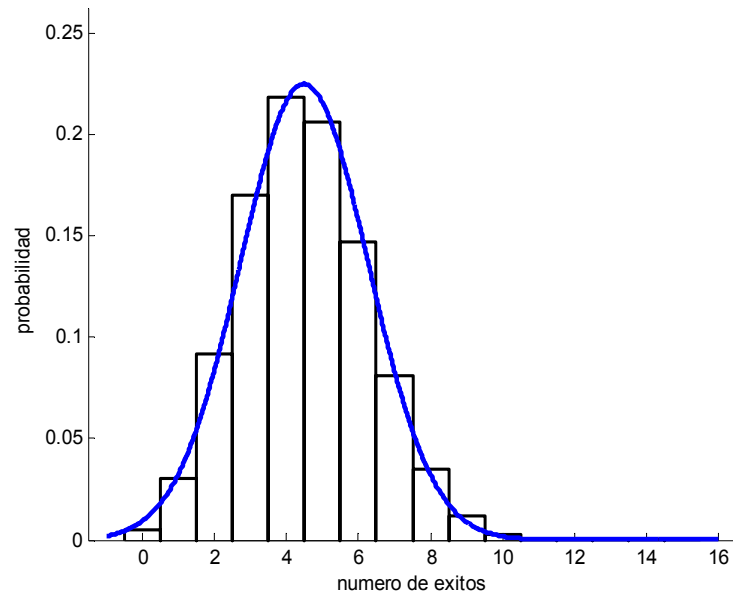


Figura. La aproximación de la binomial mediante la distribución normal no es aconsejable para este caso $n=15$ y $p=0.3$

Como se puede deducir de los dos caso anteriores si la probabilidad de éxito se aleja de 0.5 entonces para obtener una buena aproximación normal se requerirá un n mucho mayor, por ejemplo, para $n = 30$ y $p = 0.3$ se tiene que $np = 9$ y $n(1-p) = 21$, y entonces si es posible aproximar la distribución binomial mediante la normal. La siguiente figura muestra la aproximación para $n = 30$ y $p = 0.3$

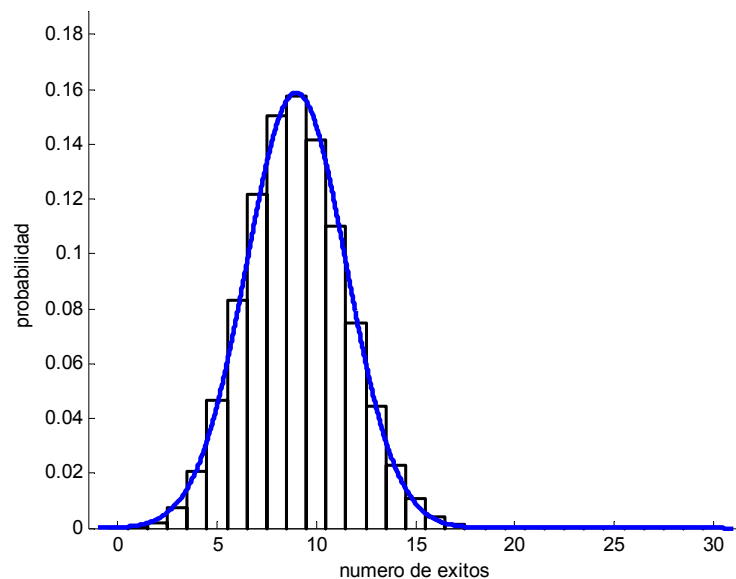


Figura. Aproximación de la distribución binomial a la normal para $n = 30$ y $p = 0.3$

EJEMPLOS

20. Supóngase que cierta medicina tiene un 80% de efectividad para curar cierto tipo de enfermedad. Es decir, en promedio de cada 100 pacientes que contraen la enfermedad y reciben la medicina, se espera que 80 se recuperen. Sea X el número de pacientes en una muestra aleatoria de 100 que se recuperan después del tratamiento. Obténganse las siguientes probabilidades mediante la aproximación normal.

- más de 80 se recuperarán o $P(X \geq 80)$;
- $P(80 < X < 90)$;
- $P(70 < X < 75)$.

SOLUCION

La probabilidad de éxito es $p = 0.8$ y el tamaño de la muestra es $n = 100$
La media y la desviación típica son

$$\mu = np = (0.8)(100) = 80$$

$$\sigma = \sqrt{npq} = \sqrt{100(0.8)(1 - 0.8)} = 4$$

Entonces

- $P(X > 80) = P(Z > (80 - 80)/4) = P(Z > 0) = 1 - P(Z < 0) = 1 - 0.5 = 0.5$
- $P(80 < X < 90) = P((80 - 80)/4 < Z < (90 - 80)/4) = P(0 < Z < 2.5) = P(Z < 2.5) - P(Z \leq 0)$
 $= 0.9938 - 0.5 = 0.4938$
- $P(70 < X < 75) = P((70 - 80)/4 < Z < (75 - 80)/4) = P(-2.5 < Z < -1.25) = P(1.25 < Z < 2.5)$
 $= P(Z < 2.5) - P(Z < 1.25) = 0.9938 - 0.8944 = 0.0994$

21. Se tira diez veces una moneda balanceada. Obténgase la probabilidad de que ocurran ya sea el seis, siete u ocho caras mediante

- la distribución binomial;
- el método de la aproximación normal con corrección por continuidad.

SOLUCION

- Puesto que la moneda es balanceada $p = 0.5$ y $n = 10$, aplicando la distribución binomial

$$P(6 \leq X \leq 8) = \binom{10}{0}(0.5)^6(0.5)^4 + \binom{10}{7}(0.5)^7(0.5)^3 + \binom{10}{8}(0.5)^8(0.5)^2$$

$$= 0.205078 + 0.11718 + 0.043945 = 0.366203$$

- Aplicando la distribución binomial y la corrección por continuidad

$$\mu = np = 10(0.5) = 5$$

$$\sigma = \sqrt{npq} = \sqrt{10(0.5)(0.5)} = \sqrt{2.5} = 1.5811$$

$$P(6 \leq X \leq 8) = P((5.5 \leq X \leq 8.5)) = P((5.5 - 5)/1.5811 \leq Z \leq (8.5 - 5)/1.5811)$$

$$= P(0.3162 \leq Z \leq 2.2136) = P(Z < 2.21) - P(Z \leq 0.32) = 0.9864 - 0.6255 = 0.3609$$

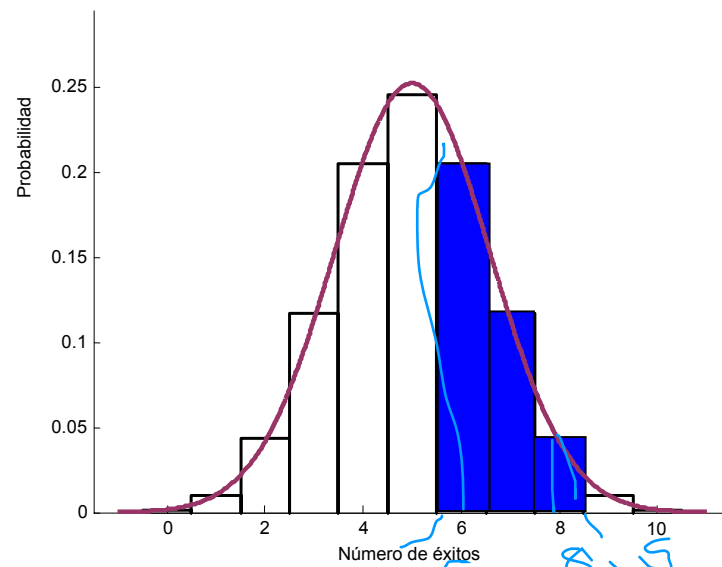


Figura representando la aproximación binomial a la normal para el ejemplo 2 $n=10$ y $p=0.5$.

UNIDAD V Inferencia estadística

INFERENCIA ESTADÍSTICA

Los conceptos básicos de probabilidad y distribuciones muestrales sirven de base para el método de **inferencia estadística**, la cual tiene como objetivo obtener información de las poblaciones a partir de las muestras obtenidas. En general se avoca a las dos siguientes áreas **prueba de hipótesis** y **estimación**.

PRUEBA DE HIPÓTESIS Y ESTIMACIÓN.

Una explicación concisa de cada una de estas áreas se da a continuación:

- **prueba de hipótesis**: aceptar o rechazar declaraciones acerca de los parámetros de la población.
- **estimación**: estimar valores de los parámetros de la población.

PLANTEAMIENTO DE LA HIPÓTESIS NULA Y ALTERNATIVA

Una hipótesis estadística consiste en realizar una declaración afirmativa o negativa acerca del valor de un parámetro o parámetros de una población. La aceptación o rechazo de la hipótesis estadística requiere de información obtenida a partir de las muestras de la población. Si la información obtenida es suficiente, la hipótesis estadística puede ser apoyada o no.

Los pasos esenciales para realizar una prueba de hipótesis se indican a continuación:

- **identificación del patrón de distribución de la variable aleatoria** (binomial, normal...)

Un procedimiento estadístico que requiere la identificación de la distribución probabilística se denomina **enfoque paramétrico**. Si no se especifica la distribución de probabilidad entonces se tiene un enfoque no paramétrico.

- **planteamiento de la hipótesis**. En general se proponen 2 hipótesis, una denominada **hipótesis nula** denotada por H_0 , la cual se propone con el objetivo de ver si puede ser rechazada y la **hipótesis alternativa** la cual se denota por H_1 y es válida si la hipótesis nula es rechazada.

Comúnmente la hipótesis nula H_0 , implica la idea de que no hay diferencia entre los parámetros, de ahí su nombre de nula.

Por ejemplo se puede proponer que el promedio no es diferente de un valor particular, esto es $H_0: \mu = \mu_0$

Las hipótesis alternativas H_1 , que pueden establecerse como complementaria para la hipótesis nula H_0 anterior, puede tomar alguna y solo una de las siguientes opciones:

PRUEBA DE DOS COLAS

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Debido a que no se especifica la dirección de la diferencia entre μ y μ_0 , la prueba se le denomina **prueba de dos colas**.

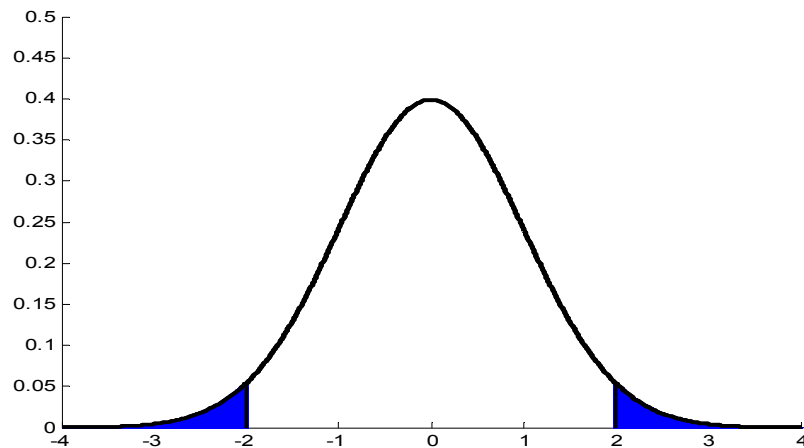


Figura. Esquema utilizando la distribución normal para mostrar la prueba de dos colas, la región sombreada representa la región de rechazo de la hipótesis nula H_0

PRUEBA DE UNA COLA DERECHA

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

Como $\mu > \mu_0$, la prueba es llamada de **una cola derecha**

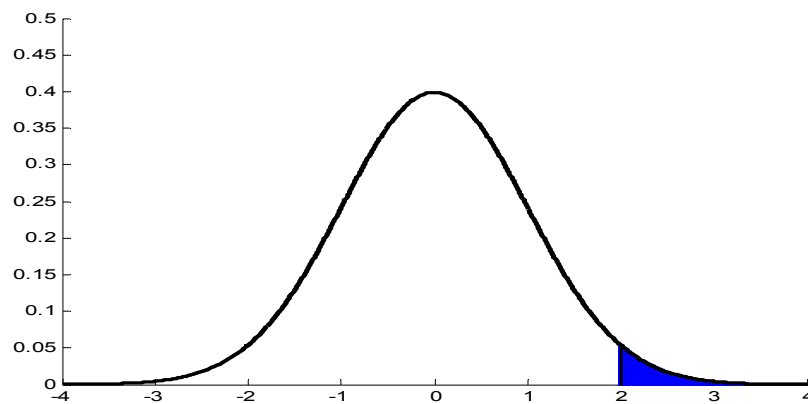


Figura. Esquema utilizando la distribución normal para mostrar la prueba de cola derecha, la región sombreada representa la región de rechazo de la hipótesis nula H_0

PRUEBA DE UNA COLA IZQUIERDA:

$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

Como $\mu < \mu_0$, la prueba es llamada de **una cola izquierda**

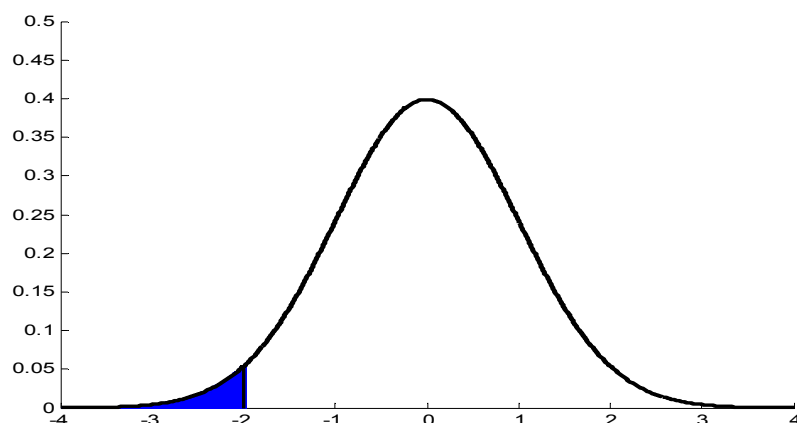


Figura. Esquema utilizando la distribución normal para mostrar la prueba de cola izquierda, la región sombreada representa la región de rechazo de la hipótesis nula H_0

ESPECIFICACION DEL NIVEL DE SIGNIFICACION α

Normalmente las muestras extraídas de una población en general no son idénticas y presentan diferentes medias y desviaciones típicas, etc., estas diferencias pueden deberse a la naturaleza aleatoria del problema, por ejemplo si se considera la prueba de hipótesis

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

La pregunta sería ¿Qué tan grande debe ser la media muestra para rechazar la hipótesis nula? De otra manera, ¿Qué tan grande debe ser la media muestral para que se considere significativamente mayor?

La respuesta a la pregunta depende directamente del **nivel de significación** elegido para realizar la prueba de hipótesis, normalmente se denota como α , por ejemplo si $\alpha = 5\%$, la hipótesis nula no se rechazará en 5 de 100 muestras lo suficientemente grandes.

Los valores comúnmente elegidos como niveles de significación son

$$\alpha=10\%, \alpha=5\%, \alpha=2.5\%, \alpha=1.0\%, \alpha=0.5\%$$

El **nivel de significación**: se puede entender también como la probabilidad de rechazar una hipótesis nula verdadera o la probabilidad de cometer un **error tipo I** que anteriormente se denotó por α . Por otra parte el error de no rechazar la hipótesis nula cuando es falsa se denomina **error tipo II**, denotado por β .

Los dos tipos de errores se resumen a continuación

	TIPO DE ERROR	PROBABILIDAD
Rechazar H_0 cuando es verdadera	I	α .
No rechazar a H_0 cuando es falsa	II	β

La relación entre los tipos de error α y β se muestra en la siguiente gráfica para la $H_0: \mu = \mu_0$ y $H_1: \mu > \mu_0$

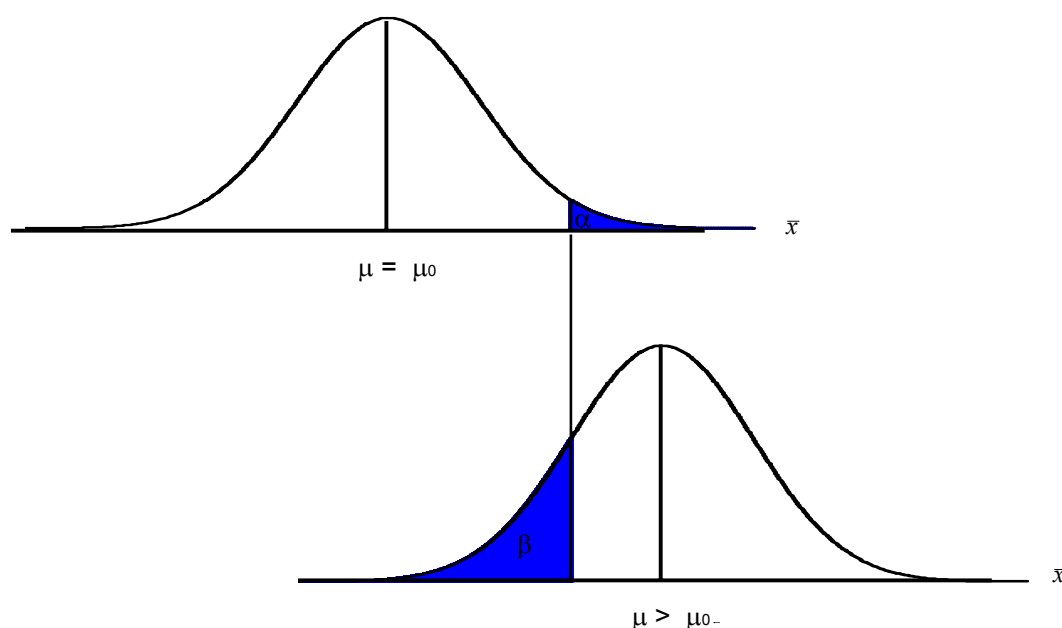


Figura. Relación entre los errores tipo I representado por el área sombreada α y el error tipo II representado por el área sombreada β

Las áreas oscuras representan la probabilidades α y β , si se disminuye la probabilidad α al desplazar la línea vertical a la derecha el valor de β aumenta, y viceversa, si la línea vertical se mueve a la izquierda aumenta α y disminuye β .

PLANTEAMIENTO DE LA REGLA DE DECISIÓN

- Elegir el estadístico de prueba el cual es una variable aleatoria cuyo valor se utiliza para aceptar o rechazar la hipótesis nula. Puedes ser un estadístico muestral tal como la media muestral, desviación típica, proporción de defectos, etc.
- Especificar el nivel de significancia de α .
- Los valores del estadístico de prueba se dividen en 2 categorías: **región de rechazo y región de aceptación**, también se conoce la región de rechazo como **región crítica**.

TOMA DE LA DECISIÓN:

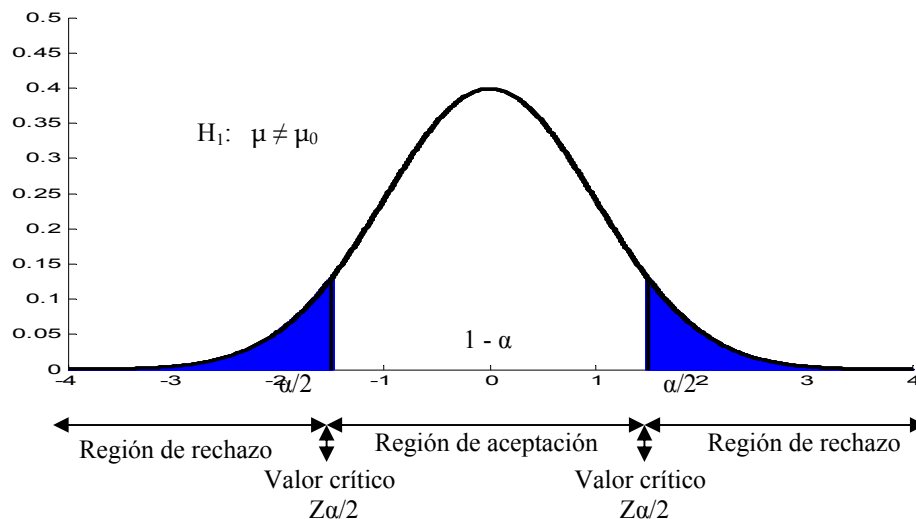
- El valor que separa las dos regiones es llamado **el valor crítico**. Se toma la decisión dependiendo en que región cae el valor del estadístico de prueba. Si el valor del estadístico de prueba cae en la región de rechazo, la hipótesis nula se rechaza, en caso contrario se acepta.

TABLA DE DECISIONES

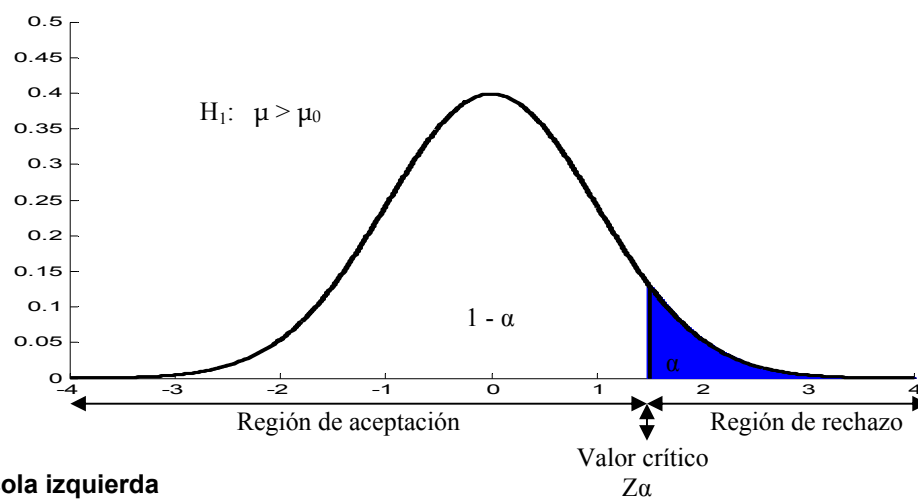
Decisión	H_0 es verdadera	H_0 es falsa
Se rechaza H_0	Error tipo I α	Decisión correcta $1-\beta$
No se rechaza H_0	Decisión correcta $1-\alpha$	Error tipo II β

Las siguientes figuras muestran el valor crítico, las regiones de aceptación y rechazo, para el caso de que se utilice a Z como estadístico de prueba, para cada una de los tres tipos de prueba de hipótesis.

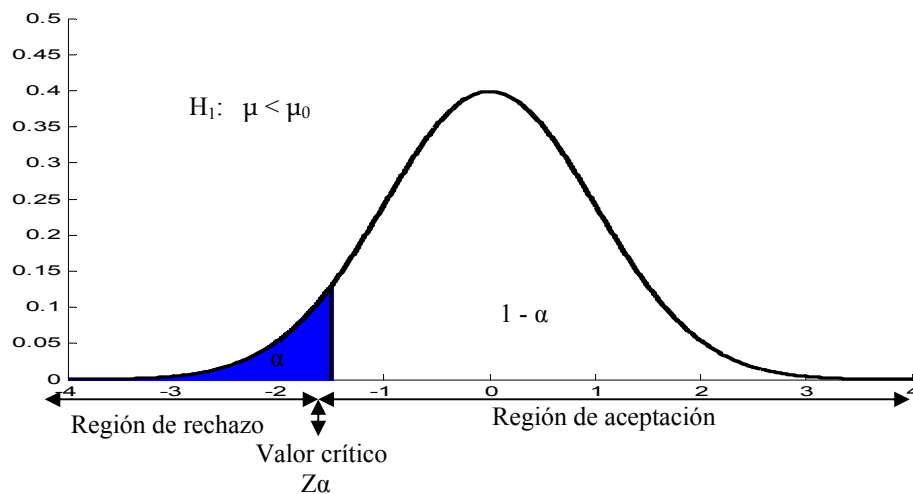
Prueba de dos colas



Prueba de cola derecha



Prueba de cola izquierda



EJEMPLOS

1. En la prueba de la hipótesis nula $\mu = 100$, la hipótesis alternativa puede ser cualquiera de las siguientes.

- | | |
|-------------------|----------------|
| a. $\mu = 110$ | b. $\mu = 90$ |
| c. $\mu > 100$ | d. $\mu < 100$ |
| e. $\mu \neq 100$ | |

¿Cuáles de estas cinco pruebas son de una cola? ¿Cuáles son de dos colas?

SOLUCION

- a) Como $\mu = 110$ y se encuentra a la derecha, es una prueba de cola derecha.
- b) En este caso $\mu = 90$ es menor a 100, por lo que es una prueba de cola izquierda.
- c) $\mu > 100$ es una prueba de cola derecha.
- d) $\mu < 100$ es una prueba de cola izquierda.
- e) $\mu \neq 100$ representa a una prueba de dos colas.

2. Supóngase que la producción promedio por hora de los trabajadores de cierta fábrica es de 60 unidades. El director de personal de la fábrica afirma que el programa de entrenamiento implantado hace algún tiempo ha aumentado la productividad de los trabajadores. Plántese las hipótesis nula y alternativa.

SOLUCION

La Hipótesis nula en general se relaciona con que el estimador no cambia, por lo tanto $H_0: \mu = 60$ y como se señala que el programa de entrenamiento ha mejorado la productividad la hipótesis alternativa se propone de cola derecha, esto es $H_1: \mu > 60$

3. Cierta proceso de producción está diseñado para dar como resultado tornillos con una longitud media de 3 plg. Plántese la regla de decisión para cada una de las siguientes situaciones:

- a. El gerente de producción desea determinar si la longitud promedio ha disminuido.
- b. Desea determinar si la longitud promedio ha aumentado.
- c. Desea determinar si la longitud promedio ha cambiado.

SOLUCION

Para el problema se debe seleccionar $\mu_0 = 3$ pulgadas y de acuerdo a cada uno de los incisos

- | | | |
|-------------------|-------------------|---------------|
| a) $H_0: \mu = 3$ | $H_1: \mu < 3$ | Ha disminuido |
| b) $H_0: \mu = 3$ | $H_1: \mu > 3$ | Ha aumentado |
| c) $H_0: \mu = 3$ | $H_1: \mu \neq 3$ | Ha cambiado |

4. Supóngase que el gasto anual en libros por parte de los estudiantes universitarios de los EUA se distribuye normalmente con media de \$ 200. Fórmese, para cada una de las siguientes pruebas, la hipótesis alternativa y plántese la regla de decisión.

- a. Pruébese si los estudiantes en la universidad a la que usted asiste han gastado más que el promedio nacional.
- b. Pruébese si el gasto anual por parte de los estudiantes de la universidad a la que usted asiste es diferente del promedio nacional.

SOLUCION

En este caso se elige $\mu_0 = 200$ y la hipótesis nula es para ambos inciso $H_0: \mu = 200$.

- a) La hipótesis alternativa es $H_1: \mu < 200$, y se rechaza H_0 para algún valor de \bar{X} lo suficientemente grande.
 b) La hipótesis alternativa es $H_1: \mu \neq 200$ y se rechaza H_0 si \bar{X} lo suficientemente grande o suficientemente pequeño.

HIPOTESIS INEXACTA

Las hipótesis se pueden clasificar como exactas e inexactas. Una **hipótesis es exacta** si se especifica en la prueba un único valor, por ejemplo, $H_0: \mu = \mu_0$, mientras que si especifica un conjunto de valores como $H_0: \mu \leq \mu_0$ ó $H_0: \mu > \mu_0$ será una **hipótesis inexacta**. Las siguientes figuras muestran los casos de la Hipótesis exacta e inexacta de manera gráfica.

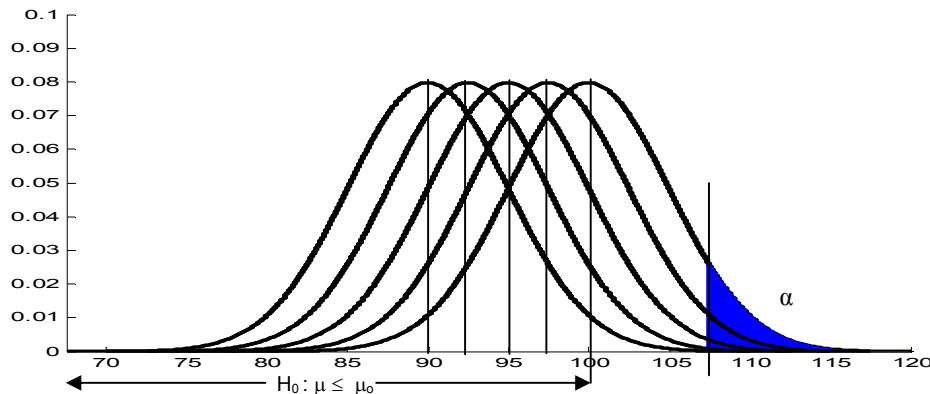


Figura. Sucesión de gráficas con media menor a 100 que muestran el caso $H_0: \mu \leq \mu_0$

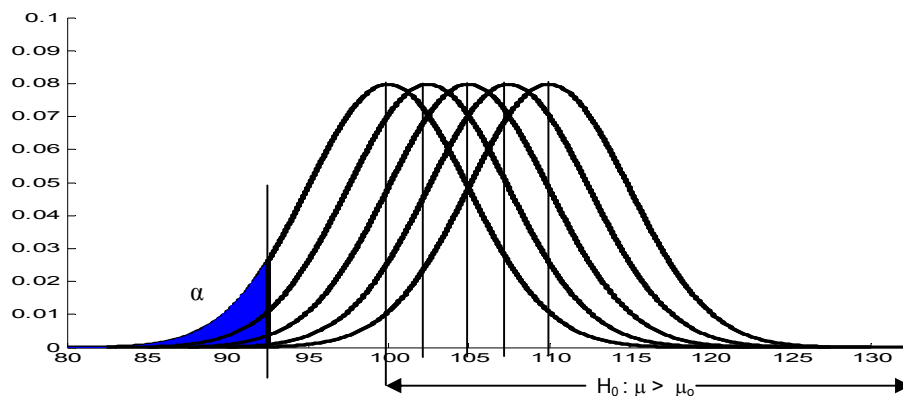


Figura. Sucesión de gráficas con media mayor a 100 que muestran el caso $H_0: \mu > \mu_0$

El área sombreada para cada una de las gráficas de las dos figuras anteriores es cada vez más pequeña conforme la media se vuelve más pequeña (ó más grande), lo anterior implica que si se rechaza la hipótesis exacta $\mu = \mu_0$ con probabilidad α entonces para todos los casos $\mu \leq \mu_0$ (ó $\mu > \mu_0$) se rechazara la hipótesis nula con una probabilidad menor a α . Por lo que los casos de hipótesis inexactas se trabajarán como hipótesis exactas $\mu = \mu_0$ con probabilidad de rechazo α .

PRUEBAS DE HIPÓTESIS PARA MUESTRAS GRANDES

PRUEBA PARA LA MEDIA DE LA POBLACION

Se utiliza la media muestral \bar{X} como variable aleatoria obtenida a partir de una muestra de tamaño n la cual se obtiene de una población con media μ y desviación típica σ . Si la muestra es grande (teorema del limite central $n > 30$) ó la población tiene una distribución normal. Entonces, la muestra tendrá una distribución normal.

Como ha sido mostrado anteriormente (distribución muestral de la media ó teorema del limite central)

$$\mu_{\bar{X}} = \mu \quad \text{y} \quad \sigma_{\bar{X}} = \frac{\sigma}{n}$$

El estadístico de prueba Z para la prueba de una media con distribución normal es

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \quad (5.1)$$

ó

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad (5.2)$$

EJEMPLOS

5. Se supone que los C.I. de los alumnos de cierto grupo étnico está en promedio ocho puntos por encima que el promedio de todos los alumnos en el país. Se sabe que para todos los alumnos la media es 100 y la desviación típica es 15. Pruebas aplicadas a una muestra de 25 alumnos seleccionados aleatoriamente entre el grupo étnico en cuestión proporcionan un C.I. medio de 104. Considerando que los C.I. Tienen una distribución normal, pruébese la hipótesis $H_0 : \mu = 100$ en contra de la hipótesis alternativa $H_1 : \mu = 108$ en $\alpha = 0.05$. Determinése también el valor de β .

SOLUCION

Los datos del problema son

La media y desviación estándar son $\mu=100$, $\sigma=15$, el nivel de significación es $\alpha=0.05$, el tamaño de la muestra es $n=25$ y la media muestral es $\bar{X}=104$

Las Hipótesis correspondientes nula y alternativa son respectivamente

$$H_0: \mu=100$$

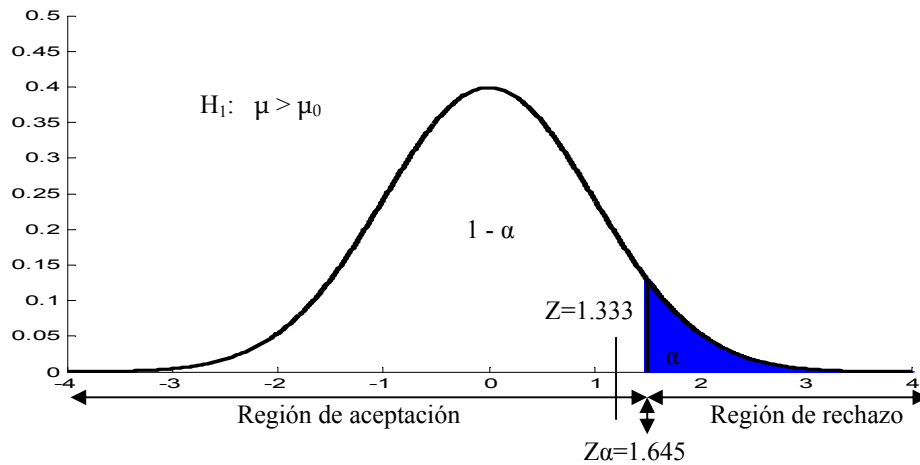
$$H_1: \mu=108$$

La prueba es de una cola derecha. A partir del nivel de significancia $\alpha=0.05$, se determina el área a la izquierda como $A=1-0.05=.95$, entonces el valor crítico Z_α se obtiene de la puntuación cuya área bajo la curva normal es igual a 0.95 este valor corresponde a $Z_\alpha = 1.645$

Calculando el estadístico de prueba correspondiente a partir de la tipificación de la media muestral \bar{x}

$$Z = \frac{104 - 100}{\frac{15}{\sqrt{25}}} = 4/3 = 1.333$$

Puesto que $1.333 < 1.645$ ($Z < Z_\alpha$) el valor cae dentro de la región de aceptación por lo que no se rechaza H_0 , ver grafica.



b) Para determinar el error tipo II ó β , se requiere determinar primero \bar{X}_α la cual se puede obtener

despejando de la relación $Z_\alpha = \frac{\bar{X}_\alpha - \mu}{\sigma/\sqrt{n}}$

$$\bar{X}_\alpha = Z_\alpha \left(\frac{\sigma}{\sqrt{n}} \right) + \mu = 1.645 \left(\frac{15}{\sqrt{25}} \right) + 100 = 1.645(3) + 100 = 104.935$$

La figura siguiente muestra la idea general para determinar el error tipo β .

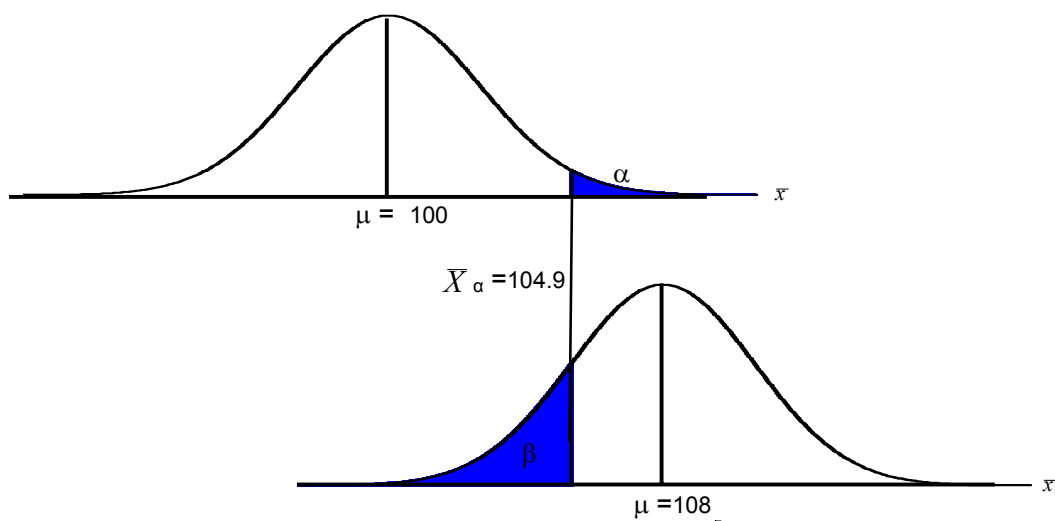


Figura. Idea general para determinar el error tipo β .

Entonces el error tipo β es igual de acuerdo a la figura anterior

$$\beta = P(X < 104.9, \mu=108, \sigma=3) = P\left(Z < \frac{104.9 - 108}{3}\right) = P(Z < -1.02166) = 1 - 0.8461 = 0.1539$$

6. Una compañía que procesa fibras naturales afirma que sus fibras tienen una resistencia media a la ruptura de 40 lb y una desviación típica de 8 lb. Un comprador sospecha que la resistencia media a la ruptura es de solamente 37 lb. Una muestra aleatoria de 64 fibras proporciona una media de 38 lb. ¿Deberá rechazar el comprador $H_0: \mu=40$ en favor de $H_1: \mu = 37$ si el nivel de significación es 0.01?

SOLUCION

Los datos del problema son

Los parámetros poblacionales son $\mu=40$, $\sigma=8$ promedio probables $\mu_1=37$, tamaño de la muestra $n = 64$ nivel de significación $\alpha=0.01$, media muestral $\bar{x} = 38$

Las Hipótesis correspondientes nula y alternativa son respectivamente

$H_0: \mu=40$

$H_1: \mu_1=37$

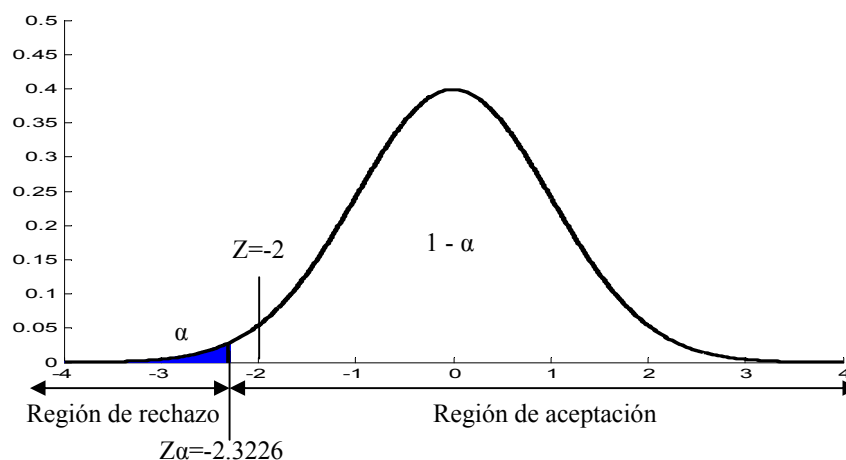
La prueba es de una cola izquierda, entonces, el área a la izquierda de la distribución debe ser

$A = 1 - \alpha = 1 - 0.01 = 0.99$ lo cual corresponde a $Z_\alpha = -2.3226$

El valor del estadístico de prueba es $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{38 - 40}{\frac{8}{\sqrt{64}}} = -2$

El cual es mayor que Z_α .

Por lo tanto no se rechaza H_0



7. Un fabricante de medias está considerando reemplazar una vieja máquina de coser por una nueva. La vieja máquina produce cuando más, un promedio de 300 pares de medias por hora, con una desviación típica de 30 pares. Se considera que la producción por hora de tales máquinas de coser tiene una distribución normal. El vendedor de la nueva máquina afirma que su producción promedio por hora es de más de 300 pares. La nueva máquina se prueba durante un periodo de 25 h y se determina su producción promedio por hora como 310 pares. si el nivel de significación es de 0.05, ¿debería rechazarse la hipótesis nula $\mu = 300$?

SOLUCION

Los datos proporcionados por el problema son

Media $\mu=300$, desviación $\sigma=30$, tamaño de la muestra $n = 25$, nivel de significancia $\alpha =0.05$, media muestral \bar{X} 310

La prueba de hipótesis se puede plantear como:

$$H_0: \mu=300$$

$$H_1: \mu>300$$

Corresponde a una prueba de una cola derecha

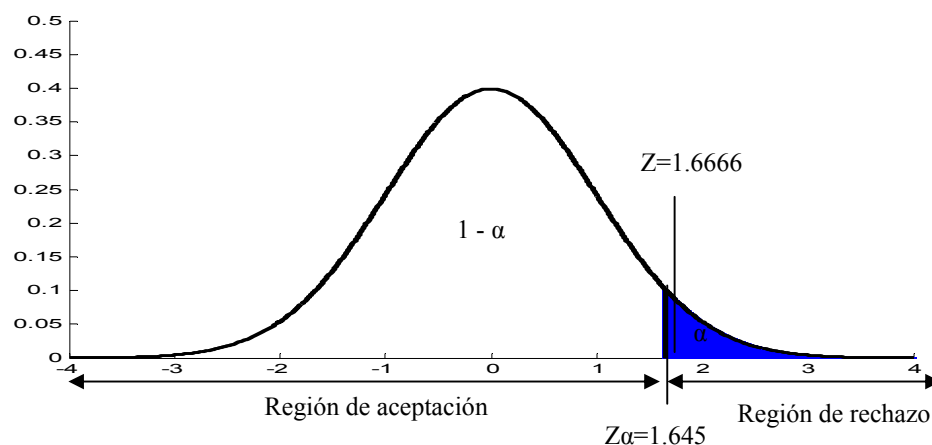
Utilizando la el nivel de significación $\alpha=0.05$, se determina el área a la izquierda de la distribución normal $A =1- \alpha=1-0.05=0.95$, el cual corresponde a una valor de puntuación crítico $Z_{\alpha}=1.645$

El valor del estadístico de prueba Z es

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{310 - 300}{\frac{30}{\sqrt{25}}} = 1.6666$$

En este caso $Z_{\alpha} < Z$, la hipótesis nula se rechaza.

Por lo tanto se rechaza H_0 a favor de de la hipótesis H_1



8. Una compañía de servicio público desea determinar si su nuevo horario de Trabajo ha reducido de manera importante el tiempo de espera de los clientes para servicio. El tiempo de espera fue de al menos 30 min en el pasado y se sabía que la desviación típica era de 12 min. Se selecciona aleatoriamente una muestra de 144 observaciones. Se obtiene una media de 28 min. ¿Debería rechazarse la hipótesis nula $\mu \geq 30$ en favor de la hipótesis alternativa $\mu < 30$ para $\alpha = 0.05$?

SOLUCION

Los datos proporcionados por el problema son

Media $\mu=30$ min, desviación $\sigma=12$ min, tamaño de la muestra $n = 144$, nivel de significancia $\alpha =0.05$, media muestral $\bar{x} = 28$ min

La prueba de hipótesis nula es inexacta se puede plantear como:

$$H_0: \mu \geq 30$$

$$H_1: \mu < 30$$

Corresponde a una prueba de una cola izquierda

Utilizando la el nivel de significación $\alpha=0.05$, se determina el área a la izquierda de la distribución normal $A = 1 - \alpha = 1 - 0.05 = 0.95$, el cual corresponde a una valor de puntuación crítico $Z_{\alpha} = -1.645$

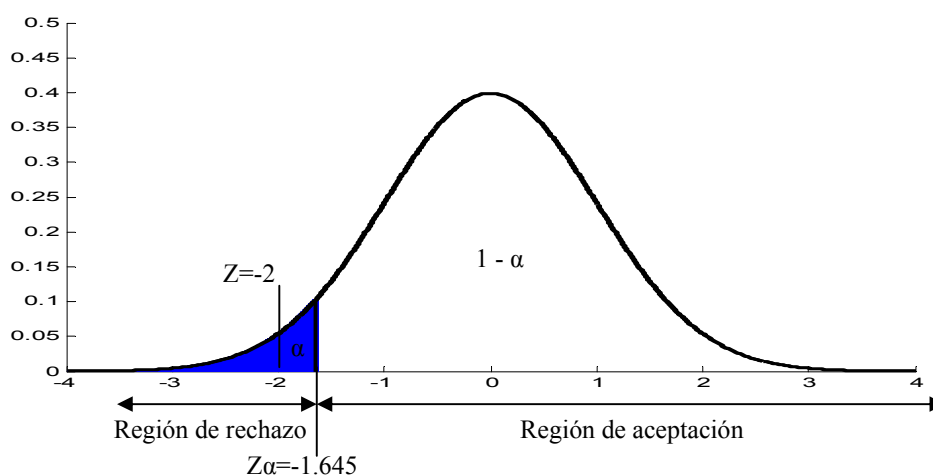
El valor del estadístico de prueba Z es

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{28 - 30}{\frac{12}{\sqrt{144}}} = -2.000$$

En este caso $Z < Z_{\alpha}$, la hipótesis nula se rechaza.

Por lo tanto se rechaza H_0 a favor de de la hipótesis H_1

Lo que se traduce en que el servicio al cliente ha mejorado.



9. Los empleados que contraen cierta enfermedad y reciben tratamiento médico normal para ella permanecen ausentes del trabajo durante un promedio de 15 días. Un equipo médico de investigación afirma que se ha desarrollado un nuevo tratamiento que reduciría el periodo promedio de ausencia del trabajo. Considérese que el periodo de ausencia del trabajo tiene una distribución normal y una desviación típica de tres días. ¿Debería rechazarse la hipótesis nula $\mu = 15$ para $\alpha = 0.1$ si una muestra de 16 pacientes que han recibido el nuevo tratamiento tiene una ausencia promedio del trabajo de exactamente 13 días?

SOLUCION

Los datos proporcionados por el problema son $\mu=15$ días, $\sigma=3$ días, $n = 16$, $\bar{X} = 13$ y $\alpha=0.1$

La prueba de hipótesis corresponde a una prueba de una cola izquierda con $A = 1 - \alpha = 1 - 0.1 = 0.9$ correspondiente a $Z_{\alpha} = -1.282$

El valor del estadístico de prueba Z es

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{13 - 15}{\frac{3}{\sqrt{16}}} = -2.666$$

En este caso $Z < Z_{\alpha}$, la hipótesis nula se rechaza.

Por lo tanto se rechaza H_0 a favor de la hipótesis H_1 , el tratamiento es mejor.

PRUEBA DE LA DIFERENCIA DE MEDIAS

En ocasiones se requiere indicar por parte de la estadística si la diferencia entre dos medias muestrales es lo suficientemente grande para asegurar que esas diferencias no se deben a efectos del azar, sino que las muestras tomadas provienen de dos poblaciones distintas. La siguiente figura muestra el caso de dos distribuciones normales con desviación típica $\sigma=10$ y medias $\mu_1 = 100$ y $\mu_2 = 120$

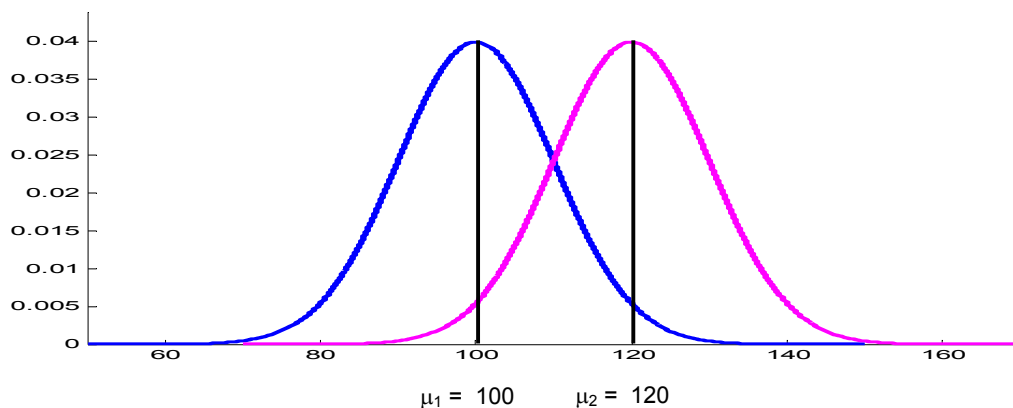


Figura. Representación de dos poblaciones con desviación típica $\sigma = 10$ y medias $\mu_1 = 100$ y $\mu_2 = 120$

Para probar la hipótesis acerca de la diferencia de medias se introduce la variable aleatoria

$$\bar{D} = \bar{X}_1 - \bar{X}_2 \quad (5.3)$$

Donde \bar{X}_1 es una muestra tomada de una población con media μ_1 y desviación típica σ_1 y \bar{X}_2 procede otra población con media μ_2 y desviación típica σ_2 . Los parámetros para variable aleatoria D se puede determinar aplicando las propiedades del valor esperado y varianza para muestras independientes

$$\delta = E(\bar{D}) = E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) + E(\bar{X}_2) = \mu_1 - \mu_2 \quad (5.4)$$

y la varianza

$$\sigma_D^2 = VAR(\bar{X}_1 - \bar{X}_2) = VAR(\bar{X}_1) + VAR(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad (5.5)$$

Entonces, la desviación típica es

$$\sigma_{\bar{D}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (5.6)$$

a la que se denomina **error típico de la diferencia** entre dos medias muestrales.

Si las muestras \bar{X}_1 y \bar{X}_2 provienen de distribuciones que son normales o si las muestras son grandes, esto es n_1 y $n_2 > 30$ la distribución de la variable aleatoria \bar{D} es normal.

La prueba de hipótesis acerca de la diferencia de medias se puede llevar a cabo bajo dos condiciones diferentes:

- 1) Cuando se conoce las varianzas poblacionales σ_1^2 y σ_2^2 ó
- 2) Cuando no se conocen las varianzas poblacionales y tienen que estimarse a partir de las varianzas muestrales s_1^2 y s_2^2 .

Primeramente los problemas que se desarrollan continuación suponen conocidas las varianzas poblacionales σ_1^2 y σ_2^2 .

La hipótesis nula para la prueba de la diferencia de medias denotada por δ es

$$H_0: \delta = 0 \quad \text{ó} \quad \mu_1 = \mu_2$$

Para la hipótesis alternativa puede tomar cualquiera de las siguientes posibilidades

$H_1: \delta < 0$ Cola izquierda	$\mu_1 < \mu_2$
$\delta > 0$ Cola derecha	$\mu_1 > \mu_2$
$\delta \neq 0$ Dos colas	$\mu_1 \neq \mu_2$

El estadístico de prueba es

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{D}}} \quad (5.7)$$

Recordando la hipótesis nula $\mu_1 = \mu_2$ y la definición de $\sigma_{\bar{D}}$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (5.8)$$

La prueba se realiza de manera semejante a la realizada anteriormente para la media, solamente que ahora para la prueba de dos medias se utiliza un estadístico diferente.

EJEMPLOS

10. Se realizó un estudio para determinar si los alumnos pertenecientes a dos grupos étnicos, I y II, tienen distintos CI., promedio. Se considera que las varianzas de los CI en los grupos I y II son respectivamente, $\sigma_1^2 = 225$ y $\sigma_2^2 = 196$. Se toma una muestra de 25 alumnos del grupo I ($n_1 = 25$) y otra de 28 del grupo II ($n_2 = 28$). En base a la diferencia entre las dos medias muestrales, $\bar{X}_1 = 102$ y $\bar{X}_2 = 98$. Pruébese la hipótesis nula de que los alumnos de los dos grupos étnicos tienen CI promedio idénticos con respecto a la hipótesis alternativa de que los dos promedios son diferentes en $\alpha = 0.05$.

SOLUCION

La lista de datos proporcionados por el problema se resume a continuación

$\bar{X}_1 = 102$	$\sigma_1^2 = 225$	$n_1 = 25$
$\bar{X}_2 = 98$	$\sigma_2^2 = 196$	$n_2 = 28$

Las hipótesis nulas y alternativas asociadas al problema son

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

La prueba es de dos colas por lo tanto $Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025}$

El valor del área para la prueba es $A = 1 - 0.025 = 0.975$

Correspondiente de acuerdo a las tablas $Z_{0.025} = 1.960$

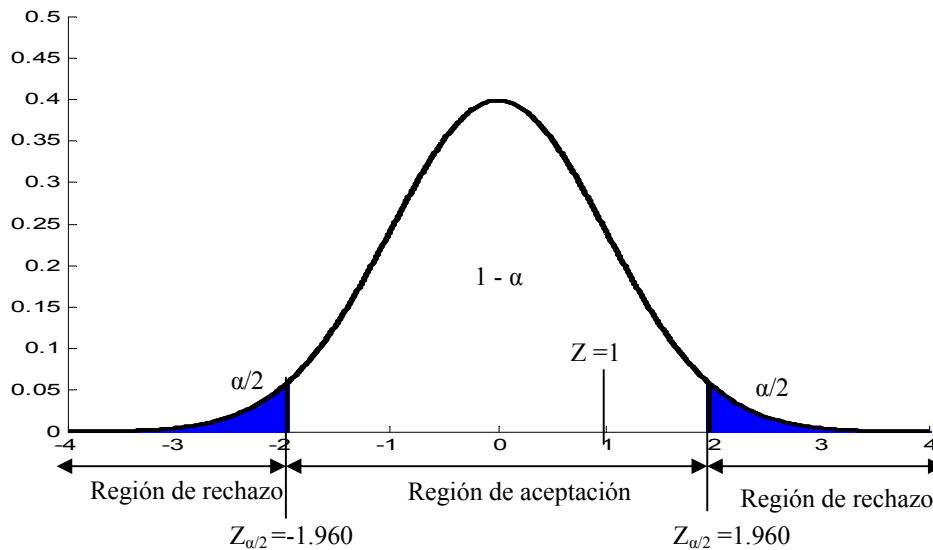
La regla de decisión es:

Rechazar H_0 si $Z \geq 1.960$ ó $Z \leq -1.960$

El estadístico de prueba Z es

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{102 - 98}{\sqrt{\frac{225}{25} + \frac{196}{28}}} = \frac{4}{4} = 1$$

Como es mayor a -1.960 y menor a 1.960 no se rechaza H_0 .



11. Cierta gran compañía emplea tanto hombres como mujeres para realizar el mismo tipo de trabajo. Se tiene la hipótesis de que la producción promedio de los hombres es menor que la de las mujeres. Supóngase que el equipo de investigación de la compañía proporciona la siguiente información.

	Hombres	Mujeres
Tamaño de la muestra	$n_1 = 36$	$n_2 = 36$
Media muestral en unidades	$\bar{X}_1 = 150$ y	$\bar{X}_2 = 153$
Varianza	$\sigma_1^2 = 70$	$\sigma_2^2 = 74$

¿Es significativamente menor la producción promedio por hora de los hombre que la de las mujeres para $\alpha = 0.05$? (Considérese que las dos muestras son independientes.)

SOLUCION

Las hipótesis nulas y alternativas son

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 < \mu_2$$

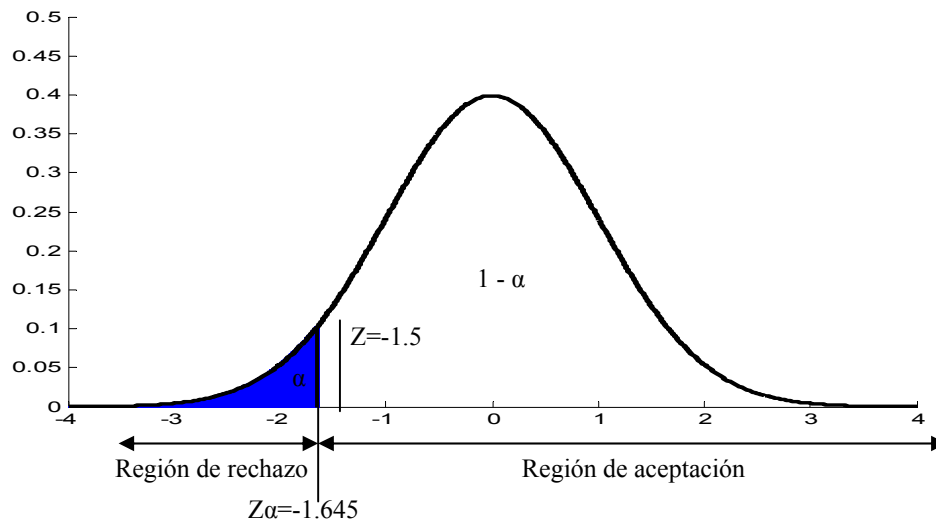
De acuerdo al nivel de significación $\alpha = 0.05$,

$A = 1 - \alpha = 1 - 0.05 = 0.95$ correspondiente al valor crítico $Z_\alpha = -1.645$

El estadístico de prueba Z es

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{150 - 153}{\sqrt{\frac{70}{36} + \frac{74}{36}}} = -\frac{3}{2} = -1.5$$

Como Z es mayor a $Z_{\alpha} = -1.645$ no se rechaza H_0 .



12. Un fabricante afirma que el cordón nylon que su compañía produce es más fuerte que el cordón de algodón. Dada la siguiente información:

	Cordón de nylon	Cordón de algodón
Tamaño de la muestra	$n_1 = 36$	$n_2 = 36$
Resistencia promedio a la ruptura	$\bar{X}_1 = 105 \text{ lb}$	$\bar{X}_2 = 101 \text{ lb}$
Varianzas	$\sigma_1^2 = 74$	$\sigma_2^2 = 70$

¿Podría llegarse a la conclusión de que en realidad el cordón de nylon es más fuerte que el de algodón para $\alpha = 0.01$?

SOLUCION

Las hipótesis nulas y alternativas son

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 > \mu_2$$

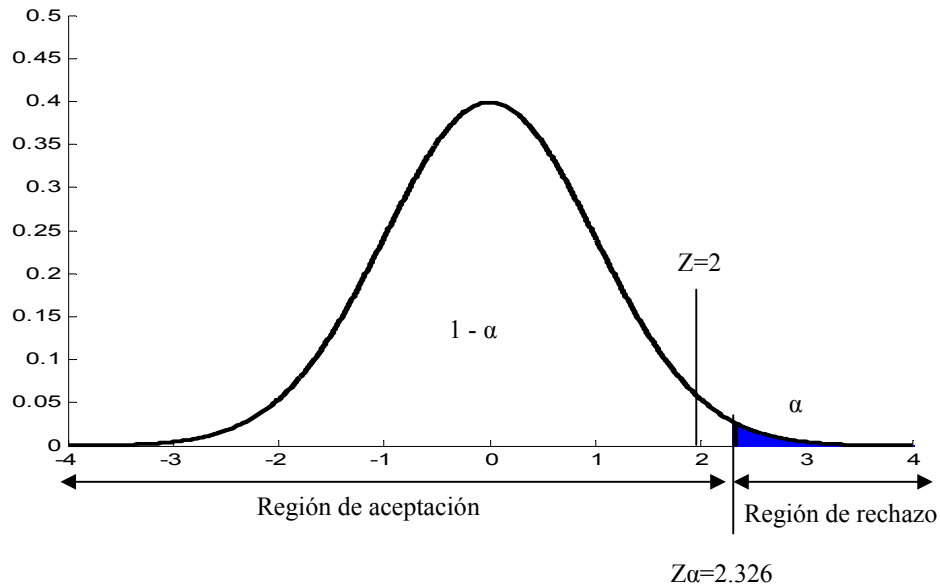
De acuerdo al nivel de significación $\alpha = 0.01$,

$A = 1 - \alpha = 1 - 0.01 = 0.99$ correspondiente al valor crítico $Z_{\alpha} = 2.326$

El estadístico de prueba Z es

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{105 - 101}{\sqrt{\frac{70}{36} + \frac{74}{36}}} = \frac{4}{2} = 2.0$$

Como Z es menor a 2.326 no se rechaza H_0 .



PUEBAS PARA LA PROPORCION DE EN LA POBLACION

En ocasiones se requiere decidir si la proporción en la población denotada por p es igual a una proporción dada p_0 , en donde, **la proporción de la muestra** o el número de éxitos en n ensayos, se utiliza para realizar la inferencia. Si el evento ha ocurrido X veces en n intentos, la proporción de la muestra es estimada es $\hat{p} = X/n$, fracción que puede utilizarse para estimar la proporción de la población o la probabilidad de éxito.

Para probar a hipótesis con respecto a la proporción p resulta más conveniente utilizar la variable aleatoria binomial X que la misma proporción p . Para valores pequeños de n (< 30) se puede utilizar las tablas binomiales acumuladas y para n grande se utilizar la aproximación normal a la binomial.

EJEMPLOS

13. Un fabricante de drogas afirma que una medicina recientemente desarrollada tiene una efectividad de más del 90% en el alivio de dolores musculares. En una muestra de 100 personas que sufren de dolores musculares, la medicina proporcionó alivio a 95. Pruébese la hipótesis nula de que la medicina tiene una efectividad de 90% contra la hipótesis alternativa de que la medicina tiene una efectividad de más del 90% para $\alpha = 0.05$.

SOLUCION

Debido a que el tamaño de la muestra es grande $n = 100$, es recomendable utilizar la aproximación normal a la binomial. Utilizando la proporción como la probabilidad de éxito, que de acuerdo a los datos proporcionados la proporción $p_0 = 0.90$, entonces el promedio es

$$\mu = np = 100(0.9) = 90$$

y la desviación típica de la población es

$$\sigma = \sqrt{npq} = \sqrt{(100)(0.9)(0.1)} = 3$$

Para $\hat{p} = 0.95$, el promedio estimado es entonces

$$\bar{X} = n \hat{p} = (0.95)(100) = 95$$

Las hipótesis nulas y alternativas del problema son

$$\begin{array}{ll} H_0: p = 0.9 & \text{o} \quad \mu = 90 \\ H_1: p > 0.9 & \text{o} \quad \mu > 90 \end{array}$$

Para el nivel de significancia $\alpha=0.05$ y la prueba de cola derecha el área a la izquierda es $A=1-\alpha=1-0.05=0.95$, correspondiente a un valor crítico para la distribución normal $Z_\alpha = Z_{0.05} = 1.645$

El valor del estadístico de prueba Z es

$$Z = \frac{\bar{X} - \mu}{\sigma} = \frac{95 - 90}{3} = 1.6666$$

como $Z > Z_\alpha$, se rechaza la hipótesis nula H_0 a favor de H_1 , esto es, la medicina tiene una efectividad mayor que el 90 %.

14. Un investigador de mercado desea determinar si las amas de casa prefieren el aceite de cocina I o el aceite de cocina II. Se entrevista a 30 amas de casa y 18 de ellas indican que prefieren el aceite I. ¿Puede llegarse a la conclusión de que las amas de casa en general prefieren el aceite I, si el nivel de significación es de 0.04937?

SOLUCION

Debido a que el tamaño de la muestra es pequeña $n = 30$, se debe utilizar preferentemente las tablas de la distribución binomial correspondientes.

Como no existe preferencia previa con respecto a la elección de los tipos de aceite, se tiene una proporción $p_0 = 0.50$, entonces el número de éxitos esperado para esta proporción es

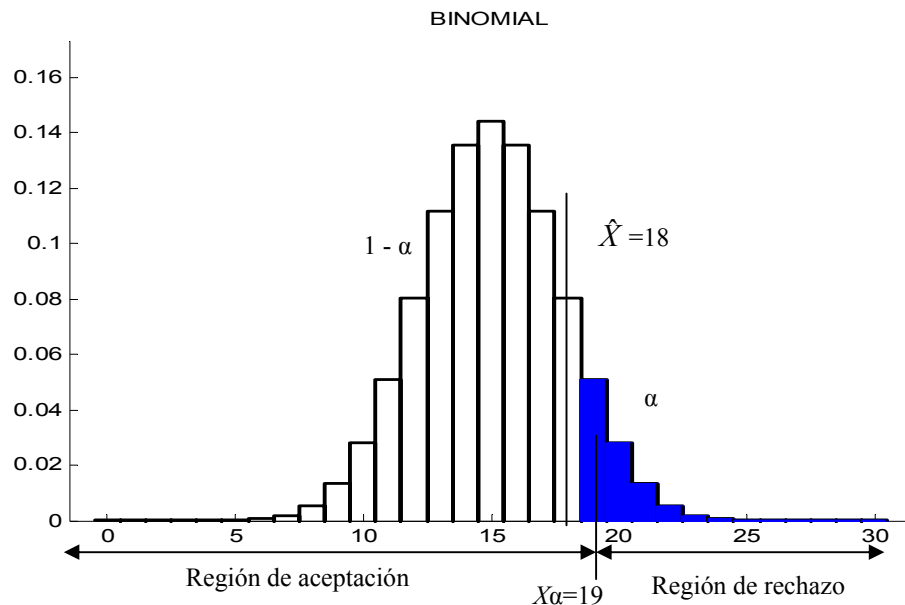
$$\mu = np = (30)(0.5) = 15$$

Las hipótesis nulas y alternativas en competencia son

$$\begin{array}{ll} H_0: p = 0.5 & \text{o} \quad \mu = 15 \\ H_1: p > 0.5 & \text{o} \quad \mu > 15 \end{array}$$

Para el nivel de significancia $\alpha=0.04937$ y considerando la prueba de cola derecha el área a la izquierda es $A=1-\alpha=1-0.04937=0.95063$, buscando en la tabla para la distribución binomial acumulada para $n=30$ y $p=0.5$ se encuentra que el número de éxitos crítico correspondiente es $X_\alpha = 19$

De acuerdo a los datos proporcionados la cantidad de éxitos ó preferencias por el aceite I es $\hat{X} = 18$, entonces, $\hat{X} < X_{\alpha}$ y no debe rechazarse la hipótesis nula.



15. Considérese p , la verdadera proporción de los votantes registrados que están en contra de la pena capital. Supóngase que en el pasado p ha sido igual a 50% menos. Actualmente existen razones para creer que p ha aumentado. Una muestra aleatoria de 20 votantes de una proporción en la muestra del 55 %, ¿Puede llegarse a la conclusión de que la verdadera proporción permanece sin cambio, es decir sin haber aumentado, para $\alpha = 0.0207$?

SOLUCION

Por el tamaño de la muestra es pequeña $n = 20$, se debe utilizar las tablas de la distribución binomial correspondientes.

La proporción previa en contra de la pena capital es $p_0 = 0.50$ lo cual corresponde a una media

$$\mu = np = (20)(0.5) = 10$$

Las hipótesis nulas y alternativas en competencia son

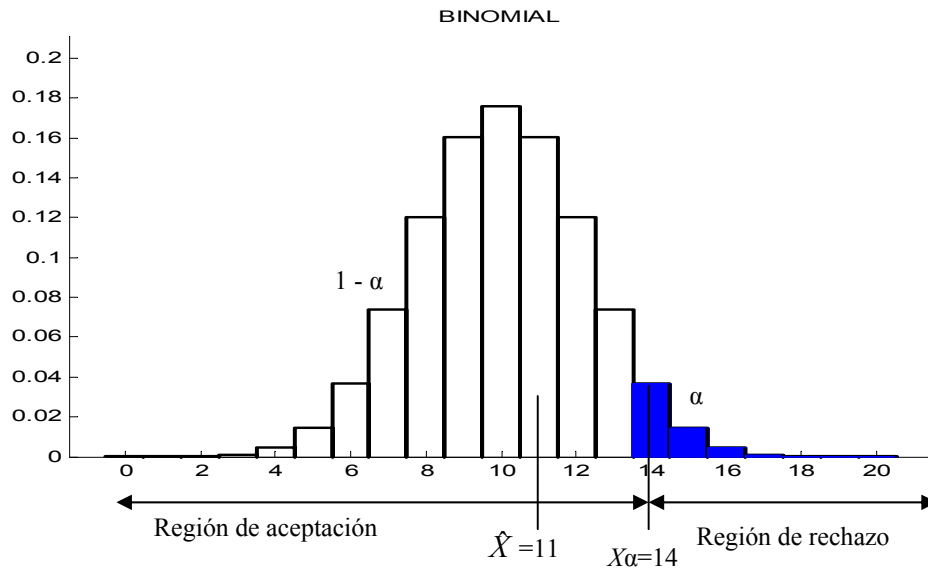
$$\begin{array}{ll} H_0: p = 0.5 & \text{o} \quad \mu = 10 \\ H_1: p > 0.5 & \text{o} \quad \mu > 10 \end{array}$$

Para el nivel de significancia $\alpha = 0.0207$ y considerando la prueba de cola derecha el área a la izquierda de la distribución binomial es $A = 1 - \alpha = 1 - 0.0207 = .9793$, buscando en la tabla para la distribución binomial acumulada para $n = 20$ y $p = 0.5$ se encuentra que el número de éxitos crítico correspondiente es $X_{\alpha} = 14$

De acuerdo a los datos la nueva proporción de votantes en contra de la pena capital es $\hat{p} = 0.55$ por lo que el valor esperado correspondiente a la cantidad de éxitos es

$$\hat{X} = n\hat{p} = (20)(0.55) = 11$$

Como $\hat{X} < X_{\alpha}$ y no debe rechazarse la hipótesis nula.



16. Se ha insinuado que los profesores se han vuelto más despreocupados al calificar a sus estudiantes. En el pasado, 80% de todos los estudiantes universitarios de primer año obtenían C o calificaciones superiores. Una encuesta de la clase más reciente de estudiantes universitarios de primer año muestra que 8100 de los 10 000 estudiantes universitarios de primer año de la muestra recibieron calificaciones de C o mayores. ¿Es verdadero que los profesores se han vuelto más despreocupados, si el nivel de significación se especifica en 0.01?

SOLUCION

La proporción previa de acuerdo a los datos es $p_0 = 0.80$

El tamaño de la muestra es $n = 10000$

Debido al tamaño de la muestra se utilizará la aproximación normal a la binomial.

Utilizando los datos anteriores se tiene que el promedio es

$$\mu = np = 10000(0.80) = 8000 \text{ estudiantes}$$

y la desviación típica de la población es

$$\sigma = \sqrt{npq} = \sqrt{(10000)(0.8)(0.2)} = 40$$

El promedio obtenido del experimento es $\hat{X} = 8100$ estudiantes

Las hipótesis nulas y alternativas del problema son

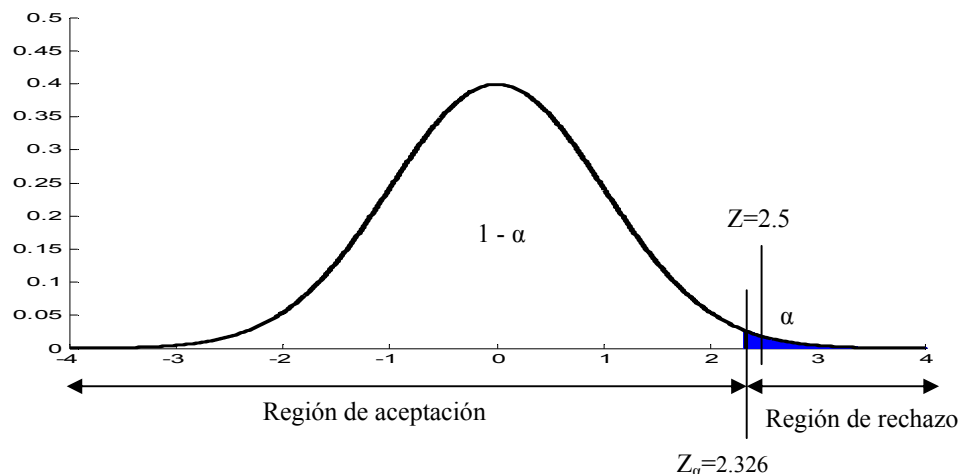
$$\begin{array}{ll} H_0: p = 0.80 & \text{o} \quad \mu = 8000 \\ H_1: p > 0.80 & \text{o} \quad \mu > 8000 \end{array}$$

Para el nivel de significancia $\alpha=0.01$ y la prueba de cola derecha el área a la izquierda es $A=1-\alpha=1-0.01=0.99$, correspondiente a un valor crítico para la distribución normal $Z_\alpha=Z_{0.01}=2.326$

El valor del estadístico de prueba Z es

$$Z = \frac{\bar{X} - \mu}{\sigma} = \frac{8100 - 8000}{40} = 2.5$$

como $Z > Z_\alpha$, se rechaza la hipótesis nula H_0 a favor de H_1 , esto es, los profesores se han vuelto más despreocupados

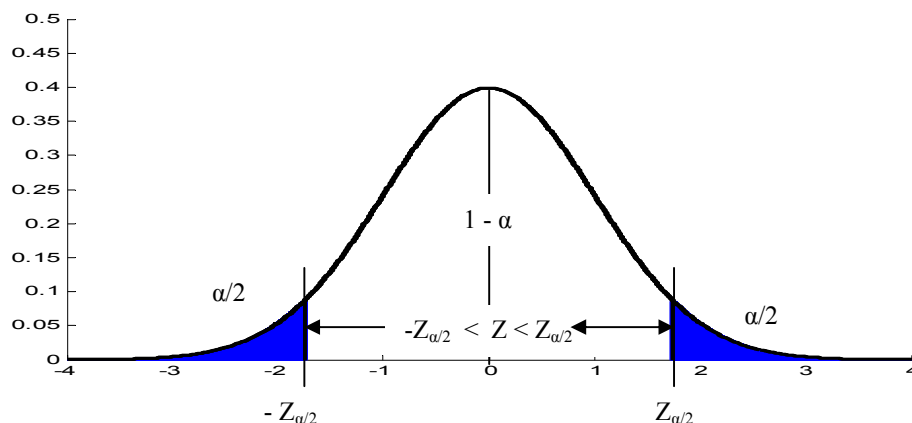


ESTIMACION MATEMATICA

El procedimiento para determinar un intervalo de valores entre los cuales se encuentre el de un parámetro de la población con una probabilidad $1-\alpha$ se conoce como **estimación del intervalo**. El parámetro α se interpreta como la probabilidad de cometer un error en la estimación, por lo que **$1-\alpha$ es la medida de la confianza para la media poblacional**, ó equivalente a la probabilidad de que el parámetro poblacional estimado se encuentre dentro de intervalo adecuado.

ESTIMACION DE LA MEDIA POBLACIONAL

Para mostrar como se obtiene el intervalo de confianza considérese a la media muestral \bar{X} para estimar a la media poblacional μ . Como ha sido mostrado anteriormente, la distribución de la media muestral puede aproximar mediante la distribución normal para el caso de muestras grandes, entonces una proporción $1-\alpha$ del área bajo la curva normal se encuentra entre el intervalo $-Z_{\alpha/2} < Z < Z_{\alpha/2}$ (ver figura siguiente).



Garantizado así que Z se encuentra en el intervalo $-Z_{\alpha/2} < Z < Z_{\alpha/2}$ con una probabilidad $1-\alpha$. Utilizando el hecho de que $Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$, se tiene que

$$-Z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} < Z_{\alpha/2}$$

Multiplicando por -1

$$Z_{\alpha/2} > \frac{-\bar{X} + \mu}{\sigma_{\bar{X}}} > -Z_{\alpha/2}$$

Cambiando el orden de la desigualdad:

$$-Z_{\alpha/2} < \frac{-\bar{X} + \mu}{\sigma_{\bar{X}}} < Z_{\alpha/2}$$

Multiplicando por $\sigma_{\bar{X}}$

$$-Z_{\alpha/2} \sigma_{\bar{X}} < \mu - \bar{X} < Z_{\alpha/2} \sigma_{\bar{X}}$$

Sumando \bar{X}

$$\bar{X} - Z_{\alpha/2} \sigma_{\bar{X}} < \mu < \bar{X} + Z_{\alpha/2} \sigma_{\bar{X}} \quad (5.9)$$

Utilizando finalmente el resultado $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (5.10)$$

ESTIMACION DE LA DIFERENCIA ENTRE DOS MEDIAS

Para obtener un intervalo de confianza de la verdadera diferencia entre dos medias poblacionales $\delta = \mu_1 - \mu_2$ se utiliza el estadístico $\bar{D} = \bar{X}_1 - \bar{X}_2$.

Si se considera que \bar{X}_1 y \bar{X}_2 son independientes y el tamaño de sus respectivas muestras es grande ($n_1, n_2 > 30$), entonces \bar{D} se distribuye normalmente, por otra parte su media y desviación típica son respectivamente

$$\mu_{\bar{D}} = \mu_1 - \mu_2 = \delta \quad \text{y} \quad \sigma_{\bar{D}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Considerando que \bar{D} se distribuye normalmente, el intervalo de confianza se puede obtener utilizando la ecuación (42) simplemente sustituyendo $\mu \rightarrow \delta$, $\bar{X} \rightarrow \bar{D}$ y $\sigma_{\bar{X}} \rightarrow \sigma_{\bar{D}}$

$$\bar{D} - Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \delta < \bar{D} + Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (5.11)$$

ESTIMACION DE LA PROPORCION DE LA POBLACION

Como se ha mencionado anteriormente la proporción p tiene una distribución binomial, pero cuando se cumple las condiciones de la aproximación normal ($np \geq 5$ y $nq \geq 5$) se puede aplicar la ecuación (42) para obtener el intervalo de confianza para la proporción de la población, simplemente realizando los siguientes cambios $\mu \rightarrow np$, $\bar{X} \rightarrow n\hat{p}$, y $\sigma_{\bar{X}} \rightarrow \sqrt{n\hat{p}(1-\hat{p})}$ donde \hat{p} es la proporción estimada a partir de una muestra y $\hat{s} = \sqrt{n\hat{p}(1-\hat{p})}$ es la desviación típica estimada de la variable aleatoria X .

Entonces

$$\bar{X} - Z_{\alpha/2} \sigma_{\bar{X}} < \mu < \bar{X} + Z_{\alpha/2} \sigma_{\bar{X}}$$

$$n\hat{p} - Z_{\alpha/2} \sqrt{n\hat{p}(1-\hat{p})} < np < n\hat{p} + Z_{\alpha/2} \sqrt{n\hat{p}(1-\hat{p})}$$

Dividiendo entre n :

$$\hat{p} - Z_{\alpha/2} \frac{\sqrt{n\hat{p}(1-\hat{p})}}{n} < p < \hat{p} + Z_{\alpha/2} \frac{\sqrt{n\hat{p}(1-\hat{p})}}{n}$$

Finalmente

$$\hat{p} - Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (5.12)$$

EJEMPLOS

17. Supóngase que un psicólogo desea realizar una estimación de intervalo de la media verdadera de los C.I. de alumno, de cierto grupo étnico. Se sabe que los C.I. se distribuyen normalmente con desviación típica de 15. Constrúyase un intervalo de confianza del 95% para la media verdadera (μ) con base en una muestra de 25 alumnos con una media muestral de 105

SOLUCION

Los datos proporcionados por el problema son

Desviación típica $\sigma=15$, media muestral $\bar{X} = 105$, tamaño de la muestra $n = 25$ y intervalo de confianza $1-\alpha=0.95$

A partir del intervalo de confianza $\alpha=1-0.95=0.05$, entonces $\alpha/2=0.025$

El área a la izquierda de la distribución normal es $A = 1-(\alpha/2)=0.975$, buscando en la tabla se obtiene que $Z_{\alpha/2}=1.960$

Sustituyendo en la ecuación 43 **5.10**

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu \leq \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$105 - 1.960 \frac{15}{\sqrt{25}} \leq \mu \leq 105 + 1.960 \frac{15}{\sqrt{25}}$$

$$99.12 < \mu < 110.88$$

18. Una compañía fabricante de harina la empaca en bolsas de papel. Se desea estimar el verdadero peso medio de las bolsas. Una muestra de 36 bolsas da media muestral de 24.5 lb. La desviación típica es de 15 lb. Obténgase el intervalo de confianza del 99 % para su verdadero peso medio de las bolsas de harina.

SOLUCION

Los datos proporcionados por le problema son

Desviación típica $\sigma=15$, media muestral $\bar{X} = 24.5$, tamaño de la muestra $n = 36$ y intervalo de confianza $1-\alpha=0.99$

A partir del intervalo de confianza $\alpha=1-0.99=0.01$, entonces $\alpha/2=0.005$

El área a la izquierda de la distribución normal es $A = 1-(\alpha/2)=0.995$, buscando en la tabla se obtiene que $Z_{\alpha/2}=2.575$

Sustituyendo en la ecuación 43

5.2

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu \leq \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$24.5 - 2.575 \frac{15}{\sqrt{36}} \leq \mu \leq 24.5 + 2.575 \frac{15}{\sqrt{36}}$$

$$18.0625 < \mu < 30.9375$$

19. Se seleccionaron aleatoriamente dos grupos de empleados de una fábrica para entrenarlos a fin de que realicen cierta operación. Cada grupo se entrenó empleando un método diferente. El tiempo promedio para que cada grupo realice la operación después del entrenamiento y otros datos importantes se proporcionan a continuación.

Método 1	Método 2
$n_1=24$	$n_2=36$
$\bar{X}_1=45$	$\bar{X}_2=55$
$\sigma_1^2=200$	$\sigma_2^2=276$

Detérmínesse el intervalo de confianza del 98% para la verdadera diferencia en la efectividad de los dos métodos de entrenamiento.

SOLUCION

A partir del intervalo de confianza $\alpha=1-0.98=0.02$, por lo tanto $\alpha/2=0.01$

El área a la izquierda de la distribución normal es $A = 1-(\alpha/2)=0.99$, buscando en la tabla se obtiene que $Z_{\alpha/2}=2.326$

Utilizando los datos proporcionados se calcula

$$\bar{D} = \bar{X}_1 - \bar{X}_2 = 45 - 55 = -10$$

$$\sigma_{\bar{D}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{200}{24} + \frac{276}{36}} = 4$$

Sustituyendo en la ecuación 44

5.3

$$\bar{D} - Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \delta < \bar{D} + Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$-10 - 2.326(4) < \delta < -10 + 2.326(4)$$

$$-19.304 < \delta < -0.696$$

20. Se realiza un experimento para estimar la verdadera diferencia en la duración promedio de dos marcas de baterías para automóviles. Con la siguiente información determínese el intervalo de confianza del 95% para la verdadera diferencia en la duración de las dos marcas de baterías para automóviles.

	Marca I	Marca II
Tamaño de la muestra	$n_1 = 36$	$n_2 = 36$
Duración promedio (meses)	$\bar{X}_1 = 38$	$\bar{X}_2 = 35$
Varianza	$\sigma_1^2 = 41$	$\sigma_2^2 = 40$

SOLUCION

El intervalo de confianza es $\alpha = 1 - 0.95 = 0.05$, por lo tanto $\alpha/2 = 0.025$

El área a la izquierda de la distribución normal es $A = 1 - (\alpha/2) = 1 - 0.025 = 0.975$, buscando en la tabla se obtiene que $Z_{\alpha/2} = 1.960$

Utilizando los datos proporcionados se calcula

$$\bar{D} = \bar{X}_1 - \bar{X}_2 = 38 - 35 = 3$$

$$\sigma_{\bar{D}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{41}{36} + \frac{40}{36}} = 1.5$$

Sustituyendo en la ecuación 44

$$\bar{D} - Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \delta < \bar{D} + Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$3 - 1.960(1.5) < \delta < 3 + 1.960(1.5)$$

$$0.06 < \delta < 5.94$$

21 Se realizó una investigación de tele audiencia. En una muestra de 900 espectadores, el número de ellos que veían un programa en particular fue de 180. Determinése el intervalo de confianza del 99% para la verdadera proporción de espectadores que ven este programa en particular.

SOLUCION

Tamaño de la muestra $n = 900$, número de espectadores que ven el programa $\bar{X} = 180$, intervalo de confianza es $1 - \alpha = 0.99$

Como el tamaño de la muestra es grande se utiliza la aproximación normal a la binomial.

A partir del intervalo de confianza $\alpha = 1 - 0.99 = 0.01$ entonces $\alpha/2 = 0.005$ y el área a la izquierda de la distribución normal es $A = 1 - 0.005 = .995$, buscando en la tabla correspondiente se obtiene que $Z_{\alpha/2} = 2.575$

La proporción estimada por los datos

$$\hat{p} = \frac{\bar{X}}{n} = \frac{180}{900} = 0.2$$

Sustituyendo los datos en la fórmula (45)

$$\hat{p} - Z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} < p < \hat{p} + Z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$0.2 - 2.575 \sqrt{\frac{0.2(1 - 0.2)}{900}} < p < 0.2 + 2.575 \sqrt{\frac{0.2(1 - 0.2)}{900}}$$

$$0.1656 < p < 0.2343$$

22. En una muestra seleccionada aleatoriamente de 64 muchachas universitarias de primer año, 32 de ellas resultan ser casadas. Determinése el intervalo de confianza del 95% para p , verdadera proporción de todas las mujeres universitarias de primer año que están casadas.

SOLUCION

Tamaño de la muestra $n = 64$, número de casadas $\bar{X} = 32$, intervalo de confianza es $1 - \alpha = 0.95$

Como el tamaño de la muestra es grande se utiliza la aproximación normal a la binomial.

A partir del intervalo de confianza $\alpha = 1 - 0.95 = 0.05$ entonces $\alpha/2 = 0.025$ y el área a la izquierda de la distribución normal es $A = 1 - 0.025 = 0.975$, buscando en la tabla correspondiente se obtiene que $Z_{\alpha/2} = 1.960$

La proporción estimada por los datos

$$\hat{p} = \frac{\bar{X}}{n} = \frac{32}{64} = 0.5$$

Sustituyendo los datos en la fórmula (45)

$$\hat{p} - Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.5 - 1.960 \sqrt{\frac{0.5(1-0.5)}{64}} < p < 0.5 + 1.960 \sqrt{\frac{0.5(1-0.5)}{64}} =$$

$$0.3775 < p < 0.6225$$

DISTRIBUCIÓN DE PROBABILIDAD PARA MUESTRAS PEQUEÑAS

En los problemas de hipótesis anteriores se supuso conocida la varianza poblacional, situación que en la mayoría de los casos no se tiene. La desviación típica de una población se puede estimar a partir de la desviación típica de una muestral, de tal forma que la razón

$$\frac{\bar{X} - \mu}{s / \sqrt{n}} \quad (5.13)$$

Se utiliza como estadístico de prueba. Sin embargo si la muestra es pequeña se tiene que la desviación típica muestral s es bastante distinta a la poblacional σ . Por lo anterior no es posible utilizar la distribución normal para el caso de muestras pequeñas.

La solución del problema anterior de la inferencia estadística acerca de un parámetro de la población utilizando muestras pequeñas y desconociendo la varianza poblacional fue resuelto por W. S. Gosset en 1908 al publicar una distribución de probabilidad la cual describe el comportamiento del estadístico dado por la ecuación (5.13), siempre y cuando la muestra sea obtenida a partir de una población con distribución de probabilidad normal.

DISTRIBUCION T-STUDENT

La distribución **t-Student** se obtiene a partir de considerar que la muestra pequeña se obtiene a partir de una población con distribución normal, si la hipótesis anterior no se cumple será necesario utilizar los métodos no paramétricos para la prueba de hipótesis.

La distribución t-student o simplemente distribución t es al igual que la distribución normal una distribución continua en forma de campana simétrica, cuyo estadístico de prueba es

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}} \quad (5.14)$$

La probabilidad acumulada para la distribución para la distribución t-student es

$$P(-\infty < T < x) = \frac{1}{\sqrt{v\pi}} \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \int_{-\infty}^x \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2} dt \quad (5.15)$$

donde $\Gamma(n) = \int_0^{\infty} t^{n-1} e^{-t} dt$ es la llamada función gamma.

Como se puede observar de la distribución t-student tiene una expresión matemática complicada, y al igual que con la distribución normal recurriremos a las tablas respectivas para la solución de los problemas.

Por otra parte la distribución t student tiene más variabilidad que la distribución normal ya que depende del número de datos n .

Esto es, a diferencia de la distribución normal en la cual el estadístico Z depende de μ y σ que son constantes e independientes del tamaño de la muestra n , en el estadístico T la desviación típica muestral s depende de el tamaño de la muestra n . en consecuencia T es más variable que Z .

La variabilidad de la distribución t-student se asocia con el concepto de **grados de libertad**, es cual es simplemente se define como

$$\nu = n - 1 \quad (5.16)$$

Así se tiene que para cada grado de libertad ν se tendría que utilizar una tabla para la distribución t-student, pero en general para las pruebas de hipótesis respectivas solo son necesarios los valores críticos correspondientes a los valores de significación α más utilizados (10%, 5%, 2.5%, 1%, etc) los cuales son reportados en una sola tabla.

Por otra parte la distribución T-student converge o se aproxima a la normal cuando el número de datos tiende a infinito. Las siguientes figuras muestran una distribución t student para $\nu = 4$ y su comparación con la distribución normal.

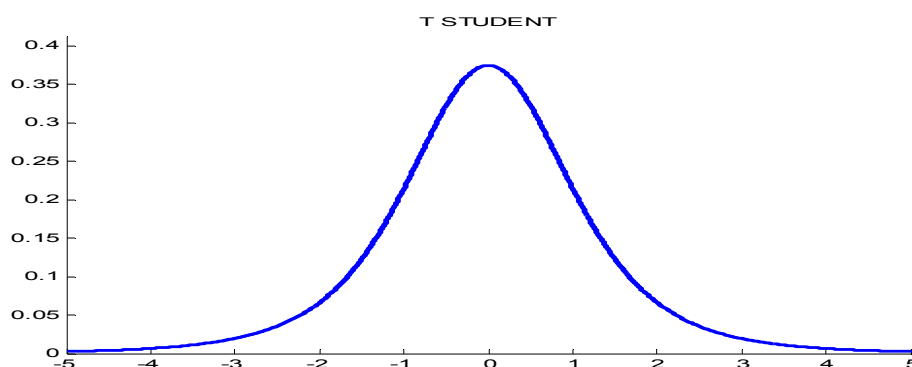


Figura Gráfica de la función t student con $\nu = 4$

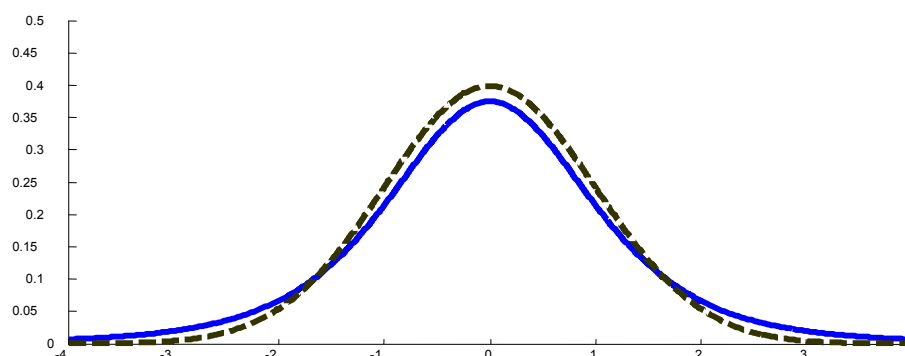


Figura Comparación de la distribución t-student con $\nu = 4$ (línea continua) y la distribución normal respectiva (línea discontinua).

EJEMPLOS

23. Para una distribución con 10 grados de libertad, obténgase el valor crítico t que corta cada una de las siguientes áreas bajo la curva.

- a. El 2.5% superior b. El 5% inferior
c. El 0.005 superior d. El 0.01 inferior

SOLUCION

Recurriendo directamente a la tabla correspondiente de la distribución t -student

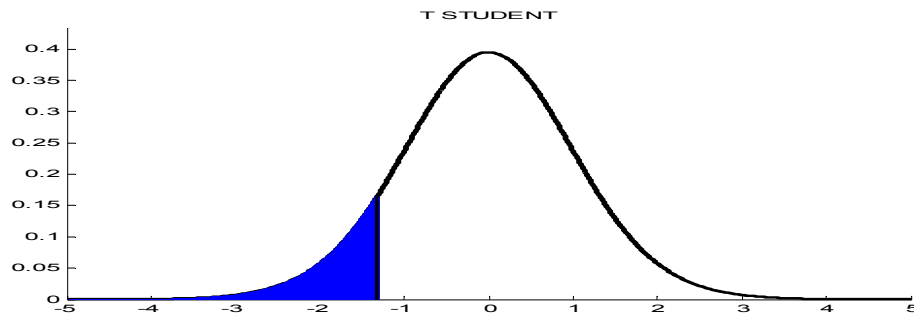
- | | | |
|----|-------------------|-------------------------|
| a) | Superior = 0.025 | $T_{10, 0.05} = 2.228$ |
| b) | El 5% inferior | $T_{10, 0.05} = -1.812$ |
| c) | El 0.005 superior | $T_{10, 0.005} = 3.169$ |
| d) | El 0.01 inferior | $T_{10, 0.01} = -2.764$ |

24. Supóngase que cierta prueba implica un nivel de significación de 0.10 y una muestra de 25 observaciones. Obténgase el valor crítico t bajo cada una de las siguientes condiciones y muéstrese gráficamente cada respuesta.

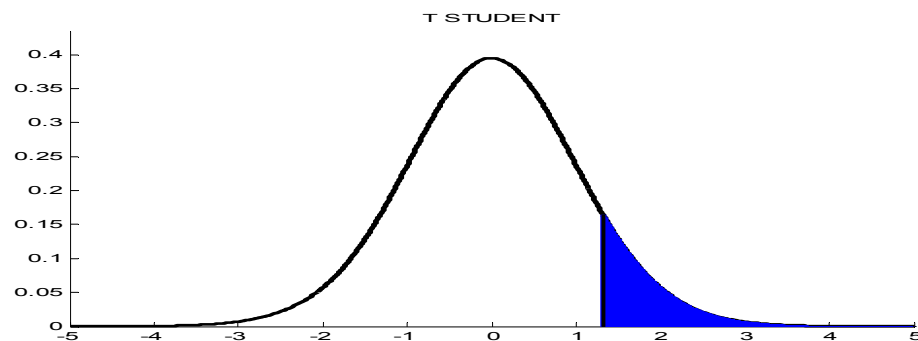
- a. Una prueba de una cola con la región de rechazo en el área de la cola superior.
b. Una prueba de una cola con la región de rechazo en el área de la cola inferior.
c. Una prueba de dos colas.

SOLUCION

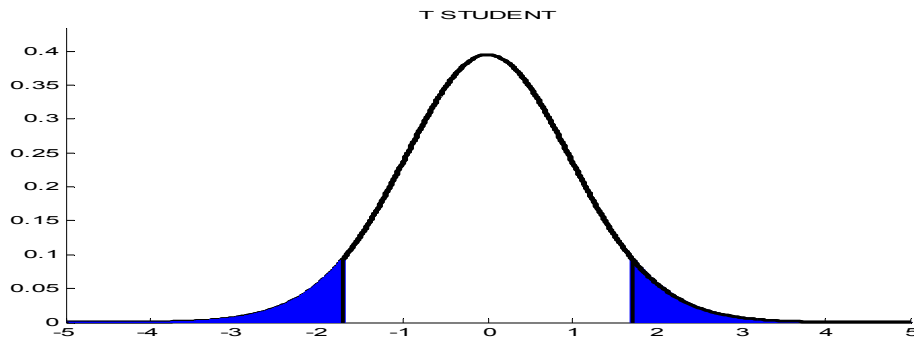
- a) Recurriendo a la tabla de la distribución t -student para $v = n-1 = 25-1 = 24$ y $\alpha = 0.1$ se tiene $T_{0.1, 24} = 1.318$



- b) El valor para el caso de cola inferior es igual al anterior pero negativo $T_{24, 0.1} = -1.318$



c) En el caso de dos colas se tiene que $\alpha/2 = 0.1/2 = 0.05$ lo cual corresponde a $T_{0.1, 24} = 1.711$



25. Sea X el salario por hora de cualquier minero seleccionado al azar y considérese que X se distribuye normalmente. Si los valores críticos t fueran 2.624, 2.492 y 2.423 para $\alpha = 0.01$ con $H_1: \mu > \mu_1$, ¿qué tan grande debería ser el tamaño de la muestra para una prueba de una cola?

SOLUCION

La prueba corresponde a una prueba de cola derecha o superior

$H_0: \mu = \mu_1$

$H_1: \mu > \mu_1$

Buscando en la tabla para la t – student, para $\alpha=0.01$ y los valores de t_α se obtienen directamente

$T_\alpha = 2.624$, entonces $v_1 = 14$ por lo tanto $n = v + 1 = 15$

$T_\alpha = 2.492$, entonces $v_2 = 24$ por lo tanto $n = 24 + 1 = 25$

$T_\alpha = 2.423$, entonces $v_3 = 40$ por lo tanto $n = 40 + 1 = 41$

PRUEBAS PARA LA MEDIA DE LA POBLACION CON MUESTRAS PEQUEÑAS

Cuando la muestra es pequeña la varianza muestral s^2 puede diferir demasiado de la poblacional σ^2 , y no es adecuado ni recomendable utilizar a la puntuación Z como estadístico de prueba, en este caso se debe utilizar a T como estadístico de prueba, esto es para obtener las fórmulas correspondientes a las pruebas de hipótesis y estimación simplemente se puede sustituir a Z por T en las fórmulas correspondientes y utilizar a la distribución t- student en lugar de la normal, siempre y cuando la distribución original de la variable aleatoria X sea normal. Siguiendo la idea anterior, el estadístico de prueba de la media poblacional es dado por la ecuación (5.14)

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

Para la estimación de un intervalo para la verdadera media población μ , con una confianza $1 - \alpha$ para muestras pequeñas se tiene

$$\bar{X} - T_{\alpha/2} \frac{s}{\sqrt{n}} < \mu \leq \bar{X} + T_{\alpha/2} \frac{s}{\sqrt{n}} \quad (5.17)$$

EJEMPLOS

26. La Federal Food and Drug Administration está realizando una prueba para determinar si una nueva medicina tiene el indeseable efecto lateral de elevar la temperatura del cuerpo. Se entiende que la temperatura del cuerpo humano se distribuye normalmente con una media de 98.6 °F. Se administra la nueva medicina a nueve pacientes, se toman las temperaturas y se obtiene una media muestral de 99°F y una desviación típica de 0.36 °F. ¿Debería permitirse a la compañía poner a la venta la nueva droga si el nivel de significación se especifica en 0.01?

SOLUCION

La hipótesis nula y alternativa de problema son

$$H_0: \mu = 98.6$$

$$H_1: \mu > 98.6$$

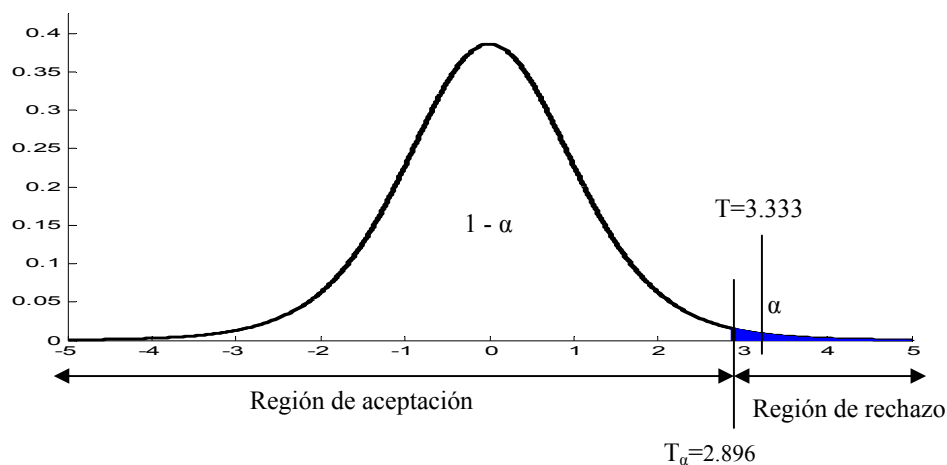
El número de datos es $n = 9$, por lo que los grados de libertad es $\nu = n - 1 = 8$.

Para el nivel de significancia $\alpha = 0.01$ y $T_{\alpha} = T_{8, 0.01} = 2.896$

La media muestral y su respectiva desviación típica es $\bar{X} = 99$, $s = 0.36$, entonces

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{99 - 98.6}{0.36/\sqrt{9}} = 3.333$$

Como $T > T_{\alpha}$ Se rechaza H_0 ya que efectivamente aumenta la temperatura, por lo que no debe salir al mercado



27. Se considera que un proceso de producción está funcionando en forma adecuada cuando la cantidad promedio de café instantáneo que se empaca en un frasco es de 6 oz. Se selecciona una muestra aleatoria de 16 frascos; se determina el promedio muestral como 6.1 oz, con una desviación típica de 0.2 oz. El nivel de significación se especifica en 0.05. Considérese que la cantidad de café en cada frasco tiene una distribución normal.

a. ¿Está funcionando adecuadamente el proceso?

b. ¿Cuáles son los límites de confianza del 95% para su promedio verdadero en vista de la información muestral?

SOLUCION

a) Los datos obtenidos del problema son $n = 16$, $\mu = 6$, $\bar{X} = 6.1$, $s = 0.2$ y $\alpha = 0.05$

El problema se puede plantear como una prueba de dos colas, con las siguientes hipótesis nula y alternativa.

$$H_0: \mu = 6$$

$$H_1: \mu \neq 6$$

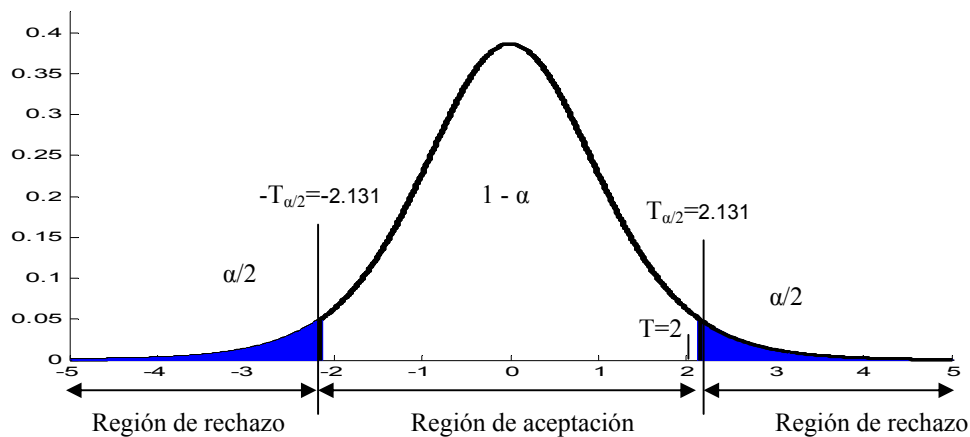
Los grados de libertad es $v = n - 1 = 16 - 1 = 15$.

Para el nivel de significancia $\alpha = 0.05$ y prueba de dos colas $T_{\alpha/2} = T_{15, 0.025} = 2.131$.

A partir de la media muestral y su respectiva desviación típica se tiene que

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{6.1 - 6}{\frac{0.2}{\sqrt{16}}} = 2$$

Como $-T_{\alpha/2} < T < T_{\alpha/2}$ No se rechaza H_0 , La maquinaria funciona adecuadamente.



b) A partir del intervalo de confianza $1 - \alpha = 0.95$, $\alpha = 0.05$ por lo tanto para dos colas $T_{\alpha/2} = 2.131$

$$\bar{X} - T_{\alpha/2} \frac{s}{\sqrt{n}} < \mu \leq \bar{X} + T_{\alpha/2} \frac{s}{\sqrt{n}}$$

$$6.1 - 2.131 \frac{0.2}{\sqrt{16}} < \mu < 6.1 + 2.131 \frac{0.2}{\sqrt{16}}$$

$$5.99345 < \mu < 6.20655$$

28. Se considera que el peso promedio de los reclutas del ejército se distribuye normalmente con una media de 160 lb. En una muestra aleatoria de 25 reclutas, la media es 150 lb y la desviación típica es 20 lb.

- Pruébese la hipótesis nula contra la hipótesis alternativa de que el peso promedio de los reclutas más recientes del ejército es diferente de 160 lb para $\alpha = 0.02$.
- Obtégase el intervalo de confianza del 98% para la media verdadera.

SOLUCION

a) Para este problema $n = 25$, $\mu = 160$, $\bar{X} = 150$, $s = 20$ y $\alpha = 0.02$

El problema plantea una prueba de dos colas, con las siguientes hipótesis nula y alternativa.

$$H_0: \mu = 160$$

$$H_1: \mu \neq 160$$

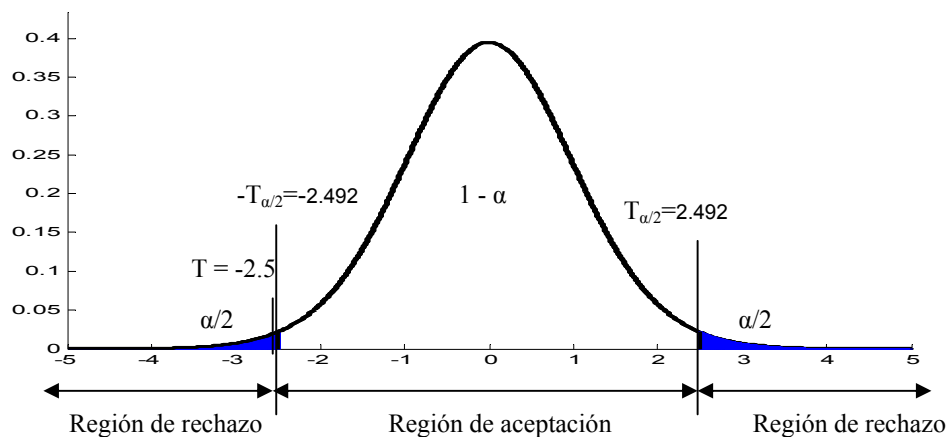
Los grados de libertad es $v = n - 1 = 25 - 1 = 24$.

Para el nivel de significancia $\alpha = 0.02$ y prueba de dos colas $T_{\alpha/2} = T_{0.01, 24} = 2.492$.

Utilizando los valores de la media muestral y su respectiva desviación típica se tiene

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{150 - 160}{20/\sqrt{25}} = -2.5$$

Como $T < -T_{\alpha/2}$, se rechaza H_0 , el peso de los reclutas es diferente.



b) A partir del intervalo de confianza $1 - \alpha = 0.98$, $\alpha = 0.02$ por lo tanto para dos colas $T_{\alpha/2} = 2.492$

$$\bar{X} - T_{\alpha/2} \frac{s}{\sqrt{n}} < \mu \leq \bar{X} + T_{\alpha/2} \frac{s}{\sqrt{n}}$$

$$150 - 2.492 \frac{20}{\sqrt{25}} < \mu < 150 + 2.492 \frac{20}{\sqrt{25}}$$

$$140.032 < \mu < 159.986$$

29. Supóngase que en una línea aérea se desea determinar si el peso promedio de las maletas llevadas por los pasajeros entre Los Angeles y New York es de más de 40 lb. Se selecciona aleatoriamente una muestra de 16 pasajeros y se obtiene una media de 42 lb y una desviación típica de 4 lb. ¿Puede llegarse a la conclusión de que el peso promedio es de más de 40 lb con $\alpha = 0.01$, considerando que los pesos de las maletas se distribuyen normalmente?

a) Los datos obtenidos del problema son $n = 16$, $\mu = 40$, $\bar{X} = 42$, $s = 4$ y $\alpha = 0.01$

El problema se puede plantear como una prueba una cola derecha, con las siguientes hipótesis nula y alternativa.

$$H_0: \mu = 40$$

$$H_1: \mu > 40$$

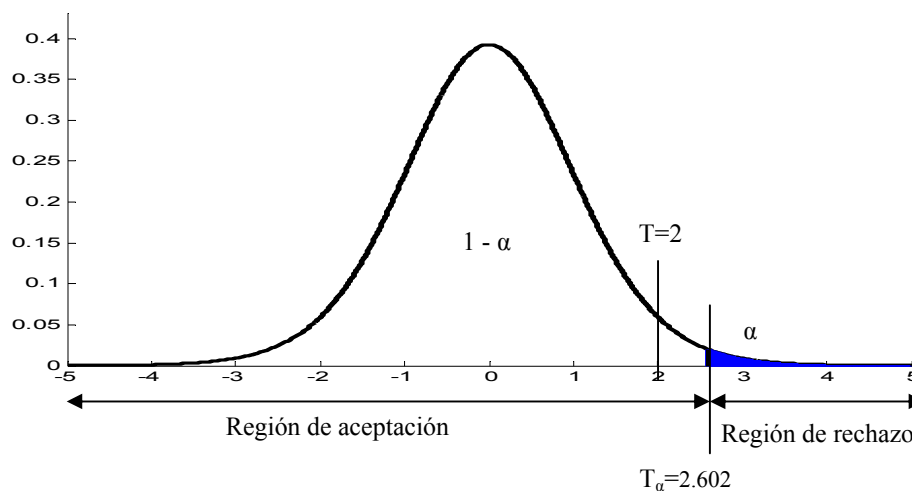
Los grados de libertad es $v = n - 1 = 16 - 1 = 15$.

Para el nivel de significancia $\alpha = 0.01$ y prueba una cola $T_\alpha = T_{15, 0.01} = 2.602$.

La media muestral y su respectiva desviación típica es $\bar{X} = 42$, $s = 4$, entonces

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{42 - 40}{\frac{4}{\sqrt{16}}} = 2$$

Como $T < T_\alpha$ No se rechaza H_0 .



PRUEBA PARA LA DIFERENCIA ENTRE DOS MEDIAS PARA MUESTRAS PEQUEÑAS.

Cuando los patrones de distribución de las poblaciones se distribuyen normalmente o de manera casi normal, y se tiene que las muestras son pequeñas ($n < 30$), se utiliza la prueba t de la distribución t-student para tomar las decisiones. Pero el proceso es diferente para muestras que se consideren independientes y/o dependientes.

En el caso de muestras independientes de tal manera que ninguna se relacione con la otra, se deberá hacer la consideración adicional de que las muestras provienen de poblaciones con idéntica desviación típica con el fin de facilitar el procedimiento, esto es, $\sigma_1 = \sigma_2$.

Como se mencionó anteriormente la varianza de la diferencia muestral $\bar{D} = \bar{X}_1 - \bar{X}_2$ es

$$\sigma_{\bar{D}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

considerando que $\sigma_1 = \sigma_2 = \sigma$ se transforma en

$$\sigma_{\bar{D}}^2 = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

La mejor estimación que se puede hacer de $\sigma_{\bar{D}}^2$ es $S_{\bar{D}}^2$ y el mejor estadístico para estimar σ^2 es s^2 , por lo tanto la expresión anterior se transforma en

$$s_{\bar{D}}^2 = s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

La mejor estimación de s^2 se puede obtener al considerar que se mezclan los datos de ambas muestras, en tal caso se obtiene que

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

por lo que el error típico de la diferencia entre dos medias para muestras pequeñas es

$$s_{\bar{D}} = \sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (5.18)$$

La hipótesis nula para la prueba de la diferencia de medias denotada por δ es

$$H_0: \delta = 0 \quad \text{ó} \quad \mu_1 = \mu_2$$

Para la hipótesis alternativa puede tomar cualquiera de las siguientes posibilidades

$H_1: \delta < 0$	Cola izquierda	$\mu_1 < \mu_2$
$\delta > 0$	Cola derecha	$\mu_1 > \mu_2$
$\delta \neq 0$	Dos colas	$\mu_1 \neq \mu_2$

El estadístico de prueba es

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{D}}} \quad (5.19)$$

Recordando la hipótesis nula $\mu_1 = \mu_2$ y la definición de $\sigma_{\bar{D}}$ dada por la ecuación (5.18)

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (5.20)$$

El valor crítico T_α se determina a partir de el nivel α de significancia, los grados de libertad

$$v = n_1 + n_2 - 2$$

Y buscando en la tabla de la distribución t-student, se realiza la comparación con T y se concluye si se acepta o rechaza la hipótesis nula H_0 .

INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS PARA MUESTRAS PEQUEÑAS

El respectivo intervalo de confianza $1 - \alpha$, para el caso de la diferencia de medias en muestras pequeñas independientes se puede determinar como

$$\bar{D} - T_{\alpha/2} s_{\bar{D}} < \delta \leq \bar{D} + T_{\alpha/2} s_{\bar{D}}$$

o utilizando la expresión (48)

$$\bar{D} - T_{\alpha/2} \sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} < \delta < \bar{D} + T_{\alpha/2} \sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (5.21)$$

EJEMPLOS

30. Se prueban dos motores distintos de automóvil para determinar si presentan diferencias en cuanto a control de contaminación. En una prueba de 16 días del Motor I, las medidas indican un índice promedio de contaminación de 60 y una desviación típica (s_1) de 9; en una prueba de 16 días del Motor II, las mediciones indican un índice promedio de contaminación de 55 y una desviación típica (s_2) de 9. Se cree que las mediciones tienen una distribución normal y varianza idéntica, y que las dos muestras son independientes. ¿Existe suficiente evidencia de que el Motor I y el Motor II tienen distinto control de contaminación para $\alpha = 0.05$?

SOLUCION

Los respectivos datos del problema son

Tamaño de muestra 1 $n_1 = 16$ Tamaño de muestra 2 $n_2 = 16$
 Promedio 1 $\bar{X}_1 = 60$, promedio 2 $\bar{X}_2 = 55$
 Desviación típica 1 $s_1 = 9$ Desviación típica 2 $s_2 = 9$ nivel de significancia $\alpha = 0.05$

Los grados de libertad para el estadístico de prueba son $v = n_1 + n_2 - 2 = 16 + 16 - 2 = 30$

La hipótesis nula y alternativa del problema son respectivamente

$H_0: \mu_1 = \mu_2$
 $H_1: \mu_1 \neq \mu_2$

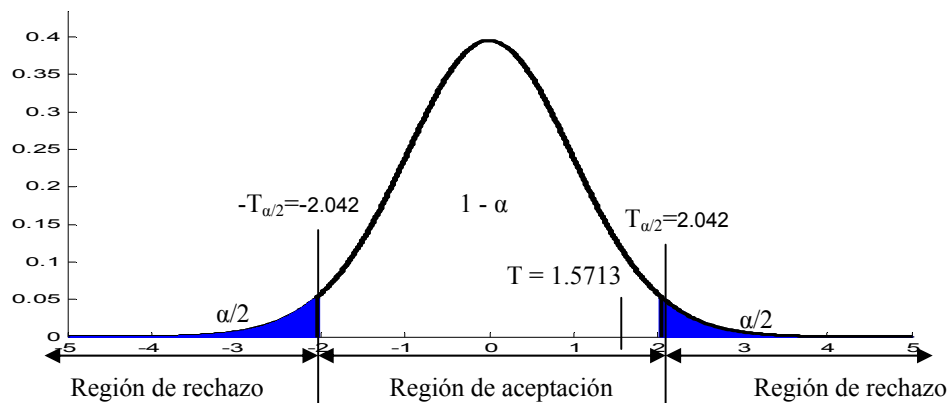
Para el nivel de significancia $\alpha=0.05$ y los grados de libertad $v=30$ y una prueba de dos colas $T_{\alpha/2}=2.042$

Sustituyendo los datos en la ecuación

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$T = \frac{60 - 55}{\sqrt{\frac{(16-1)9^2 + (16-1)9^2}{16+16-2} \left(\frac{1}{16} + \frac{1}{16} \right)}} = 1.5713$$

Como $T_{\alpha/2} < T < T_{\alpha/2}$, no se rechaza H_0 .



31. Se desea determinar si los promedios de puntos de calificación (PPC) son diferentes para niños y niñas. Se considera que el PPC se distribuye normalmente con varianza idéntica para ambos sexos. Dos muestras independientes de cinco estudiantes cada una proporcionan lo siguiente:

PPC para niños: 2.9 3.1 2.7 3.3 3.0

PPC para niñas: 3.6 2.8 3.6 3.2 2.8

a. Utilizando $\alpha = 0.05$, pruébese la hipótesis de que el PPC medio para niños es el mismo que el PPC medio para niñas, contra la hipótesis alternativa de que las dos medias son diferentes.

b. Obténganse los límites de confianza del 95% para la verdadera diferencia entre las dos medias poblacionales.

SOLUCION.

a) Para la solución de problema primero es necesario calcular la media y la desviación típica insesgada para cada uno de los datos dados.

Para los niños la media y la varianza son

$$\bar{X}_1 = \frac{2.9 + 3.1 + 2.7 + 3.3 + 3.0}{5} = 3$$

$$s_1^2 = \frac{(2.9-3)^2 + (3.1-3)^2 + (2.7-3)^2 + (3.3-3)^2 + (3-3)^2}{5-1} = 0.05$$

para las niñas

$$\bar{X}_2 = \frac{3.6 + 2.8 + 3.6 + 3.2 + 2.8}{5} = 3.2$$

$$s_2^2 = \frac{(3.6-3.2)^2 + (2.8-3.2)^2 + (3.6-3.2)^2 + (3.2-3.2)^2 + (2.8-3.2)^2}{5-1} = 0.16$$

Los grados de libertad para el estadístico de prueba son $v = n_1 + n_2 - 2 = 5 + 5 - 2 = 8$

La hipótesis nula y alternativa del problema son respectivamente

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Para el nivel de significancia $\alpha = 0.05$ y los grados de libertad $v = 8$ y una prueba de dos colas $T_{\alpha/2} = 2.306$

Sustituyendo los datos en la ecuación

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$T = \frac{3 - 3.2}{\sqrt{\frac{(5-1)(0.05) + (5-1)(0.16)}{(5+5-2)} \left(\frac{1}{5} + \frac{1}{5} \right)}} = -\frac{0.2}{0.2049} = -0.9760$$

Como $T_{\alpha/2} < T < T_{\alpha/2}$, no se rechaza H_0 .

b) Para el nivel de significancia $1 - \alpha = 0.95$ y una prueba de dos colas con $\alpha = 0.05$ y $v = 8$, se tiene que $T_{\alpha/2} = 2.306$

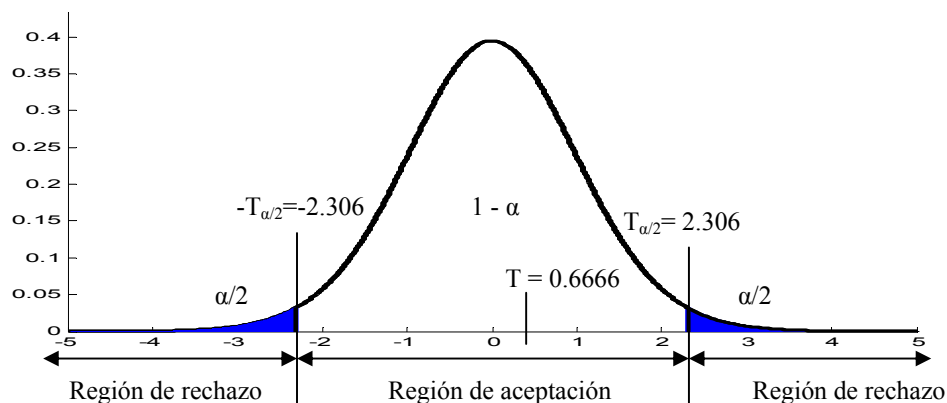
Conviene primero conviene evaluar

$$s_D = \sqrt{\frac{(5-1)(0.05) + (5-1)(0.16)}{(5+5-2)} \left(\frac{1}{5} + \frac{1}{5} \right)} = 0.2049$$

Finalmente evaluado la expresión

$$\bar{D} - T_{\alpha} \sigma < \delta < D + T_{\alpha} \sigma$$

$$-0.2 - (2.306)(0.2049) < \delta < -0.2 + (2.306)(0.2049) = -0.67273 < \delta < 0.27273$$



32. Supóngase que se desea determinar si una dieta completada con una hormona de crecimiento puede aumentar significativamente la ganancia en peso de los cerditos. Con este fin, se seleccionan aleatoriamente dos grupos independientes de cerditos. A un grupo se le alimenta con la dieta acostumbrada y al otro con una dieta con la hormona de crecimiento. Las ganancias de peso para los dos grupos se registran un mes después de que se han estado utilizando las dietas respectivas. a continuación se muestran los datos de importancia.

	Grupo 1 (Dieta acostumbrada)	Grupo 11 (Dieta con hormonas)
Tamaño de la muestra	$n_1 = 21$	$n_2 = 21$
Media muestral (en libras)	$\bar{X}_1 = 16$	$\bar{X}_2 = 19$
Varianza	$s_1^2 = 35$	$s_2^2 = 45$

¿Es posible que la dieta completada con una hormona de crecimiento aumente la ganancia en peso de los cerditos para $\alpha = 0.05$? (Considérese que las ganancias en peso se distribuyen normalmente.)

SOLUCION

Las hipótesis respectivas del problema son:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_2 > \mu_1$$

El número de grados de libertad es $v = n_1 + n_2 - 2 = 21 + 21 - 2 = 40$

Para el nivel de significancia $\alpha = 0.05$ y $v = 40$ y una prueba de cola izquierda $T_{v,\alpha} = -1.684$

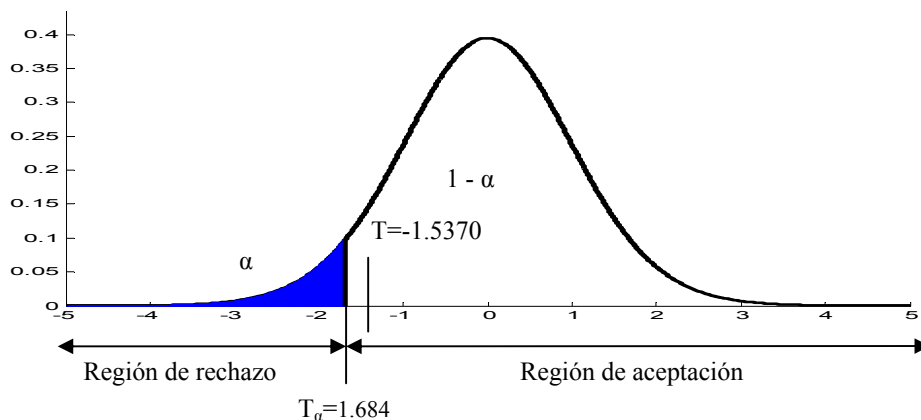
$$s_{\bar{D}} = \sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$s_{\bar{D}} = \sqrt{\left(\frac{(21-1)35 + (21-1)45}{21 + 21 - 2} \right) \left(\frac{1}{21} + \frac{1}{21} \right)} = \sqrt{\left(\frac{700 + 900}{40} \right) \left(\frac{2}{21} \right)} = 1.9518$$

El estadístico de prueba es

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{D}}} = \frac{16 - 19}{1.9518} = -1.5370$$

Puesto que $T < T_{\alpha}$ no se rechaza H_0 .



APROXIMACIÓN NORMAL A LA DISTRIBUCIÓN T-STUDENT

En general en la mayoría de los casos no se conoce la desviación típica de la población. Una forma de solventar esta carencia es observar que la distribución t-student tiende a la distribución normal cuando n es grande, la aproximación se puede aplicar a partir de que $n \geq 30$. La aproximación se realiza simplemente sustituyendo en los estadísticos de prueba de las pruebas de hipótesis la desviación típica o desviaciones típicas por sus correspondientes desviaciones típicas muestrales.

Para la prueba de una media

$$Z = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

Y para la de la diferencia de medias

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

EJEMPLOS

33. Sea Y una variable aleatoria que se sabe tiene una media de 500. Una muestra aleatoria de 900 observaciones para Y proporciona una media $\bar{Y} = 550$ y una varianza $s^2 = 562\,500$.

a. Pruébese la hipótesis de que la media de Y permanece siendo 500 contra la hipótesis alternativa de que es diferente de 500 con $\alpha = 0.01$.

b. Determinése el intervalo de confianza del 99% para la verdadera media.

SOLUCION

a) Los datos que se tienen del problema son

Media poblacional $\mu=500$, número de datos $n = 900$, media muestral $\bar{X} = 550$, varianza muestral $s^2 = 562\,500$ y nivel de significancia $\alpha=0.01$

La hipótesis nula y alternativa es

$$H_0: \mu=500$$

$$H_1: \mu \neq 500$$

Para la prueba de dos colas con $\alpha=0.01$ se tiene que $\alpha/2=0.005$ y $A = 1-\alpha/2 = 0.995$ lo que corresponde de acuerdo a la tabla respectiva de la distribución normal $Z_{\alpha/2} = 2.575$

El estadístico de prueba es

$$Z = \frac{\bar{X} - \mu}{s / \sqrt{n}} = \frac{550 - 500}{750 / \sqrt{900}} = 2$$

Puesto que $-Z_{\alpha/2} < Z < Z_{\alpha/2}$ No se rechaza H_0 .

b) A partir del intervalo de confianza solicitado $1-\alpha = 0.99$, se tiene que, $\alpha = 0.01$, y $\alpha/2 = 0.005$ por lo que $A = 1-\alpha/2 = 0.995$ lo que corresponde $Z_{\alpha/2} = 2.575$

Utilizando la expresión siguiente

$$\bar{X} - Z_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2} \frac{s}{\sqrt{n}}$$

$$550 - 2.575 \frac{750}{\sqrt{900}} < \mu < 550 + 2.575 \frac{750}{\sqrt{900}}$$

$$485.625 < \mu < 614.375$$

34. Un productor de azúcar la empaca en bolsas de papel, cada una de las cuales debe contener 10 lb ó 160 oz. Algunos clientes se han quejado de que las bolsas contienen solamente 9.5 lb ó 152 oz. Se realiza una prueba para determinar si la queja es razonable. Una muestra aleatoria de 49 bolsas proporciona una media de 156 oz y una desviación típica (s) de 10.5 oz. ¿Deberá rechazarse la hipótesis nula de que el peso promedio es de 160 oz en oposición a la hipótesis alternativa a de que es de 152 oz para $\alpha = 0.01$?

SOLUCION

Los datos que se tienen del problema son los siguientes

Media poblacional $\mu=160$, número de datos $n = 49$, media muestral $\bar{X} = 156$, varianza muestral $s^2 = 10.5$ y nivel de significancia $\alpha=0.01$

La hipótesis nula y alternativa es

$$H_0: \mu=160$$

$$H_1: \mu<160$$

La prueba es de cola izquierda, para $\alpha=0.01$ se tiene que $A=1-\alpha = 0.99$, por lo que $Z_\alpha = -2.326$
El estadístico de prueba es

$$Z = \frac{\bar{X} - \mu}{s / \sqrt{n}} = \frac{156 - 160}{10.5 / \sqrt{49}} = -2.666$$

Puesto que $Z < Z_\alpha$ se rechaza H_0 .

35. Un nutriólogo desea comparar la efectividad de dos dietas para reducir de peso. Los siguientes datos se obtienen a partir de dos muestras independientes.

Con $\alpha = 0.10$, ¿existe suficiente evidencia de que la Dieta I produce una pérdida menor de peso que la Dieta II?

	Dieta I	Dieta II
Tamaño de la muestra	$n_1=40$	$n_2=60$
Pérdida promedio de peso en libras	$\bar{X}_1=9$	$\bar{X}_2=11$
Varianza muestral	$s_1^2=20$	$s_2^2=30$

SOLUCION

La hipótesis nula y alternativa del problema son

$$H_0: \mu=\mu$$

$$H_1: \mu_1<\mu_2$$

Correspondiendo a una prueba de una cola izquierda

Para el nivel de significancia $\alpha=0.10$, se tiene que $A = 1-\alpha=0.90$ por lo que $Z_\alpha = -1.282$

El estadístico de prueba es en este caso

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{9-11}{\sqrt{\frac{20}{40} + \frac{30}{60}}} = \frac{9-11}{\sqrt{1}} = -2$$

Puesto que $Z < Z_\alpha$ se rechaza H_0 , la dieta I produce una pérdida de peso que la dieta II

DISTRIBUCION χ^2 (chi cuadrada)

La distribución χ^2 (chi cuadrada) también es conocida como Ji – cuadrada y surge como distribución reprobabilidad de la variable aleatoria $X^2 = \frac{(n-1)s^2}{\sigma^2}$ la cual es utilizada como estadístico de prueba para algunas pruebas de hipótesis, por ejemplo para la prueba de una sola varianza de la población.

La probabilidad acumulada para la distribución χ^2 es

$$P(0 < X^2 < x) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} \int_0^x t^{(\nu-2)/2} e^{-t/2} dt \quad (5.22)$$

De manera semejante a la distribución t-student, la distribución χ^2 depende solamente de un parámetro, que es el número de grados de libertad ($\nu = n-1$), La gráfica de χ^2 para algunos grados de libertad es mostrada a continuación,

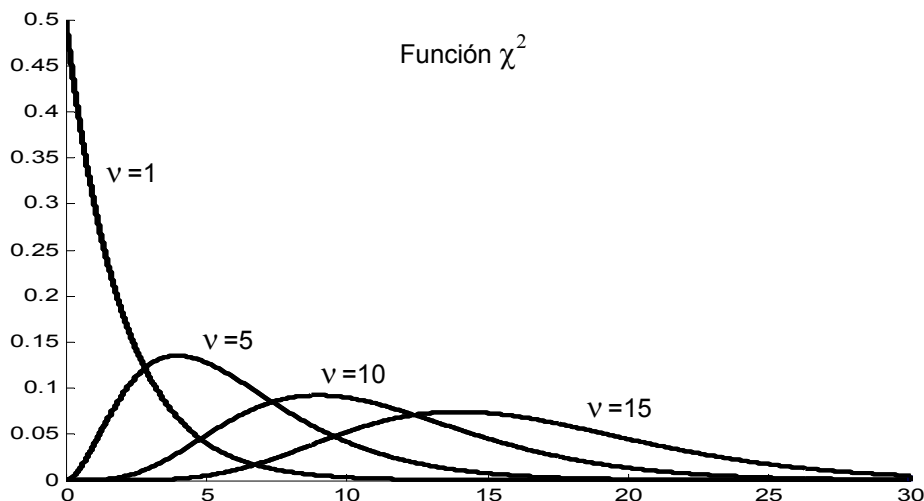


Figura. Gráfica de algunas funciones χ^2 con $\nu=1$, $\nu=5$, $\nu=10$ y $\nu=15$

Observándose que la distribución no tiene simetría para valores pequeños de ν , tendiendo a la simetría respecto a una recta perpendicular que pasa por su valor máximo para valores grandes de ν , además, el valor de χ^2 nunca es negativo pudiendo tomar solamente valores positivos o cero.

Al igual que para las anteriores distribuciones existen tablas de probabilidad acumulada para los valores de significación α más utilizados en la práctica que permiten localizar los valores críticos de χ^2 denotados en ocasiones como $\chi^2_{\nu, \alpha}$, el primer subíndice indica los grados de libertad y el segundo la significancia, como la distribución no tiene valores negativos los valores de para una prueba de cola izquierda es totalmente diferente que el requerido de cola derecha, por ejemplo, para una distribución chi cuadrado con χ^2 grados de libertad para una significancia $\alpha = 0.05$ de cola izquierda se localiza en la tabla respectiva el valor de $\nu = 10$ y $\alpha = 0.95$, esto es debido a que el área bajo la curva reportada en la tabla para la distribución chi cuadrada se calcula de manera inversa a la reportada en las anteriores distribuciones de probabilidad, obteniéndose un valor crítico $\chi^2_{10, 0.95} = 3.9403$ y correspondiente valor para una significancia $\alpha = 0.05$ de cola derecha se localiza directamente $\chi^2_{10, 0.05} = 18.307$. La figura siguiente muestra los valores críticos anteriores para la distribución chi cuadrada con $\nu = 10$.

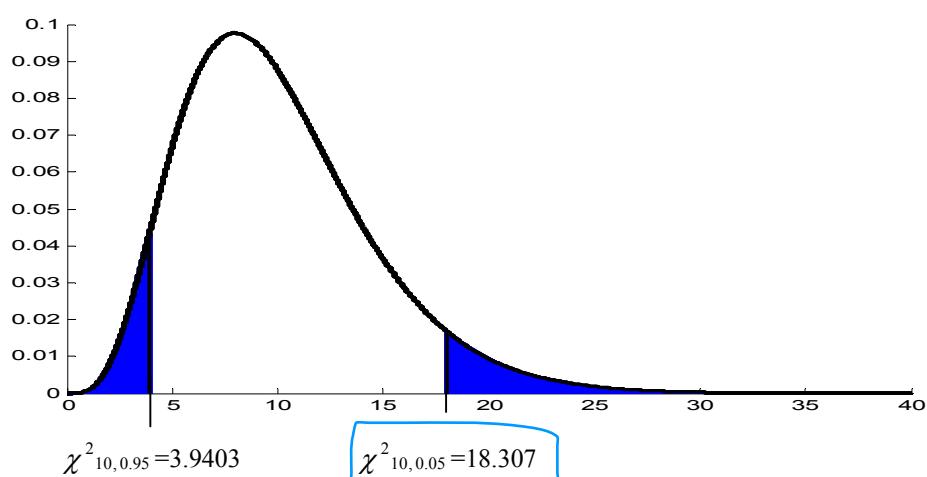


Figura. Representación gráfica de los valores críticos para la distribución chi cuadrada para $\nu = 10$ y $\alpha = 0.95$, para una prueba de cola izquierda y cola derecha.

PRUEBA PARA UNA SOLA VARIANZA

Esta prueba permite comparar la varianza de una población que tiene una distribución normal, con tales condiciones se puede mostrar que el estadístico

$$X^2 = \frac{(n-1)s^2}{\sigma^2} \quad (5.23)$$

tiene una distribución χ^2 con $\nu = n-1$ grados de libertad. En la prueba de la varianza se considera que σ^2 y n son constantes para cada problema particular, por lo que la distribución de s^2 de acuerdo a la ecuación (53) tiene una distribución X^2 . Por lo tanto se puede utilizar la expresión (53) como el estadístico de prueba para realizar la prueba de hipótesis para una sola varianza poblacional. Como en todos los casos de prueba de hipótesis la hipótesis nula se define como

$$H_0: \sigma^2 = \sigma_0^2$$

Y las correspondientes hipótesis alternativas

$$H_1: \begin{array}{l} \sigma^2 > \sigma_0^2 \\ \sigma^2 \neq \sigma_0^2 \\ \sigma^2 < \sigma_0^2 \end{array}$$

Dependiendo de la elección de la hipótesis alternativa y el nivel de significancia α se tomará la decisión, por ejemplo, si $H_1: \sigma^2 > \sigma_0^2$, la hipótesis nula se rechazará solamente cuando $X^2 > \chi^2_{v, \alpha}$.

DETERMINACION DEL INTERVALO DE CONFIANZA PARA LA VERDADERA VARIANZA POBLACIONAL

Para obtener el respectivo intervalo de confianza $1 - \alpha$, para la varianza poblacional se procede como en los casos anteriores utilizando el estadístico de prueba y los respectivos valores críticos $\chi^2_{\alpha/2 \text{ inf}}$ y $\chi^2_{\alpha/2 \text{ sup}}$.

Esto es

$$\chi^2_{v, \alpha/2 \text{ inf}} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{v, \alpha/2 \text{ sup}}$$

Invirtiendo la desigualdad

$$\frac{1}{\chi^2_{v, \alpha/2 \text{ inf}}} > \frac{\sigma^2}{(n-1)s^2} > \frac{1}{\chi^2_{v, \alpha/2 \text{ sup}}}$$

Multiplicando por $(n-1)s^2$

$$\frac{(n-1)s^2}{\chi^2_{v, \alpha/2 \text{ inf}}} > \sigma^2 > \frac{(n-1)s^2}{\chi^2_{v, \alpha/2 \text{ sup}}}$$

Finalmente

$$\frac{(n-1)s^2}{\chi^2_{v, \alpha/2 \text{ sup}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{v, \alpha/2 \text{ inf}}} \quad (5.24)$$

EJEMPLOS

36. Dada una distribución χ^2 con 20 grados de libertad, obténgase el valor χ^2 que corta cada una de las siguientes áreas bajo la curva.

- | | | |
|-----------------|-----------------|-----------------|
| a) 2.5 superior | b) 10% superior | c) 90% superior |
| d) 5% interior | e) 1% interior | |

SOLUCION

Buscando en la tabla $v = 20$ y los correspondientes puntos porcentuales o niveles de significación

- a) $\chi^2_{10, 0.025} = 34.1696$
 b) $\chi^2_{10, 0.10} = 28.4120$
 c) $\chi^2_{10, 0.90} = 12.4426$
 d) $\chi^2_{10, 0.95} = 10.8508$ se busca el 0.95 ya que el área a la izquierda es 0.05.
 e) $\chi^2_{10, 0.99} = 8.2604$ procediendo como en el inciso anterior el área a la izquierda es 0.99

37. Obténganse los puntos porcentuales bajo la cola superior de la distribución con 16 grados de libertad, que estén cortados por los siguientes valores chi cuadrada
 a. 23.5418 b. 26.2962 c. 31.9999

SOLUCION

Buscando en la tabla de la χ^2 y en el número de grados de libertad $v = 16$ los respectivos valores de área se tiene directamente que

- | | | | | | |
|----|---------|---|------|---|-----|
| a) | 23.5418 | → | 0.10 | → | 10% |
| b) | 26.2962 | → | 0.05 | → | 5% |
| c) | 31.999 | → | 0.01 | → | 1% |

38. En una muestra de 10 observaciones tornadas a partir, de una población normal, se encuentra que la varianza s^2 es 15. ¿Cuáles son los límites de confianza del 90% para la varianza de la población?

SOLUCION

Los datos proporcionados en el problema son

Varianza muestral $s^2 = 15$ número de datos $n = 10$ $1 - \alpha = 0.9$

A partir de los datos se tiene que el número de grados de libertad es $v = 10 - 1 = 9$

Del intervalo de confianza $1 - \alpha = 0.9$, el área a la derecha $\alpha/2 = 0.05$, y para el área a la izquierda de la distribución chi -cuadrado $1 - 0.05 = 0.95$, buscando estos valores en la tabla correspondiente para $v = 10$ se tiene

$$\chi^2_{\alpha/2 \text{ inf}} = 3.32511 \quad \chi^2_{\alpha/2 \text{ sup}} = 16.9190$$

Sustituyendo en la ecuación (54)

$$\frac{(10-1)(15)}{3.32511} < \sigma^2 < \frac{(10-1)(15)}{16.9190}$$

$$\rightarrow \quad \bar{x} \pm t_{\alpha/2}$$

39. Cuando un proceso de producción está funcionando adecuadamente, la varianza de las partes producidas es cuatro. Las medidas de las partes se distribuyen normalmente. Se sugiere que el proceso de producción en la actualidad se encuentra fuera de control. Se selecciona aleatoriamente una muestra de nueve partes producidas y se obtienen las siguientes medidas.

9 10 12 13 12 8 6 11 9

- a. Obténgase la varianza s^2
 b. Pruébese la hipótesis de que el proceso de producción sigue funcionando adecuadamente, con $\alpha = 0.10$.
 c. Establézcase el intervalo de confianza del 90% para la verdadera varianza (s^2 , con base en la información muestral.

SOLUCION

a) se puede determinar la varianza muestral insesgada a partir de la ecuación $s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$

x	9	10	12	13	12	8	6	11	9	total
										$\sum x_i = 90$
x^2	81	100	144	169	144	64	36	121	81	$\sum x_i^2 = 940$

sustituyendo

$$s^2 = \frac{940 - \frac{(90)^2}{9}}{9-1} = 5$$

El número de muestras es $n = 9$, por lo tanto el numero de grados de libertad es $v = 9 - 1 = 8$

b) La varianza poblacional es $\sigma^2 = 4$ y el número total de datos es $n = 9$, entonces los grados de libertad son $v = 9 - 1 = 8$

Debido a que el proceso no funciona adecuadamente si la varianza es muy grande a pequeña, la prueba de hipótesis es de dos colas, con las hipótesis nula y alternativa

$$H_0: \sigma^2 = 4$$

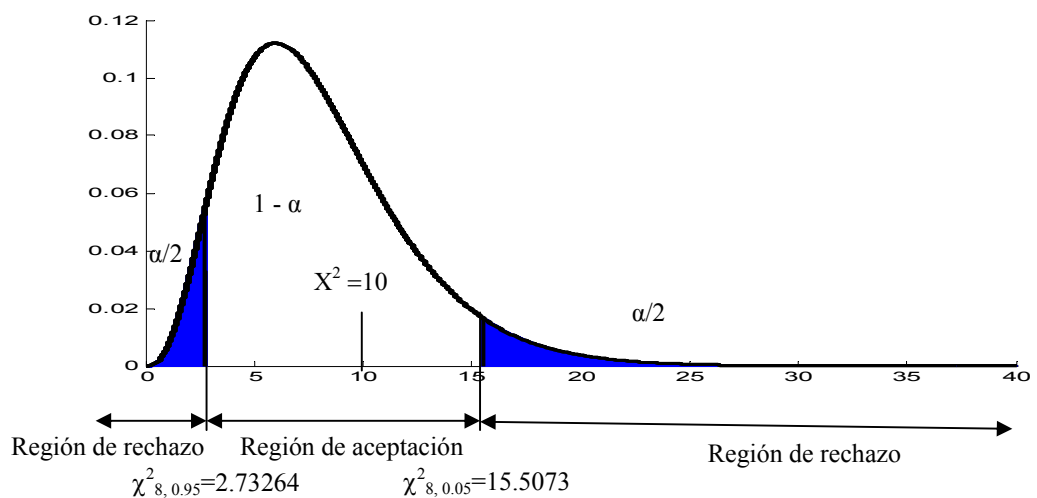
$$H_1: \sigma^2 \neq 4$$

Para el nivel de significancia $\alpha = 0.10$ se tiene para el área a la derecha $\alpha/2 = 0.05$ y el área a la izquierda $1 - \alpha/2 = 1 - 0.05 = 0.95$, por lo que los valores críticos correspondientes para estos valores con $v = 8$, son $\chi^2_{8,9.5 \text{ inf}} = 2.73264$ $\chi^2_{8,0.5 \text{ sup}} = 15.5073$

Evaluando el estadístico de prueba

$$X^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(9-1)(5)}{4} = 10$$

puesto que 15.5073, no se rechaza H_0 , el sistema funciona adecuadamente.



c) Evaluando la ecuación

$$\frac{(n-1)s^2}{\chi^2_{v,\alpha/2\sup}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{v,\alpha/2\inf}}$$

$$\frac{(9-1)(5)}{2.73264} < \sigma^2 < \frac{(9-1)(5)}{15.5073}$$

$$2.5794 < \sigma^2 < 14.6378$$

40. Se sugiere que después de firmar un contrato laboral, la producción por hora de los trabajadores mostrará una variación mayor que antes de firmar el contrato. Se sabe que la varianza de las producciones por hora antes del contrato laboral era de $\sigma^2 = 80$. Considérese que las producciones por hora se distribuyen normalmente. Se selecciona una muestra aleatoria de 30 trabajadores y se obtienen sus producciones por hora después de la firma del contrato. Se encuentra que la varianza de la muestra es 90 ($s^2 = 90$). ¿Debe llegarse a la conclusión de que la dispersión de las producciones por hora ha aumentado significativamente, con $\alpha = 0.05$?

SOLUCION

La varianza poblacional es $\sigma^2 = 80$, la varianza muestral es $s^2 = 90$, el tamaño de muestra es 30 y el nivel de significancia es $\alpha = 0.05$, entonces los grados de libertad son $v = 30 - 1 = 29$. Las hipótesis de la prueba son

$$H_0: \sigma^2 = 80$$

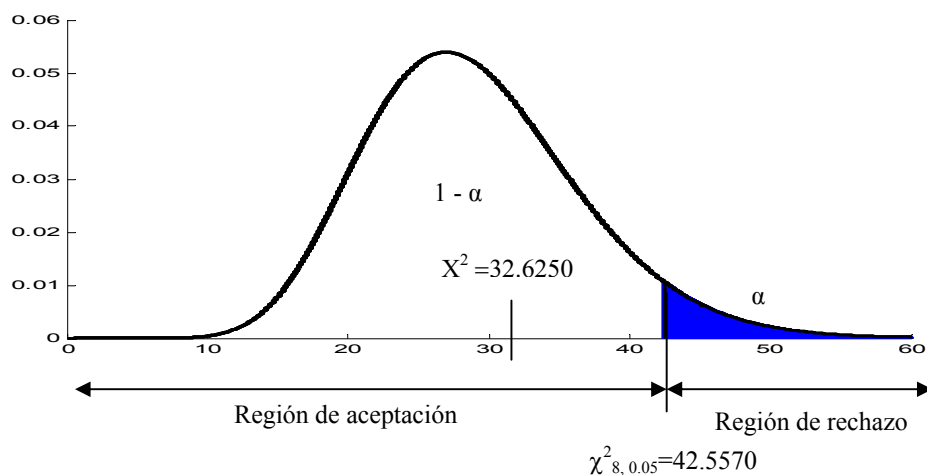
$$H_1: \sigma^2 > 80$$

Situación correspondiente a una de cola derecha.

Para estas condiciones el valor crítico es $\chi^2_{29, 0.05} = 42.5570$ y en valor del estadístico de prueba

$$X^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(30-1)(90)}{80} = 32.6250$$

Como $X^2 = 32.6250 < 42.5570$ no se rechaza H_0 .



PRUEBA DE BONDAD DE AJUSTE

Mediante esta prueba se puede verificar si los datos obtenidos de un experimento particular siguen alguna distribución particular, por ejemplo, una distribución uniforme, distribución binomial, distribución normal, etc. La prueba necesita la clasificación de los datos muestrales en una tabla de distribución de frecuencia denominada **frecuencias observadas** y esta se compara con **las frecuencias esperadas** obtenidas utilizando alguna distribución elegida, las frecuencias observadas se denotan por la letra O y las correspondientes esperadas con la letra E tal como se muestra a continuación.

I	E ₁	E ₂	E ₃	E _J
II	O ₁	O ₂	O ₃	O _J

El estadístico de prueba X^2 está definido como

$$X^2 = \sum_{k=1}^J \frac{(O_k - E_k)^2}{E_k} \quad (5.25)$$

Donde la sumatoria se lleva a cabo sobre todas las frecuencias ó clases (J) en que han sido dividido los datos. Cuando el tamaño de la muestra es grande de tal manera que ninguna frecuencia esperada es menor a 5, X^2 se distribuye aproximadamente siguiendo una distribución chi cuadrada con $v = J - 1$, grados de libertad.

Por la definición dada al estadístico de prueba en la ecuación (55), la prueba de hipótesis es de una cola derecha, que indica que el ajuste o comparación con la distribución esperada es bueno si la diferencia entre los valores observados son muy parecidos a los esperados dando por resultado un valor de X^2 pequeño, pero cuando el valor de X^2 es más grande que un valor especificado (valor crítico $\chi^2_{v,\alpha}$), la hipótesis nula se rechaza indicando que no existe suficiente evidencia para decir que los datos propuestos tienen la distribución propuesta.

EJEMPLOS

41. Se supone que una tabla de dígitos aleatorios es no sesgada; esto es, cada uno de los 10 dígitos debe tener la misma probabilidad de aparecer. Para probar si éste es o no en realidad el caso, se selecciona una muestra de 100 dígitos y se obtienen los siguientes resultados.

Dígito:	0	1	2	3	4	5	6	7	8	9	Total
Número de veces:	8	11	10	14	7	12	6	9	13	10	100

¿Debería rechazarse la hipótesis de que los dígitos de la tabla están arreglados aleatoriamente, con $\alpha = 0.05$?

SOLUCION

El número de clases es $J = 10$, por lo tanto, los grados de libertad son $v = J - 1 = 10 - 1 = 9$.

Para el nivel de significancia $\alpha = 0.05$ y 9 grados de libertad el valor crítico es $\chi^2_{v,\alpha} = \chi^2_{9,0.05} = 16.9190$

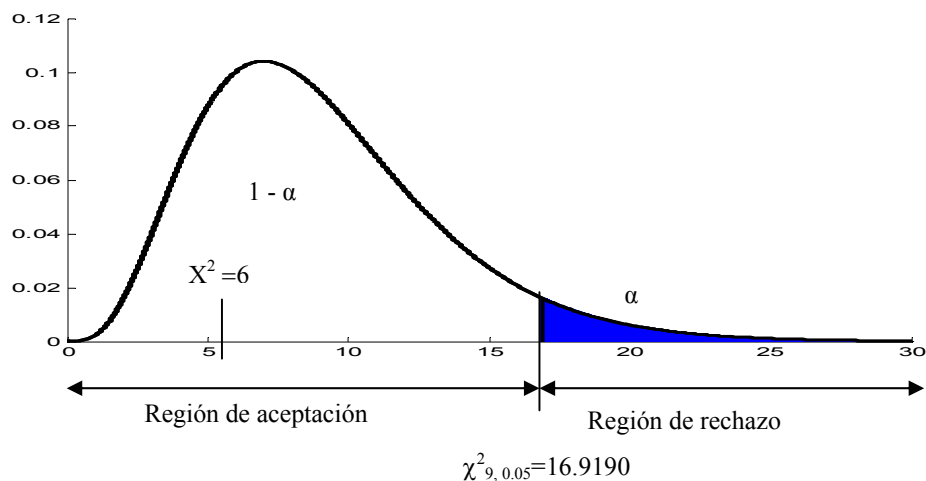
Considerando la distribución uniforme, se tiene que el valor esperado correspondiente es

Dígito:	0	1	2	3	4	5	6	7	8	9	Total
Frecuencia esperada	10	10	10	10	10	10	10	10	10	10	100

A partir de las tablas anteriores se calcula el estadístico de prueba

$$X^2 = \sum_{k=1}^J \frac{(O_k - E_k)^2}{E_k} = (8-10)^2/10 + (11-10)^2/10 + (10-10)^2/10 + (14-10)^2/10 + (7-10)^2/10 + (8-10)^2/10 + (6-10)^2/10 + (9-10)^2/10 + (13-10)^2/10 + (10-10)^2/10 = 6$$

Como $6 < 16.9190$ no se rechaza H_0 , La distribución si es uniforme.



42. Se arrojan simultáneamente cuatro monedas balanceadas 160 veces. A continuación se muestran los resultados.

Número de caras:	0	1	2	3	4	Total
Frecuencia observada:	16	35	55	48	6	160

Con $\alpha = 0.05$, pruébese la hipótesis nula de que las cuatro monedas están todas bien balanceadas y fueron arrojadas aleatoriamente.

SOLUCION

La distribución de probabilidad para el experimento de arrojar cuatro monedas balanceadas se muestra a continuación

x	0	1	2	3	4
f(x)	1/16	4/16	6/16	4/16	1/16

Por lo que las frecuencias esperadas para el experimento

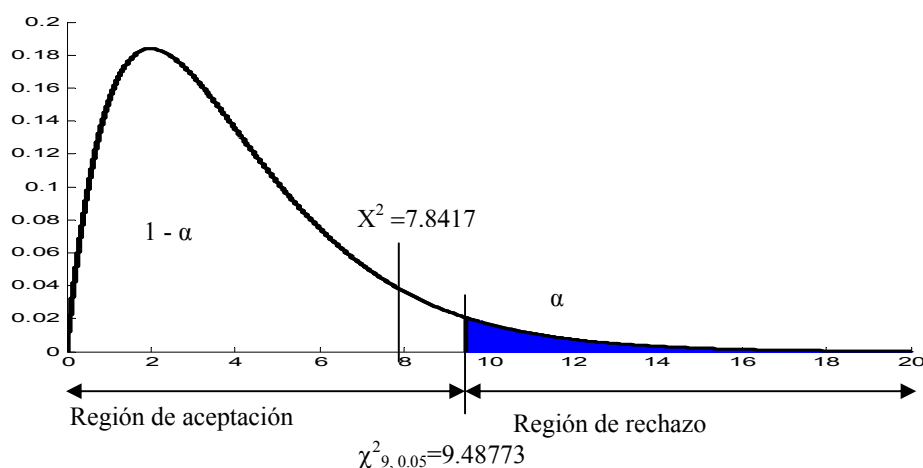
Número de caras:	0	1	2	3	4	Total
Frecuencia esperada:	10	40	60	40	10	160

El número de clases es $J = 5$, por lo que $v = J - 1 = 4$, el valor crítico es para el nivel de significancia $\alpha = 0.05$ es $\chi^2_{v,\alpha} = \chi^2_{4, 0.05} = 9.48773$.

El estadístico de prueba es

$$X^2 = \sum_{k=1}^J \frac{(O_k - E_k)^2}{E_k} = \frac{(16-10)^2}{10} + \frac{(35-40)^2}{40} + \frac{(55-60)^2}{60} + \frac{(48-40)^2}{40} + \frac{(6-10)^2}{10} = 7.8417$$

Como $7.8417 < 9.48773$ no se rechaza H_0 , las monedas se encuentran bien balanceadas.



43. En un experimento con chícharos, un biólogo observa 186 plantas altas y coloridas, 66 altas y sin color, 54 bajas y coloridas, y 14 bajas y sin color. De acuerdo a la teoría de la herencia de Mendel, sería de esperarse que las diferentes categorías tuvieran las siguientes proporciones: 9:3:3:1. ¿Existe suficiente evidencia para apoyar la teoría de Mendel, al nivel de significación del 0.01?

SOLUCION

La información de la frecuencia observada del experimento se resume en la siguiente tabla

Clases	Altas y color	Altas sin color	Bajas con color	Bajas sin color	Total
Frecuencia observada	186	66	54	14	320

Las proporciones del problema son 9:3:3:1, lo cual se puede traducir en términos de la probabilidad en

$9x + 3x + 3x + x = 1$, de donde $x = 1/16$, por lo que las frecuencias esperadas son

$$9/16 \times 320 = 180 \quad 3/16 \times 320 = 60 \quad 3/16 \times 320 = 60 \quad 1/16 \times 320 = 20$$

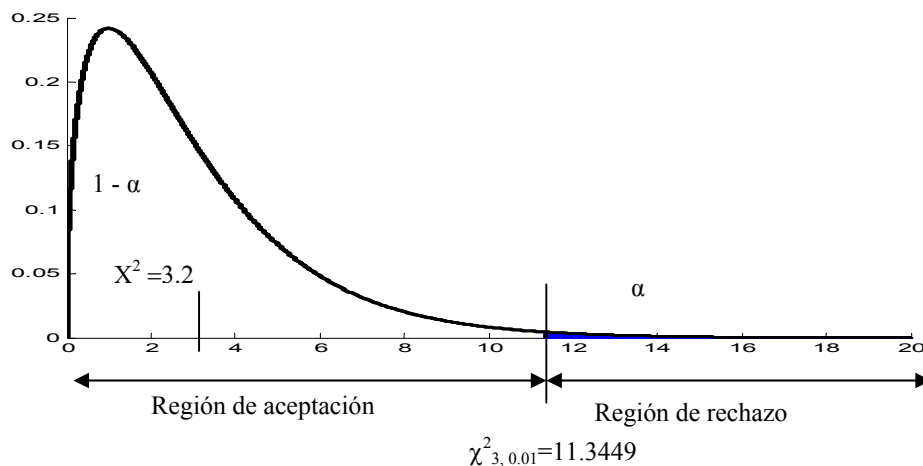
Clases	Altas y color	Altas sin color	Bajas con color	Bajas sin color	Total
Frecuencia esperada	180	60	60	20	320

El número de clases es $J = 4$, por o que $v = J - 1 = 3$, el valor crítico es para el nivel de significancia $\alpha = 0.01$ es $\chi^2_{v,\alpha} = \chi^2_{3, 0.01} = 11.3449$.

El estadístico de prueba es

$$X^2 = \sum_{k=1}^J \frac{(O_k - E_k)^2}{E_k} = \frac{(186 - 180)^2}{180} + \frac{(66 - 60)^2}{60} + \frac{(54 - 60)^2}{60} + \frac{(14 - 20)^2}{20} = 3.2$$

Como $3.2 < 11.3449$, no se rechaza H_0 , el experimento cumple las leyes de Mendel.



PRUEBA DE INDEPENDENCIA

Otro tipo de prueba donde se puede aplicar la distribución chi cuadrado en la prueba de independencia donde se toma la decisión acerca de si una variable es independiente de la otra de otra variable. La hipótesis nula se establece suponiendo que son independientes. Los datos se acomodan en una tabla llamada **tabla de contingencia**, en la cual existe N clases o categorías de renglón y M clases o categorías de columna. Al final de cada una de las filas o columnas se escriben los **totales marginales** de fila R_j o columna C_k . La intersección de cada columna y fila da una celda C_{jk} que es la frecuencia observada. A continuación se muestra una tabla de contingencia general.

C_{11}	C_{12}	---	C_{1k}	---	---	C_{1M}	
C_{21}	C_{22}	---	C_{2k}	---	---	C_{2M}	R_1
C_{31}	---	---	---	---	---	---	R_2
---	---	---	---	---	---	---	---
C_{j1}	C_{j2}		C_{jk}			C_{jM}	R_j
---	---	---	---	---	---	---	---
C_{N1}	C_{N2}	---	C_{Nk}	---		C_{NM}	R_N
C_1	C_2	---	C_i	---	C_k	C_M	

El estadístico de prueba es una generalización del utilizado en la prueba de bondad de ajuste, por lo que es necesario calcular primero los valores esperados E_{jk} , los cuales se pueden obtener a partir de los

totales marginales de fila R_j , los totales marginales de columna C_k y el número total de datos n , mediante la siguiente ecuación.

$$E_{jk} = \frac{R_j \cdot C_k}{n} \quad (5.26)$$

El estadístico de prueba para probar la independencia de dos variables es:

$$X^2 = \sum_{j=1}^N \sum_{k=1}^M \frac{(C_{jk} - E_{jk})^2}{E_{jk}} \quad (5.27)$$

La cual tiene una distribución chi cuadrado con $v = (N - 1)(M - 1)$ número de grados de libertad.

La prueba es una prueba de cola derecha, y se rechazará la hipótesis nula H_0 si el valor del estadístico de prueba es lo suficientemente grande para superar el valor crítico establecido a partir de la significancia α y de el número de grados de libertad v . El rechazo de la hipótesis nula implicará que las variables son dependientes, en caso contrario serán independientes.

EJEMPLOS

44. Supóngase que la siguiente es la distribución de frecuencias observada de 1000 votantes clasificados según el partido al que están afiliados y su preferencia al votar con respecto a cierto asunto.

Pref. al votar	Demócratas	Republicanos	Total
En contra	250	200	450
A favor	400	150	550
Total	650	350	1000

Pruébese la hipótesis de que la preferencia al votar no esta relacionada con la afiliación de partido, con $\alpha = 0.05$.

SOLUCION

A partir de los totales marginales y el total de datos se obtienen los valores esperados E_{ij} utilizando la ecuación $E_{jk} = \frac{R_j \cdot C_k}{n}$. Los resultados esperados son acomodados en la siguiente tabla

Pref. al votar	Demócratas	Republicanos	Total
En contra	292.5	157.5	450
A favor	357.5	192.5	550
Total	650	350	1000

A partir de las dos tablas anteriores se calcula el estadístico de prueba

$$X^2 = \sum_{j=1}^N \sum_{k=1}^M \frac{(C_{jk} - E_{jk})^2}{E_{jk}} = \frac{(250 - 292.5)^2}{292.5} + \frac{(200 - 157.5)^2}{157.5} + \frac{(400 - 357.5)^2}{357.5} + \frac{(150 - 192.5)^2}{192.5}$$

$$= 32.079$$

El número de grados de libertad para el problema es $v = (2 - 1)(2 - 1) = 1$, Por lo que el valor crítico es $\chi^2_{v,\alpha} = \chi^2_{1,0.05} = 3.84146$

Puesto que $3.84146 < 32.079$ se rechaza H_0 , por lo que si hay dependencia en las variables,

45. Se realiza una investigación para determinar si la calificación de desempeño en el trabajo es independiente de los logros académicos en universidad. Se selecciona aleatoriamente una muestra de 100 empleados y su clasificación en una tabla de 3 por 3 se muestra a continuación.

Calificación de desempeño	Nivel académico en universidad			Total
	A	B	C o menos	
Excelente	10	5	5	20
Promedio	20	12	8	40
Malo	20	13	7	40
Total	50	30	20	100

Especificando el nivel de significación en 0.01, ¿debe llegarse a la conclusión de que la calificación de desempeño en el trabajo no está relacionada con los logros académicos en universidad?

SOLUCION

Primero se construye la tabla de continencia de los valores esperados utilizando la ecuación

$$E_{jk} = \frac{R_j \cdot C_k}{n}$$

Calificación de desempeño	Nivel académico en universidad			Total
	A	B	C o menos	
Excelente	10	6	4	20
Promedio	20	12	8	40
Malo	20	12	8	40
Total	50	30	20	100

Procediendo a calcular el estadístico de prueba

$$X^2 = \sum_{j=1}^N \sum_{k=1}^M \frac{(C_{jk} - E_{jk})^2}{E_{jk}} = \frac{(10 - 10)^2}{10} + \frac{(20 - 20)^2}{20} + \frac{(20 - 20)^2}{20} + \frac{(5 - 6)^2}{6} + \frac{(12 - 12)^2}{12} +$$

$$+ \frac{(13-12)^2}{12} + \frac{(5-4)^2}{4} + \frac{(8-8)^2}{8} + \frac{(8-7)^2}{8} = 0.54166$$

El número de grados de libertad para el problema es $v = (3 - 1)(3 - 1) = 4$, Por lo que el valor crítico para $v = 4$ y $\alpha = 0.01$ es $\chi^2_{v,\alpha} = \chi^2_{4,0.01} = 13.2767$

Puesto que $0.54166 < 13.2767$ no se rechaza H_0 , por lo que las variables son independientes.

dependientes

46. Un psicólogo realizó un experimento para determinar si el desempeño de los estudiantes está relacionado con el método utilizado en cierto tema. Se están considerando tres métodos de enseñanza: I, II, y III, y el desempeño de los estudiantes se clasifica como A, B o C. Los resultados fueron los siguientes.

Pruébese la hipótesis nula de que el desempeño de los estudiantes no está relacionado con el método de enseñanza, con $\alpha = 0.01$.

INCOMPLETO

SOLUCION

Construyendo primero la tabla de contingencia de los valores esperados utilizando la ecuación

$$E_{jk} = \frac{R_j \cdot C_k}{n}$$

Desempeño	METODOS DE ENSEÑANZA			Total
	I	II	III	
A	7.5	15	7.5	30
B	10	20	10	40
C	7.5	15	7.5	30
Total	25	50	25	100

Calculando el estadístico de prueba

$$X^2 = \sum_{j=1}^N \sum_{k=1}^M \frac{(C_{jk} - E_{jk})^2}{E_{jk}} = \frac{(5-7.5)^2}{7.5} + \frac{(15-10)^2}{10} + \frac{(5-7.5)^2}{7.5} + \frac{(20-15)^2}{15} + \frac{(15-20)^2}{20} + \frac{(15-15)^2}{15} + \frac{(5-7.5)^2}{7.5} + \frac{(10-10)^2}{10} + \frac{(10-7.5)^2}{7.5} = 8.73$$

El número de grados de libertad para el problema es $v = (3 - 1)(3 - 1) = 4$, Por lo que el valor crítico es

$$\chi^2_{v,\alpha} = \chi^2_{4,0.01} = 13.2767$$

Puesto que $8.73 < 13.2767$ no se rechaza H_0 , por lo que no hay dependencia en las variables,

PRUEBA DE FISHER

R. A. Fisher, quien fue el primero en obtener la distribución y desarrollar la prueba, de ahí el nombre de la distribución. La prueba F se utiliza principalmente para probar la igualdad entre dos varianzas poblacionales que provienen de poblaciones que tiene una distribución normal, también se ha desarrollado un procedimiento basado en esta prueba para investigar la igualdad entre tres ó más medias poblacionales, procedimiento que comúnmente se denomina análisis de varianza (ANOVA).

El estadístico de prueba para la prueba F es la razón de los estimadores insesgados de de dos varianzas poblacionales

$$F = \frac{s_1^2}{s_2^2} > 1 \quad (5.28)$$

Se debe cumplir siempre que $s_1^2 > s_2^2$ para que la razón sea mayor que uno ($F \geq 1$).

La probabilidad acumulada para la distribución Fisher se obtiene de la siguiente ecuación

$$P(0 < F < x) = \frac{\Gamma\left(\frac{v_1 + v_2}{2}\right) v_1^{v_1/2} v_2^{v_2/2}}{\Gamma\left(v_1/2\right)\Gamma\left(v_2/2\right)} \int_0^x t^{(v_1/2)-1} (v_2 + v_1 t)^{-(v_1+v_2)/2} dt \quad (5.29)$$

La distribución F tiene 2 variables v_1 y v_2 que son los grados de libertad de cada una de las poblaciones.

$v_1 = n_1 - 1$ grados de libertad de la población 1

$v_2 = n_2 - 1$ grados de libertad de la población 2

Entonces, para cada pareja de valores v_1 y v_2 se tendrá una tabla correspondiente a los valores porcentuales de α más utilizados. En general los valores críticos F_{α, v_1, v_2} es diferente de F_{α, v_2, v_1} , esto es, si se intercambian los valores de v_1 y v_2 no se obtiene el mismo valor crítico, por lo que hay que tener cuidado al utilizar las tablas y recordar que v_1 se asocia la población que tiene la mayor varianza y v_2 a la que tiene la menor varianza. Algunas gráficas de la distribución F se muestran a continuación. Se observa que la distribución no tiene simetría en ningún caso mostrado.

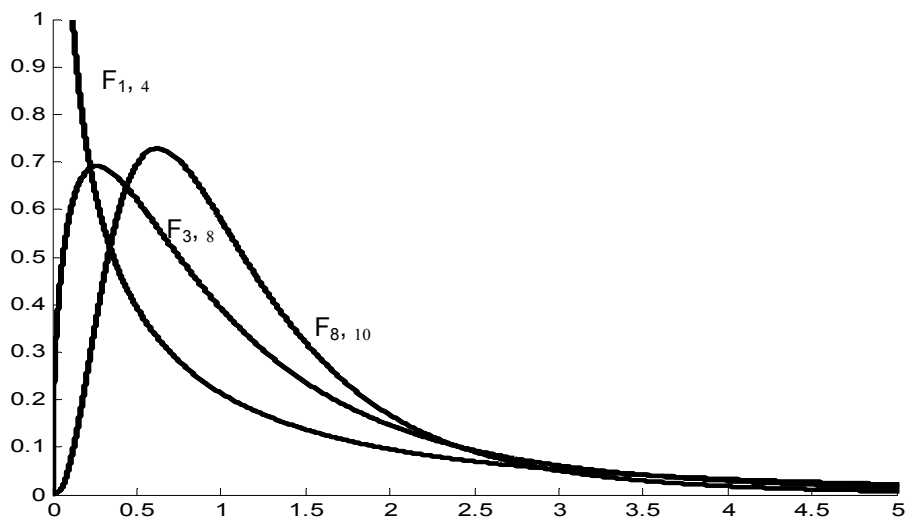


Figura. Gráfica de algunas de distribuciones Fisher, $F_{1,4}$, $F_{3,8}$ y $F_{8,10}$.

PRUEBA DE LA DIFERENCIA DE DOS VARIANZAS

Al igual que en las pruebas anteriores, la hipótesis nula H_0 se asocia con la igualdad entre los estadísticos de prueba poblacionales y la hipótesis alternativa H_1 solamente tiene dos posibles opciones, una prueba de cola derecha y una prueba de dos colas

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \begin{matrix} \sigma_1^2 > \sigma_2^2 \\ \sigma_1^2 \neq \sigma_2^2 \end{matrix}$$

El estadístico de prueba a utilizar es $F = \frac{s_1^2}{s_2^2}$, el cual cumple con la distribución Fisher. La hipótesis nula se rechazará si el valor de F es lo suficientemente grande para que sea mayor que el valor crítico F_{α, v_1, v_2} .

EJEMPLOS

47. Supóngase que se comparan las materias primas suministradas por dos proveedores. En apariencia los dos proveedores proporcionan materiales distribuidos normalmente con el mismo promedio, pero existe preocupación en cuanto a la variabilidad de los materiales. Una muestra de 16 lotes del Proveedor I proporciona una varianza de 150 ($s_1^2 = 150$), mientras que una muestra de 21 lotes provenientes del Proveedor II proporciona una varianza de 225 ($s_2^2 = 225$). Pruébese la hipótesis nula de que sus varianzas verdaderas son iguales contra la hipótesis alternativas de que son diferentes, con $\alpha = 0.05$.

SOLUCION

Los datos de cada uno de los proveedores se resumen a continuación (reacuérdesse que $s_1^2 > s_2^2$)

Proveedor I	Proveedor II
$s_1^2 = 150$	$s_2^2 = 225$
$n_1 = 16$	$n_2 = 21$

La hipótesis nula y alternativa de problema son respectivamente

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Utilizando el número de datos de cada muestra, $v_1 = 21 - 1 = 20$ y $v_2 = 16 - 1 = 15$.

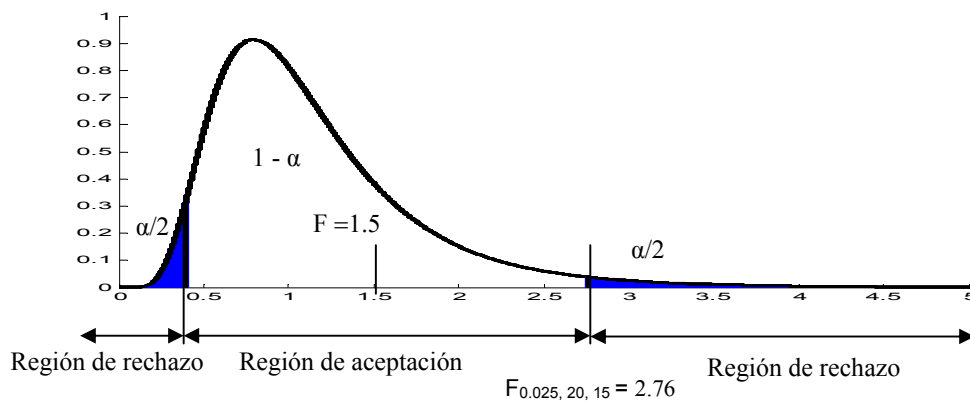
Por otra parte, puesto que la prueba es de dos colas y $\alpha = 0.05$, $\alpha/2 = 0.025$

El valor crítico para la prueba es $F_{0.025, 20, 15} = 2.76$.

EL estadístico de prueba es

$$F = \frac{s_1^2}{s_2^2} = \frac{225}{150} = 1.5$$

Como $1.5 < 2.76$, no se rechaza H_0 , las varianzas son estadísticamente iguales.



48. Se emplean dos métodos de enseñanza de la lectura a dos grupos seleccionados aleatoriamente de niños de nueve años. Se desea determinar si los resultados de los dos métodos, en términos de las puntuaciones obtenidas en una prueba estándar de lectura, tienen la misma variabilidad. Supóngase que se obtienen los siguientes datos de las dos poblaciones consideradas como normales:

	Método I	Método II
Tamaño de la muestra	$n_1 = 25$	$n_2 = 30$
Varianza muestral	$s_1^2 = 108$	$s_2^2 = 95$

Con un nivel de significación de 0.05, ¿debería llegarse a la conclusión de que las puntuaciones de prueba de los dos grupos tienen la misma varianza poblacional?

SOLUCION

En este caso la hipótesis nula y alternativa de problema son

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

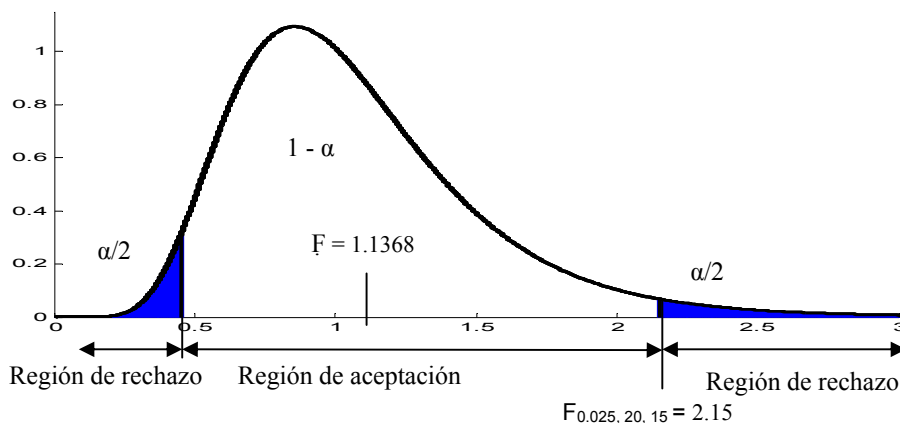
Utilizando el número de datos de cada muestra, $v_1 = 25 - 1 = 24$ y $v_2 = 30 - 1 = 29$.

La prueba es de dos colas, entonces como $\alpha = 0.05$, $\alpha/2 = 0.025$

El valor crítico para la prueba es $F_{0.025, 24, 29} = 2.15$, por otra parte estadístico de prueba es

$$F = \frac{s_1^2}{s_2^2} = \frac{108}{95} = 1.1368$$

Como $1.1368 < 2.15$, no se rechaza H_0 , las varianzas son estadísticamente iguales.



49. Un psicólogo desea determinar si la inteligencia de las niñas más variable que la de los niños. Se sabe que los C.I. tanto de niños como de niñas se distribuyen normalmente. Supóngase que una muestra aleatoria de los C.I. de 61 niñas proporciona una varianza de $s_1^2 = 240$, y una muestra aleatoria de los C.I. de 61 niños proporciona una varianza de $s_2^2 = 200$. Con $\alpha = 0.01$, pruébese la hipótesis nula de que la variabilidad de los C.I. de las niñas es igual que la de los niños, contra la hipótesis alternativa de que la primera es mayor que la segunda.

SOLUCION

Los datos para el grupo de niños y niñas se resumen a continuación

Niñas	Niños
$s_1^2=240$	$s_2^2=200$
$n_1=61$	$n_2=61$

La hipótesis nula y alternativa de problema son respectivamente

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

Los grados de libertad para cada muestra son respectivamente

$$v_1 = n_1 - 1 = 61 - 1 = 60 \text{ y } v_2 = n_2 - 1 = 61 - 1 = 60$$

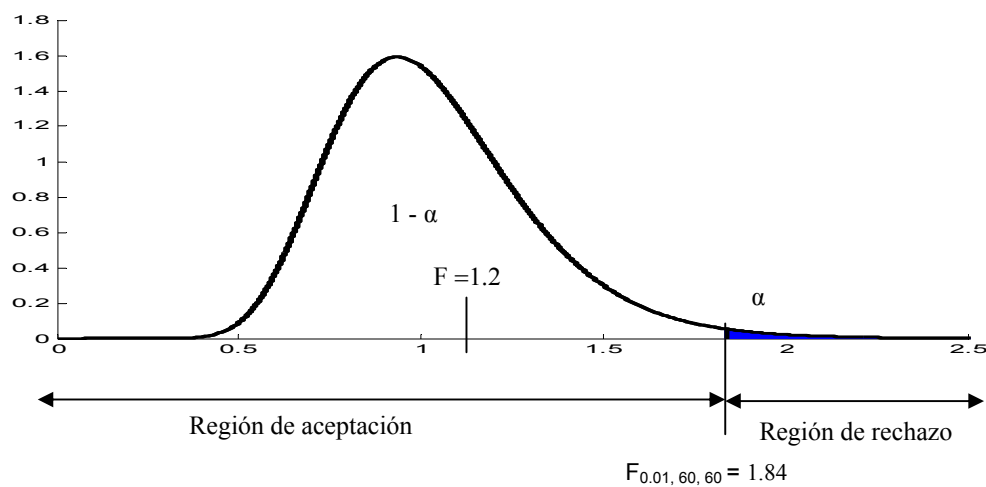
La prueba de hipótesis es de cola derecha con $\alpha=0.01$.

El valor crítico para la prueba es $F_{0.01, 60, 60} = 1.84$.

EL estadístico de prueba es

$$F = \frac{s_1^2}{s_2^2} = \frac{240}{200} = 1.2$$

Como $1.2 < 1.84$, no se rechaza H_0 , las varianzas son estadísticamente iguales.



50. Se emplean dos máquinas, I y II, para producir pernos idénticos cuyas longitudes se cree que se distribuyen normalmente. Una muestra aleatoria de 41 pernos producidos por la máquina I da una $s_1^2=0.5$, una muestra de 61 pernos producidos por la máquina II da una $s_2^2=0.3$. Pruebe la hipótesis nula de que pernos producidos por las dos máquinas tienen variabilidad idéntica, contra la hipótesis alternativa de que tiene varianza diferente, con $\alpha=0.10$.

SOLUCION

Las varianzas y número de datos se resumen a continuación para cada máquina

Maquina I	Maquina II
$s_1^2=0.5$	$s_2^2=0.3$
$n_1=41$	$n_2=61$

Para este problema la hipótesis nula y alternativa de problema son

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

A partir del número de datos de cada muestra se determina los grados de libertad

$$v_1=41 - 1 = 40 \text{ y } v_2=61 - 1 = 60.$$

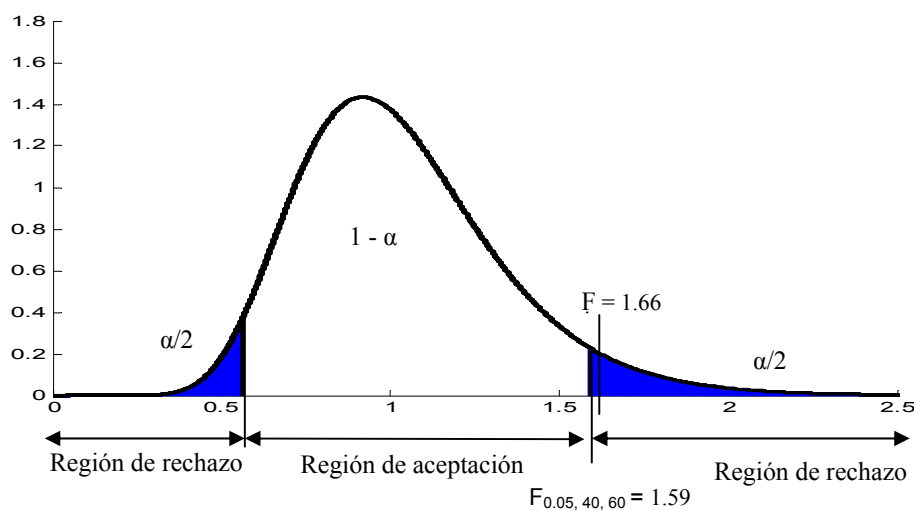
Como $\alpha=0.1$ y la prueba es de dos colas, se tiene que $\alpha/2=0.05$

El valor crítico para la prueba es $F_{0.05, 40, 60} = 1.59$,

El estadístico de prueba es

$$F = \frac{s_1^2}{s_2^2} = \frac{108}{95} = 1.66$$

Como $1.59 < 1.66$, se rechaza H_0 , las varianzas son estadísticamente diferentes.



ANALISIS DE VARIANZA (ANOVA)

El análisis realizado mediante la distribución t-student permite entre otras cosas realizar la comparación entre dos medias muestrales que provienen de poblaciones con distribución normal y tiene la misma varianza, pero si se desea generalizar el problema anterior, esto es, comparar entre tres o más medias muestrales provenientes de poblaciones con distribución normal y varianza idéntica, la distribución t-student no sería el método más adecuado para llevar a cabo tal comparación, ya que esta prueba solo se aplica a parejas de medias, afortunadamente se ha desarrollado un método conocido como **análisis de varianza** (ANOVA) que permite de una manera directa realizar la comparación, esta prueba utiliza a la distribución F o Fisher como base, ya que el estadístico de prueba se define como la razón de dos cantidades positivas que se relacionan con la varianza total de los datos y con la varianza de las medias respecto de la media total, más adelante se da una descripción del método utilizando un ejemplo numérico.

La prueba ANOVA tiene como hipótesis nula H_0 de que todas las medias $\mu_1, \mu_2, \mu_3, \dots, \mu_k$ son iguales y la hipótesis H_1 que alguna de ellas es diferente, lo anterior se indica a continuación

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \dots \neq \mu_k$$

La descripción del método se realizará mediante el siguiente ejemplo, en donde cada columna muestra las calificaciones obtenidas al aplicar un método de aprendizaje, hay tres métodos diferentes, por lo que la hipótesis nula es que los tres métodos producen resultados idénticos y la hipótesis alternativa es que producen resultados diferentes.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3$$

	METODO I	METODO II	METODO III
	74	84	83
	78	77	85
	73	79	86
	73	79	87
	72	81	89
Total	370	400	470

A partir de la suma total de cada método se determina las medias para cada uno de los métodos

utilizando la fórmula para el promedio $\bar{x} = \sum \frac{x_i}{n_i}$, donde n_i es el número de datos en cada método o clase.

$$\bar{x}_1 = 370/5 = 74$$

$$\bar{x}_2 = 400/5 = 80$$

$$\bar{x}_3 = 470/5 = 94$$

Las respectivas varianzas insesgadas de cada método se pueden calcular aplicando $s^2 = \sum \frac{(x_i - \bar{x})^2}{n_i - 1}$

$$s_1^2 = \frac{(74 - 74)^2 + (78 - 74)^2 + (73 - 74)^2 + (73 - 74)^2 + (72 - 74)^2}{5 - 1} = 5.5$$

$$s_2^2 = \frac{(84-80)^2 + (77-80)^2 + (79-80)^2 + (79-80)^2 + (81-80)^2}{5-1} = 7$$

$$s_3^2 = \frac{(83-86)^2 + (85-86)^2 + (86-86)^2 + (87-86)^2 + (89-86)^2}{5-1} = 5$$

La media de las medias o media total es

$$\bar{x} = \frac{370 + 400 + 470}{15} = 80$$

La varianza de las medias muestrales se puede calcular como

$$s_{\bar{x}}^2 = \frac{\sum (\bar{x}_i - \bar{x})^2}{k - 1} = \frac{(74-80)^2 + (80-80)^2 + (86-80)^2}{3-1} = 36$$

K C L N J E S

$s_{\bar{x}}^2$ (varianza de la media muestral) es un estimador de $\sigma_{\bar{x}}^2$ (varianza de la media poblacional), esto es $\sigma_{\bar{x}}^2 = s_{\bar{x}}^2 = 36$

Por otra parte recordando el teorema del límite central $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ y tomando como $n_i = 5$ ya que es el número de datos en cada muestra, se tiene que

$$\sigma^2 = n \sigma_{\bar{x}}^2 = 5(36) = 180$$

Lo anterior muestra como la varianza para las medias se transforma en un estimador de la varianza de una población.

Como σ^2 se obtiene a partir de las 3 medias que representan a cada uno de los métodos, por lo que sus grados de libertad son $v_1 = 3 - 1 = 2$.

Para un caso más general donde existan K clases se tendrá que los grados de libertad para σ^2 son general.

$$v_1 = K - 1$$

La estimación de σ^2 mejora si se utiliza toda la información disponible de las muestras, por lo que un mejor estimador sería el promedio de cada una de las varianzas individuales s_1^2 , s_2^2 y s_3^2 ,

$$s_{\bar{w}}^2 = \sum_{i=1}^K \frac{s_i^2}{K-1} = \frac{5.5 + 7 + 5}{3-1} = 5.83$$

Los grados de libertad de $s_{\bar{w}}^2$ para un caso general se puede obtener mediante

$$v_2 = n_1 + n_2 + n_k - K = N - k$$

Donde n_k , es el número de datos en la clase k y N es el número total de datos.

$$\text{Para el presente ejemplo } v_2 = 5 + 5 + 5 - 3 = 12$$

El estadístico de prueba se define como

$$F = \frac{s_{\bar{x}}^2}{s_{\bar{w}}^2}$$

por lo tanto, para el ejemplo

$$F = \frac{180}{5.83} = 30.9$$

Para aceptar o rechazar la hipótesis nula, se requiere de un valor crítico, por ejemplo si $\alpha=0.05$

$$F_{\alpha, v_1, v_2} = F_{0.05, 2, 12} = 3.89$$

Puesto que $3.89 < 30.9$ Se rechaza H_0 , lo que se traduce en que los métodos de aprendizaje son diferentes.

Método general

En general si se tiene una tabla con K muestras o clases y cada muestra tiene n_k datos como se muestra a continuación

	Muestra I	Muestra II	...	Muestra K
	X_{11}	X_{21}	...	X_{k1}
	X_{12}	X_{22}		X_{k2}
	.	.		.
	.	.		.
	X_{1n}	X_{2n}		X_{kn}
Tamaño de la muestra	n_1	n_2		n_k
Total de la muestra	T_1	T_2		T_k

Las siguientes definiciones permiten simplificar los resultados

Total de la muestra k
$$T_k = \sum_{i=1}^{n_k} X_{i,k}$$

Suma total de la muestras
$$T = \sum_{j=1}^K \sum_{i=1}^{n_k} X_{i,j}$$

Total de las observaciones
$$N = n_1 + n_2 + \dots + n_k = \sum_{i=1}^k n_i$$

Recordando que el estadístico de prueba se definió como la razón de la varianza entre las medias muestrales y la varianza dentro de cada una de las muestras.

La **suma externa de cuadrados** se define como

$$SSB = \sum_{k=1}^K \frac{T_k^2}{n_k} - \frac{T^2}{N} \quad (5.30)$$

La cual tiene $v_1 = K - 1$ grados de libertad.

La **suma interna de cuadrados** calcula la varianza dentro de cada una de las muestras.

$$SSW = \sum_{j=1}^K \sum_{i=1}^{n_k} x_{i,j}^2 - \sum_{k=1}^K \frac{T_k^2}{n_k} \quad (5.31)$$

La cual tiene $v_2 = N - K$ grados de libertad.

La **suma total de cuadrados** se define como la suma

$$SST = SSB + SSW \quad (5.32)$$

Utilizando las definiciones anteriores, la suma total de cuadrados es

$$SST = \sum_{j=1}^K \sum_{i=1}^{n_k} x_{i,j}^2 - \frac{T^2}{N} \quad (5.33)$$

La varianza entre las medias muestrales se determina como

$$S_B^2 = \frac{SSB}{K - 1} \quad (5.34)$$

La varianza dentro de cada una de las muestras es

$$S_W^2 = \frac{SSW}{N - K} \quad (5.35)$$

La razón o estadístico de prueba se define como

$$F = \frac{S_B^2}{S_W^2} \quad (5.36)$$

El procedimiento de análisis de varianza se resume en la siguiente tabla

Fuentes de variación	Suma de cuadrados	Grados de libertad	Varianza	Razón F
Entre grupos	$SSB = \sum_{k=1}^K \frac{T_k^2}{n_k} - \frac{T^2}{N}$	$v_1 = K - 1$	$S_B^2 = \frac{SSB}{K - 1}$	$F = \frac{S_B^2}{S_W^2}$
Dentro de los grupos	$SSW = \sum_{j=1}^K \sum_{i=1}^{n_k} x_{i,j}^2 - \sum_{k=1}^K \frac{T_k^2}{n_k}$	$v_2 = N - K$	$S_W^2 = \frac{SSW}{N - K}$	
Total	$SST = \sum_{j=1}^K \sum_{i=1}^{n_k} x_{i,j}^2 - \frac{T^2}{N}$	$N - 1$		

EJEMPLOS

51. Utilizando los datos del ejemplo anterior y las fórmulas (60) y (61) obtenga: S_B^2 y S_W^2 y F.

SOLUCION

	Método I	Método II	Método III	Método I	Método II	Método III
	X_1	X_2	X_3	X_1^2	X_2^2	X_3^2
	74	84	83	5476	7056	6889
	78	77	85	6084	5929	7225
	73	79	86	5329	6241	7396
	73	79	87	5329	6241	7589
	72	81	89	5184	6561	7921
Total	370	400	470	27402	32028	37000

Numero de clases $K = 3$.

Número total de datos $N = n_1 + n_2 + \dots + n_k = 5 + 5 + 5 = 15$

La suma de cada muestra es $T_1=370$ $T_2=400$ $T_3=430$

Total de las observaciones $T = 370 + 400 + 430 = 1200$

Suma externa de cuadrados

$$SSB = \sum_{k=1}^K \frac{T_k^2}{n_k} - \frac{T^2}{N} = \frac{370^2}{5} + \frac{400^2}{5} + \frac{430^2}{5} - \frac{1200^2}{15} = 360$$

Grados de libertad $v_1 = K - 1 = 3 - 1 = 2$

$$S_B^2 = \frac{SSB}{K - 1} = \frac{360}{3 - 1} = 180$$

Suma interna de cuadrados

$$SSW = \sum_{j=1}^K \sum_{i=1}^{n_k} x_{i,j}^2 - \sum_{k=1}^K \frac{T_k^2}{n_k} = 27402 + 32028 + 37000 - \left[\frac{370^2}{5} + \frac{400^2}{5} + \frac{430^2}{5} \right] = 70$$

Grados de libertad $v_2 = N - K = 15 - 3 = 12$

$$S_W^2 = \frac{SSW}{N - K} = \frac{70}{15 - 3} = 5.833$$

El estadístico de prueba es

$$F = \frac{S_B^2}{S_W^2} = \frac{180}{5.833} = 30.86$$

Obteniéndose los mismos resultados descritos en el ejemplo anterior.

52. A tres grupos de pollos seleccionados aleatoriamente se les alimenta con tres dietas diferentes. Cada grupo consta de cinco pollos. Sus aumentos de peso durante un periodo específico de tiempo son los siguientes:

Dieta I	Dieta II	Dieta III
4	3	6
4	4	7
7	5	7
7	6	7
8	7	8

Utilícese $\alpha = 0.05$ para probar la hipótesis nula de que las tres dietas tienen el mismo efecto en el aumento de peso de los pollos, contra la hipótesis alternativa de que tienen distintos efectos.

SOLUCION

Un resultado interesante es que la suma externa de cuadrados y la suma interna de cuadrados no se ven alteradas si a cada dato de la tabla se le suma o resta un número fijo.

Haciendo uso de la idea anterior conviene restarle a cada dato el número 7

	Dieta I	Dieta II	Dieta III	Dieta I	Dieta II	Dieta III
	X_1	X_2	X_3	X_1^2	X_2^2	X_3^2
	-3	-4	-1	9	16	1
	-3	-3	0	9	9	0
	0	-2	0	0	4	0
	0	-1	0	0	1	0
	1	0	1	1	0	1
Total	-5	-10	0	19	30	2

El número de clases es $K = 3$ y el número total de datos es $N = 15$

La hipótesis nula y alternativa del problema es

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3$$

Los respectivos grados de libertad son $v_1 = K - 1 = 3 - 1 = 2$ y $v_2 = N - K = 15 - 3 = 12$

Como $\alpha = 0.05$ el valor crítico para la prueba es $f_{\alpha, v_1, v_2} = f_{0.05, 2, 12} = 3.89$

La suma de cada muestra es $T_1 = -5$ $T_2 = -10$ $T_3 = 0$

Total de las observaciones $T = -5 - 10 + 0 = -15$

Calculando la suma externa de cuadrados

$$SSB = \sum_{k=1}^K \frac{T_k^2}{n_k} - \frac{T^2}{N} = \frac{(-5)^2}{5} + \frac{(-10)^2}{5} + \frac{0^2}{5} - \frac{(-15)^2}{15} = 10$$

por lo tanto

$$S_B^2 = \frac{SSB}{K-1} = \frac{10}{3-1} = 5$$

La suma interna de cuadrados es

$$SSW = \sum_{j=1}^K \sum_{i=1}^{n_k} x_{i,j}^2 - \sum_{k=1}^K \frac{T_k^2}{n_k} = 19 + 30 + 2 - \left[\frac{(-5)^2}{5} + \frac{(-10)^2}{5} + \frac{0^2}{5} \right] = 26$$

$$S_W^2 = \frac{SSW}{N-K} = \frac{26}{15-3} = 13/6 = 2.1667$$

El estadístico de prueba es

$$F = \frac{s_B^2}{s_W^2} = \frac{5}{2.1667} = 2.307$$

Como $2.307 < 3.89$, no se rechaza H_0 , las dietas son igualmente efectivas.

53. Una compañía manufacturera tiene cuatro máquinas idénticas en un proceso específico de producción. Cada máquina es operada por un trabajador distinto.

Se toma de cada máquina una muestra de los productos obtenidos durante un periodo de cinco horas y se obtiene el número de partes defectuosas producidas cada hora. Los resultados son los siguientes:

Máquina I	Máquina II	Máquina III	Máquina IV
10	7	2	3
9	7	3	3
9	8	3	6
9	8	3	6
8	5	4	7

Utilizando $\alpha = 0.01$, pruébese la hipótesis nula de que las máquinas producen el mismo promedio de partes defectuosas por hora, contra la hipótesis alternativa de que los cuatro promedios son diferentes.

SOLUCION

Restando el numero 6 a cada elemento de tabla

	M I	M II	M III	M IV	M I	M II	M III	M IV
	X ₁	X ₂	X ₃	X ₄	X ₁ ²	X ₂ ²	X ₃ ²	X ₄ ²
	4	1	-4	-3	16	1	16	9
	3	1	-3	-3	9	1	9	9
	3	2	-3	0	9	4	9	0
	3	2	-3	0	9	4	9	0
	2	-1	-2	1	4	1	4	1
Total	15	5	-15	-5	47	11	47	19

El

número de clases es $K = 4$ y el número total de datos es $N = 20$

La hipótesis nula y alternativa del problema es

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3$$

Los respectivos grados de libertad son $v_1 = K - 1 = 4 - 1 = 3$ y $v_2 = N - K = 20 - 4 = 16$

Como $\alpha = 0.01$ el valor crítico para la prueba es $f_{\alpha, v_1, v_2} = f_{0.01, 3, 16} = 5.29$

La suma de cada muestra es $T_1 = 15$ $T_2 = 7$ $T_3 = 15$ $T_4 = 47$

Total de las observaciones $T = 15 + 5 + 15 + 5 = 0$

Calculando la suma externa de cuadrados

$$SSB = \sum_{k=1}^K \frac{T_k^2}{n_k} - \frac{T^2}{N} = \frac{(15)^2}{5} + \frac{(5)^2}{5} + \frac{(-15)^2}{5} + \frac{(-5)^2}{5} - \frac{(0)^2}{20} = 100$$

por lo tanto

$$S_B^2 = \frac{SSB}{K - 1} = \frac{100}{4 - 1} = 33.3333$$

La suma interna de cuadrados es

$$SSW = \sum_{j=1}^K \sum_{i=1}^{n_k} x_{i,j}^2 - \sum_{k=1}^K \frac{T_k^2}{n_k} = 47 + 11 + 47 + 19 - \left[\frac{(15)^2}{5} + \frac{(5)^2}{5} + \frac{(-15)^2}{5} + \frac{(-5)^2}{5} \right] = 24$$

$$S_W^2 = \frac{SSW}{N - K} = \frac{24}{20 - 4} = 1.5$$

El estadístico de prueba es

$$F = \frac{S_B^2}{S_W^2} = \frac{33.3333}{1.5} = 22.222$$

Como $5.29 < 22.222$, se rechaza H_0 , los promedios de producción son diferentes.

UNIDAD VI Regresión y correlación

REGRESIÓN

Existen problemas experimentales en los cuales se trata de establecer si existe una relación entre dos conjuntos de datos X y Y , por ejemplo se desea establecer la cantidad de lluvia (X) se relaciona con la producción de trigo (Y), o si la experiencia en años (X) se relaciona con las ventas obtenidas (Y), etc.

Si la relación existe entonces se puede estimar que tan fuerte es esta relación o dependencia, además es posible determinar el valor posible de una variable a partir del valor de la otra.

Dependiendo del problema es posible determinar la relación entre las variables X y Y , mediante la **técnica de regresión**. La fuerza de la relación entre las variables X y Y se determina mediante el **coeficiente de correlación**.

Si en un problema se tienen solamente dos variables, se dice que la técnica es una regresión o correlación **simple**. Cuando existen más variables involucradas se dice que el problema es de regresión o correlación **múltiple**.

En caso de regresión simple la variable que es utilizada para estimar a la otra se llama **variable independiente** y se denota por X , mientras que la otra es conocida como variable dependiente y se denota por la letra Y . La regresión múltiple involucra dos o más variables independientes y una variable dependiente.

REGRESION LINEAL

La regresión lineal se refiere a determinar la “mejor ecuación lineal” de la forma: $y = mx + b$ que es posible establecer entre las variables X y Y . En muchas ocasiones la relación entre las variables es no lineal lo cual complica el problema, pero en muchos casos es posible determinar una relación entre las variables de la forma: $y = f(x)$, donde $f(x)$ puede ser una relación polinomial, potencial, exponencial. etc.

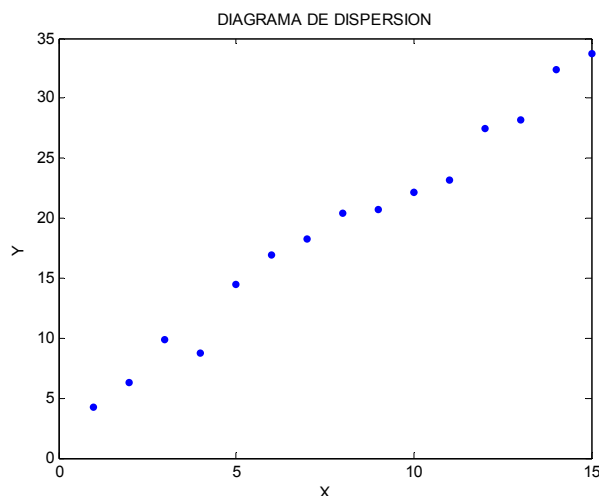
El trabajo de aplicar la regresión lineal a un problema consiste en determinar los valores ó parámetros a y b de la recta $y = mx + b$ a partir del conjunto de datos X y Y

DIAGRAMA DE DISPERSIÓN

Como primer paso para la obtención de una regresión primero se grafican los datos, lo cual es conocido como diagrama de dispersión. En la figura A siguiente se muestran una tabla de datos y su respectivo diagrama de dispersión.

TABLA DE DATOS

X	Y
x_1	y_1
x_2	y_2
.	.
.	.
.	.
x_n	y_n

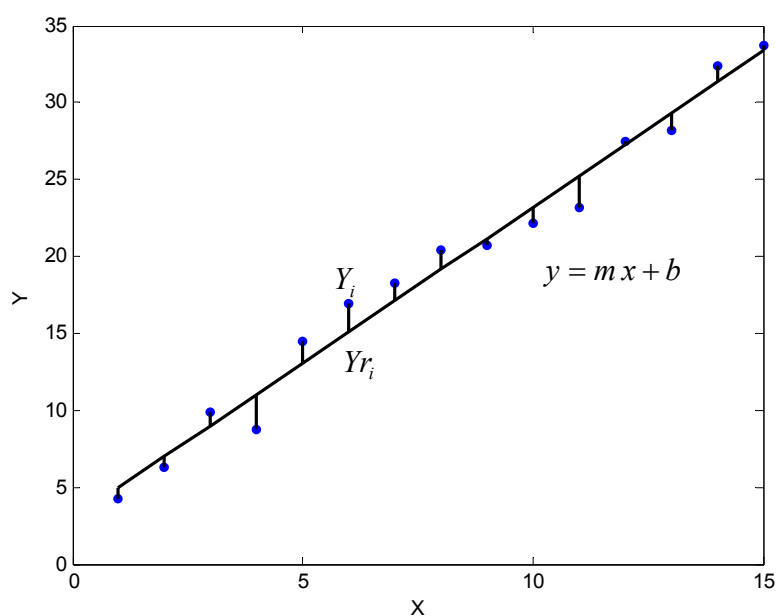


MÉTODO DE MÍNIMOS CUADRADOS

Como se puede observar del diagrama de dispersión anterior los datos no se encuentran exactamente en una línea recta.

El criterio que más se utiliza para determinar la mejor recta de ajuste se conoce como recta **método de mínimos cuadrados**, consiste en buscar los parámetros a y b de la recta $y = mx + b$ de tal manera que la suma de los cuadrados de las distancias verticales entre los puntos de la recta y del diagrama de dispersión sea lo más pequeña posible.

La figura siguiente muestra la idea general del método de mínimos cuadrados, cada uno de los 15 puntos graficados muestra representa a cada uno de los pares ordenados (X_i, Y_i) donde $i = 1, 2, 3, \dots, n$. Al sustituir el valor de la abscisa X_i de cada uno de los puntos en la ecuación de la recta $y = mx + b$ se obtienen un conjunto de valores $Yr_i = mX_i + b$, donde $i = 1, 2, 3, \dots, n$, los cuales se encuentran sobre la recta.



La diferencia $D_i = Y_i - Yr_i$ se denomina desviación, por lo que la idea básica del método de mínimos cuadrados se puede expresar matemáticamente como:

$$S(m, b) = \sum_{i=1}^n D_i^2 = \sum_{i=1}^n (Y_i - Yr_i)^2 \quad (6.1)$$

Para el caso de la línea recta la ecuación anterior toma la forma siguiente

$$S(m, b) = \sum_{i=1}^n D_i^2 = \sum_{i=1}^n (Y_i - mX_i - b)^2 \quad (6.2)$$

La función debe $S(m, b)$ se considera como una función de dos variables m y b para la cual debe de existir al menos un par de valores (m, b) tales que sean un mínimo de la función.

La condición que debe de cumplir la función $S(m, b)$ para tener un mínimo (o máximo) es que sus derivadas parciales con respecto a los parámetros m y b sean cero, esto es:

$$\frac{\partial S}{\partial m} = 0 \quad (6.3)$$

$$\frac{\partial S}{\partial b} = 0 \quad (6.4)$$

Aplicando la condición dada por la ecuación (6.3)

$$\frac{\partial S}{\partial m} = \frac{\partial}{\partial m} \left[\sum_{i=1}^n (Y_i - mX_i - b)^2 \right] = \sum_{i=1}^n 2(Y_i - mX_i - b)(-X_i)$$

Utilizando las propiedades de la sumatoria se tiene que

$$S(m, b) = 2 \sum_{i=1}^n (-Y_i X_i + mX_i^2 + bX_i) = 2 \left[- \sum_{i=1}^n Y_i X_i + m \sum_{i=1}^n X_i^2 + b \sum_{i=1}^n X_i \right]$$

Posteriormente igualando a cero

$$2 \left[- \sum_{i=1}^n Y_i X_i + m \sum_{i=1}^n X_i^2 + b \sum_{i=1}^n X_i \right] = 0$$

Despejando se obtiene la ecuación

$$m \sum_{i=1}^n X_i^2 + b \sum_{i=1}^n X_i = + \sum_{i=1}^n Y_i X_i \quad (6.5)$$

Ahora si se aplica la condición dada por la ecuación (6.4)

$$\frac{\partial S}{\partial b} = \frac{\partial}{\partial b} \left[\sum_{i=1}^n (Y_i - mX_i - b)^2 \right] = \sum_{i=1}^n 2(Y_i - mX_i - b)(-1)$$

Aplicando nuevamente las propiedades de la sumatoria

$$S(m, b) = 2 \sum_{i=1}^n (-Y_i + mX_i + b) = 2 \left[- \sum_{i=1}^n Y_i + m \sum_{i=1}^n X_i + b n \right]$$

Igualando a cero

$$2 \left[- \sum_{i=1}^n Y_i + m \sum_{i=1}^n X_i + b n \right] = 0$$

Reacomodando términos se obtiene la ecuación

$$m \sum_{i=1}^n X_i^2 + b \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i X_i \quad (6.6)$$

Las ecuaciones 5.41 y 5.42 forman un sistema de de ecuaciones donde m y b son las incógnitas,

$$m \sum_{i=1}^n X_i^2 + b \sum_{i=1}^n X_i = \sum_{i=1}^n X_i Y_i$$

$$m \sum_{i=1}^n X_i + b n = \sum_{i=1}^n Y_i$$

La solución del sistema de ecuaciones anterior se puede resolver mediante determinantes, a continuación se evalúan los determinantes requeridos para el cálculo

$$\Delta = \begin{vmatrix} \sum X_i^2 & \sum X_i \\ \sum X_i & n \end{vmatrix} = n \sum X_i^2 - (\sum X_i)^2$$

$$\Delta_1 = \begin{vmatrix} \sum X_i Y_i & \sum X_i \\ \sum Y_i & n \end{vmatrix} = n \sum X_i Y_i - \sum X_i \sum Y_i$$

$$\Delta_2 = \begin{vmatrix} \sum X_i^2 & \sum X_i Y_i \\ \sum X_i & \sum Y_i \end{vmatrix} = \sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i$$

De donde se obtiene las ecuaciones que permiten obtener los parámetros para la mejor recta de mínimos cuadrados.

$$m = \frac{\Delta_1}{\Delta} = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} \quad (6.7)$$

$$b = \frac{\Delta_2}{\Delta} = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2} \quad (6.8)$$

Como se puede observar de las ecuaciones anteriores, para obtener los parámetros m y b es necesario realizar las sumatorias indicadas a partir de los datos (X_i, Y_i) donde $i = 1, 2, 3, \dots, n$.

EJEMPLOS

1. En una compañía de seguros se desea determinar la relación entre la experiencia en ventas y el volumen de las mismas. Se selecciona una muestra aleatoria de nueve vendedores. Se encuentra que sus años de experiencia (X) y ventas anuales normales (Y) son los siguientes:

X 1 2 3 4 5 6 7 8 9
Y : 2 1 3 3 4 5 6 5 7 (en \$100 000)

- Constrúyase un diagrama de dispersión y trácese la recta de regresión de Y sobre X en el diagrama.
- Estímese el volumen de ventas anuales para un vendedor que tiene una experiencia en ventas de diez años.

SOLUCION

a) Es conveniente primero el fin de determinar las cálculo de m y b

	X	Y	X ²	XY
	1	2	1	2
	2	1	4	2
	3	3	9	9
	4	3	16	12
	5	4	25	20
	6	5	36	30
	7	6	49	42
	8	5	64	40
	9	7	81	63
Σ	45	36	285	220

construir la tabla siguiente, con sumatorias necesarias para el

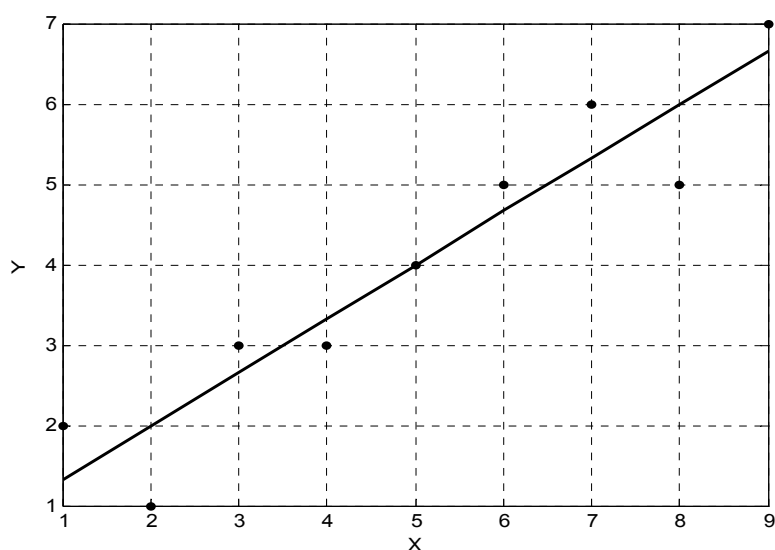
Evaluando en las expresiones

$$m = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{(9)(220) - (45)(36)}{(9)(285) - (45)^2} = \frac{2}{3} = 0.6667$$

$$b = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{(285)(36) - (45)(220)}{(9)(285) - (45)^2} = \frac{2}{3} = 0.6667$$

Entonces, la recta de regresión tiene la ecuación $y = \frac{2}{3}x + \frac{2}{3}$

El diagrama de dispersión y la recta de regresión se muestran en la siguiente gráfica



b) El volumen de ventas anuales para un vendedor que tiene una experiencia en venta de 10 años se obtiene al evaluar la recta de regresión obtenida para $x = 10$.

$$y = \frac{2}{3}x + \frac{2}{3} = y = \frac{2}{3}(10) + \frac{2}{3} = 7.33$$

el resultado anterior se multiplica por 10 000 para obtener el total de ventas.

Ventas = $7.33(100000) = \$ 733\,000$.

2. Se tiene un registro de los costos de mantenimiento para seis máquinas idénticas de distintas edades. Por parte de la gerencia se desea determinar si existe una relación funcional entre la edad de la máquina (X) y el costo de mantenimiento (Y) Se obtienen los siguientes datos.

Máquina	.X	Y
1	2	\$ 70
2	1	40
3	3	100
4	2	80
5	1	30
6	3	100

Obtégase la ecuación de regresión con X como variable independiente y Y como variable dependiente. ¿Cuál sería el costo de mantenimiento para una máquina de cuatro años?

SOLUCION

La tabla siguiente resume los cálculos necesarios para las sumatorias

X	Y	X Y	X ²
2	70	140	4
1	40	40	1
3	100	300	9
2	80	160	4
1	30	30	1
3	100	300	9
Σ	12	420	28

Evaluando en las expresiones para calcular m y b

$$m = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{(6)(970) - (12)(420)}{(6)(28) - (12)^2} = 32.5$$

$$b = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{(28)(420) - (12)(970)}{(6)(28) - (12)^2} = 5$$

Así se tiene la recta de regresión $y = 32.5x + 5$, evaluado para $x = 4$

$$y = (32.5)(4) + 5 = 135$$

Por lo que el costo de reparación de la maquina de 4 años es \$135.

CORRELACIÓN

Como ya se ha señalado anteriormente, la **correlación** es la fuerza de la relación entre las variables X y Y , y se determina mediante el **coeficiente de correlación**.

COEFICIENTE DE CORRELACIÓN

A partir de la ecuación de mínimos cuadrados se puede realizar una predicción de el valor de Y sustituyendo el valor respectivo X , pero el grado de exactitud de la predicción depende de el grado de correlación entre las variables X y Y . Cuando la correlación es pequeña se tiene poca precisión en la determinación del valor Y , pero cuando la correlación es grande se tiene una gran exactitud en la determinación del valor Y .

La medida del grado de correlación utilizando los n pares de datos (X_i, Y_i) es llamado **coeficiente de correlación**, normalmente denotado por r . Para determinar a r se considera primero que Y es una variable aleatoria cuya desviación respecto de la recta de mínimos cuadrados es la menor posible, esto quiere decir que la variabilidad se divide en dos partes, la primera es la eliminada por la recta de mínimos cuadrados y la cantidad que permanece a pesar de de la recta de regresión. Si $Y_r = mX + b$

(valor calculado a partir de la recta de regresión) y $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$, entonces la variación total se puede separar de la forma.

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_r - \bar{Y})^2 + \sum (Y_i - Y_r)^2 \quad (6.9)$$

Variación total

Variación eliminado
por regresión

Variación restante

Mientras más variación se elimine mediante la recta de regresión más cercana será la relación entre X y Y y se volverá más precisa la estimación del valor Y .

Dividiendo ambos lados de la ecuación 68 entre $\sum (Y - \bar{Y})^2$ se obtiene

$$\frac{\sum (Y_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\sum (Y_r - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} + \frac{\sum (Y_i - Y_r)^2}{\sum (Y_i - \bar{Y})^2}$$

Entonces, la expresión anterior se puede escribir como.

$$1 = r^2 + \frac{\sum (Y_i - Y_r)^2}{\sum (Y_i - \bar{Y})^2}$$

Donde r es el coeficiente de correlación, así se tiene que

$$r = \sqrt{1 - \frac{\sum (Y_i - Y_r)^2}{\sum (Y_i - \bar{Y})^2}} \quad (6.10)$$

En lugar de usar la ecuación anterior para determinar el coeficiente de correlación se utiliza para el caso de la línea recta la fórmula siguiente

$$r = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n \sum X_i^2 - (\sum X_i)^2} \sqrt{n \sum Y_i^2 - (\sum Y_i)^2}} \quad (6.11)$$

Si la correlación entre las variables X y Y es fuerte, la mayor parte de la variabilidad de Y puede atribuirse a la relación con X y r será cercana a 1 o -1, en particular se $r = 1$ o -1 se dirá que hay un ajuste perfecto a la recta. En general el valor de r varía de -1 a 1, y cuando la correlación es débil su valor es cercano a 0. Si $r = 0$, se dice que no existe correlación entre X y Y .

Cuando r se encuentra entre 0 y 1 existe correlación positiva y cuando está entre -1 y 0 hay correlación negativa.

PRUEBA DE HIPÓTESIS PARA EL COEFICIENTE DE CORRELACIÓN

Existe una prueba de hipótesis para determinar si el un coeficiente de correlación (r) es lo suficientemente grande para afirmar que hay correlación entre los pares de valores X y Y . o si el valor r corresponde al azar. Dicho de otra manera, se desea probar la hipótesis de que el coeficiente de correlación poblacional ρ es igual a cero contra la hipótesis alternativa de que no lo es. Si la distribución de las dos variables involucradas es normal entonces, el estadístico de prueba T empleado se define como

$$T = r \sqrt{\frac{n-2}{1-r^2}} \quad (6.12)$$

El cual se distribuye de acuerdo a una distribución T-Student con $\nu = n - 2$ grados de libertad. Si no es clara la idea de que las variables se distribuyan normalmente se pueden aplicar métodos no paramétricos a la prueba de hipótesis como la prueba de correlación de rangos.

EJEMPLOS

4. Por parte de una compañía de seguros se desea determinar la relación entre los años de experiencia en ventas de sus vendedores y su volumen de ventas. Se selecciona una muestra aleatoria de nueve vendedores y se encuentra que sus años de experiencia (X) y ventas anuales actuales (Y) son los siguientes:

X	1	2	3	4	5	6	7	8	9
Y	2	1	3	4	3	5	6	7	5 (en \$100 000)

- Obtégase el coeficiente de correlación r .
- Pruébese la hipótesis de que el coeficiente de correlación de la población es cero con $\alpha = 0.05$.

SOLUCION

a) La siguiente tabla muestra los cálculos requeridos para determinar las sumatorias que permiten determinar el coeficiente de correlación

X	Y	X ²	Y ²	XY	
1	2	1	4	2	
2	1	4	1	2	
3	3	9	9	9	
4	4	16	16	16	
5	3	25	9	15	
6	5	36	25	30	
7	6	49	36	42	
8	7	64	49	56	
9	5	81	25	45	
Σ	45	36	285	174	217

$$r = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n \sum X_i^2 - (\sum X_i)^2} \sqrt{n \sum Y_i^2 - (\sum Y_i)^2}} = \frac{9(217) - (45)(36)}{\sqrt{9(285) - (45)^2} \sqrt{9(174) - (36)^2}} = 0.8721$$

- b) La prueba de hipótesis del problema se plantea como $H_0: \rho = 0$ $H_1: \rho \neq 0$

El estadístico de prueba es

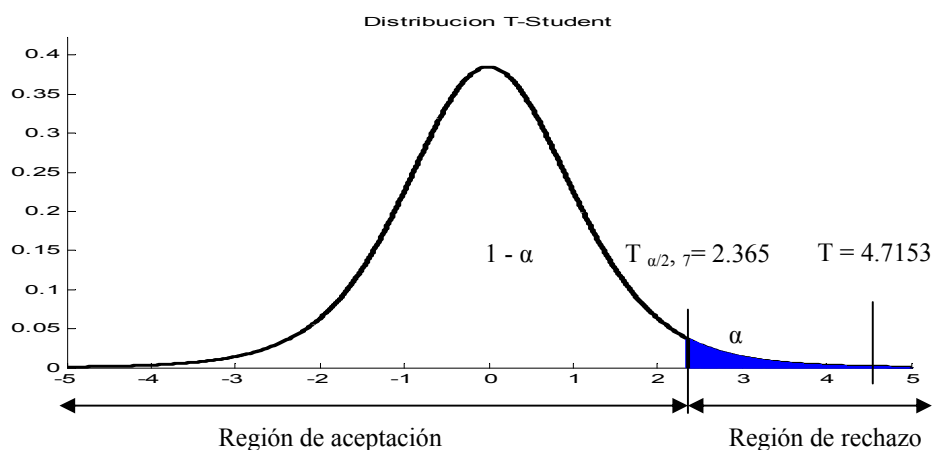
$$T = r \sqrt{\frac{n-2}{1-r^2}} = 0.8721 \sqrt{\frac{9-2}{1-(0.8721)^2}} = 4.7153$$

El cual tiene una distribución como T-student con $v = n-2 = 9 - 2 = 7$ grados de libertad.

El planteamiento de la Hipótesis conduce a una prueba de dos colas, como $\alpha = 0.05$ entonces $T_{\alpha/2,7} = 2.365$

Comparando el valor crítico con el estadístico de prueba se tiene que $T > T_{\alpha/2,7}$ ($4.7153 > 2.365$).

Se rechaza H_0 , sí hay correlación



5. Se realiza un experimento para determinar la relación entre la precipitación pluvial y el rendimiento del trigo. Supóngase que se obtienen los siguientes datos.

Precipitación pluvial en pulgadas: X 1 2 3 4 5 5 6 7 8 9
 Rendimiento de trigo en bushel: Y 1 3 2 5 5 4 7 6 9 8

- Ajústese una recta de mínimos cuadrados a los datos con X como variable independiente y grafíquese después la recta sobre un diagrama de dispersión.
- Estímese el rendimiento de trigo si la precipitación pluvial es de 10 pulg.
- Obténgase el coeficiente de correlación r.
- Pruébese la hipótesis nula de que no existe relación entre la precipitación pluvial y el rendimiento del trigo, con $\alpha = 0.05$.

SOLUCION

a) La siguiente tabla muestra los cálculos requeridos para determinar las sumatorias

X	Y	XY	X ²	Y ²
1	1	1	1	1
2	3	6	4	9
3	2	6	9	4
4	5	20	16	25
5	5	25	25	25
5	4	20	25	16
6	7	42	36	49
7	6	42	49	36
8	9	72	64	81
9	8	72	81	64
Σ	50	50	306	310

Evaluando en las expresiones para calcular m y b

$$m = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{(10)(306) - (50)(50)}{(10)(310) - (50)^2} = 0.9333$$

$$b = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{(310)(50) - (50)(306)}{(6)(28) - (12)^2} = 0.3333$$

Así se tiene la recta de regresión $y = 0.9333x + 0.3333$, la gráfica siguiente muestra el diagrama de dispersión y la recta de regresión.

b) Evaluado en la ecuación de regresión el valor de $x = 10$ pulg se obtiene

$$y = (0.9333)(10) + 0.3333 = 9.6667 \text{ bushel:}$$

c) El coeficiente de correlación es

$$r = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n \sum X_i^2 - (\sum X_i)^2} \sqrt{n \sum Y_i^2 - (\sum Y_i)^2}} = \frac{(10)(306) - (50)(50)}{\sqrt{(10)(310) - (50)^2} \sqrt{(10)(310) - (50)^2}} = 0.9333$$

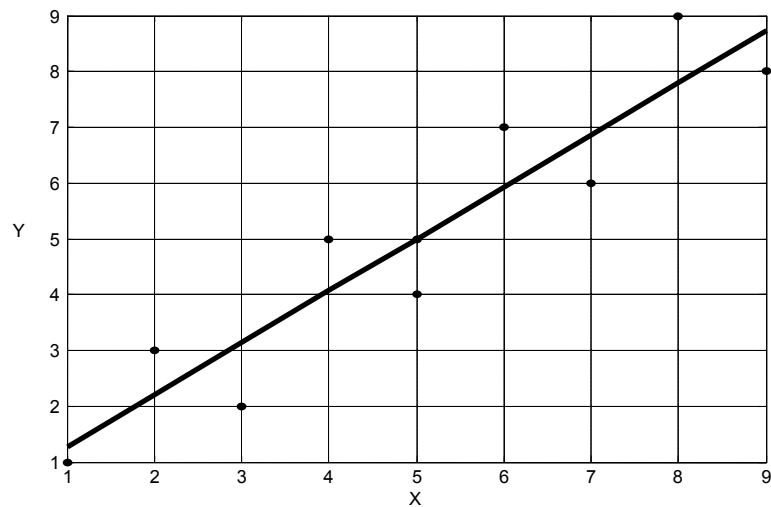


Diagrama de dispersión y recta de mínimos cuadrados del problema

d) La prueba de hipótesis del problema se plantea como $H_0: \rho = 0$ $H_1: \rho \neq 0$

El estadístico de prueba es

$$T = r \sqrt{\frac{n-2}{1-r^2}} = 0.9333 \sqrt{\frac{10-2}{1-(0.9333)^2}} = 7.3532$$

El cual tiene una distribución como T-student con $v = n-2 = 10 - 2 = 8$ grados de libertad.

El planteamiento de la Hipótesis conduce a una prueba de dos colas, como $\alpha = 0.05$ entonces $T_{\alpha/2, 8} = 2.306$

Comparando el valor crítico con el estadístico de prueba se tiene que $T > T_{\alpha/2, 8}$ ($7.3532 > 2.306$).

Se rechaza H_0 , sí hay correlación

