

SERIE UNIVERSITARIA
PATRIA
Teoría, ejemplos
y problemas



PROBABILIDAD y ESTADÍSTICA

Victor M. Alvarado Verdin



PROBABILIDAD Y ESTADÍSTICA

Víctor Manuel Alvarado Verdín

PRIMERA EDICIÓN EBOOK
MÉXICO, 2014

GRUPO EDITORIAL PATRIA

Para establecer comunicación
con nosotros puede hacerlo por:



correo:
Renacimiento 180, Col. San Juan
Tlihuaca, Azcapotzalco,
02400, México, D.F.



fax pedidos:
(01 55) 5354 9109 • 5354 9102



e-mail:
info@editorialpatria.com.mx



home page:
www.editorialpatria.com.mx

Dirección editorial: Javier Enrique Callejas
Coordinadora editorial: Estela Delfín Ramírez
Supervisor de preensa: Gerardo Briones González
Diseño de portada: Juan Bernardo Rosado Solís
Fotografías: © Thinkstockphoto

Revisión técnica:
Carlos Gerardo Velázquez Rodríguez
UVM-Coacalco/ESIQIE-Instituto Politécnico Nacional
Ana Elizabeth García Hernández
Instituto Politécnico Nacional

Probabilidad y estadística

Derechos reservados:
© 2014, Víctor Manuel Alvarado Verdín
© 2014, Grupo Editorial Patria, S.A. de C.V.
Renacimiento 180, Colonia San Juan Tlihuaca,
Delegación Azcapotzalco, Código Postal 02400, México, D.F.

Miembro de la Cámara Nacional de la Industria Editorial Mexicana
Registro núm. 43

ISBN ebook: 978-607-438-930-2

Queda prohibida la reproducción o transmisión total o parcial del contenido
de la presente obra en cualesquiera formas, sean electrónicas o mecánicas,
sin el consentimiento previo y por escrito del editor.

Impreso en México
Printed in Mexico

Primera edición ebook: 2014

Agradecimientos

A familiares y amigos por el apoyo prestado durante el desarrollo de este contenido; espero que disculpen los momentos de abstracción.

A Tere, Geo, Osiris y Doris por su apoyo al recopilar y ordenar las notas y apuntes de quien suscribe.

A mi editora Estela Delfín R., por su orientación y paciencia durante el desarrollo del texto.

A la Dra(c). Sánchez Arriola por su paciencia y apoyo durante el desarrollo del libro.

Prefacio

A últimas fechas, tanto la Probabilidad como la Estadística se han erigido como las materias o asignaturas académicas que ofrecen las herramientas de orden objetivo, a fin de generar procesos de pensamiento lógico-rationales basados en el análisis de la información cuantitativa relativa a todo tipo de actividades y fenómenos desarrollados por personas y organizaciones.

El panorama propuesto exige a estudiantes y académicos mantenerse familiarizados con los principios, métodos y técnicas matemáticos que les faciliten organizar, sintetizar, computar y analizar información cuantitativa, con el objetivo de poder tomar las decisiones más adecuadas, así como para la optimización del desarrollo de aplicaciones con propósitos específicos.

En atención a lo expuesto antes, la estructura de los contenidos de este libro ha sido diseñada con el propósito de:

- Exponer de manera resumida los principales temas relativos a la Probabilidad y a la Estadística, en términos y palabras sencillos y simples, evitando un lenguaje matemático complejo.
- Destacar notas al margen donde se detallan aquellos conceptos que se consideran trascendentes, a fin de que el estudiante se familiarice con estos, relacionándolos con los desarrollos temáticos.
- Desarrollar los procesos de solución de problemas basados en aplicaciones de orden práctico, los cuales permitan al estudiante repasar los métodos de cálculo, con el objetivo de que este refuerce su entendimiento temático.
- Proponer un amplio catálogo de problemas resueltos y para resolver; así, al final de cada unidad se proporcionan ejercicios y problemas cuyos planteamientos exponen retos teórico-conceptuales para quienes intentan resolverlos.

En términos generales, este libro se considera un texto útil para el complemento de las exposiciones teóricas dictadas en el aula, quedando en manos de los estudiantes y los académicos la disertación temática y la orientación práctica del presente contenido.

El Autor

Contenido



Unidad 1 La estadística y la estadística descriptiva 1

1.1 Introducción	2
1.2 Definición de estadística	2
1.3 Clasificación de la estadística	2
1.4 La estadística descriptiva	2
1.5 Análisis de datos por estadística descriptiva para datos no agrupados	3
1.6 Análisis de estadística descriptiva para datos agrupados	8
1.7 Gráficas descriptivas: histograma, polígono de frecuencias y ojiva	13
1.8 Medidas de tendencia central para datos agrupados	20
1.9 Determinación de la ubicación y el valor de la moda de manera gráfica mediante el histograma y el polígono de frecuencias	22
1.10 Determinación de la ubicación y el valor de la mediana de manera gráfica mediante la ojiva	24
1.11 Cálculo de cuantiles para datos agrupados	25
1.12 Momentos estadísticos	26
1.13 Determinación del número de intervalos de clase	29
1.14 La media geométrica	31
Problemas para resolver	33
Problemas reto	35
Referencias	37
Direcciones electrónicas	37



Unidad 2 Teoría de la probabilidad y distribuciones de probabilidad 39

2.1 Introducción	40
2.2 Concepto de probabilidad	40
2.3 Los espacios muestrales y la teoría de conjuntos	40
2.4 Probabilidad clásica	42
2.5 Probabilidad de eventos mutuamente exclusivos o excluyentes	43
2.6 Probabilidad de eventos comunes	44
2.7 Probabilidad de eventos simultáneos o sucesivos	45
2.8 Los diagramas de árbol	46
2.9 Eventos consecutivos con y sin reemplazo	47
2.10 Probabilidad condicional	48
2.11 Teorema de Bayes	48
2.12 El principio de multiplicación	50
2.13 Permutaciones	51
2.14 Combinaciones	52
2.15 Variables aleatorias	53
2.16 Distribuciones de probabilidad	53
2.17 Curva de distribución de probabilidad discreta	58
2.18 Valor esperado de una variable aleatoria discreta o esperanza matemática	59
2.19 Distribuciones de probabilidad continuas	60
2.20 Teorema del límite central	63
Problemas para resolver	64
Problemas reto	68
Referencias	69
Direcciones electrónicas	69



Unidad 3 Estadística inferencial 71

3.1 Introducción	72
3.2 Concepto y propósito de la estadística inferencial	72

3.3 Estimadores	73
3.4 Distribución de las medias muestrales	75
3.5 Distribución muestral de las proporciones	80
3.6 Intervalos de confianza	83
3.7 Determinación del tamaño muestral para estimar una media poblacional	86
3.8 Determinación del tamaño muestral para estimar la proporción poblacional	88
3.9 Grados de libertad	89
3.10 Intervalos de confianza para muestras pequeñas	
Problemas para resolver	92
Problemas reto	95
Referencias	95
Direcciones electrónicas	96



Unidad 4 Análisis estadístico de experimentos 97

4.1 Introducción	98
4.2 Concepto de experimento	98
4.3 Hipótesis estadísticas	98
4.4 Aplicación de la distribución t de Student en el análisis de experimentos	99
4.5 Estudios comparativos simples	99
4.6 Estudios comparativos basados en pruebas de hipótesis sustentados en el análisis de la varianza (ANOVA) de un factor con una muestra por grupo	102
4.7 Estudios comparativos basados en ANOVA de dos factores	105
4.8 Estudios comparativos basados en ANOVA de dos factores de varias muestras por grupo	109
4.9 Discriminantes para pruebas de hipótesis basadas en ANOVA	114
4.10 Método de discriminación de la R de Duncan	120
Problemas para resolver	123
Problema reto	127
Referencias	128
Direcciones electrónicas	128



La estadística y la estadística descriptiva

OBJETIVOS

- Entender la importancia de la estadística como la herramienta que facilita el análisis y la interpretación de la información cuantitativa relacionada con los fenómenos de interés que acontecen en la realidad.
- Aplicar el proceso de análisis de la estadística descriptiva.
- Entender la diferencia entre datos agrupados y no agrupados.
- Entender el concepto de medidas de tendencia central.
- Entender el concepto de medidas de dispersión.
- Entender el concepto de medidas de posición.
- Desarrollar los elementos gráficos básicos de la estadística descriptiva.
- Desarrollar el proceso de análisis por momentos estadísticos.

¿QUÉ SABES?

- ¿Qué es la estadística y para qué sirve?
- ¿Cuál es la diferencia entre la estadística descriptiva y la inferencial?
- ¿Cuál es la diferencia entre los datos agrupados y los no agrupados?
- ¿Cuáles son las principales medidas de tendencia central y qué significan?
- ¿Qué son los intervalos de clase?
- ¿Cuántos tipos de frecuencia existen?
- ¿Cuáles son las principales gráficas del análisis de la estadística descriptiva?
- ¿Qué es una curva de distribución de frecuencias?
- ¿Cuáles son los principales cuantiles?
- ¿Para qué sirven los momentos estadísticos?

1.1 Introducción

La estadística proporciona la metodología para el tratamiento de la información cuantitativa relacionada con hechos y sucesos de la realidad, fundamentada en la organización y cómputo de la misma, con el propósito de generar conocimiento sobre los mismos.

1.2 Definición de estadística

Por estadística debe entenderse el proceso sistemático aplicado al análisis y la interpretación de numéricos con la intención de comprender los hechos de la realidad pudiendo apoyar la toma de decisiones racionales.

La definición anterior expone que la estadística es un conjunto de actividades interrelacionadas, las cuales se desarrollan bajo principios bien definidos y en estricto orden, a efecto de coleccionar, organizar, computar, presentar e interpretar los conjuntos de cifras numéricas (datos) que guardan relaciones significativas con un fenómeno en particular.

Asimismo, a los conjuntos de datos se les denomina colección de datos, mismos que se pueden clasificar en:

- **No agrupados**, cuando los datos que las conforman a lo más guardan un orden secuencial de acuerdo con su valor.
- **Agrupados**, cuando los datos que las conforman han sido catalogados dentro de un grupo de rangos denominados intervalos de clase en atención a que representan al grupo de rangos en los que se puede subdividir la colección, permitiendo clasificar los datos de acuerdo con su valor dentro de los mismos.

1.3 Clasificación de la estadística

La estadística, en atención a su aplicación y propósitos, se clasifica en:

Descriptiva, la cual tiene por objeto organizar y presentar conjuntos de datos numéricos con la intención de facilitar el análisis y la caracterización de un fenómeno.

Inferencial, la cual tiene como finalidad la validación de los parámetros de una población mediante los estadísticos de una o varias muestras.

1.4 La estadística descriptiva

El análisis por estadística descriptiva se fundamenta en el cálculo de las llamadas medidas descriptivas, las cuales recapitulan la información de una colección de datos permitiendo describir el comportamiento de un fenómeno.

Las medidas descriptivas de uso común se clasifican de la siguiente forma:

- Medidas de centralización.** Son el grupo de valores más representativos de una colección ordenada de datos, los cuales tienden a ubicarse al centro de la misma. Entre las medidas más representativas se pueden citar:
 - La media.
 - La mediana.
 - La moda.
 - El promedio ponderado.
 - El promedio geométrico.
- Medidas de dispersión.** Señalan qué tan alejados están los valores de una colección de datos con respecto a un valor de centralización, que por lo general es la media. Entre las medidas más comunes se encuentran:

- El rango.
- La varianza.
- La desviación estándar.
- El coeficiente de variación.

c) **Medidas de posición o cuantiles.** Son los valores que permiten dividir la colección ordenada de datos en partes iguales con el mismo número de datos en cada segmento. Los cuantiles más comunes son:

- Los percentiles, los cuales dividen la colección en 100 partes iguales, considerando que existen 99 percentiles ($P_1, P_2, P_3 \dots P_{99}$).
- Los deciles, los cuales dividen la colección en 10 partes iguales, considerando que existen 9 deciles ($D_1, D_2, D_3 \dots D_9$).
- Los cuartiles, los cuales dividen la colección en 4 partes iguales, considerando que existen 3 cuartiles (Q_1, Q_2 y Q_3).

Es de notar que el P_{50} , el D_5 y el Q_2 representan el valor de la mediana.

d) **Medidas de forma.** Son los valores que permiten establecer cómo están distribuidos los valores de una colección de datos. Las medidas principales son:

- Sesgo o asimetría.
- Apuntamiento.



Alerta

Las medidas descriptivas son de centralización y de forma.

1.5 Análisis de datos por estadística descriptiva para datos no agrupados

Este tipo de análisis se recomienda cuando el número de datos que estructuran una colección de datos permite su manejo y cómputo de manera ágil.

Para que las cifras ofrezcan un significado es conveniente ordenarlas, sugiriendo en este caso de menor a mayor, de acuerdo con sus valores.



Alerta

Los datos no agrupados no se encuentran clasificados por categorías o intervalos.

Problema resuelto

Procede a ordenar de menor a mayor la siguiente colección de datos.

18 5 11 52 35 52 72

Solución

Ordenando de menor a mayor:

5 11 18 35 52 52 72

Ordenar los datos permite contar con una mejor perspectiva de los mismos, pudiendo establecer las diferencias entre los diferentes valores.

■ Medidas de tendencia central para datos no agrupados

Debe recordarse que de manera básica son tres las medidas de tendencia central: media, mediana y moda.

a) Determinación del valor de la media

De manera concreta, la media o promedio representa el valor más representativo de una colección de datos tendiendo a ubicarse al centro de la misma, cuyo valor permite establecer un equilibrio

La estadística y la estadística descriptiva

en cuanto a las diferencias existentes con el resto de los valores.

Matemáticamente, la media queda definida como

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

donde

\bar{x}_i = Representa el valor i ($i = 1, 2, 3, \dots, n$).

n = Número total de datos de la colección.

Problema resuelto

Considerando la colección

5 11 18 35 52 52 72

Determina el valor de la media.

Solución

El valor de la media del ejemplo en cuestión es:

$$\bar{X} = \frac{5+11+18+35+52+52+72}{7} = 35$$

Nótese que la media dentro de la colección ordenada se ubica exactamente en el centro:

5 11 18 35 52 52 72
 ↑
 \bar{X}



Alerta

La media es el valor más representativo de una colección de datos.

b) Determinación del valor de la mediana

La siguiente medida de tendencia central es la mediana (M_d), la cual representa el valor central de la colección.

La determinación de la mediana debe cumplir con las siguientes reglas:

- I. Si el número de datos de la colección es impar, el valor de la mediana es el valor central de la misma dividiendo en dos segmentos iguales a la colección.
- II. Si el número de datos de la colección es par, el valor de la mediana es el promedio aritmético de los valores centrales.



Alerta

La mediana es el valor central de una colección de datos.

Problema resuelto

Considerando la colección

5 11 18 35 52 52 72

Determina el valor de la mediana.

Solución

En el caso de la colección en análisis, esta cuenta con un número impar de datos, por lo que el valor de la mediana es el que cumple con la regla I, antes mencionada, coincidiendo con la posición y el valor de la media.

Solución (continuación)

5	11	18	35	52	52	72
			↑			
			\bar{X}			
			M_d			

c) Determinación del valor de la moda

En el caso de la moda (M_o), es el valor o valores que tienen la mayor frecuencia, o sea, son los que más se repiten.

Con referencia a lo anterior, debe considerarse que en una colección de datos puede haber más de una moda, por lo que una colección puede ser:

- **Modal**, cuando cuenta con un valor con mayor frecuencia.
- **Bimodal**, cuando cuenta con dos valores con la misma frecuencia.
- **Multimodal**, cuando cuenta con más de dos valores con la misma frecuencia.

Problema resuelto

Considerando la siguiente colección

5 11 18 35 52 52 72

Determina el valor de la moda.

Solución

Puede observarse que la colección es modal ya que el valor que más se repite es el 52, considerando que cuenta con una frecuencia con valor 2.

5	11	18	35	52	52	72
				↓		
				M_o		

Alerta

La moda es el valor o valores que más se repiten en una colección de datos; una colección puede tener más de una moda.

■ Medidas de dispersión para datos no agrupados

La determinación de los valores de las medidas de dispersión es relativamente simple ya que el valor de referencia es la media de la colección, por lo que se deben establecer las diferencias o desviaciones existentes entre los distintos valores de la colección con respecto a ella.

Si esto se realiza, se encuentra que el valor promedio es cero, ya que la media equilibra las desviaciones tanto por arriba como por debajo de la misma, tal como se muestra en el siguiente problema resuelto.

Problema resuelto

Considerando la siguiente colección

5 11 18 35 52 52 72

Comprueba que la suma de las diferencias con respecto a la media es cero.

**Alerta**

Las medidas de dispersión permiten establecer el grado de alejamiento entre un valor específico en relación con el resto de los demás datos.

Solución**Tabla 1.1**

x	$x - \bar{x}$
5	$5 - 35 = -30$
11	$11 - 35 = -24$
18	$18 - 35 = -17$
35	$35 - 35 = 0$
52	$52 - 35 = 17$
52	$52 - 35 = 17$
72	$72 - 35 = 37$
$\Sigma =$	0

Por tanto, no se puede establecer la diferencia o desviación promedio de los valores de la colección con respecto a la media, pero para evitar esto se procede a elevar las diferencias al cuadrado a efecto de evitar los números negativos y obtener un promedio de las desviaciones; a este parámetro se le denomina varianza (S^2), pero ha de observarse que la varianza expone el promedio de las desviaciones al cuadrado por lo que un valor más significativo lo propone la desviación estándar, la cual es la raíz cuadrada del valor de la varianza. Para ejemplificar lo anterior obsérvese el siguiente problema resuelto.

Problema resuelto

Considerando la colección

5 11 18 35 52 52 72

Determina el valor de la varianza y de la desviación estándar.

Solución**Tabla 1.2**

x	$(x - \bar{x})^2$
5	$(5 - 35)^2 = 900$
11	$(11 - 35)^2 = 576$
18	$(18 - 35)^2 = 289$
35	$(35 - 35)^2 = 0$
52	$(52 - 35)^2 = 289$
52	$(52 - 35)^2 = 289$
72	$(72 - 35)^2 = 1369$
$\Sigma =$	3712

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = 530.29$$

Sin embargo, la varianza expone un promedio de cuadrados, por lo que se procede a calcular la raíz cuadrada para obtener un valor más significativo, al cual se le denomina desviación estándar.

$$S = \sqrt{S^2} = 23.03$$

■ Medidas de posición para datos no agrupados

En el caso de las colecciones de datos no agrupados, por lo regular se procede a determinar los valores de las principales medidas de posición como lo son los cuartiles. Para determinar estos valores se utiliza la siguiente fórmula:

$$PS_p = (n + 1) \left(\frac{P}{100} \right)$$

Donde

PS_p = Posición del percentil P .

n = Número de elementos de la colección.

La fórmula anterior arroja la posición del percentil de interés, por lo que se deberá determinar el valor del mismo mediante diferencias y proporciones, como se explicará más adelante. Para ejemplificar lo anterior se formula el siguiente problema.

Alerta

Las medidas de posición establecen los valores que permiten dividir la colección en segmentos iguales.

Problema resuelto

Considerando la colección en análisis, determina los valores del primer y tercer cuartiles, así como del 80avo percentil.

5	11	18	35	52	52	72
---	----	----	----	----	----	----

Solución

5	11	18	35	52	52	72
1	2	3	4	5	6	7

En la primera línea de la serie anterior se muestra de manera ordenada la colección y en la segunda línea la posición de cada valor dentro del arreglo ordenado.

Para determinar el primer cuartil se procede como sigue:

$$P = 25$$

$$n = 7$$

por tanto,

$$PS_{25} = (7 + 1) \left(\frac{25}{100} \right) = 2$$

Donde se interpreta que el valor del primer cuartil corresponde al del dato ubicado en la posición 2 de la colección ordenada, que en este caso es $Q_1 = PS_{25} = 11$.

Para determinar el tercer cuartil se procede de igual forma si $P = 75$,

$$PS_{75} = (7 + 1) \left(\frac{75}{100} \right) = 6$$

Donde se interpreta que el valor del tercer cuartil corresponde al del dato ubicado en la posición 6 de la colección ordenada, que en este caso es $Q_3 = PS_{75} = 52$.

En el caso del 80avo percentil $P = 80$:

$$PS_{80} = (7 + 1) \left(\frac{80}{100} \right) = 6.4$$

El valor de $P = 80$ se encuentra en la posición 6.4, la cual no se encuentra de manera explícita, por lo que se procede a determinar la distancia entre los valores de las posiciones 6 y 7,

$$\text{Dif} = 72 - 52 = 20$$

Solución (continuación)

Entonces se puede señalar que el valor en cuestión se encuentra a 40% de la distancia a partir del valor de la posición 6:

$$X = \text{Dif} \times 40\% = 20 \times 0.4 = 8$$

Por tanto,

$$P_{80} = 52 + X = 60$$

Para completar los referentes sobre las colecciones no agrupadas considérese el caso de una colección de datos con número de elementos par tal como se expone en el siguiente problema resuelto.

Problema resuelto

Considerando la siguiente colección de datos, procede a determinar los valores y la ubicación de las medidas de tendencia central, así como de las medidas de dispersión.

23 40 54 69 69 69

Solución

Se procede a calcular las medidas de tendencia central, comenzando por la media (\bar{X}):

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{23 + 40 + 54 + 69 + 69 + 69}{6} = 54$$

En este caso en particular el valor de la mediana (M_d) se determina a través del promedio de los datos centrales, mismos que permiten dividir la colección en dos segmentos iguales.

$$M_d = \frac{54 + 69}{2} = 61.5$$

Se observa que el valor de la moda (M_o) es 69 debido a que el valor de su frecuencia es de 3.

$$M_o = 69$$

Adicionalmente, se calcula el valor de las medidas de dispersión, comenzando con el rango.

$$R = \text{Valor máximo} - \text{Valor mínimo} = 69 - 23 = 46$$

Donde el valor de la varianza (S^2) es:

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n} = 305.33$$

En consecuencia el valor de la desviación estándar es:

$$S = \sqrt{S^2} = 17.47$$

1.6 Análisis de estadística descriptiva para datos agrupados

Cuando el número de datos de una colección es grande puede generar dificultades para su manejo y cómputo, por lo que se aconseja organizarlos a través de intervalos de clase. Los intervalos de clase, como se mencionó antes, son un conjunto de clases en las que se puede dividir el rango de una colección de datos.

■ La tabla de frecuencias

Tabla 1.3 Tabla de frecuencias

Intervalos nominales			f
LI		LS	
5	-	9	3
10	-	14	7
15	-	19	14
20	-	24	4
25	-	29	2
Σf	=	n	= 30



Alerta

Los datos agrupados se distinguen por estar catalogados en categorías o intervalos.

La tabla anterior se denomina tabla de frecuencias y en ella se observa que la primera columna es la de los intervalos nominales, los cuales son los rangos que permiten clasificar los datos de la colección de manera *a priori*, donde cada intervalo de clase cuenta con un límite inferior (LI) y un límite superior (LS); la segunda columna es la de las frecuencias, cuyos valores representan el número de datos de la colección que se ubican dentro de los límites de cada intervalo; si se realiza la suma de las frecuencias en los intervalos se obtiene el número total de datos de la colección.

■ Cálculo de las frecuencias acumulada y relativa

No cabe duda que el análisis descriptivo de datos agrupados es interesante, ya que ofrece la oportunidad de detallar cada uno de los parámetros estadísticos. Para iniciar se pueden calcular algunas variaciones sobre la frecuencia, como son:

- **Frecuencia acumulada (f_a)**, la cual consiste en acumular la frecuencia de los intervalos, de manera que se aprecie la cantidad de datos que se van acumulando conforme se recorren los intervalos de la colección.
- **Frecuencia relativa (f_r)**, la cual indica la proporción porcentual sobre el total del número de datos con los que cuenta cada intervalo.
- **Frecuencia acumulada relativa (f_{ar})**, que representa la proporción porcentual de datos que se va acumulando conforme se recorren los intervalos.

Para ejemplificar lo anterior revisa con detalle el siguiente problema resuelto.

Problema resuelto

Teniendo en cuenta los datos que se muestran en la siguiente tabla de frecuencias, procede a calcular la frecuencia acumulada (f_a), la frecuencia relativa (f_r) y la frecuencia acumulada relativa (f_{ar}) de cada intervalo.

Tabla 1.4

Intervalos nominales			f
LI		LS	
5	-	9	3
10	-	14	7
15	-	19	14
20	-	24	4
25	-	29	2
Σf	=	n	= 30

Solución

A partir de la columna de frecuencias se determinan los valores de las frecuencias solicitadas en cada intervalo.

Tabla 1.5

Intervalos nominales			f	f_u	fr	f_{ur}
LI		LS				
5	-	9	3	3	$\frac{3}{30} = 10.00\%$	10.00%
10	-	14	7	10	$\frac{7}{30} = 23.33\%$	33.33%
15	-	19	14	24	$\frac{14}{30} = 46.67\%$	80.00%
20	-	24	4	28	$\frac{4}{30} = 13.33\%$	93.33%
25	-	29	2	30	$\frac{2}{30} = 6.67\%$	100.00%
$\Sigma f = n =$			30		100.00%	

La amplitud de clase

Por otro lado, existen algunos parámetros importantes por determinar dentro de las tablas de frecuencias, como lo es la **amplitud de clase** (c), la cual indica qué tan ancho o amplio es un intervalo, o sea, cuántos valores componen el intervalo, donde la amplitud de clase se determina por la diferencia entre el límite inferior de un intervalo "A" con respecto al límite inferior de un intervalo "B" o en su caso entre los límites superiores.

$$c = LI_B - LI_A \quad \text{o} \quad c = LS_B - LS_A$$

**Alerta**

La amplitud de clase indica el número de elementos entre los límites de los intervalos.

Problema resuelto

Teniendo en cuenta los datos que se muestran en la siguiente tabla de frecuencias, procede a calcular la amplitud de clase.

Tabla 1.6

Intervalos nominales			f	f_u	fr	f_{ur}
LI		LS				
5	-	9	3	3	$\frac{3}{30} = 10.00\%$	10.00%
10	-	14	7	10	$\frac{7}{30} = 23.33\%$	33.33%
15	-	19	14	24	$\frac{14}{30} = 46.67\%$	80.00%
20	-	24	4	28	$\frac{4}{30} = 13.33\%$	93.33%

Problema resuelto (continuación)

Tabla 1.6						
Intervalos nominales			f	fu	fr	fur
LI		LS				
25	-	29	2	30	$\frac{2}{30} = 6.67\%$	100.00%
Σf	=	n	=	30		100.00%

Solución

En el caso particular del ejemplo la amplitud de clase es 5, para comprobarlo considérense los valores de los límites inferiores del segundo y tercer intervalos, por lo que,

$$c = 15 - 10 = 5$$

Para verificar tómense en cuenta los valores de los límites superiores del cuarto y quinto intervalos:

$$c = 29 - 24 = 5$$

■ Las marcas de clase

Otro parámetro primordial dentro de las tablas de frecuencias son las denominadas marcas de clase (MC), las cuales son los valores que representan los puntos medios de cada intervalo, donde su valor se determina calculando el promedio de los límites superior e inferior de cada intervalo:

$$MC = \frac{LI + LS}{2}$$

Para ejemplificar lo anterior se muestra el siguiente problema.

Problema resuelto

Teniendo en cuenta los datos que se muestran en la siguiente tabla de frecuencias, procede a calcular las marcas de clase de cada intervalo.

Tabla 1.7						
Intervalos nominales			f	fu	fr	fur
LI		LS				
5	-	9	3	3	$\frac{3}{30} = 10.00\%$	10.00%
10	-	14	7	10	$\frac{7}{30} = 23.33\%$	33.33%
15	-	19	14	24	$\frac{14}{30} = 46.67\%$	80.00%
20	-	24	4	28	$\frac{4}{30} = 13.33\%$	93.33%
25	-	29	2	30	$\frac{2}{30} = 6.67\%$	100.00%
Σf	=	n	=	30		100.00%

Alerta

Las marcas de clase son el punto medio de cada intervalo.

Solución

En el caso particular del ejemplo, los valores de las marcas de clase quedan expuestos en la columna identificada como MC.

Tabla 1.8

Intervalos nominales			<i>f</i>	<i>MC</i>	<i>fr</i>	<i>fur</i>
<i>LI</i>		<i>LS</i>				
5	-	9	3	7	10.00%	10.00%
10	-	14	7	12	23.33%	33.33%
15	-	19	14	17	46.67%	80.00%
20	-	24	4	22	13.33%	93.33%
25	-	29	2	27	6.67%	100.00%
$\Sigma f = n =$			30		100.00%	

■ Los intervalos de clase reales

Debe notarse que entre los límites superiores e inferiores de los intervalos nominales no se muestra una continuidad, o sea que "existen espacios entre ellos", lo que indica que los datos con valores que estén dentro de estos segmentos no estarían catalogados; para evitar lo anterior se requiere calcular los llamados límites reales, los cuales permiten tener una continuidad en el recorrido de los intervalos, evitando así la indefinición.

Los límites reales se calculan promediando el límite nominal superior de un intervalo "A" con el límite nominal inferior de un intervalo "B", donde los valores del límite inferior real del primer intervalo de clase, así como el límite superior del último intervalo de clase, se pueden determinar restando y sumando la amplitud de clase a los límites reales de los intervalos de clase inmediatos posterior y anterior a los mismos.

Para aclarar lo anterior se expone el siguiente problema resuelto.

Problema resuelto

Teniendo en cuenta los datos que se muestran en la siguiente tabla de frecuencias, procede a calcular los valores de los límites reales de cada intervalo.

Tabla 1.9

Intervalos nominales			<i>f</i>	<i>MC</i>	<i>fr</i>	<i>fur</i>
<i>LI</i>		<i>LS</i>				
5	-	9	3	7	10.00%	10.00%
10	-	14	7	12	23.33%	33.33%
15	-	19	14	17	46.67%	80.00%
20	-	24	4	22	13.33%	93.33%
25	-	29	2	27	6.67%	100.00%
$\Sigma f = n =$			30		100.00%	

Solución

En el caso particular del ejemplo, los valores de los límites de clase reales se muestran en la columna correspondiente.

Tabla 1.10

Intervalos nominales			Intervalos reales			F	fa	fr	far
LI		LS	LIR		LSR				
5	—	9	4.5	—	9.5	3	3	$\frac{3}{30} = 10.00\%$	10.00%
10	—	14	9.5	—	14.5	7	10	$\frac{7}{30} = 23.33\%$	33.33%
15	—	19	14.5	—	19.5	14	24	$\frac{14}{30} = 46.67\%$	80.00%
20	—	24	19.5	—	24.5	4	28	$\frac{4}{30} = 13.33\%$	93.33%
25	—	29	24.5	—	29.5	2	30	$\frac{2}{30} = 6.67\%$	100.00%
						$\Sigma f = n = 30$		100.00%	

1.7 Gráficas descriptivas: histograma, polígono de frecuencias y ojiva

La estadística descriptiva se distingue por sus elementos gráficos, siendo los principales:

- Histograma o diagrama de barras**, el cual permite representar la magnitud de la frecuencia de cada uno de los intervalos mediante rectángulos, barras o prismas cuyo largo es el valor de la frecuencia, mientras el ancho es la amplitud de clase delimitado por los límites reales de cada intervalo.
- Polígono de frecuencias**, es la línea discontinua que une a las marcas de clase cada intervalo y que permite visualizar la forma en que se han distribuido los datos en correspondencia al recorrido de los intervalos de clase; el polígono de frecuencias da origen a la curva de distribución de frecuencias.
- Ojiva**, es una línea discontinua que representa los valores de la frecuencia acumulada relativa (*far*) en razón del recorrido de los intervalos reales de clase de 0% a 100%, expone cómo se acumulan de manera proporcional los datos de una colección, donde los cambios de pendiente representan los cambios en las proporciones acumuladas.

■ Histograma

Para construir un histograma se utiliza un plano X-Y, en cuyas abscisas se representan los límites reales de clase mientras que en las ordenadas se representan las frecuencias. Para desarrollarlo basta con marcar los límites reales de cada intervalo levantando sobre los mismos un rectángulo con una altura que represente el valor de la frecuencia en dicho intervalo.

Para ejemplificar lo anterior considérese el siguiente problema resuelto.



Alerta

Las gráficas descriptivas permiten concentrar información para interpretarla con más facilidad.

Problema resuelto

Teniendo en cuenta los datos de la siguiente tabla de frecuencias, procede a construir el histograma correspondiente.

Tabla 1.11

Intervalos nominales			Intervalos reales			f	f_a	fr	far
LI	LS		LIR	LSR					
5	—	9	4.5	—	9.5	3	3	$\frac{3}{30} = 10.00\%$	10.00%
10	—	14	9.5	—	14.5	7	10	$\frac{7}{30} = 23.33\%$	33.33%
15	—	19	14.5	—	19.5	14	24	$\frac{14}{30} = 46.67\%$	80.00%
20	—	24	19.5	—	24.5	4	28	$\frac{4}{30} = 13.33\%$	93.33%
25	—	29	24.5	—	29.5	2	30	$\frac{2}{30} = 6.67\%$	100.00%
						$\Sigma f = n = 30$		100.00%	



Alerta

La construcción del histograma se fundamenta en los límites reales de clase y las frecuencias.

Solución

En el caso particular del ejemplo se ubican los valores de los límites reales en el eje de las x , procediendo a levantar las barras de acuerdo con su correspondiente frecuencia.

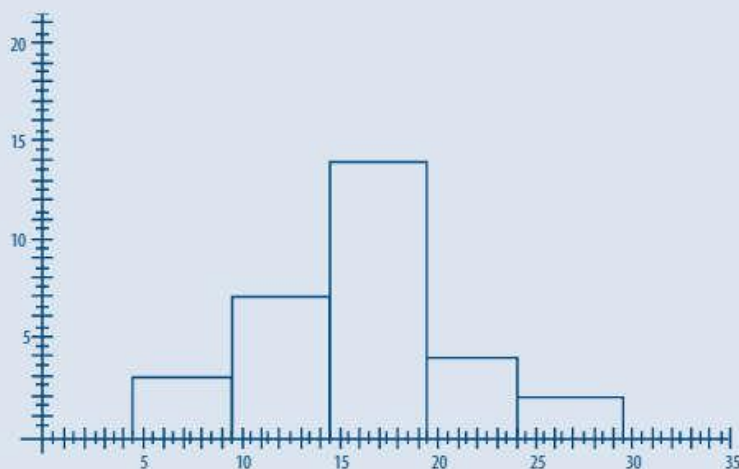


Figura 1.1

■ Polígono de frecuencias

El trazo del polígono de frecuencias se fundamenta en el histograma, ya que basta con ubicar las marcas de clase de cada intervalo, representado por el centro de cada barra, ubicando este en la parte superior de cada una, para que posteriormente se unan con una línea discontinua.

A continuación se expone el siguiente problema resuelto para mostrar lo anterior.

Problema resuelto

Considerando el polígono de frecuencias que se muestra, procede a ubicar las marcas de clase dentro de cada barra de manera que permita el trazo del polígono de frecuencias.

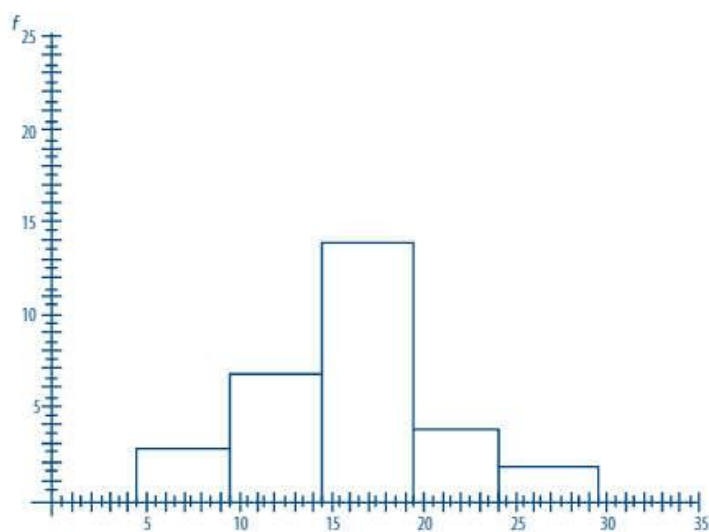


Figura 1.2

Solución

Para ubicar geoméricamente las marcas de clase en cada barra se trazan las diagonales en cada rectángulo de manera que se ubiquen los centros de cada una.

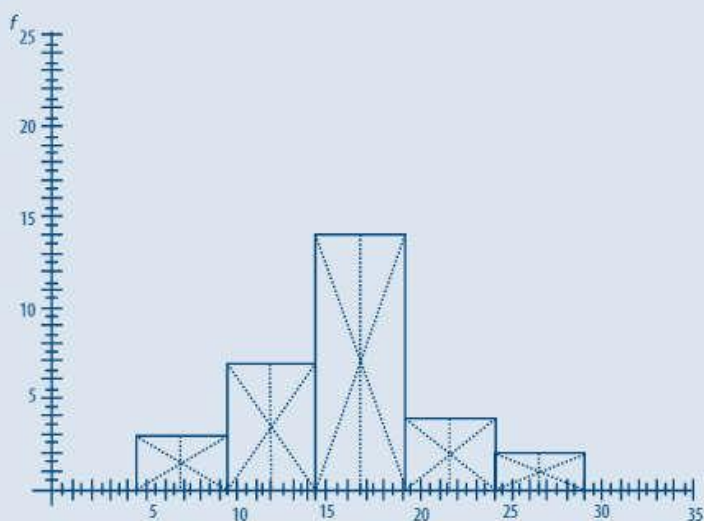


Figura 1.3

Solución (continuación)

Uniendo las marcas de clase:

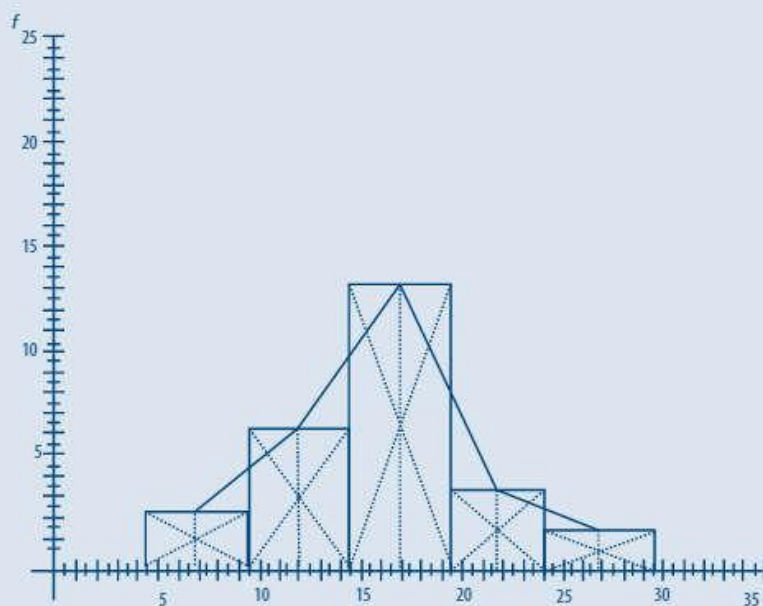


Figura 1.4

Alerta

El polígono de frecuencias permite establecer el perfil de la curva de distribución de frecuencias.

Sin embargo, nótese que parece que la línea se encuentra suspendida sobre el histograma, además de que la mitad del primer y último intervalos no se encuentran cubiertas por el polígono, por lo que se deben ubicar unas marcas de clase ficticias en los extremos a efecto de "aterrizar" al polígono procediendo a restar y sumar la amplitud de clase a la primera y última marcas de clase correspondientemente.

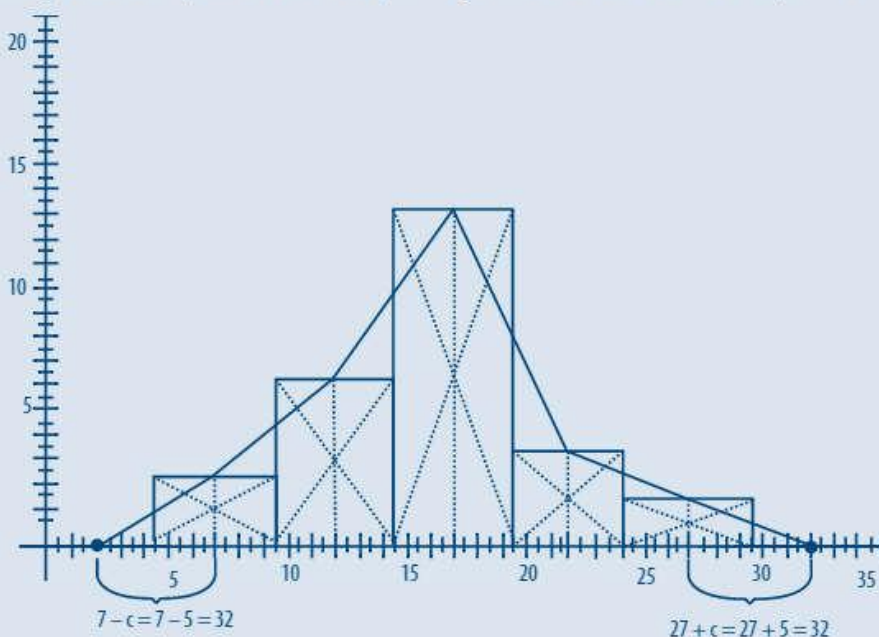


Figura 1.5

■ La curva de distribución de frecuencias

La importancia del polígono de frecuencias es que expone de manera gráfica cómo se distribuyen los datos de la colección a través de los intervalos de confianza; en realidad, si el trazo del polígono de frecuencias se suaviza da lugar a la llamada curva de distribución de frecuencias. De hecho se cuenta con modelos bien definidos que exponen las diferentes formas en que se distribuyen los datos y que determinan la ubicación de los valores de las medidas de tendencia central, tal como se expone a través de los siguientes casos.

- I. **Curva de distribución simétrica.** En este tipo de distribución el intervalo central cuenta con la mayor frecuencia (mayor cantidad de datos), existiendo las mismas frecuencias tanto por arriba como por debajo de este intervalo, además de que la media, la mediana y la moda se concentran al centro de este intervalo compartiendo el mismo valor.

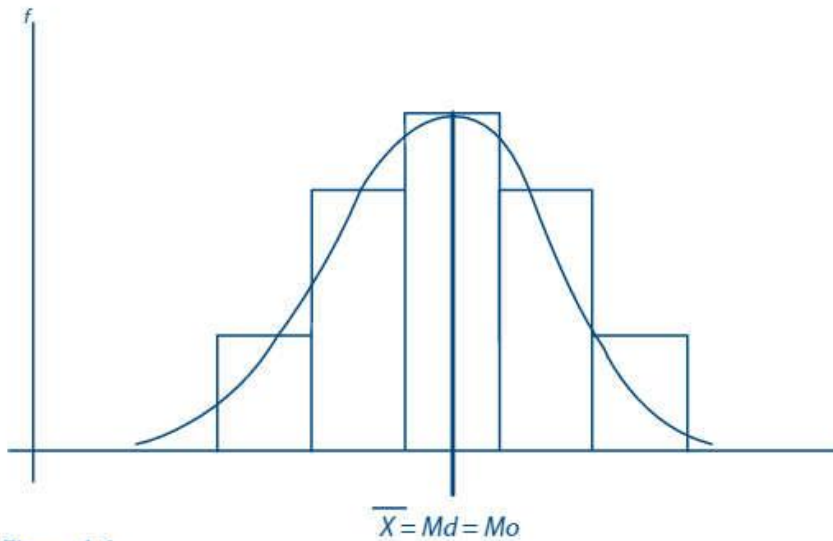


Figura 1.6

- II. **Curva de distribución asimétrica a la derecha o con sesgo positivo.** En este tipo de distribución la mayoría de los datos se concentran en los primeros intervalos, donde se ubican los valores bajos de la colección, mientras que el área de asimetría se ubica en los intervalos donde se concentran los valores altos de la colección (hacia la derecha). En este esquema el orden de las medidas de tendencia central es el siguiente: moda, mediana y media.

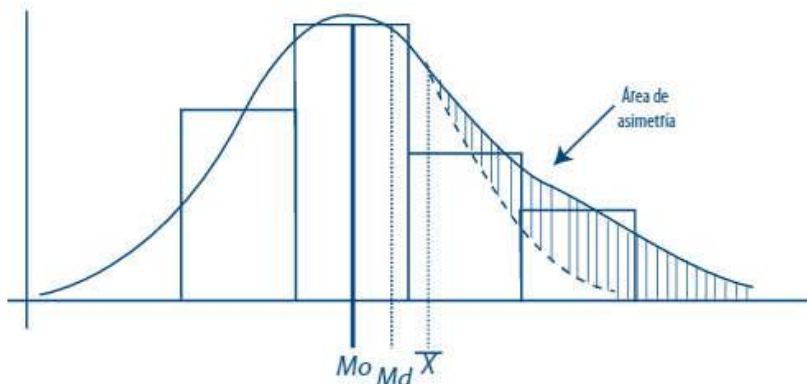


Figura 1.7

- III. **Curva de distribución asimétrica a la izquierda o sesgo negativo.** En este tipo de distribución la mayoría de los datos se concentran en los últimos intervalos donde se ubican los valores altos de la colección, mientras que el área de asimetría se ubica en los intervalos donde se concentran los valores bajos de la colección (a la izquierda). En este esquema el orden de las medidas de tendencia central es el siguiente: media, mediana y moda.

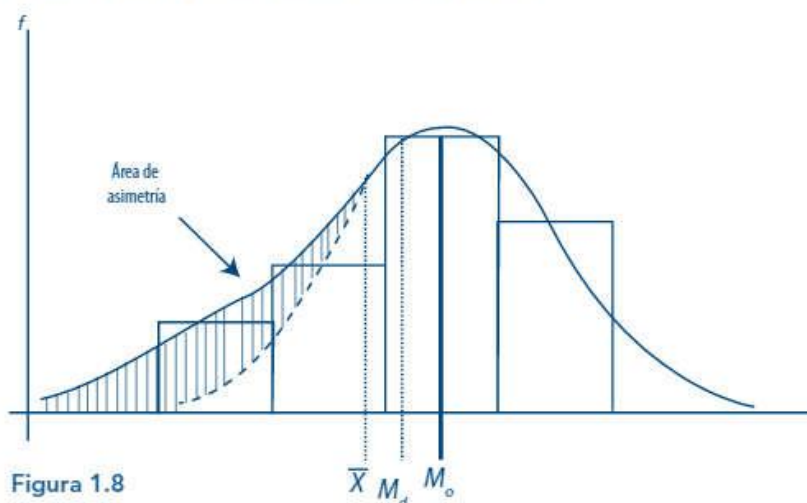


Figura 1.8

Además, el polígono de frecuencias también permite establecer *a priori* qué tan dispersos se encuentran los datos con respecto a la media, ya que con base en esto la curva de distribución tomará ciertas configuraciones, las cuales se denominan de apuntamiento y se exponen a continuación:

Tipo de apuntamiento	Interpretación
<p>Leptocúrtica</p>	Los valores de los datos están cercanos a la media, habiendo desviaciones mínimas y, en consecuencia, se presenta una baja dispersión, haciendo que la curva de distribución sea esbelta y erguida.
<p>Mesocúrtica</p>	Los valores de los datos se disponen de manera moderada alrededor de la media; están dispersos de manera moderada.
<p>Platocúrtica</p>	Los valores de los datos se encuentran bastante alejados de la media, por lo que se presentan grandes desviaciones, por lo que la curva de distribución es amplia y de baja altura.

■ La ojiva

Es la gráfica que permite establecer los valores acumulados de frecuencia de manera porcentual, o sea, se fundamenta en la **far**. La ojiva se compone de segmentos de recta correspondientes a cada intervalo de clase, donde la pendiente de estos segmentos indica el aumento o la disminución de la frecuencia acumulada relativa.

Considerando los valores de los límites reales así como de la **far**, la ojiva se construye marcando los valores de la **far** sobre los límites superiores de cada intervalo, tal como se muestra a continuación.

Problema resuelto

Considerando la siguiente tabla de frecuencias, procede a construir la ojiva correspondiente.

Tabla 1.12

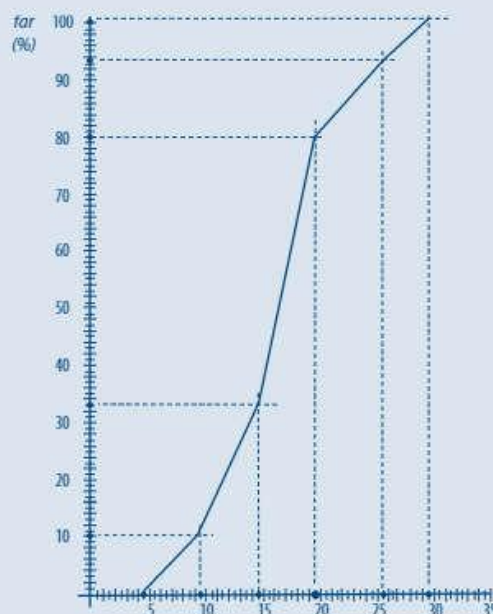
Intervalos nominales			Intervalos reales			<i>f</i>	<i>fa</i>	<i>fr</i>	<i>far</i>
LI		LS	LIR		LSR				
5	—	9	4.5	—	9.5	3	3	$\frac{3}{30} = 10.00\%$	10.00%
10	—	14	9.5	—	14.5	7	10	$\frac{7}{30} = 23.33\%$	33.33%
15	—	19	14.5	—	19.5	14	24	$\frac{14}{30} = 46.67\%$	80.00%
20	—	24	19.5	—	24.5	4	28	$\frac{4}{30} = 13.33\%$	93.33%
25	—	29	24.5	—	29.5	2	30	$\frac{2}{30} = 6.67\%$	100.00%
						$\Sigma f = n = 30$		100.00%	

Solución

Como se mencionó anteriormente, se marcan los límites de clase de los intervalos de clase reales, procediendo a ubicar las **far** correspondientes tal como se muestra.

Nótese que el cambio en la pendiente del segmento del tercer intervalo se debe al marcado cambio en el valor de la **far**, ya que pasa de 33.33% a 80%.

Figura 1.9



Alerta

La ojiva permite establecer la **far** de valores de interés dentro del rango de la colección.

Alerta

El cálculo de las medidas de tendencia central se fundamenta en los límites de clase real, la frecuencia, la frecuencia acumulada y la amplitud de clase.

1.8 Medidas de tendencia central para datos agrupados

El cálculo de los valores de las medidas de tendencia para datos agrupados puede causar algo de dificultad en el entendido de que estos se encuentran implícitos dentro de los límites de los intervalos que los contienen, ya que tan solo se cuenta con el número de datos que se han catalogado en cada intervalo, o sea las frecuencias; sin embargo, con base en las mismas es posible determinar sus valores.

De hecho, los procesos de cálculo consisten en determinar la proporción de la amplitud de clase que separa el valor de la medida de centralización con respecto al límite inferior real del intervalo donde se encuentra; la distancia se determina mediante el cálculo de una razón compuesta por las frecuencias reales y acumuladas del intervalo.

■ Cálculo de la media para datos agrupados

Para determinar el valor de la media se aplicará el principio: "el promedio de los promedios es el promedio", de manera que al determinar el promedio de las veces en que las marcas de clase se repiten dentro de los intervalos se obtendrá la media:

$$\bar{X} = \frac{\sum_{i=1}^n f_i MC_i}{n}$$

donde:

f_i = Valor de la frecuencia en el intervalo i ($i = 1, 2, 3 \dots n$).

MC_i = Valor de la marca de clase del intervalo i .

N = Número de datos de la colección.

Alerta

La media de datos agrupados se fundamenta en el promedio de las marcas de clase.

Problema resuelto

Considerando la siguiente tabla de frecuencias, procede a determinar el valor de la media.

Tabla 1.13

Intervalos nominales			Intervalos reales			f	fa	fr	far
LI	LS		LIR	LSR					
5	—	9	4.5	—	9.5	3	3	$\frac{3}{30} = 10.00\%$	10.00%
10	—	14	9.5	—	14.5	7	10	$\frac{7}{30} = 23.33\%$	33.33%
15	—	19	14.5	—	19.5	14	24	$\frac{14}{30} = 46.67\%$	80.00%
20	—	24	19.5	—	24.5	4	28	$\frac{4}{30} = 13.33\%$	93.33%
25	—	29	24.5	—	29.5	2	30	$\frac{2}{30} = 6.67\%$	100.00%
						$\Sigma f = n = 30$		100.00%	

Solución

Por tanto, de acuerdo con el ejemplo en desarrollo, el valor de la media es:

$$\bar{X} = \frac{\sum_{i=1}^n f_i MC_i}{n} = \frac{485}{30} = 16.67$$

■ Cálculo de la mediana para datos agrupados

El cálculo de la mediana se fundamenta en el límite inferior real y las frecuencias donde se sitúa el valor de la frecuencia acumulada relativa del 50%, debido a que el valor de la mediana de una colección de datos se ubica al centro de la misma.

$$M_d = LIR_{M_d} + \frac{\frac{n}{2} - fa_{antM_d}}{f_{M_d}} \cdot c$$

donde

LIR_{M_d} = Límite inferior real de la clase mediana.

fa_{antM_d} = Frecuencia acumulada de la clase anterior a la clase mediana.

f_{M_d} = Frecuencia de la clase mediana.

c = Amplitud de clase de la clase mediana.

n = Número total de datos de la colección.

Por tanto, para ejemplificar se presenta el siguiente problema.

Problema resuelto

Considerando la siguiente tabla de frecuencias, procede a determinar el valor de la mediana.

Tabla 1.14

Intervalos nominales			Intervalos reales			f	fa	fr	far
LI	LS		LIR	LSR					
5	—	9	4.5	—	9.5	3	3	$\frac{3}{30} = 10.00\%$	10.00%
10	—	14	9.5	—	14.5	7	10	$\frac{7}{30} = 23.33\%$	33.33%
15	—	19	14.5	—	19.5	14	24	$\frac{14}{30} = 46.67\%$	80.00%
20	—	24	19.5	—	24.5	4	28	$\frac{4}{30} = 13.33\%$	93.33%
25	—	29	24.5	—	29.5	2	30	$\frac{2}{30} = 6.67\%$	100.00%
						$\Sigma f = n = 30$		100.00%	

Solución

Considerando los datos de la tabla de frecuencias nos damos cuenta de que, con base en los valores de la **far**, 50% de la colección se encuentra ubicado en el tercer intervalo, por lo que aplicando la fórmula para el cálculo de la mediana:

$$M_d = LIR_{M_d} + \frac{\frac{n}{2} - fa_{antM_d}}{f_{M_d}} \cdot c = 14.5 + \frac{\frac{30}{2} - 10}{14} (5) = 16.28$$



Alerta

La mediana es el valor que representa una **far** igual a 50%.

■ Cálculo de la moda para datos agrupados

En relación con la moda (M_o), se fundamenta en ubicar el intervalo o intervalos que tengan la frecuencia mayor entre todas las frecuencias, en consideración a que en esa o esas clases es donde se puede repetir uno o varios valores. El cálculo de la moda se logra a través de la siguiente fórmula:

$$M_o = LIR_{M_o} + \frac{(f_{M_o} - f_1)}{(f_{M_o} - f_1) + (f_{M_o} - f_2)} \cdot c = LIR_{M_o} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot c$$

donde

LIR_{M_o} = Límite inferior real de la clase modal.

f_{M_o} = Frecuencia de la clase modal.

f_1 = Frecuencia anterior a la clase modal.

f_2 = Frecuencia posterior a la clase modal.

c = Amplitud de clase de la clase modal.

Problema resuelto

Considerando la siguiente tabla de frecuencias, procede a determinar el valor de la moda.

Tabla 1.15

Intervalos nominales			Intervalos reales			f	f_u	f_r	f_{ur}
LI	LS		LIR	LSR					
5	—	9	4.5	—	9.5	3	3	$\frac{3}{30} = 10.00\%$	10.00%
10	—	14	9.5	—	14.5	7	10	$\frac{7}{30} = 23.33\%$	33.33%
15	—	19	14.5	—	19.5	14	24	$\frac{14}{30} = 46.67\%$	80.00%
20	—	24	19.5	—	24.5	4	28	$\frac{4}{30} = 13.33\%$	93.33%
25	—	29	24.5	—	29.5	2	30	$\frac{2}{30} = 6.67\%$	100.00%
						$\Sigma f = n = 30$		100.00%	

Solución

En atención a la tabla de frecuencias del ejemplo en desarrollo se determina que la colección de datos es modal, considerando que la frecuencia con mayor valor se ubica en el tercer intervalo, por lo que, sustituyendo valores en la fórmula, se tiene que:

$$M_o = LIR_{M_o} + \frac{(f_{M_o} - f_1)}{(f_{M_o} - f_1) + (f_{M_o} - f_2)} \cdot c = 14.5 + \frac{(14 - 7)}{(14 - 7) + (14 - 4)} \cdot (5) = 16.56$$

1.9 Determinación de la ubicación y el valor de la moda de manera gráfica mediante el histograma y el polígono de frecuencias

Mediante el histograma y el polígono de frecuencias se puede determinar la ubicación y el valor aproximado de la moda. En este caso particular se debe ubicar el intervalo de la clase mediana de manera que se trace una diagonal que una los vértices superiores derechos de las barras del intervalo inmediato anterior con el de la clase modal, procediendo de igual manera con los vértices superiores izquierdos del intervalo inmediato posterior y el de la clase modal, de manera que la línea perpendicular a la intersección de las diagonales permitirá ubicar el valor, tal como se muestra a continuación en el siguiente problema resuelto.

Alerta

La moda se ubica en el intervalo o intervalos con la frecuencia mayor.

Problema resuelto

Considerando el histograma y polígono de frecuencias que se muestran, procede a determinar el valor de la moda de manera gráfica.

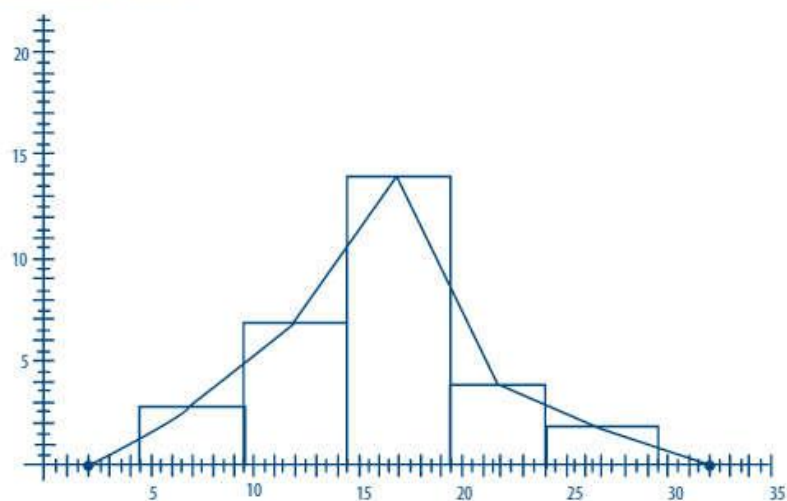


Figura 1.10

Solución

Desarrollando el método descrito se determina el valor de la moda dentro del intervalo con la mayor frecuencia.

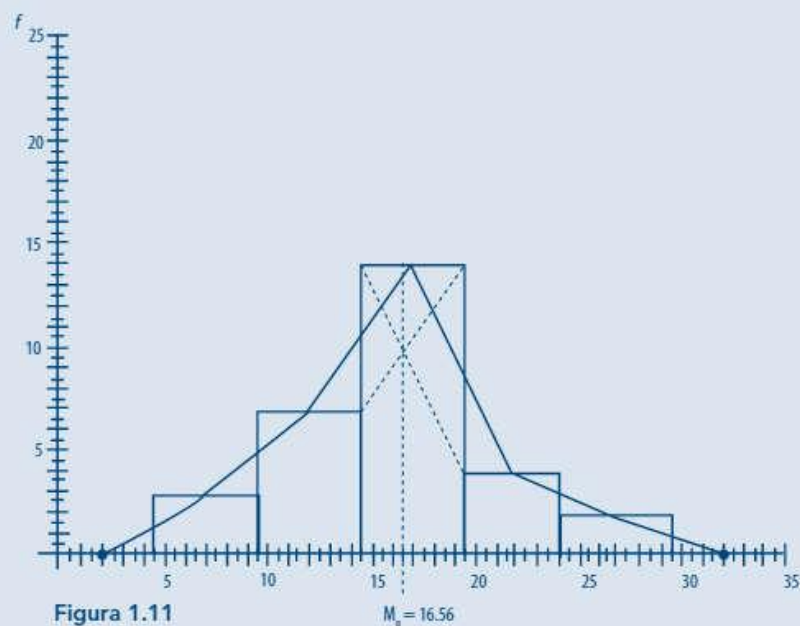


Figura 1.11



Alerta

El método gráfico permite verificar el resultado del cálculo de la moda.

1.10 Determinación de la ubicación y el valor de la mediana de manera gráfica mediante la ojiva

Para determinar el valor de la mediana por medio de la ojiva es necesario trazar lo que algunos denominan la ojiva descendente, donde los valores de la *far* para la misma están definidos por $1 - far$. El valor de la mediana corresponde a la abscisa de la intersección de las dos ojivas, tal como se demuestra en el siguiente problema.

Problema resuelto

Considerando la información que se proporciona en la siguiente tabla de frecuencias, procede a determinar el valor de la mediana de manera gráfica.

Tabla 1.16					
Intervalos reales			<i>f</i>	<i>fa</i>	<i>fr</i>
LIR		LSR			
4.5	—	9.5	3	3	$\frac{3}{30} = 10.00\%$
9.5	—	14.5	7	10	$\frac{7}{30} = 23.33\%$
14.5	—	19.5	14	24	$\frac{14}{30} = 46.67\%$
19.5	—	24.5	4	28	$\frac{4}{30} = 13.33\%$
24.5	—	29.5	2	30	$\frac{2}{30} = 6.67\%$
			$\Sigma f = n = 30$		100.00%

Solución

Desarrollando el método descrito se determina el valor de la mediana.

Tabla 1.17						
Intervalos reales			<i>f</i>	<i>fa</i>	<i>fr</i>	<i>far I</i>
LIR		LSR				
4.5	—	9.5	3	3	$\frac{3}{30} = 10.00\%$	10.00%
9.5	—	14.5	7	10	$\frac{7}{30} = 23.33\%$	33.33%
14.5	—	19.5	14	24	$\frac{14}{30} = 46.67\%$	80.00%
19.5	—	24.5	4	28	$\frac{4}{30} = 13.33\%$	93.33%
24.5	—	29.5	2	30	$\frac{2}{30} = 6.67\%$	100.00%
			$\Sigma f = n = 30$		100.00%	

Solución (continuación)

Trazando las ojivas de acuerdo con los límites de clase correspondientes:

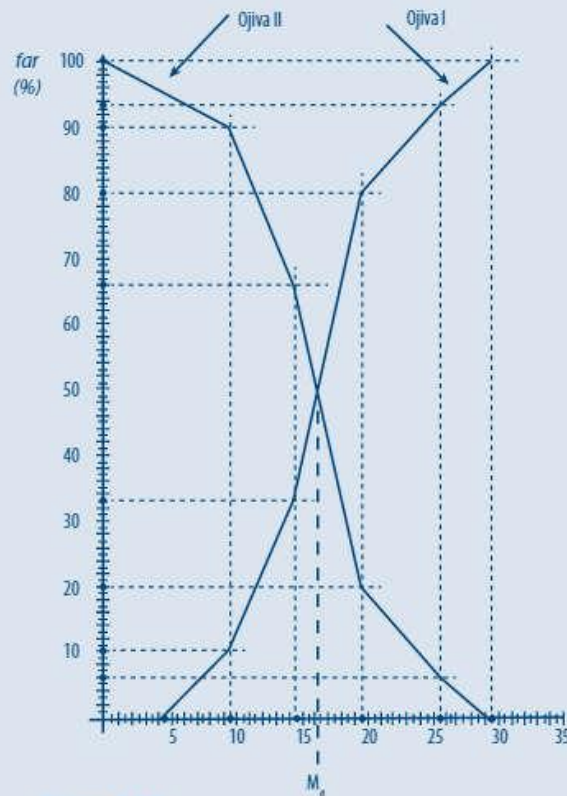


Figura 1.12

**Alerta**

Debe observarse la construcción de las ojivas a efecto de lograr la correcta ubicación del valor de la mediana.

1.11 Cálculo de cuantiles para datos agrupados

El cálculo de los cuantiles para colecciones de datos agrupados se fundamenta en la determinación del valor de los percentiles, por lo que para determinar el valor de un cuartil o decil en particular se deberá considerar su equivalencia en percentiles. El valor de cualquier percentil se puede calcular por medio de la fórmula que da origen a la fórmula para la determinación del valor de la mediana.

$$P_p = LIR_p + \frac{\left(\frac{P}{100} \cdot n\right) - fa_{antP}}{f_p} \cdot c$$

donde

LIR_p = Límite inferior real de la clase que contiene al percentil P .

fa_{antP} = Frecuencia acumulada de la clase anterior a la clase percentil.

f_p = Frecuencia de la clase percentil.

c = Amplitud de clase de la clase percentil, recordando que se trata de una constante.

n = Número total de datos de la colección.

Para ejemplificar la aplicación de la fórmula anterior considérese el siguiente problema.

Problema resuelto

Considerando la siguiente tabla de frecuencias, procede a determinar los valores del tercer cuartil y del segundo decil.

Tabla 1.18 Frecuencias					
Intervalos reales			<i>f</i>	<i>f_a</i>	<i>f_r</i>
LIR		LSR			
4.5	—	9.5	3	3	$\frac{3}{30} = 10.00\%$
9.5	—	14.5	7	10	$\frac{7}{30} = 23.33\%$
14.5	—	19.5	14	24	$\frac{14}{30} = 46.67\%$
19.5	—	24.5	4	28	$\frac{4}{30} = 13.33\%$
24.5	—	29.5	2	30	$\frac{2}{30} = 6.67\%$
			$\Sigma f = n = 30$		100.00%

Solución

Considerando los valores de la **far** se puede determinar que el tercer cuartil (75%) se ubica en el tercer intervalo, mientras que el segundo decil (20%) se ubica en el segundo intervalo, por lo que, sustituyendo en la fórmula:

$$Q_3 = P_{75} = LIR_p + \frac{\left(\frac{P}{100} \cdot n\right) - fa_{antp}}{f_p} \cdot c = 14.5 + \frac{\left(\frac{75}{100} \cdot 30\right) - 10}{14} (5) = 18.96$$

$$D_2 = P_{20} = LIR_p + \frac{\left(\frac{P}{100} \cdot n\right) - fa_{antp}}{f_p} \cdot c = 9.5 + \frac{\left(\frac{20}{100} \cdot 30\right) - 3}{7} (5) = 11.64$$

Alerta

La fórmula de los cuantiles para datos agrupados ofrece el valor del percentil de interés.

1.12 Momentos estadísticos

Considerando que la configuración de la curva de distribución de frecuencias ofrece la disposición de los datos en razón del apuntamiento y simetría, mismo que depende de los valores de las frecuencias y de la dispersión, es posible establecer la caracterización de la curva sin necesidad de desarrollar los elementos gráficos; esta opción la ofrecen los llamados momentos estadísticos con respecto a la media.

Los momentos estadísticos se fundamentan en las diferencias entre las marcas de clase y la media, considerando que las marcas de clase son los valores representativos de cada intervalo, como lo es también la media pero de toda la colección. Estas diferencias bajo condiciones promedio permitirán establecer las características de la curva tal como se expone a continuación.

El cálculo fundamental es la diferencia de las marcas de clase con respecto a la media:

$$Y = MC - \bar{X}$$

De manera que los momentos estadísticos (ME) se definen e interpretan como sigue:

- a) **Momento estadístico de primer grado**, el cual expone que las diferencias entre las marcas de clase con respecto a la media deben ser cero o próximas a él, ya que los valores de las desviaciones por arriba y por debajo de la media deben ser iguales.

$$ME_1 = \frac{\sum_{i=1}^n f_i \cdot Y_i}{n} = 0$$

donde

f_i = Frecuencia del intervalo i ($i = 1, 2, 3, \dots, n$).

Y_i = Diferencia de marca de clase del intervalo i con respecto a la media.

n = Número total de datos de la colección.

- b) **Momento de segundo grado**, que expone el promedio de las desviaciones al cuadrado de las marcas de clase con respecto a la media, lo que significa que se está calculando la varianza, por lo que es posible determinar el valor de la desviación estándar (S).

$$ME_2 = \frac{\sum_{i=1}^n f_i \cdot Y_i^2}{n} = S^2$$

donde

f_i = Frecuencia del intervalo i ($i = 1, 2, 3, \dots, n$).

Y_i = Diferencia de marca de clase del intervalo i con respecto a la media elevada al cuadrado.

n = Número total de datos de la colección.

- c) **Momento estadístico de tercer grado**, en el que las desviaciones de las marcas de clase con respecto a la media se elevan al cubo promediándolas con respecto al número de elementos de la colección, lo que permite establecer el tipo de sesgo que guarda la curva de distribución a través del coeficiente de asimetría (K_3).

$$ME_3 = \frac{\sum_{i=1}^n f_i \cdot Y_i^3}{n}$$

donde

f_i = Frecuencia del intervalo i ($i = 1, 2, 3, \dots, n$).

Y_i = Diferencia de marca de clase del intervalo i con respecto a la media elevada al cubo.

n = Número total de datos de la colección.

Con base en el momento de tercer grado se calcula el coeficiente de asimetría:

$$k_3 = \frac{ME_3}{S^3}$$

donde

Si $k_3 > 0$, la curva de distribución tiene una asimetría derecha o sesgo positivo.

Si $k_3 < 0$, la curva de distribución tiene una asimetría izquierda o sesgo negativo.

Si $k_3 = 0$, la curva de distribución es simétrica.

- d) **Momento de cuarto grado**, en el que las desviaciones de las marcas de clase con respecto a la media se elevan a la cuarta potencia, promediándose con respecto al número de elementos de la colección, lo que permite establecer el tipo de apuntamiento de la curva de distribución de frecuencias a través del coeficiente de apuntamiento de curtosis.

$$ME_4 = \frac{\sum_{i=1}^n f_i \cdot Y_i^4}{n} = S^4$$

Alerta

El momento de primer grado permite corroborar que el promedio de las diferencias con respecto a la media es cero.

Alerta

El momento de segundo grado permite determinar el valor de la desviación estándar con base en la suma de los cuadrados de las variaciones entre las marcas de clase y la media.

Alerta

El momento de tercer grado determina el sesgo de la curva de la distribución de frecuencias.

Alerta

El momento de cuarto grado determina el apuntamiento de la curva de la distribución de frecuencias.

donde

f_i = Frecuencia del intervalo i ($i = 1, 2, 3, \dots, n$).

Y_i = Diferencia de marca de clase del intervalo i con respecto a la media elevada a la cuarta potencia.

n = Número total de datos de la colección.

Con base en el momento de cuarto grado se calcula el coeficiente de apuntamiento o de curtosis:

$$k_4 = \frac{ME_4}{S^4}$$

De manera que

Si $k_4 - 3 > 0$, la curva es leptocúrtica.

Si $k_4 - 3 = 0$, la curva es mesocúrtica.

Si $k_4 - 3 < 0$, la curva es platocúrtica.

Problema resuelto

Considerando la siguiente tabla de frecuencias, procede a determinar la caracterización de la curva de distribución de frecuencias por medio del cálculo de momentos estadísticos.

Tabla 1.19

Intervalos nominales		Intervalos reales		MC	f	f_u	f_{ur}	$f \cdot MC$	$Y = MC - \bar{X}$
5	9	4.5	9.5	7	3	3	10.00%	21.00	-9.17
10	14	9.5	14.5	12	7	10	33.33%	84.00	-4.17
15	19	14.5	19.5	17	14	24	80.00%	238.00	0.83
20	24	19.5	24.5	22	4	28	93.33%	88.00	5.83
25	29	24.5	29.5	27	2	30	100.00%	54.00	10.83
				$\Sigma f = n =$	30		$\Sigma =$	485	

Solución

Tabla 1.20

Intervalos nominales		Intervalos reales		f	$Y = MC - \bar{X}$	$f \cdot Y$	$f \cdot Y^2$	$f \cdot Y^3$	$f \cdot Y^4$
5	9	4.5	9.5	3	−9.17	−27.50	252.08	−2310.76	21 182.00
10	14	9.5	14.5	7	−4.17	−29.17	121.53	−506.37	2 109.86
15	19	14.5	19.5	14	0.83	11.67	9.72	8.10	6.75
20	24	19.5	24.5	4	5.83	23.33	136.11	793.98	4 631.56
25	29	24.5	29.5	2	10.83	21.67	234.72	2 542.82	27 547.26
$\Sigma f = n =$				30	$\Sigma =$	0.00	754.17	527.78	55 477.43

Donde las características de la curva de distribución de frecuencias a través de los momentos se exponen a continuación.

$$ME_1 = \frac{\sum_{i=1}^n f_i \cdot Y_i}{n} = \frac{0}{30} = 0$$

$$ME_2 = S_2 = \frac{\sum_{i=1}^n f_i \cdot Y_i^2}{n} = \frac{754.17}{30} = 25.14$$

$$S = \sqrt{25.14} = 5.014$$

$$ME_3 = \frac{\sum_{i=1}^n f_i \cdot Y_i^3}{n} = \frac{527.78}{30} = 17.59$$

$$k_3 = \frac{ME_3}{S^3} = \frac{17.59}{5.014^3} = 0.134$$

Solución (continuación)

Si $k_3 > 0$, la curva de distribución tiene una asimetría derecha o sesgo positivo.

$$ME_4 = \frac{\sum_{i=1}^n f_i \cdot Y_i^4}{n} = \frac{55477.43}{30} = 1849.25$$

$$k_4 = \frac{ME_4}{S^4} = \frac{1849.25}{5.014^4} = 2.92$$

Si $k_4 - 3 = -0.08$, la curva es ligeramente platocúrtica.

1.13 Determinación del número de intervalos de clase

La determinación del número de intervalos para realizar un estudio de estadística descriptiva puede generar puntos de vista encontrados y ser motivo de discusión. Sin embargo, es posible determinarlo de manera simple si se observan el cumplimiento del proceso de cálculo y el uso del buen criterio.

Debe tenerse presente que un número muy amplio de intervalos propiciaría la obtención de demasiados valores que no ofrecerían una buena perspectiva de análisis; tampoco favorece al análisis tener un número muy limitado de intervalos.

En esta obra se presentan dos métodos simples, mismos que se exponen a continuación. El primero de ellos es el método empírico, el cual consiste en que una vez que se haya determinado el valor del rango de una colección se buscará un par de números cuyo producto se aproxime al mismo, considerando que el primer número represente el número de intervalos y el segundo la amplitud de clase; para ejemplificar este método considérese el siguiente problema resuelto.

Problema resuelto

Considerando la colección de datos que se muestra, procede a determinar el número de intervalos por el método empírico.

169	128	140	176	161	119	164	154
140	136	154	136	148	144	175	148
136	162	130	141	123	159	151	150
146	139	160	146	136	148	144	152
145	142	156	145	128	150	135	161

Solución

En la colección de datos el valor máximo es 176 mientras que el valor mínimo es 119, por lo que,

$$R = 176 - 119 = 57$$

Revisando las tablas de multiplicar:

Número de intervalos	Amplitud de clase
1	56
56	1
2	28
28	2
4	14
14	4
7	8
8	7

Solución (continuación)

Analizando las opciones se considera que las dos últimas representan las mejores; por optar por alguna se elige la primera de ellas, por lo que se calcula la amplitud de clase con el propósito de ajustarla de ser necesario.

$$\text{Amplitud de clase} = c = \frac{\text{Rango}}{\text{No. de intervalos}} = \frac{57}{7} = 8.1 \approx 9$$

Se sube a la siguiente unidad debido a que $7 \times 8 = 56$ que es menor que 57; de manera que $7 \times 9 = 63$ excede el rango de la colección propuesta.

Por ello, los valores de los límites nominales de los intervalos de clase se determinan considerando que el límite inferior del primer intervalo debe empezar en 119, pero quedaría muy justo, por lo que se puede sugerir restar una unidad, de modo que empezaría en 118 y partiendo de este número se le sumará la amplitud de clase para definir los valores de los límites inferiores de clase. Para determinar el valor de los límites superiores de cada intervalo se aplica la siguiente fórmula:

$$LS = (LI + c) - 1$$

La razón de restar una unidad es que el intervalo empieza en el valor del límite inferior, por lo que si $c = 9$,

Tabla 1.22

Límite inferior		Límite superior
118	-	126
127	-	135
136	-	144
145	-	153
154	-	162
163	-	171
172	-	180

El segundo método de cálculo para determinar el número de intervalos es aplicando la fórmula de Sturges:

Problema resuelto

Considerando la colección de datos que se muestra, se procede a determinar el número de intervalos por el método de la fórmula de Sturges.

$$\text{Número de intervalos} = 1 + 3.3 \cdot \log(n)$$

169	128	140	176	161	119	164	154
140	136	154	136	148	144	175	148
136	162	130	141	123	159	151	150
146	139	160	146	136	148	144	152
145	142	156	145	128	150	135	161

Solución

Considerando que el número de elementos de la colección del ejemplo anterior es 40.

$$\text{Número de intervalos} = 1 + 3.3 \cdot \log(40) = 6.28 \approx 7$$

Como se aprecia, hay cierta similitud entre los resultados de los métodos expuestos.

Solución (continuación)

Una vez que se han definido los intervalos se procede a catalogar los datos de la colección en los mismos.

169	128	140	176	161	119	164	154
140	136	154	136	148	144	175	148
136	162	130	141	123	159	151	150
146	139	160	146	136	148	144	152
145	142	156	145	128	150	135	161

Tabla 1.23

Límite inferior	Límite superior		f	
118	-	126		2
127	-	135		4
136	-	144		11
145	-	153		11
154	-	162		8
163	-	171		2
172	-	180		2
				$\Sigma =$ 40

Una vez completado este paso la tabla de frecuencias está dispuesta para desarrollar un análisis por estadística descriptiva, tal como se ha propuesto a lo largo de esta unidad.

**Alerta**

Los métodos para la determinación de los intervalos de clase pueden coincidir, considerando la importancia del criterio del analista en la decisión.

1.14 La media geométrica

Es un valor de tendencia central representativo de una colección de datos cuyos valores guardan estrecha relación unos con otros y observan un orden secuencial bien definido, como lo son la tasa de inflación, la tasa de crecimiento, los factores de devaluación y la tasa de interés, por citar algunos entre los más importantes.

El valor de la media geométrica puede calcularse tanto para datos no agrupados como para datos agrupados.

En el caso de datos no agrupados la media geométrica es la raíz n -ésima del producto de los valores del conjunto de datos.

$$\bar{X}_G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \dots x_n}$$

Considérese el siguiente ejemplo de aplicación.

Problema resuelto

A un jubilado le proponen invertir parte de su pensión mensual en un producto de inversión a cuatro meses que le ofrece pagar 8% el primer mes, 12% el segundo mes, 11% el tercer mes y 8% el cuarto mes. ¿Cuál sería el valor de la tasa promedio que recibiría?

Solución

Si se considera que el fundamento de las matemáticas financieras en el manejo de la tasa de interés a futuro es: $(1+i)^n$, donde i es la tasa de interés y n el número de periodos de aplicación de la tasa, que en este caso es "1", por tanto:

$$\bar{X}_G = \sqrt[4]{1.08 \cdot 1.12 \cdot 1.11 \cdot 1.08} = 1.0974$$

Por lo que restando 1 al resultado la tasa promedio es de 9.74%.

**Alerta**

La media geométrica ofrece un criterio adicional para el análisis estadístico, dependiendo del criterio que se pretenda cubrir.

La estadística y la estadística descriptiva

En el caso de datos agrupados el cálculo del valor de la media geométrica se determina mediante la siguiente fórmula:

$$\ln(\bar{X}_G) = \frac{\sum_{i=1}^n f_i \cdot \ln(MC_i)}{n}$$

Considérese el siguiente ejemplo de aplicación.

Problema resuelto

Considerando la siguiente tabla de frecuencias, procede a determinar el valor de la media geométrica.

Tabla 1.24

Intervalos nominales	f	MC
2-4	2	3
5-7	5	6
8-10	3	9
Σ	10	

Solución

Se complementan los cálculos necesarios para determinar el valor de la media geométrica.

Tabla 1.25

Intervalos nominales	f	MC	$\ln(MC)$	$f \times \ln(MC)$
2-4	2	3	1.0987	2.1974
5-7	5	6	1.7918	8.9590
8-10	3	9	2.1972	6.5916
Σ	10			17.748

Aplicando la fórmula de la media geométrica para datos agrupados:

$$\ln(\bar{X}_G) = \frac{\sum_{i=1}^n f_i \cdot \ln(MC_i)}{n} = \frac{17.748}{10} = 1.7748$$

Obteniendo el antilogaritmo:

$$\text{antiln}(1.7748) = 5.90$$

1.1 Ordena la siguiente colección de datos y calcula el valor de las medidas de tendencia central.

98 23 56 37 78 34 84 16 67

1.2 Ordena la siguiente colección de datos y calcula las medidas de tendencia central.

142 161 150 145 130 184 176 154 149 178
112 90 162 77 198 181 116 100 165 199

1.3 Ordena la siguiente colección de datos y calcula las medidas de tendencia central.

1234 1178 1340 1189 1567
1645 1934 1937 1756 1111
1023 1404 1023 987 1032

1.4 Considerando la colección del problema 1.1, calcula las medidas de dispersión: rango, varianza y desviación estándar.

1.5 Considerando la colección de datos del problema 1.2, calcula las medidas de dispersión: rango, varianza y desviación estándar.

1.6 Considerando la colección del problema 1.3, calcula las medidas de dispersión: rango, varianza y desviación estándar.

1.7 Considerando la colección de datos del problema 1.1, calcula el valor de Q_1 , Q_2 y Q_3 .

1.8 Considerando la colección de datos del problema 1.2, calcula el valor de los deciles: D_1 , D_4 , D_6 y D_8 .

1.9 Con base en la colección de datos del problema 1.3, calcula y comprueba el valor de la mediana en razón de: $M_d = Q_2 = D_5$.

1.10 Con base en la colección de datos del problema 1.2, calcula los siguientes cuantiles: P_{15} , P_{40} y P_{80} .

1.11 Con base en la colección de datos del problema 1.3, calcula los siguientes cuantiles: Q_1 , Q_2 , Q_3 , P_{32} , P_{54} y P_{95} .

1.12 Considerando la colección de datos del problema 1.1, determina el valor de la mediana de manera directa y verifica su valor con la fórmula de los cuantiles para datos agrupados.

1.13 Una empresa de mantenimiento de equipo industrial hace un recuento de los rodamientos que fueron sustituidos en equipos de bombeo clasificándolos por su diámetro en milímetros. Considerando la información de la siguiente tabla de frecuencias, calcula los intervalos reales de clase y las marcas de clase, así como la amplitud de clase.

Tabla 1.26

Intervalos	f
221-230	4
231-240	6
241-250	9
251-260	6
261-270	9
271-280	5
281-290	2
291-300	9
301-310	4

1.14 Una fábrica de herrajes para construcción desarrolla un proceso de estadística descriptiva sobre la longitud de su inventario de tensores para tubería de $\frac{1}{2}$ " de diámetro, tal como se expone en la siguiente tabla.

Tabla 1.27

Longitud (cm)	Cantidad
60-62	5
63-65	18
66-68	42
69-71	27
72-74	8

Calcula los intervalos reales y la amplitud de clase, adicionalmente determina las marcas de clase.

1.15 Una empresa importadora de instrumentos de medición y control de procesos químicos clasifica a sus proveedores por el monto de las operaciones de importación anuales, tal como se muestra en la siguiente tabla de frecuencias.

Tabla 1.28

Monto de las operaciones (millones de pesos)	Número de proveedores
De \$2 a \$5	6
De \$5 a \$8	13
De \$8 a \$11	20
De \$11 a \$14	10
De \$14 a \$17	1

Calcula la amplitud de clase y las marcas de clase.

1.16 Una empresa de elementos prefabricados de madera procedió a clasificar los excedentes de bastidores de una pulgada de espesor por su longitud en centímetros. Donde el detalle del inventario se expone a continuación.

25	37	47	60	74
34	38	52	63	66
27	38	49	61	72
46	64	70	44	45
41	53	62	67	42
45	59	71	72	60
50	58	51	56	52
45	49	52	53	57

Como analista de operaciones industriales le ha sido encomendado el desarrollo de un análisis por estadística descriptiva de manera que clasifique el inventario en 5 intervalos, donde el valor del límite inferior nominal del primer intervalo sea el valor mínimo de la colección y el valor del límite superior nominal del quinto intervalo sea el valor máximo de la colección; para que una vez cubierta esta condición se elaboren la tabla de frecuencias, los intervalos de clase reales y las marcas de clase.

1.17 Al término de un proyecto de construcción una empresa de proyectos electromecánicos realiza el levantamiento del inventario de los tramos de cable de 500 MCM considerando su largo en pulgadas. Los responsables del departamento de Ingeniería de Procura (compras) estructuran la muestra que se detalla a continuación.

37	38	38	64	53	59	58	49
74	66	72	45	42	60	52	57
60	63	61	44	67	72	56	53
25	34	27	46	41	45	50	45
65	73	39	70	54	73	71	61
47	52	49	70	62	71	51	52

Elabora la tabla de frecuencias y los intervalos de clase reales, así como las marcas de clase.

1.18 Con base en los datos de la tabla de frecuencias resultante del problema 1.13, determina las medidas de tendencia central.

1.19 Con base en los resultados de la tabla de frecuencias del problema anterior, construye el histograma y el polígono de frecuencias, ubicando y comprobando el valor de la moda.

1.20 Con base en los resultados del problema 1.15, calcula los valores de las medidas de tendencia central.

1.21 Con base en los resultados del problema 1.15, construye el histograma y el polígono de frecuencia.

1.22 Con base en los resultados del problema 1.15, construye las ojivas con el propósito de ubicar y comprobar el valor de la mediana.

1.23 Con base en los resultados del problema 1.17, determina las medidas de tendencia central.

1.24 Con base en los resultados de los problemas 1.17 y 1.23, construye el histograma y el polígono de frecuencias y ubica y comprueba el valor de la moda.

1.25 Con base en los resultados del problema 1.17, construye las ojivas, procediendo a ubicar y comprobar el valor de la mediana.

1.26 Con base en los resultados del problema 1.14, construye el histograma y el polígono de frecuencias, procediendo a ubicar y comprobar el valor de la moda.

1.27 Con base en los resultados del problema 1.15, elabora la tabla de frecuencias de manera que determine las medidas de tendencia central.

1.28 Con base en los resultados del problema 1.27, construye el histograma y el polígono de frecuencias, procediendo a ubicar y comprobar el valor de la moda.

1.29 Con base en los resultados del problema 1.28, desarrolla las ojivas de manera que ubiques y compruebes el valor de la mediana.

1.30 A partir de la colección de datos que se muestra, realiza un análisis de estadística descriptiva para datos no agrupados determinando el valor de las medidas de tendencia central.

169	128	140	176	161	119	164	154
140	136	154	136	148	144	175	148
136	162	130	141	123	159	151	150
146	139	160	146	136	148	144	152
145	142	156	145	128	150	135	161

1.31 Con base en la colección de datos del problema 1.30, determina el sexto decil y el 85avo percentil.

1.32 Con base en la colección de datos del problema 1.30, determina el número de intervalos aplicando la fórmula de Sturges, así como la amplitud de clase y la tabla de frecuencias.

1.33 Con base en los resultados del problema anterior, construye el histograma, el polígono de frecuencias y las ojivas.

1.34 Con base en la tabla de frecuencias del problema 1.30, calcula los valores del sexto decil y el 85avo percentil, determinando si existe diferencia o no con respecto a los resultados del problema 1.31.

1.35 Con base en la tabla de frecuencias del problema 1.30, determina la caracterización de la curva de distribución de frecuencias a través de un análisis descriptivo por momentos estadísticos.

1.36 Con base en la colección de datos que se muestra, calcula las medidas de tendencia central con base en datos no agrupados.

49	121	98	21	56	77
68	34	87	45	72	66
89	46	65	78	37	118
45	78	41	120	98	90
95	90	32	114	45	81

1.37 Con base en la colección del problema anterior, calcula Q_1 , Q_2 , Q_3 , D_2 , D_7 , P_{48} y P_{85} .

1.38 Con base en la colección de datos que se muestra en el problema 1.36, determina el número de intervalos por regla empírica, procediendo a diseñar los mismos.

1.39 Con base en la colección de datos del problema 1.36, determina el número de intervalos por la fórmula de Sturges, procediendo a diseñar los mismos, y argumenta: ¿existe diferencia en el número de intervalos y diseño con los propuestos por el método empírico?

1.40 Con base en los resultados del problema 1.36, desarrolla la tabla de frecuencias de manera que determine las medidas de tendencia central.

1.41 Con base en los resultados del problema 1.40, construye el histograma y el polígono de frecuencias.

1.42 Con base en los resultados del problema 1.40, determina el valor de los siguientes cuantiles: D_1 , D_4 , D_7 , Q_1 , Q_3 , P_{50} , P_{67} y P_{72} .

1.43 Con base en los resultados del problema 1.40, construye la ojiva ascendente y corrobora gráficamente los cuantiles: D_4 , Q_3 , P_{50} y P_{72} .

1.44 Con base en la tabla de frecuencias correspondiente a la colección de datos del problema 1.36, obtén la caracterización de la curva de distribución de frecuencias mediante un análisis por momentos estadísticos.

1.45 En relación con la tabla de frecuencias que se presenta en el problema 1.13, calcula el valor de la media geométrica.

1.46 Con base en la tabla de frecuencias que se presenta en el problema 1.14, calcula el valor de la media geométrica.

1.47 Con base en la colección de datos del problema 1.17, compara el valor de la media aritmética con el valor de la media geométrica.

1.48 Considerando la colección de datos del problema 1.36, calcula la media geométrica por medio de la fórmula para datos no agrupados.

1.49 Compara el valor de la media aritmética con el valor de la media geométrica de datos agrupados del problema 1.36.

1.50 Usa la siguiente tabla de frecuencias a continuación para contestar lo que se te pide.

Tabla 1.29

Intervalos nominales		f
Límite inferior	Límite superior	
18	26	14
27	35	23
36	44	56
45	53	32
54	62	45

Continua tabla

Intervalos nominales		f
Límite inferior	Límite superior	
63	71	56
72	80	56
81	89	40
90	98	22
99	107	56

- Determina el valor de las medidas de tendencia central.
- Elabora el histograma y el polígono de frecuencias, así como la comprobación gráfica del valor de la mediana mediante el cruce de ojivas.
- Determina las características de la curva de distribución de frecuencias mediante momentos estadísticos.
- Determina el valor de los cuartiles y de todos los deciles.
- Determina la proporción de datos que se ubican entre el 3er cuartil y el 95avo percentil; así como entre el 2º decil y la mediana.
- Determina el valor de la media geométrica.



PROBLEMAS RETO

En una importante empresa de alimentos se realizan diferentes pruebas de laboratorio para cuidar la calidad de sus productos, una de ellas es verificar el peso de cada uno de sus productos. Cada lunes en el laboratorio se reciben cinco lotes y cada uno de ellos consta de 25 paquetes de 250 gr. Los registros que se obtienen se encuentran capturados en la siguiente tabla de frecuencias. El gerente del área de Control de Calidad le pide a su analista que le proporcione la siguiente información.

Tabla 1.30

Peso	Frecuencia
239	4
240	6
245	15
247	7
248	11
250	65
252	16
254	1

- Determina el valor de las medidas de tendencia central.
- Elabora el histograma y el polígono de frecuencias, así como la comprobación gráfica del valor de la mediana por medio del cruce de ojivas.
- Determina las características de la curva de distribución de frecuencias mediante momentos estadísticos.

d) Determina el valor de los cuartiles y de todos los deciles.

- Determina el valor de la media geométrica.
- ¿Crees que es necesario revisar las básculas?

Una empresa de cosméticos le encargó a una encuestadora determinar el grado de satisfacción de sus clientes por una crema desmaquillante que tiene en el mercado más de tres años. La respuesta se catalogó en: Excelente (E), Buena (B), Regular (R) o Mala (M). Los resultados obtenidos son los siguientes:

B	E	B	E	E	E	R	E	M
B	M	M	B	M	B	M	E	B
E	R	E	B	E	R	E	E	B
R	E	B	B	E	B	E	B	E
B	R	E	E	M	R	B	B	R
M	R	M	B	B	R	B	R	B
M	R	B	B	R	R	M	B	R
R	B	B	B	E	M	M	M	R
B	E	E	M	B	B	B	R	M
E	E	B	E	B	E	M	M	R
B	R	M	E	B	E	E	E	R
E	B	E	B	B	M	B	E	E
M	E	B	B	R	R	R	E	M
B	B	E	B	B	M	E	E	E
E	R	R	E	R	M	E	M	B
B	B	M	E	B	B	B	B	M

B	R	B	E	B	E	B	B	E
E	B	E	E	B	M	E	R	E
E	E	M	E	R	R	E	R	M
B	B	M	E	B	M	B	M	E
E	B	E	B	B	E	E	M	M
R	E	M	B	E	B	E	E	B
E	R	M	B	M	E	B	E	B
R	E	M	E	M	E	B	E	M
E	B	B	E	B	B	B	B	E
B	M	B	E	B	E	B	B	E
R	B	B	R	E	E	R	B	E
E	E	E	E	B	B	B	R	B
R	B	R	B	E	B	B	M	B
M	M	R	R	E	E	E	E	R
E	R	B	B	B	M	B	B	B
E	E	B	E	E	R	M	R	E

- Construye una tabla de frecuencias, si gustas puedes utilizar Excel.
- ¿Cuál es la ojiva?
- Elabora su histograma y su polígono de frecuencias.
- Determina la caracterización de la curva de distribución de frecuencias mediante un análisis descriptivo por momentos estadísticos.
- ¿Qué porcentaje de personas considera que la crema es mala?
- ¿Cuántas clientes consideran que la crema es excelente?
- ¿Cuál es la imagen general de la crema que tienen las personas?
- ¿Cuáles son tus conclusiones?

Petrolax es una importante empresa petrolera, que desea contratar el suministro de tubos de acero para sus instalaciones, solo que le piden que la asignación del proveedor la tiene que realizar por medio de una licitación abierta. Para la licitación se presentaron tres empresas (Tubular, Tubos de acero y Acero y tubos), las tres empresas venden la unidad al mismo precio y con las mismas especificaciones técnicas del material. Petrolax solicita que el proveedor mantenga un diámetro promedio por cada 20 tubos entregados de 46 pulgadas; para lo cual solicitó a cada empresa un lote muestra de este tamaño. El área de metrología de Petrolax obtuvo los siguientes resultados (las unidades están dadas en pulgadas).

Tabla 1.31

Tubular	Tubos de acero	Acero y tubos
46	45	47
46	47	46
45	47	46
46	46	46
45	46	45
46	47	46
44	47	46
46	45	45
45	46	46
44	46	46
46	46	46
45	46	45
45	45	46
46	46	47
44	46	46
46	46	45
45	46	46
45	45	46
44	45	46
45	45	45

¿Qué proveedor seleccionarías? Es muy importante justificar tu respuesta con un buen análisis.



REFERENCIAS

Berenson, Mark L. y David M. Levine (1996). *Estadística básica en administración* (6a. ed.). México: Prentice Hall Hispanoamericana.

Hines, William W., Douglas C. Montgomery, David M. Goldsman y Connie M. Borror. *Probabilidad y estadística para ingenieros* (4a. ed.). México: Patria, 3a. reimpresión.

Kohler, Heinz (1996). *Estadística para negocios y economía* (1a. ed.). México: Compañía Editorial Continental.

Lind, Douglas, William G. Marchal y Samuel A. Wathen (2005). *Estadística aplicada a los negocios y a la economía* (12a. ed.). México: McGraw-Hill.

Mendenhall, William (1999). *Estadística para administradores* (2a. ed.). México: Grupo Editorial Iberoamérica.



DIRECCIONES ELECTRÓNICAS

<http://www.youtube.com/watch?v=pFotql8CRLU>

<http://www.slideshare.net/jeffertyni/histograma>

Teoría de la probabilidad y distribuciones de probabilidad

OBJETIVOS

- Aplicar el concepto de probabilidad y su trascendencia en la vida cotidiana y profesional.
- Entender los experimentos como actividades que dan origen a los eventos.
- Conocer y aplicar los diferentes tipos de probabilidad.
- Representar y operar los espacios muestrales a través de diagramas de Venn.
- Entender el concepto de variable aleatoria y su relación con la realidad.
- Distinguir las diferencias entre las variables aleatorias, discretas y continuas.
- Comprender y distinguir las principales diferencias entre distribución de probabilidad y función de densidad.
- Conocer y distinguir las diferencias entre las distribuciones de probabilidad discretas.
- Entender y distinguir el concepto y las características de una distribución continua.
- Entender y aplicar el teorema del límite central.

¿QUÉ SABES?

- ¿Cuál es la diferencia entre un experimento y un evento?
- ¿Qué es el espacio muestral?
- ¿Cómo apoya la teoría de conjuntos al análisis de la probabilidad?
- ¿Qué es una permutación y qué es una combinación?
- ¿Cuál es la diferencia entre una distribución de probabilidad y una función de densidad?
- ¿Cuáles son las principales diferencias entre las distribuciones binomial y de Poisson?
- ¿Cuáles son las principales características de la curva de distribución normal?
- ¿Qué son las unidades estandarizadas?

2.1 Introducción

A lo largo de la historia, el hombre siempre ha deseado tener la certeza de los hechos que son de su interés. Por esa razón, ha recurrido a diversos medios, que involucran desde las predicciones, las cuales ofrecen criterios subjetivos, hasta la interpretación sobrenatural de los hechos. Sin embargo, el desarrollo de la teoría de la probabilidad ofrece el análisis de los hechos basados en la objetividad, los cuales se exponen a lo largo de este capítulo.

2.2 Concepto de probabilidad

Probabilidad se define como la cuantificación de la posibilidad de que un hecho ocurra.

El análisis de la probabilidad se fundamenta en los experimentos, los cuales se definen como: *el conjunto de actividades interrelacionadas que permite el desarrollo de un fenómeno de manera parcial o total, y donde al resultado de un experimento se le denomina evento*.

Con base en la definición anterior, debe tenerse presente que un experimento puede tener más de un evento (resultado). Por tanto, al conjunto de eventos se le denomina *espacio muestral*, al cual, por lo general, se le denota con la letra S (del inglés *space*, que en español significa espacio).

A continuación se presenta una serie de ejemplos de experimentos y sus correspondientes espacios muestrales.

Ejemplos

- **Experimento:** Lanzar un dado $S = \{1, 2, 3, 4, 5, 6\}$
- **Experimento:** Lanzar una moneda $S = \{\text{águila}, \text{sol}\}$
- **Experimento:** Observación del género de los recién nacidos en un servicio de maternidad $S = \{\text{varón}, \text{mujer}\}$

2.3 Los espacios muestrales y la teoría de conjuntos

Es importante resaltar que los espacios muestrales de un experimento se consideran conjuntos, donde los eventos son subconjuntos de los mismos.

Con base en lo expuesto antes, a los eventos de los espacios muestrales se les representa de acuerdo con la notación propia de los conjuntos, como se muestra en los ejemplos siguientes:

- a) Denotación con una letra mayúscula, para identificar el conjunto, y listar entre llaves los eventos.
 - **Experimento:** Lanzar un dado.
 - **Espacio muestral:** $S = \{1, 2, 3, 4, 5, 6\}$
- b) Denotación con una letra mayúscula, para identificar el conjunto y exponer entre llaves un argumento lógico.
 - **Experimento:** Lanzar un dado.
 - **Espacio muestral:** $S = \{X: 1 \leq X \leq 6\}$
- c) Empleo de los llamados diagramas de Venn o Euler.

En este caso, el área delimitada por el rectángulo expresa al espacio muestral, mientras que el círculo adentro del rectángulo representa los eventos (véase figura 2.1).

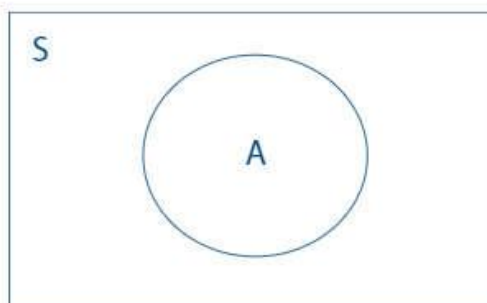


Figura 2.1

Alerta

La probabilidad es la cuantificación de la posibilidad de que un hecho ocurra.

Alerta

Los eventos son los diferentes resultados de un experimento.

Alerta

Se denomina **espacio muestral** al conjunto de eventos.

Considerando el ejemplo en desarrollo:

- **Experimento:** Lanzar un dado.
- **Espacio muestral:** Está compuesto por seis eventos, por lo que el diagrama de Venn del espacio muestral queda expresado de la siguiente manera:

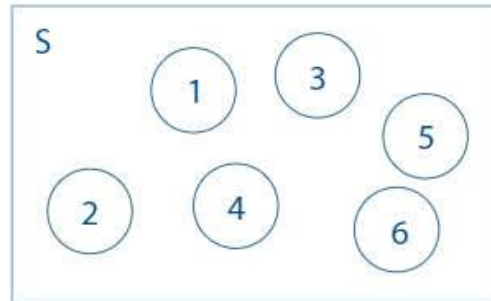


Figura 2.2

■ Operaciones con conjuntos

De manera clásica, algunas operaciones se identifican con los conjuntos. Estas operaciones y los conjuntos se aplican en el análisis de la probabilidad. A continuación se relacionan los más importantes:

- a) **Unión de conjuntos.** Consiste en agrupar los elementos de los conjuntos en estudio en un solo bloque, en el cual simbólicamente se representan por $A \cup B$ (véase figura 2.3).

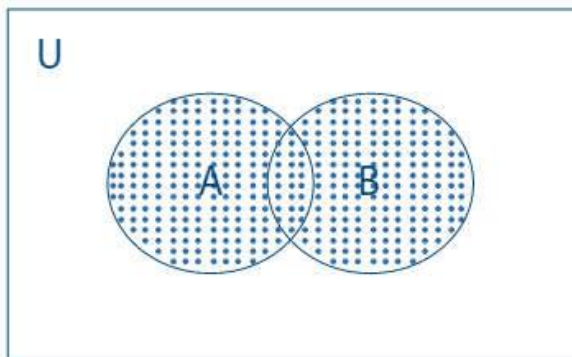


Figura 2.3

- b) **Intersección de conjuntos.** Se refiere a considerar solo aquellos elementos que son compartidos por conjuntos en análisis; simbólicamente, la intersección de conjuntos se representa por $A \cap B$ (véase figura 2.4).

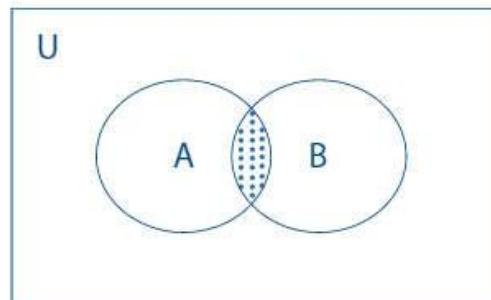


Figura 2.4

Alerta

Los espacios muestrales se pueden representar mediante diagramas de Venn, donde los conjuntos simbolizan los eventos de un experimento.

- c) **Conjunto complemento.** Este se compone por todos aquellos elementos que no forman parte de un conjunto. Simbólicamente, se representa como A^c o A' (véase figura 2.5).

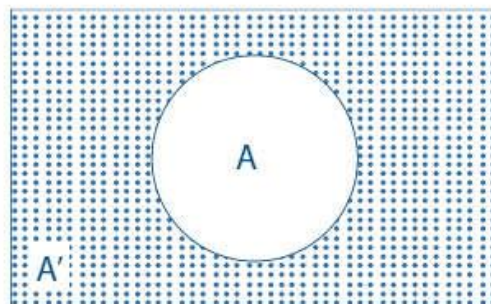


Figura 2.5

Problema resuelto

En favor de una campaña contra la obesidad, la cooperativa de una escuela secundaria ofrece jugos naturales de naranja y zanahoria. Al encuestar a 40 alumnos, 16 de ellos dijeron tomar jugo de zanahoria, 20 afirmaron tomar jugo de naranja, y 8 tomaron de los dos jugos.

En razón a las preferencias, determina:

- ¿Cuántos alumnos toman solo jugo de zanahoria?
- ¿Cuántos alumnos toman solo jugo de naranja?
- ¿Cuántos alumnos no toman ningún jugo?

Solución

Con base en el número de alumnos que dijeron tomar ambos jugos, primero se realiza la siguiente distribución:

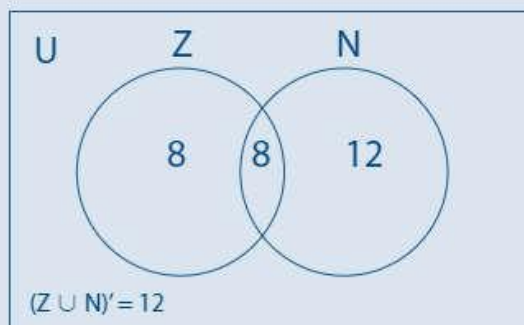


Figura 2.6

De la figura, obsérvese que los 8 alumnos que toman de los dos jugos forman parte de los conjuntos Z y N, de manera que el número de alumnos del conjunto Z es 16 ($Z = 8 + 8$). Lo mismo ocurre con el conjunto N, en donde el número de alumnos es 20 ($N = 8 + 12$). Sin embargo, al realizar el recuento de los alumnos, estos suman 28 entre los dos conjuntos, por lo que el complemento de la unión de ellos es la diferencia de los 40 alumnos, que es 12 [$(Z \cup N)' = 12$].

2.4 Probabilidad clásica

La probabilidad clásica se representa mediante una razón, la cual guarda la siguiente estructura:

$$P(A) = \frac{\text{Núm. éxitos}}{\text{Núm. total de eventos}}$$

Donde $P(A)$ se lee como:

La probabilidad (P) de que ocurra el evento 'A'.

Aquí, cabe resaltar que las probabilidades se pueden expresar en fracciones, en números decimales o en porcentajes. Pero, por lo común, el valor de la probabilidad se ubica entre 0 y 1; es decir, el valor máximo que puede tomar una probabilidad es 1.



Alerta

Las probabilidades se pueden expresar en decimales, porcentajes o fracciones dentro del rango entre 0 y 1.

Problema resuelto

Supón que se lanza un dado balanceado. Determina: ¿cuál es la probabilidad de que salga el 4?

Solución

Primero, se debe definir el evento:

a) 4

Luego, se establece el espacio muestral:

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

Por tanto, como se puede observar, existe un solo evento que cumple con lo propuesto. Esto es, solo hay un 4 en seis posible eventos.

Entonces:

$$P(A) = \frac{1}{6} = 0.1666 = 16.66\%$$

2.5 Probabilidad de eventos mutuamente exclusivos o excluyentes

La probabilidad de eventos mutuamente exclusivos o excluyentes se presenta cuando la ocurrencia de un evento no afecta la ocurrencia de otro u otros eventos.

De acuerdo con la teoría de conjuntos, estos eventos se expresan de la siguiente manera:

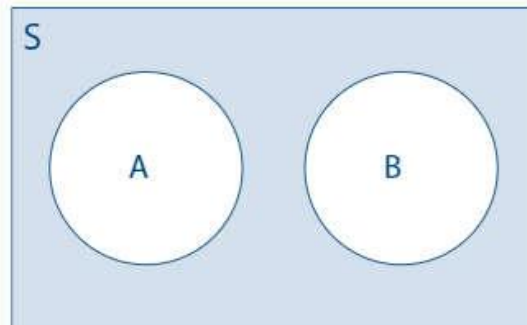


Figura 2.7

Donde el modelo matemático es el siguiente:

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

Para ejemplificar lo anterior, planteamos el siguiente problema.

Problema resuelto

Supón que se lanza un dado. ¿Cuál es la probabilidad de que se obtenga 4 o 6?

Solución

De acuerdo con lo expuesto antes, primero se definen los eventos, que en este caso son dos:

A: Que se obtenga un 4.

B: Que se obtenga un 6.

Siendo el espacio muestral:

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

Por tanto, las probabilidades de los eventos son:

$$P(A) = \frac{1}{6} \quad \text{y} \quad P(B) = \frac{1}{6}$$

En consecuencia:

$$P(A \cup B) = P(A) + P(B) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Alerta

La diferencia entre eventos mutuamente excluyentes y comunes, con respecto a su operación, solo se establece al definir los espacios muestrales correspondientes, con el fin de determinar la cantidad de eventos en común.

2.6 Probabilidad de eventos comunes

La probabilidad de eventos comunes, se refiere a la probabilidad donde parte de un evento A es parte de un evento B.

De acuerdo con la teoría de conjuntos, estos eventos quedan expresados de la siguiente manera:

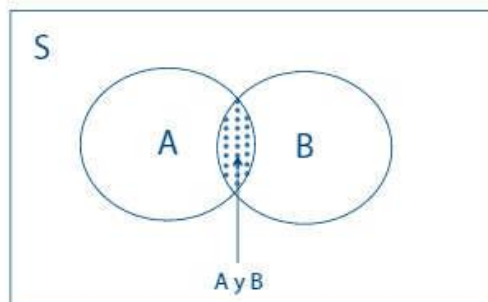


Figura 2.8

Donde el modelo matemático es el siguiente:

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

Nótese que el evento en común se encuentra tanto en A como en B, es decir está duplicado, por lo que debe restarse en una ocasión, de manera que solo sea contabilizado una sola vez.

Con el objeto de ejemplificar lo expuesto antes, se presenta el siguiente problema resuelto.

Problema resuelto

Como es sabido, una baraja cuenta con 52 cartas. Si en un juego de destreza se extrae una carta, determina: ¿cuál es la probabilidad de que esta sea una pica o un rey?

Solución

Como en los casos anteriores, primero se definen los eventos:

A: Carta de pica.

B: Carta de rey.

Luego, se definen los espacios muestrales:

$A = \{2\spadesuit, 3\spadesuit, 4\spadesuit, 5\spadesuit, 6\spadesuit, 7\spadesuit, 8\spadesuit, 9\spadesuit, 10\spadesuit, J\spadesuit, Q\spadesuit, K\spadesuit, A\spadesuit\}$

$B = \{K\clubsuit, K\heartsuit, K\spadesuit, K\diamondsuit\}$

Como se puede observar en los espacios muestrales, aquí también coexiste la carta del rey de picas ($K\spadesuit$), por lo que las probabilidades de cada evento son las siguientes:

$$P(A) = \frac{13}{52} \quad P(B) = \frac{4}{52} \quad P(A \text{ y } B) = \frac{1}{52}$$

Por tanto:

$$P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

2.7 Probabilidad de eventos simultáneos o sucesivos

Este tipo de probabilidad se refiere al hecho de que, durante el desarrollo de un experimento, los eventos suceden al mismo tiempo o uno después del otro.

El modelo matemático que define la probabilidad de eventos simultáneos o sucesivos es el siguiente:

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

A continuación se expone un planteamiento clásico en el siguiente problema resuelto, para ejemplificar lo anterior.

Problema resuelto

En un experimento de azar, se lanza una moneda dos veces seguidas. Determina: ¿cuál es la probabilidad de que la moneda caiga dos veces consecutivas en águila?

Solución

En este caso, lo primero que debe hacerse es señalar que en este experimento, en específico, cada vez que se lanza la moneda, el espacio muestral es el mismo, independientemente del resultado del evento inmediato anterior. Esto se representa de la siguiente manera:

$$S = \{\text{Sol (S), Águila (A)}\}$$

Por tanto, en el primer lanzamiento, la probabilidad de que la moneda caiga sol es:

$$P(S) = \frac{1}{2}$$

Por su parte, en el segundo lanzamiento, la probabilidad de que la moneda caiga sol es:

$$P(S) = \frac{1}{2}$$

En consecuencia, el valor de la probabilidad es:

$$P(SS) = \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) = \frac{1}{4}$$

Alerta

La probabilidad de eventos consecutivos o simultáneos se obtiene multiplicando las probabilidades de los eventos.

2.8 Los diagramas de árbol

Alerta

Los diagramas de árbol exponen de manera gráfica la secuencia de eventos, así como las probabilidades de ocurrencia de los mismos.

Una de las herramientas que se utiliza más comúnmente para el análisis de la ocurrencia de los eventos y, en su caso, para estudiar aquellos problemas que requieren del análisis de una secuencia de eventos, son los *diagramas de árbol*, debido a que por medio de estos se pueden listar y organizar todas las opciones viables acerca de un problema o condición.

Los diagramas de árbol son un despliegue gráfico que expone la secuencia de los eventos, así como las probabilidades de ocurrencia de los mismos. En general, en el diagrama de árbol es posible representar los diversos eventos que pueden suceder, como ramas que parten de una bifurcación. Por tanto, el diagrama arbóreo reúne conjuntamente una secuencia de eventos y probabilidades. Para una mejor comprensión de este tema, en la siguiente figura se presenta el modelo general de un diagrama de árbol.

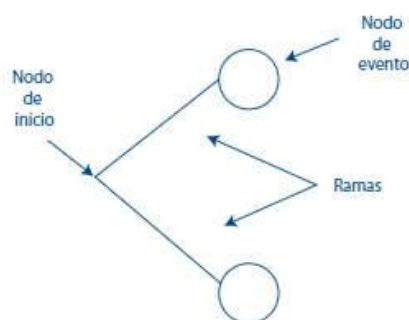


Figura 2.9

Para ejemplificar lo anterior considérese el siguiente problema resuelto.

Problema resuelto

En un experimento al azar, se lanza una moneda dos veces de manera continua. Construye el diagrama de árbol correspondiente, de tal manera que expongas el espacio muestral correspondiente.

Solución

Primero, es indispensable recordar que el espacio muestral del lanzamiento de una moneda es:

$$S = (\text{Sol}, \text{Águila})$$

Por consiguiente, después de lanzar por primera vez la moneda, se tienen dos resultados posibles: sol y águila; no obstante, al lanzar la moneda de nuevo, independientemente del resultado de la primera experiencia, cada evento tiene dos eventos posibles: sol y águila (véase figura 2.10).

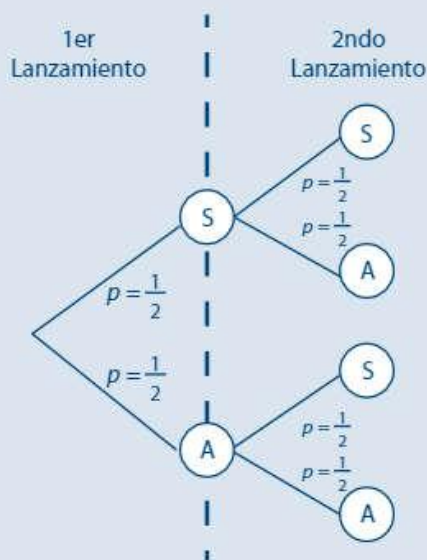


Figura 2.10

2.9 Eventos consecutivos con y sin reemplazo

El desarrollo de algunos experimentos propone que estos se puedan repetir sucesivamente bajo ciertas condiciones que determinen que el espacio muestral cambie en cuanto al número de eventos.

De manera explícita, es posible considerar que los elementos que generan los eventos de un experimento puedan ser reconsiderados para la siguiente réplica del mismo o no. Esto se conoce como condiciones con reemplazo o sin reemplazo, como se observa en el siguiente ejemplo.

Ejemplo

Experimento: Seleccionar una carta de una baraja.

Con reemplazo: La carta se vuelve a integrar al mazo.

Sin reemplazo: Una vez seleccionada la carta es retirada y, por tanto, no es considerada en las réplicas posteriores.

Con base en las condiciones expuestas, es posible señalar que las probabilidades se conservan en el caso con reemplazo, lo cual no sucede en el caso sin reemplazo, donde estas cambian.

Para ejemplificar lo anterior, se retoma el caso clásico del bote con canicas.

Alerta

Los espacios muestrales de eventos consecutivos con reemplazo permiten conservar las probabilidades de ocurrencia a lo largo de las réplicas de las experiencias.

Problema resuelto

En un bote se colocan cinco canicas (tres blancas y dos negras); inmediatamente después, se agita el bote, con el fin de sortear su contenido. Determina: ¿cuál es la probabilidad de sacar de manera consecutiva una canica negra y una canica blanca con y sin reemplazo?

Solución

El problema plantea un experimento que se puede desarrollar bajo dos condiciones. De esta manera, en primera instancia se determina la probabilidad con reemplazo (véase figura 2.11).

En este caso, luego se definen los eventos:

A: Canica negra (CN)

B: Canica blanca (CB)

Considerando que en la primera extracción se obtiene la canica negra, la probabilidad de este evento es:

$$P(CN) = \frac{2}{5}$$

Para la réplica del experimento, se devuelve la canica negra al bote y se sacude para sortear nuevamente. Considérese que en esta segunda extracción se obtiene una canica blanca; en ese caso, la probabilidad asociada a este evento es:

$$P(CB) = \frac{3}{5}$$

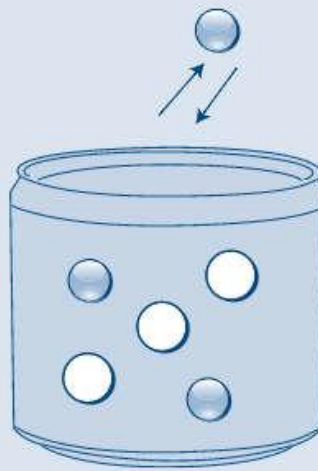


Figura 2.11

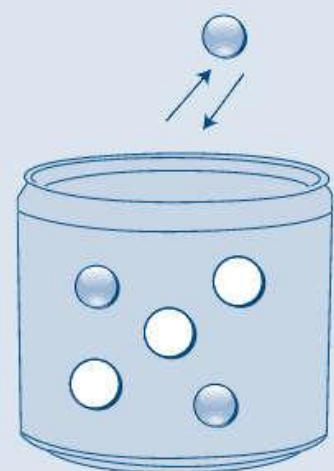


Figura 2.12

i Alerta

Los espacios muestrales de eventos consecutivos sin reemplazo son diferentes a lo largo de las réplicas de las experiencias, por lo que el valor de las probabilidades cambia.

Solución (continuación)

Por tanto, la probabilidad de estos eventos sucesivos es:

$$P(CN \cap CB) = \left(\frac{2}{5}\right)\left(\frac{3}{5}\right) = \frac{6}{25}$$

Para el caso sin reemplazo, se demuestra que las probabilidades cambian de un evento a otro, como se estudia más adelante.

2.10 Probabilidad condicional

Este tipo de probabilidad expone la ocurrencia de un evento B , una vez que ha ocurrido un evento A .

El modelo matemático que define lo anterior es el siguiente:

$$P(A \cap B) = P(A) \cdot P(B/A)$$

Donde $P(B/A)$ se lee como: "La probabilidad de B dada la ocurrencia de A ".

Para ejemplificar lo anterior, retomamos el problema del bote con canicas en el siguiente problema resuelto.

Problema resuelto

Determina el valor de la probabilidad de extraer de manera consecutiva una canica blanca (CB) y posteriormente una canica negra (CN), en riguroso orden sin reemplazo.

Solución

En primer lugar se definen los eventos:

A : canica blanca CB

B : canica negra CN

De manera que: B/A

Como se puede observar, se señala la ocurrencia de B si y solo si ha ocurrido A ; por consiguiente, las probabilidades de cada evento son:

$$P(A) = \frac{3}{5}$$

Considerando que se haya obtenido la canica blanca en la primera extracción, entonces el espacio muestral se ajusta y, en consecuencia, la probabilidad del evento condicional es:

$$P(B/A) = \frac{2}{4}$$

Por tanto:

$$P(A \cap B) = P(A) \cdot P(B/A) = \frac{3}{5} \cdot \frac{2}{4} = \frac{6}{20}$$

2.11 Teorema de Bayes

El teorema de Bayes fue desarrollado por el clérigo y matemático inglés Thomas Bayes, cuyos fundamentos fueron publicados *post mortem*. De manera concreta, el teorema que lleva su nombre expone la reformulación de un conjunto de probabilidades previas (probabilidades *a priori*) como consecuencia

de contar con información adicional acerca de los eventos del fenómeno estudiado, dando origen a nuevas probabilidades denominadas probabilidades *a posteriori*.

El modelo matemático del teorema de Bayes se fundamenta en la probabilidad condicional, como se muestra a continuación:

$$P(A/B) = \frac{P(A) \cdot P(B/A)}{P(B)}$$

De acuerdo con esta expresión matemática, se puede interpretar que el evento condicional (B/A) es parte de un evento A ; por tanto, se procede a establecer la proporción de la probabilidad que ocupa el evento B/A al interior del evento A .

Para ejemplificar lo anterior considérese el siguiente problema resuelto.

Problema resuelto

Una planta productora de gelatinas cuenta con tres máquinas empacadoras. Así, la distribución del volumen de empaque se realiza de la siguiente manera:

- Máquina 1: 38%
- Máquina 2: 32%
- Máquina 3: 30%

De esta manera, la probabilidad de que el empaque salga defectuoso es de 11%, 15% y 14%, respectivamente por cada máquina.

La gerencia de producción de la planta está interesada en conocer cuál es la probabilidad de que si se selecciona una unidad al azar y es defectuosa, esta se haya empacado en la máquina 2.

Solución

Para la resolución de este problema, primero se precisa definir los eventos. Así, de estos, el primero lo constituye la máquina y el segundo es que el producto sea defectuoso. Por tanto, el evento condicional es que sea defectuoso dado que haya sido producido por la máquina 2.

Para ejemplificar de manera gráfica lo anterior, se sugiere desarrollar el diagrama de árbol correspondiente.

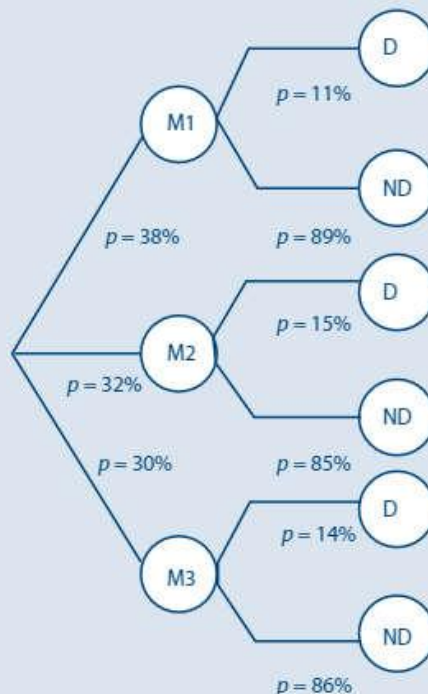


Figura 2.13

Solución (continuación)

Luego, debe considerarse que las unidades defectuosas pueden provenir de $M1$, $M2$ o $M3$, por lo que a estas se les considera eventos mutuamente exclusivos; esto es que el que una unidad haya sido producida en una máquina en específico no afecta el hecho de que provenga de otras máquinas. Entonces:

PM: Máquina

D: Unidad defectuosa

$$\therefore P(D) = P(PM1) \cdot P(D/PM1) + P(PM2) \cdot P(D/PM2) + P(PM3) \cdot P(D/PM3)$$

Tabla 2.1

Máquina	Producción (PM)	Defectuosa (D)	(PM) (D)
1	38%	11%	0.0418
2	32%	15%	0.0480
3	30%	14%	0.0420
		$P(D) = \Sigma$	0.1318

$$\therefore P(M2/D) = \frac{P(PM2) \cdot P(D/PM2)}{P(D)} = \frac{(0.32)(0.15)}{0.1318} = 0.3642 = 36.42\%$$

2.12 El principio de multiplicación

El principio de multiplicación sucede cuando, en primera instancia, "algo" se puede llevar a cabo en "a" formas, en una segunda ocasión en "b" formas y en una tercera en "c" formas, y así sucesivamente. Por tanto, las "n" formas hechas en conjunto pueden ser expresadas como:

$$a \times b \times c \times \dots \text{ (n factores) formas}$$

Lo anterior queda ejemplificado en el siguiente problema resuelto.

Problema resuelto

Una fábrica de juguetes, la cual es la responsable de producir la muñeca de moda, ha diseñado un kit de guardarropa para esta muñeca, el cual está compuesto de tres vestidos: un azul, un gris y uno negro; así como también de dos pares de zapatos: un par de color rojo y un par de color amarillo.

¿Cuántas combinaciones de ropa para esta muñeca se pueden lograr con este kit de guardarropa?

Solución

Como primer paso, se sugiere desarrollar el diagrama de árbol, con el fin de mostrar todas las opciones del guardarropa desarrollado.

Como se puede observar del diagrama de árbol, existen seis formas posibles de combinar el guardarropa de la muñeca. Así, aplicando el principio de multiplicación, se sabe que existen tres formas de seleccionar el vestido (a) y 2 formas de seleccionar los zapatos (b). Por tanto:

$$a \times b = 3 \times 2 = 6 \text{ formas}$$

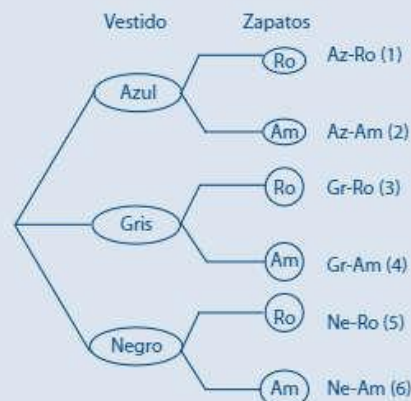


Figura 2.14

2.13 Permutaciones

Una permutación es un arreglo donde los elementos que lo integran y su orden no importan.

Considérese el conjunto {a, b, c, d}. ¿Cuántas permutaciones de tres elementos pueden obtenerse de este conjunto?

La respuesta a esta pregunta puede responderse desarrollando el diagrama de árbol correspondiente:

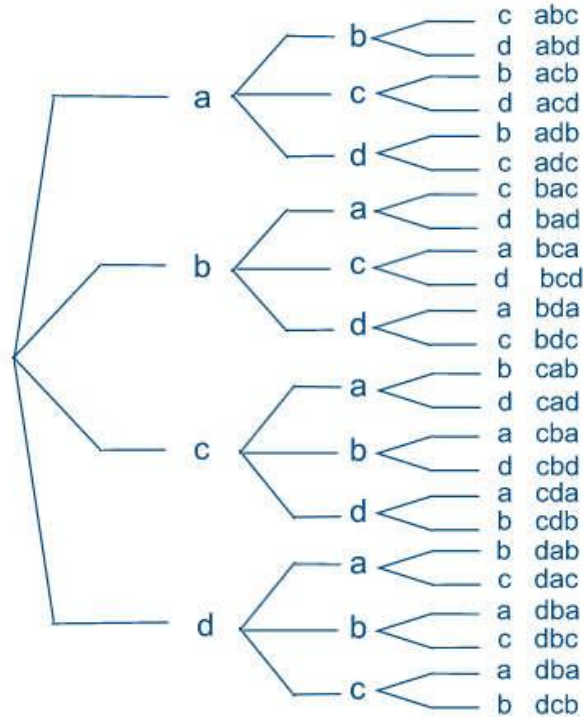


Figura 2.15

Como se puede observar del diagrama de árbol, si $k = 4$, $l = 3$ y $m = 2$; entonces:

$$k \cdot l \cdot m = (4)(3)(2) = 24 \text{ arreglos}$$

Nótese que el valor de $k = (n - 1)$ formas, $l = (n - 2)$ formas y $m = (n - 3)$ formas.

Por tanto, de manera general, se puede decir que el número de permutaciones de " n " elementos, tomados a la vez como permutaciones de un conjunto con " n " elementos, queda matemáticamente definido como:

$${}_nP_n = n!$$

Aunque, si se considera que se pueden formar arreglos de " r " elementos, tomados a la vez a partir de un conjunto de " n " elementos, se tiene que la serie referida al cálculo queda expresada como:

$$(n)(n-1)(n-2)(n-3)\dots(n-(r-1))$$

Así, la fórmula para determinar el número de permutaciones se expresa de la siguiente manera:

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

Para ejemplificar lo antes expuesto, a continuación se presenta el siguiente problema resuelto.

Alerta

Una permutación es un arreglo donde los elementos y su orden no importan.

Alerta

El cálculo del número de permutaciones se fundamenta en el principio de la multiplicación.

Considerando el planteamiento del conjunto $\{a, b, c, d\}$, comprueba que el número de permutaciones de tres elementos tomados a la vez es 24.

Primero, se aplica la fórmula para el cálculo del número de permutaciones.

Así, si $n = 4$ y $r = 3$, entonces:

$${}_4P_3 = \frac{4!}{(4-3)!} = \frac{4 \times 3 \times 2 \times 1}{1} = 24 \text{ arreglos}$$

Una combinación es un arreglo donde los elementos que lo integran y su orden sí importan.

Considérese de nueva cuenta al conjunto $\{a, b, c, d\}$. ¿Cuántas combinaciones de tres elementos pueden obtenerse del conjunto?

Con base en el diagrama de árbol desarrollado antes, se tiene:

Una combinación es un arreglo donde los elementos y su orden **sí** importan.

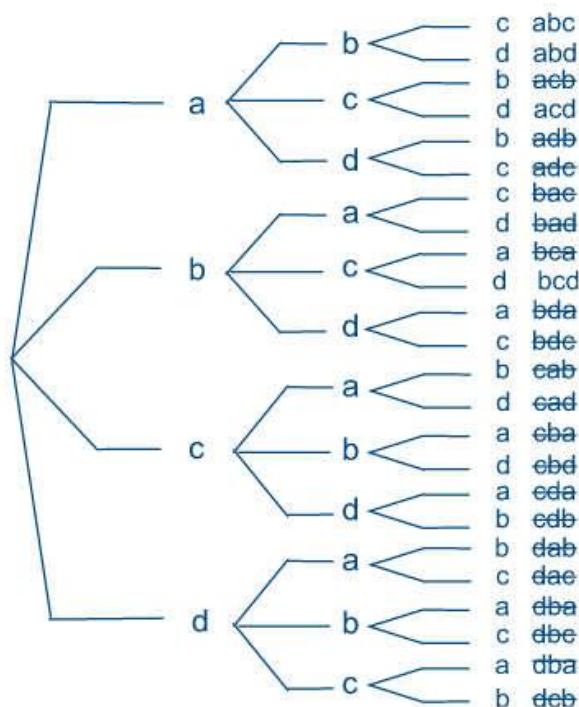


Figura 2.16

Obsérvese que el arreglo "abc" es igual a los arreglos "bac", "bca", "cab" y "acb", por lo que únicamente se considera el primer arreglo. Entonces, siguiendo con el criterio anterior, las combinaciones a considerar son cuatro: "abc", "abd", "acd" y "bcd".

Para calcular el número de combinaciones, la fórmula de las permutaciones debe ajustarse considerando el número de arreglos de " r " elementos:

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

Una combinación siempre es diferente a una permutación.

Para demostrar lo anterior, se expone el siguiente problema resuelto.

Problema resuelto

Considerando nuevamente el planteamiento del conjunto {a, b, c, d}, comprueba que el número de combinaciones de tres elementos, tomados a la vez, es cuatro.

Solución

Aplicando la fórmula para el cálculo del número de combinaciones, si $n = 4$ y $r = 3$, entonces:

$${}_4C_3 = \frac{4!}{3! \cdot (4-3)!} = \frac{4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} = 4 \text{ arreglos}$$

2.15 Variables aleatorias

En general, una variable es un hecho cuantificable relacionado con un fenómeno; en particular, una variable aleatoria constituye un hecho cuantificable asociado a un experimento, cuyos valores están determinados por el azar dentro de los números reales. En otras palabras, se puede señalar que existe una relación funcional entre los eventos que conforman el espacio muestral del experimento y los números reales.

De esta manera, las variables aleatorias se clasifican en:

- **Variables aleatorias discretas.** Si los valores asociados a los eventos de un experimento se obtienen mediante conteo, dichos valores se determinan de manera puntual dentro del terreno de los enteros. Es decir, sus posibles valores son numerables dentro de espacios finitos o infinitos.
- **Variables aleatorias continuas.** Si los valores asociados a los eventos de un experimento se obtienen, principalmente, de procesos de medición, sus valores se determinan dentro de un intervalo perteneciente a los reales.



Alerta

Las variables aleatorias son hechos medibles relacionados con un experimento y cuyo valor está determinado por el azar.

2.16 Distribuciones de probabilidad

Una distribución de probabilidad es una función matemática real, la cual permite calcular el valor de la probabilidad de que ocurra un evento propio del espacio muestral de un experimento; en otras palabras, una distribución de probabilidad permite establecer las probabilidades de que se den los diferentes valores de una variable aleatoria.

Así, de manera concreta, las distribuciones de probabilidad se definen como un espacio de probabilidad y un espacio medible, donde una aplicación entre el espacio de probabilidad y el espacio medible es una variable aleatoria, siempre y cuando esta sea una aplicación medible.

Las distribuciones de probabilidad se dividen en discretas y continuas, las cuales se analizan con detalle a continuación.



Alerta

Las variables aleatorias se clasifican en discretas y continuas.

■ Distribuciones de probabilidad discretas

En este caso, es importante tener presente que las **distribuciones de probabilidad discretas** están relacionadas con fenómenos cuyos eventos (resultados) pueden tener origen en procesos de conteo, como los que se citan a continuación:

- Número de piezas defectuosas.
- Número de retardos.
- Número de empleados con incapacidad médica.

Alerta

Las variables aleatorias discretas cuentan con una función de densidad, la cual permite calcular un valor puntual de probabilidad.

Las distribuciones de probabilidad discreta se distinguen, entre otras cosas, por:

- Contar con una función de densidad de probabilidades, la cual permite determinar el valor puntual de la probabilidad para un valor X de la variable aleatoria.
- Contar con una función de distribución, la cual permite calcular la probabilidad acumulada; esto es, que permita obtener el valor de la probabilidad de un rango de valores de la variable aleatoria.

■ Distribución binomial o de Bernoulli

Este tipo de distribución se diferencia por tener un espacio muestral basado en el éxito o el fracaso de generar un evento en un determinado número de ensayos.

Este tipo de distribución se fundamenta en el proceso de Bernoulli, el cual expone los siguientes puntos:

- Cada prueba tiene un número concreto de resultados.
- La probabilidad de éxito permanece constante en cada prueba.
- Las pruebas son independientes, por lo que se pueden acumular sus resultados.

El modelo matemático que identifica este tipo de distribución es el siguiente:

$$P(r, n, p) = \frac{n!}{r! \cdot (n-r)!} p^r \cdot q^{n-r}$$

donde

r = Número de éxitos.

n = Número de ensayos.

p = Probabilidad de éxito.

q = Probabilidad de fracaso: $q = 1 - p$.

Para ejemplificar lo anterior, considérese el siguiente problema resuelto.

Alerta

Las variables aleatorias discretas cuentan con una función de distribución, la cual permite calcular la probabilidad acumulada.

Alerta

Las variables aleatorias discretas de aplicación común son del tipo binomial y de Poisson.

Problema resuelto

En cierta región de nuestro país, se observa que las mujeres en edad fértil tienen, en promedio, tres hijos. De estos, la probabilidad de que alguno de ellos sea mujer es de 60%. Un estudio social desea determinar la probabilidad de que alguno de los tres hijos de cada una de estas mujeres sea varón.

Solución

En este caso, es posible resolver el problema por medio de dos métodos de cálculo. En primera instancia, se puede solucionar a través del árbol de decisión, en el cual se aprecian de manera explícita los eventos que cumplen con la condición de que uno de los tres hijos sea varón; es decir, eventos consecutivos e independientes. El segundo método que se utiliza para resolver el problema es por medio de la fórmula de la distribución binomial.

El árbol de decisión de este problema se observa como el que aparece a continuación:

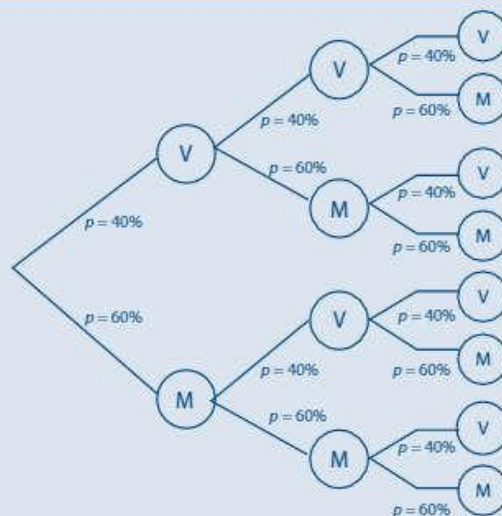


Figura 2.17

Solución (continuación)

Como se puede observar de la figura, VMM, MVM o MMV son eventos independientes uno del otro; por tanto, el valor de la probabilidad es:

$$\begin{aligned} P(VMM \cup MVM \cup MMV) &= P(VMM) + P(MVM) + P(MMV) \\ &= (0.4 \times 0.6 \times 0.6) + (0.6 \times 0.4 \times 0.6) + (0.6 \times 0.6 \times 0.4) \\ &= 0.432 = 43.2\% \end{aligned}$$

Por otra parte, aplicando la fórmula de la distribución binomial, donde:

$$\begin{aligned} r &= 1 \\ n &= 3 \\ p &= 0.4 \\ q &= 1 - 0.4 = 0.6 \end{aligned}$$

Por tanto:

$$P(1, 3, 0.4) = \frac{3!}{1!(3-1)!} (0.4)^1 \cdot (0.6)^{3-1} = 0.432 = 43.2\%$$

Ahora, para entender el concepto de función de distribución considérese el siguiente problema resuelto.

Problema resuelto

De acuerdo con el problema anterior (acerca de las mujeres en edad fértil), los responsables del estudio social requieren conocer los siguientes aspectos:

- ¿Cuál es el valor de la probabilidad de que una mujer tenga cuando menos 2 varones?
- ¿Cuál es la probabilidad de que una mujer tenga a lo más 2 varones?
- ¿Cuál es la probabilidad de que una mujer tenga entre 2 y 3 varones?

Solución

Para la resolución de este problema, primero se calculan los valores de la probabilidad de que nazcan 0, 1, 2 y 3 varones y se transcriben en una tabla:

R	$P(r, n, p)$
0	21.60%
1	43.20%
2	28.80%
3	6.40%

Acto seguido, se procede a solucionar cada uno de los incisos planteados.

- a) El término cuando menos 2, debe interpretarse como $X \geq 2$; esto es:

$$P(X \geq 2) = 1 - [P(X=0) + P(X=1)] = 1 - [21.60\% + 43.20\%] = 35.20\%$$

- b) El término a lo más 2, debe interpretarse como $X \leq 2$; esto es:

$$P(X \leq 2) = P(X=0) + P(X=1) = 21.60\% + 43.20\% = 64.80\%$$

- c) El término entre dos y tres varones significa que $2 \leq X \leq 3$; por tanto:

$$\begin{aligned} P(2 \leq X \leq 3) &= P(X \leq 3) - P(X \leq 2) = \\ &= [P(X=0) + P(X=1) + P(X=2)] - [P(X=0) + P(X=1)] = \\ &= 93.60\% - 64.80\% = 28.80\% \end{aligned}$$

Alerta

En la distribución de Poisson, los éxitos por unidad de observación son aleatorios; los eventos independientes se distribuyen de manera uniforme en la unidad de observación.

Distribución de Poisson

Este tipo de distribución de probabilidad discreta se distingue por considerar el número promedio de éxitos por unidad de tiempo, volumen o área. Su fórmula general es:

$$P(x) = \frac{\mu^x \cdot e^{-\mu}}{X!} = \frac{\lambda^x \cdot e^{-\lambda}}{X!}$$

donde

$\mu = \lambda$: Número promedio de éxitos.

X: Número de éxitos requerido.

Es importante destacar que la distribución de Poisson se fundamenta en las siguientes premisas:

- La probabilidad de que ocurra un evento por unidad de tiempo, área o volumen es pequeña.
- La probabilidad de que ocurran dos o más eventos de manera simultánea es prácticamente cero.
- El número de eventos que ocurren por unidad de tiempo, área o volumen es independiente de otras observaciones, ya que no se encuentran relacionadas; es decir, son independientes.

De hecho, de manera similar a la distribución binomial, en la distribución de Poisson se pueden determinar las probabilidades de manera puntual o acumulada.

Para ejemplificar y exponer su aplicación se presenta el siguiente problema resuelto.

Problema resuelto

Una fábrica dedicada a la fabricación de duela de madera de cedro encuentra, durante un proceso de control de calidad, que en promedio hay dos defectos en la duela por cada metro cuadrado. La gerencia de control de calidad está interesada en saber las probabilidades de obtener:

- Cuatro defectos exactamente.
- Más de tres defectos.
- A lo más tres defectos.

Solución

En primera instancia, se calculan las probabilidades de uno a cuatro defectos, aplicando la fórmula de Poisson:

X	P(X)
0	13.53%
1	27.07%
2	27.07%
3	18.04%
4	9.02%

De acuerdo con los resultados de la tabla, se procede a realizar los cálculos de cada inciso.

- Para $X = 4$ la probabilidad es:

$$P(X=4) = 9.02\%$$

- Para el caso de más de tres defectos, debe interpretarse una probabilidad cuando $P(X > 3)$:

$$P(X > 3) = 1 - [P(X=0) + P(X=1) + P(X=2) + P(X=3)] = 1 - 0.8571 = 14.29\%$$

- Para el caso de a lo más tres defectos, se debe entender que $P(X \leq 3)$. Por tanto:

$$P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) = 85.71\%$$

■ Aproximación a la distribución binomial a través de la distribución de Poisson

Toda distribución de probabilidad discreta tiene una media y una desviación estándar, las cuales se definen como:

$$\mu = np$$

y

$$\sigma = \sqrt{npq}$$

Con base en lo anterior, es posible señalar que en algunos problemas propuestos de distribución binomial se puede calcular la probabilidad ajustando la función de Poisson; aun cuando la probabilidad de éxito solo sea muy pequeña y se cuente con un número de ensayos grande.

$$P(r, n, p) = \frac{n!}{r!(n-r)!} p^r \cdot q^{n-r} = \frac{(n \cdot p)^x \cdot e^{-(n \cdot p)}}{X!}$$

Alerta

Las distribuciones de probabilidad discreta tienen una media y una desviación estándar definidas por:

$$\mu = np \text{ y } \sigma = \sqrt{npq}$$

Problema resuelto

De acuerdo con un estudio realizado con antelación, una compañía dedicada a la fabricación de equipos de cómputo encontró que de cada 20 000 discos duros, solo 19 880 discos cumplen con las especificaciones de control de calidad establecidas por la compañía. Si los lotes de embarque están compuestos por 600 unidades, la gerencia de control de calidad precisa saber si los valores de las probabilidades puntuales de hallar 0, 1, 2, 3, 4, 5 y 6 discos defectuosos en un lote son iguales entre la binomial y la aproximación de la binomial por Poisson.

Solución

$$\text{Discos no defectuosos} = \frac{19\,880}{20\,000} = 99.40\%$$

Como en otros casos, primero se determina la probabilidad de que un disco sea defectuoso:

$$\therefore P(\text{Disco defectuoso}) = 1 - P(\text{Disco no defectuoso}) = 1 - 0.9940 = 0.006 = 0.6\%$$

donde

$$p = 0.006$$

$$q = 0.994$$

$$n = 600$$

Luego, se aplica la binomial:

R	P(r,n,p)
0	2.70%
1	9.79%
2	17.70%
3	21.29%
4	19.18%
5	13.80%
6	8.26%

Sin embargo, aplicando la de Poisson, se tiene:

$$\mu = np = 600 \times 0.006 = 3.6$$

Solución (continuación)

X	$P(X)$
0	2.73%
1	9.84%
2	17.71%
3	21.25%
4	19.12%
5	13.77%
6	8.26%

Como se puede observar, las diferencias son mínimas entre los valores.

**Alerta**

Las curvas de distribución de probabilidades muestran el comportamiento, en cuanto al cambio en los valores de las probabilidades, en razón del número de éxitos.

2.17 Curva de distribución de probabilidad discreta

Los valores de la probabilidad de los eventos en análisis, a través de una distribución de probabilidad discreta, pueden exponerse de manera gráfica. De esta manera, al unir los valores que representan las probabilidades, se obtiene la curva de probabilidad correspondiente. Esta curva permite mostrar el comportamiento que sucede respecto del cambio en los valores de las probabilidades, en razón del número de éxitos.

Así, una distribución de probabilidad indica toda la gama de valores que pueden representarse como resultado de un experimento, si este se llevara a cabo.

Para mayor comprensión del tema, a continuación se expone el siguiente problema resuelto:

Problema resuelto

Con base en los datos del problema expuesto acerca de la probabilidad de nacimientos de varones en cierta región del país, desarrolla la curva de probabilidad correspondiente, sabiendo que el número mínimo y máximo de hijos varones por madre es de 0 y 3, respectivamente.

Solución

Primero, se calculan los valores de la probabilidad de que nazcan niños varones, que es de: 0, 1, 2 y 3. Enseguida, se procede a graficar estos datos, con el propósito de definir su curva de distribución.

r	$P(r, n, p)$
0	21.60%
1	43.20%
2	28.80%
3	6.40%

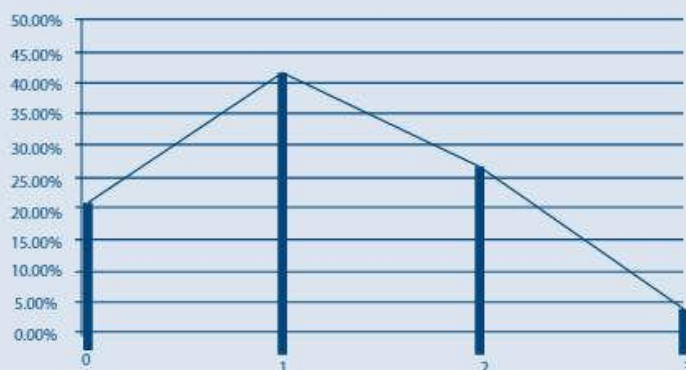


Figura 2.18

2.18 Valor esperado de una variable aleatoria discreta o esperanza matemática

El valor que puede tomar una variable aleatoria discreta depende de su probabilidad de ocurrencia; por ello, algunos valores tienen menos probabilidades de ocurrir que otros. De acuerdo con lo expuesto antes, puede surgir la inquietud de saber: ¿cuál es el valor promedio de una distribución de probabilidades? Lo anterior da origen al concepto del valor esperado de una variable aleatoria discreta o esperanza matemática o promedio ponderado, la cual se estructura como el acumulado del producto de los valores que puede tomar la variable aleatoria y su probabilidad de ocurrencia correspondiente:

$$E(X) = \sum_{i=1}^n P_i \cdot X_i$$

donde

$E(X)$: Valor esperado o esperanza matemática ($E(X) \approx \mu$).

P_i : Probabilidad de ocurrencia del evento X_i .

X_i : Valor del evento "i".

Es preciso destacar aquí que la esperanza matemática se aproxima al promedio aritmético. De esta manera, al considerar un promedio es conveniente considerar la medición del grado de dispersión con respecto a la esperanza matemática, donde también resulta importante el hecho de considerar la probabilidad de ocurrencia de cada evento, por lo que la fórmula de la desviación estándar se modifica, como se muestra a continuación:

$$\sigma_p = \sqrt{\sum_{i=1}^n P_i \cdot [X_i - E(X)]^2}$$

En algunos cálculos, la probabilidad de ocurrencia se sustituye por el grado de importancia que un valor de la variable aleatoria tiene dentro de la distribución de probabilidad; como es el caso del costo porcentual promedio de capital o cuando en algún sorteo un cupón se repite varias veces en la urna.

Para la comprensión de lo anterior, obsérvese con cuidado el siguiente problema resuelto.

Problema resuelto

El encargado de una tienda que vende al mayoreo estima la probabilidad de ocurrencia de los diferentes niveles de venta de cajas de leche en polvo durante los cinco días de la semana. ¿Cuál es el valor esperado del nivel de ventas semanal? ¿Cuál el valor de la desviación estándar ponderada?

Tabla 2.2

Día	Ventas (Cajas)	Probabilidad
Lunes	180	20%
Martes	220	25%
Miércoles	260	20%
Jueves	240	15%
Viernes	200	20%

Solución

De acuerdo con la fórmula de la esperanza matemática, en primera instancia se procede como sigue:

Tabla 2.3

Día	Ventas (X)	Probabilidad (P)	
Lunes	180	20%	36
Martes	220	25%	55
Miércoles	260	20%	52
Jueves	240	15%	36
Viernes	200	20%	40
	$\Sigma =$	100%	219

Solución (continuación)

$$\therefore E(X) = 219 \text{ cajas}$$

Donde el valor de la media aritmética es de 219 cajas, con lo cual se confirma que la esperanza matemática se aproxima a la media aritmética.

De esta manera, el valor de la desviación estándar ponderada es:

Tabla 2.4				
Día	Ventas (X)	Probabilidad (P)		$P[X - E(x)]^2$
Lunes	180	20%	36	304.20
Martes	220	25%	55	0.25
Miércoles	260	20%	52	336.20
Jueves	240	15%	36	66.15
Viernes	200	20%	40	72.20
	$\Sigma =$	100%	219	779.00

$$\sigma_p = \sqrt{779} = 27.91$$

2.19 Distribuciones de probabilidad continuas

Como se refirió antes, las distribuciones de probabilidad continuas son aquellas cuyos eventos se generan por procesos de medición y cuyos valores pertenecen a los números reales.

Las distribuciones de probabilidad continuas se relacionan con la denominada curva de *distribución normal* o *campana de Gauss*, la cual constituye una curva, cuyas principales características son las siguientes:

- Su media vale 0 y el valor de su desviación estándar es 1.
- Es simétrica alrededor de la media, por lo que la mediana y la moda tienen el mismo valor y ubicación que la media.
- Su rango en unidades estandarizadas (Z) se encuentra entre -3 y $+3$; de hecho, los puntos de inflexión de la curva se dan para $\mu \pm \sigma$.
- Su amplitud cubre $-\infty \leq x \leq +\infty$, de manera asintótica al eje x.
- El valor del área bajo la curva es 1, ya que cubre todos los posibles eventos relacionados con una variable aleatoria.

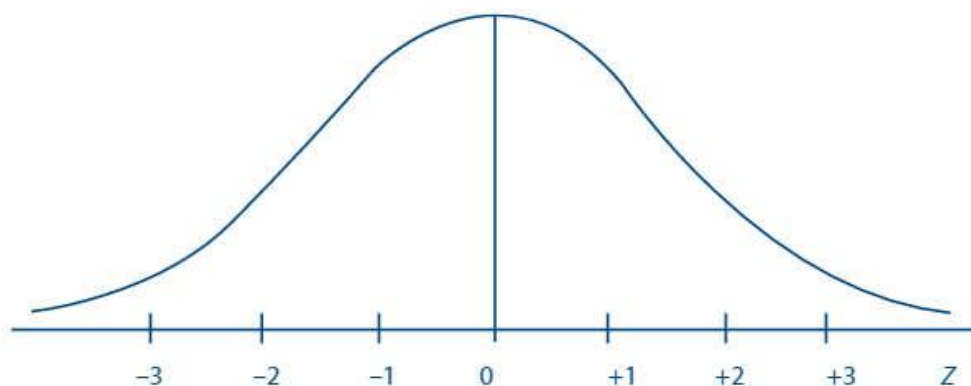


Figura 2.19

La curva de distribución normal tiene la siguiente función de densidad:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Por esta razón, el cálculo de las probabilidades corresponde a la determinación del área bajo la curva delimitada por dos valores. Donde, para mayor referencia, se considera a partir de la media ($\mu = 0$) y un valor x , haciéndose necesario integrar la función de densidad de la distribución:

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

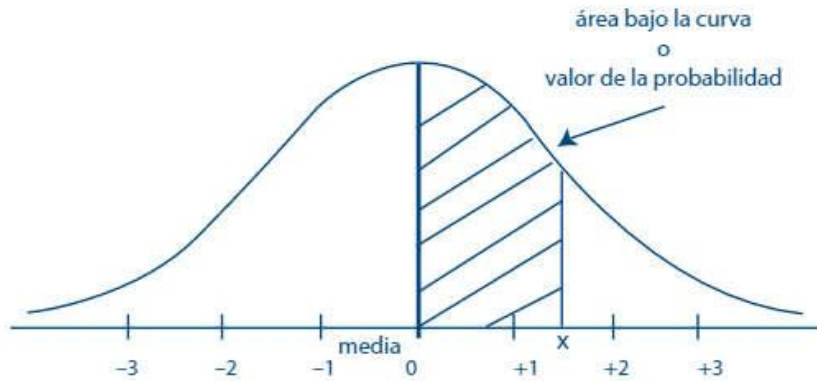


Figura 2.20

No obstante, resulta necesario hacer hincapié en que las áreas bajo la curva normal ya se encuentran tabuladas y que tan solo se requiere convertir las unidades de la variable aleatoria a unidades estandarizadas (Z), mediante de la siguiente fórmula:

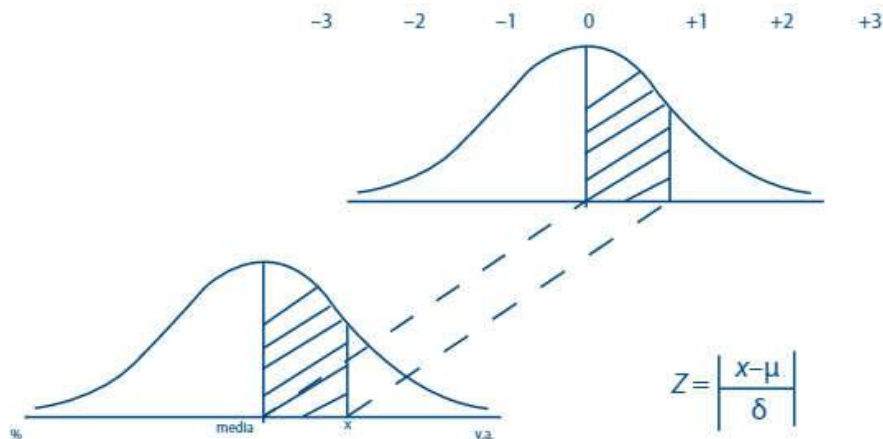


Figura 2.21

Si se desarrolla el cálculo, resulta posible apreciar las proporciones de la distribución que cubre, de acuerdo con la siguiente tabla y gráfica:

$\mu \pm \sigma$	68.20%
$\mu \pm 2\sigma$	95.40%
$\mu \pm 3\sigma$	99.70%

Alerta

La fórmula de Z expresa la proporción de σ que es $X - \mu$.

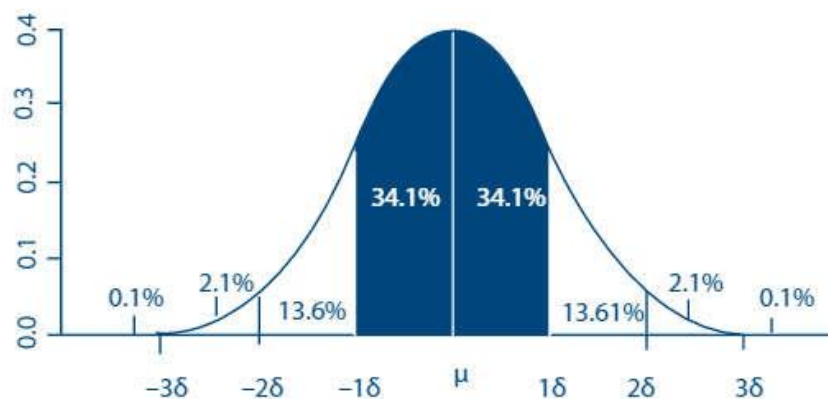


Figura 2.22

Fuente: [http://commons.wikimedia.org/wiki/File:Standard_deviation_diagram_\(decimal_comma\).svg](http://commons.wikimedia.org/wiki/File:Standard_deviation_diagram_(decimal_comma).svg)

Para ejemplificar lo antes descrito, considérese el siguiente problema resuelto.

Problema resuelto

Un almacén realiza en promedio 500 entregas a la semana, más/menos 50. El administrador del negocio desea saber la probabilidad de:

- Realizar 440 entregas a la semana.
- Realizar 550 entregas a la semana.
- Realizar entre 440 y 550 entregas a la semana.

Solución

- En este caso, primero se calcula el valor de Z . Donde:

$$\mu = 500$$

$$\sigma = 50$$

$$Z = \left| \frac{440 - 500}{50} \right| = 1.2$$

Enseguida, se ubica la probabilidad en tablas.

$$P(Z = 1.2) = 0.3849 = 38.49\%$$

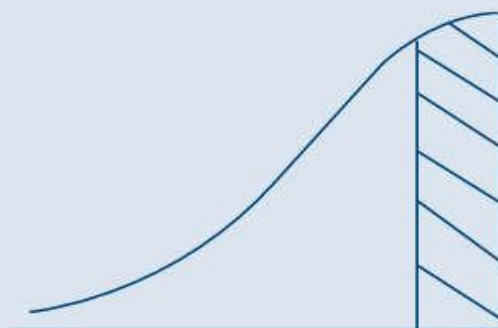


Figura 2.23

Solución (continuación)

b) En este caso, primero se calcula el valor de Z :

$$Z = \frac{|550 - 500|}{50} = 1.0$$

Luego, en tablas se ubica la probabilidad:

$$P(Z = 1.0) = 0.3413 = 34.13\%$$

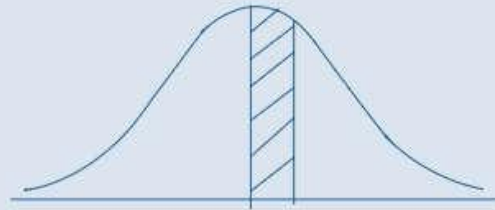


Figura 2.24

c) En este caso, la probabilidad es la suma de las áreas. Por tanto:

$$38.49\% + 34.13\% = 72.62\%$$

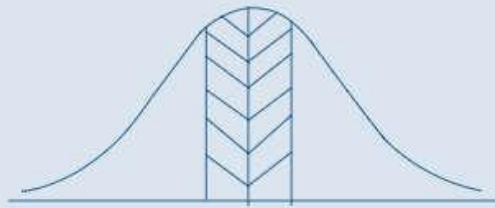


Figura 2.25

2.20 Teorema del límite central

El teorema del límite central tiene como función señalar que las variables aleatorias tienden a distribuirse como la normal; en otras palabras, señala que la distribución normal puede usarse como aproximación a otras funciones de distribución. Por ejemplo:

- La distribución binomial con parámetros n y p se aproxima a la normal para grandes valores de n y p , no demasiado cercano a 1 o 0. De tal manera que la normal aproximada tiene los siguientes parámetros: $\mu = np$, $\sigma^2 = np(1 - p)$.
- La distribución de Poisson con parámetro λ se aproxima a la normal para grandes valores de λ . Por lo que la distribución normal aproximada tiene parámetros $\mu = \sigma^2 = \lambda$.



Alerta

El teorema del límite central expone que algunas distribuciones de probabilidad tienden a la normal si se cumplen ciertos valores en sus parámetros.

2.1 Desarrolla el espacio muestral correspondiente al lanzamiento de dos dados.

2.2 Desarrolla el espacio muestral de dos monedas lanzadas simultáneamente.

SR **2.3** En un estudio sobre preferencias de lectura de diarios, se realizó una encuesta a 45 lectores, en la cual se obtuvieron los siguientes resultados:

- a) 29 señalaron leer el diario *El Planeta*.
- b) 25 dijeron leer *El Mercurio*.
- c) 30 mencionaron leer *Atlas*.

Sin embargo, a su vez, 10 lectores señalaron leer *El Planeta* y *El Mercurio*; en tanto, 20 afirmaron leer *El Planeta* y *Atlas*; mientras que 18 dijeron leer *El Mercurio* y *Atlas*; por último, 5 señalaron leer los tres.

Desarrolla el diagrama de Venn correspondiente a la información obtenida de la encuesta.

2.4 En una clase en la que todos los alumnos practican algún deporte, 60% de los estudiantes juega fútbol o baloncesto, de los cuales 10% practica ambos deportes y 40% solo juega baloncesto. Determina: ¿cuál será la probabilidad de que, escogido al azar, un alumno de la clase:

- a) juegue solo fútbol?
- b) juegue solo baloncesto?
- c) practique uno solo de los deportes?
- d) no juegue ni fútbol ni baloncesto?

2.5 Un club de lectura de estudiantes de la carrera de Ingeniería aeronáutica recibe tres revistas: *Alas* (A), *Aeronaves* (B) y *Aviones* (C). El número de miembros del club es de 100 estudiantes, cuya preferencia de lectura es la siguiente:

- a) 66 leen *Alas*.
- b) 38 leen *Aeronaves*.
- c) 59 leen *Aviones*.
- d) 17 leen *Alas* y *Aeronaves*.
- e) 43 leen *Alas* y *Aviones*.
- f) 13 leen *Aeronaves* y *Aviones*.

El tutor y el presidente del club están interesados en saber, si se selecciona a un miembro del club al azar, cuál es la probabilidad de que...

- a) Lea las tres revistas.
- b) Lea *Alas* pero no *Aviones*.
- c) Lea *Alas* o *Aviones*, pero no *Aeronaves*.
- d) Lea *Alas*, pero no *Aviones* ni *Aeronaves*.

2.6 El departamento de Ciencias Sociales de una universidad cuenta con 800 estudiantes, por lo que decidió realizar un estudio sobre el número de estudiantes que durante el actual semestre cursarán las asignaturas de Metodología de

la investigación, Administración y Estadística. A través de una encuesta, se obtuvieron los siguientes resultados:

Tabla 2.5

Metodología	490	Metodología y Administración	90
Administración	160	Metodología y Estadística	22
Estadística	320	Administración y Estadística	78

Con base en esta información y si se selecciona a un alumno al azar, calcula las siguientes probabilidades:

- a) De que estudie las tres asignaturas.
- b) De que estudie solo Estadística.
- c) De que estudie Metodología y Administración.
- d) De que estudie Administración y Estadística.

2.7 Como es sabido, un dominó cuenta con 28 fichas, ordenadas en siete fichas por cada serie de números: 0, 1, 2, 3, 4, 5, 6. Halla la probabilidad de que al levantar las fichas de dominó se obtenga:

- a) Un número de puntos mayor a 25 puntos.
- b) Un número de puntos que sea múltiplo de 6.
- c) Un número de puntos mayor a 20, pero menor a 36.

2.8 En la realización de un experimento se utilizan tres cajas. En la primera hay 3 bolas blancas y 2 bolas rojas; en la segunda caja, hay 4 bolas blancas y 3 bolas verdes, y en la tercera caja hay 5 bolas blancas y una bola amarilla. Si se extraen dos bolas de cada caja sin reemplazo, determina: ¿cuál es la probabilidad de obtener una bola roja, una bola verde y una bola amarilla?

2.9 En la realización de un experimento se utilizan 4 cajas, cada una de las cuales tiene en su interior 3 bolas azules y 5 bolas blancas. Si se extrae una bola de cada caja sin reemplazo, ¿cuál es la probabilidad de obtener dos bolas azules y 2 bolas blancas en cualquier orden?

2.10 Considerando el planteamiento del problema 2.9, determina: ¿cuál es la probabilidad de obtener 3 bolas azules y una bola blanca en cualquier orden?

2.11 Considerando el planteamiento del problema 2.9, si todas las bolas se juntan en una sola caja y se extraen 4 bolas de manera consecutiva, sin reemplazo; establece: ¿cuál es la probabilidad de obtener una bola blanca, 2 bolas azules y una bola blanca, en ese orden?

2.12 Se lanzan dos dados al aire y se suman los puntos obtenidos. Determina:

- a) La probabilidad de que el número obtenido sea 7.
- b) La probabilidad de que el número obtenido sea par.
- c) La probabilidad de que el número obtenido sea múltiplo de tres.

2.13 Para la realización de un experimento se emplean dos monedas. Con la primera de estas, la probabilidad de obtener "sol" es de $\frac{2}{3}$, mientras que con la segunda es de $\frac{5}{8}$. Determina la probabilidad de que al lanzar al aire dos monedas salgan:

- a) Dos veces sol.
- b) Dos veces águila.
- c) Una vez águila y una vez sol.

2.14 En una clase universitaria de ciencias hay inscritos 30 alumnos, de ellos 5 estudian física, 15 matemáticas y 10 biología. De estos mismos estudiantes, 22 son mujeres y el resto son hombres. Determina: ¿cuál es la probabilidad de que al escoger un estudiante al azar para pasar al pizarrón, este sea hombre y estudiante de matemáticas?

2.15 Una fábrica que produce coples para tubería de $\frac{3}{4}$ " de diámetro reúne los excedentes de su producción en una caja, para su clasificación en el almacén. De esta manera, en la caja hay 100 coples, tanto de pared gruesa como de pared delgada. De estos, 55 coples son cortos, de los cuales 25 son de pared delgada, mientras que 45 son coples largos, de los cuales 15 son de pared gruesa. Si se selecciona un cople al azar de la caja, establece cuál es la probabilidad de que este sea...

- a) Un cople corto o de pared delgada.
- b) Un cople largo de pared delgada.
- c) Un cople corto.
- d) Un cople largo o de pared gruesa.

2.16 Como es sabido, el juego del cubilete es un juego de azar que cuenta con 5 dados, normalmente marcados con palos de la baraja americana. Determina: ¿cuál es la probabilidad de obtener cuatro ases en dos tiradas consecutivas?

2.17 Una empresa dedicada al diseño de juegos de mesa, hoy día se encarga de desarrollar las reglas de uno de sus nuevos productos, proponiendo que el jugador deba lanzar dos dados, uno después del otro obligadamente. El ganador será aquel jugador que acumule más puntos, siempre y cuando la suma de los puntos sea de por lo menos 7, para que cuente la tirada. Determina lo siguiente:

- a) ¿Cuál es la probabilidad de que la suma de los dados sume al menos 7?
- b) ¿Cuál es la probabilidad de que la suma de los dados sea al menos 7, pero que en el segundo dado aparezca el número 4?
- c) ¿Cuál es la probabilidad de que la suma de los dados sea al menos 7 y que en el primer dado aparezca el número 2?

2.18 Durante la época de invierno, en un aeropuerto ubicada en la frontera norte del país, la probabilidad de que un aterrizaje se realice bajo lluvia es de 40%, que aterrice bajo nieve es de 25%, y de que aterrice bajo clima despejado es de 35%. En atención a las probabilidades climáticas, los riesgos de despiste son de: 15%, 15% y 7%, respectivamente.

Los analistas del aeropuerto están interesados en conocer las siguientes probabilidades:

- a) Que un aterrizaje se desarrolle bajo condiciones de riesgo.
- b) Que un aterrizaje se desarrolle libre de riesgo.
- c) Que un aterrizaje se desarrolle con riesgo bajo lluvia.
- d) Que un aterrizaje se desarrolle libre de riesgo bajo nieve.

2.19 En la actualidad, una cadena de restaurantes de comida rápida cuenta con cuatro sucursales en la ciudad y zona metropolitana. Debido a sus pretensiones de extensión, está interesada en captar la preferencia del consumidor con base en la calidad en el servicio. Se sabe que el nivel de concurrencia de sus clientes es de 30%, 24%, 28% y 16%, respectivamente en cada sucursal. Como resultado de su evaluación del SERVQUAL (Service Quality Evaluation) se sabe que 84%, 87%, 80% y 92% de los clientes de cada sucursal, respectivamente, están de acuerdo con el servicio. La dirección de la cadena está interesada en conocer la probabilidad de que la opinión de un cliente escogido al azar sea...

- a) Que no esté de acuerdo con el servicio y que sea cliente de la sucursal "1".
- b) Que esté de acuerdo con el servicio y que sea cliente de la sucursal "3".
- c) Que no esté de acuerdo y que sea cliente de la sucursal "4".

2.20 En una oficina, 70% de los empleados son originarios del Distrito Federal. De ellos, 50% son hombres. En tanto, de los empleados que son originarios de otras partes del país, solo 20% son hombres. Con base en estos datos:

- a) Calcula la probabilidad de que un empleado no capitalino sea mujer.
- b) Calcula la probabilidad de que un empleado de la oficina sea mujer.

2.21 En un supermercado, 65% de las compras son realizadas por mujeres. De las compras realizadas por estas, 80% supera los 200 pesos, mientras que de las compras realizadas por hombres solo 30% supera esa cantidad. Elegido un comprobante de compra al azar, determina:

- a) ¿Cuál es la probabilidad de que este supere los 200 pesos?
- b) Si se sabe que el comprobante de compra no supera los 200 pesos, ¿cuál es la probabilidad de que la compra haya sido hecha por una mujer?

2.22 El equipo directivo de una empresa perteneciente al sector hotelero está constituido por 25 personas. Del número total de empleados directivos, 60% son mujeres. De ellos, el gerente tiene que seleccionar a una persona de dicho equipo, para que represente a la empresa en un certamen internacional. Para la selección, decide lanzar una moneda al aire. Si sale sol, selecciona a una mujer y si sale águila, elige a un hom-

bre. Sabiendo que 5 mujeres y 3 hombres del equipo directivo no hablan inglés, determina:

- La probabilidad de que la persona seleccionada hable inglés.
- La probabilidad de que la persona seleccionada no hable inglés.

2.23 Una empresa fabricante de bandas para motores de equipo de construcción cuenta con cuatro plantas de producción en el país (A, B, C, D). La producción total (en las cuatro plantas) de bandas para bulldozer es de 5 000 unidades al mes. De la producción total, cada planta manufactura 1 000, 1 150, 1 500 y 1 350 unidades, respectivamente. Se sabe que el porcentaje de unidades defectuosas generadas por cada planta es de 15%, 18%, 12% y 16%, respectivamente. Considerando la información expuesta antes, la gerencia de producción está interesada en conocer, en caso de que se seleccione una banda al azar, cuál es la probabilidad de que...

- Sea manufacturada sin defectos.
- Sea manufacturada por la planta D sin defectos.
- Sea manufacturada por la planta B con defectos.

2.24 En el proceso de producción, una fábrica utiliza líneas de producción. La primera línea produce 30% de la producción total; la segunda línea produce 50% y la tercera línea contribuye con 20%. La tabla siguiente muestra las probabilidades de que se seleccionen productos sin defectos de cada una de las líneas.

Línea 1	0.8
Línea 2	0.7
Línea 3	0.9

Con base en la información anterior, el gerente de control de calidad desea conocer las siguientes probabilidades:

- Seleccionar un artículo defectuoso de la tercera línea.
- Seleccionar un artículo de calidad de la segunda línea.
- Seleccionar un artículo defectuoso de la primera línea.
- Seleccionar un artículo de calidad de cualquier línea.

2.25 Una empresa de servicios por cable determinó que la probabilidad de contratación de los servicios en una zona de clase media de reciente creación es la siguiente:

Cliente	Probabilidad de contratación
Casa	65%
Departamento	35%

De acuerdo con los resultados del estudio, estima la probabilidad de contratación de los servicios, como se presentan a continuación:

Servicio	Probabilidad de servicio
Teléfono	50%
TV	30%
Internet	20%

De acuerdo con los datos anteriores, determina lo siguiente:

- La probabilidad de que se contraten servicios de telefonía.
- La probabilidad de que un contrato de Internet sea de un departamento.
- La probabilidad de que un contrato de televisión sea de una casa.
- La probabilidad de que un contrato sea de telefonía, televisión e Internet.

2.26 La compañía de juguetes responsable de la fabricación de la muñeca de moda, hace poco tiempo lanzó al mercado otro kit de guardarropa. Este incluye: cuatro vestidos (azul, amarillo, rojo y blanco), tres pares de zapatos (rojo, blanco y negro), tres bolsas (blanca, roja y rosa). Desarrolla el árbol de decisión, con el fin de determinar las diversas formas de combinación del guardarropa.

2.27 Una cadena de hamburguesas tomó la decisión de rediseñar la composición de sus productos. En esta nueva etapa, el cliente puede escoger tres de los siguientes ingredientes extras:

- Tocino
- Queso
- Salsa especial
- Lechuga y pepinillos
- Aderezo agrícolce

Con base en estos datos, desarrolla el árbol de decisión, para determinar el número de formas en que se pueden combinar los ingredientes extras.

2.28 El estudio demográfico de un país arrojó los siguientes datos:

- Tres de cada siete nacimientos son varones.
- En promedio, las parejas procrean cuatro hijos.

Desarrolla el árbol de decisión, para determinar las posibles combinaciones de los hijos de cada pareja.

2.29 Con base en los datos del problema 2.26, aplica el principio de la multiplicación para confirmar el número de formas posibles de combinar el nuevo guardarropa de la muñeca de moda.

2.30 De acuerdo con los datos del problema 2.27, la gerencia de servicio está interesada en determinar el número de combinaciones posibles de hamburguesas que los clientes pueden formar.

2.31 Una agencia de automóviles desarrolla una estrategia de promoción para vender todos los autos modelo 2010, que aún se hallan en la agencia, la cual consiste en ofrecer un paquete que consiste en escoger tres accesorios, de cuatro posibles, a un precio especial, los cuales son:

Accesorio	Descripción
A	Rines deportivos
B	Manos libres para celular en el tablero
C	Quemacocos
D	Equipo de sonido para iPod integrado

El gerente de servicio desea saber cuántas y cuáles serían las combinaciones de los paquetes. Determina todas las probabilidades.

2.32 Una fábrica de alarmas para autos establece que la probabilidad de que una de sus alarmas falle es de 0.005. Si cada uno de los lotes de producción consta de 2000 unidades, la gerencia de calidad está interesada en saber las siguientes probabilidades:

- De que fallen exactamente 10 unidades.
- De que fallen más de 10 unidades.
- De que falle 5% de las unidades.
- De que falle 1% de las unidades.

2.33 Con respecto al estudio demográfico, referido en el problema 2.28, los responsables del mismo están interesados en conocer las siguientes probabilidades:

- De que conciban un varón.
- De que conciban cuatro varones.
- De que no conciban ningún varón.
- De que conciban dos varones y dos mujeres.

2.34 Durante la realización del examen para obtener la licencia de manejo, se conoció que 3 de cada 10 candidatos son rechazados la primera vez que realizan dicho examen. Determina la distribución de probabilidad de que se acepten 5, 6, 7, 8, 9, 10 candidatos.

2.35 La probabilidad de que un artículo producido por una fábrica no sea defectuoso es de 99.80%. Si se envía un cargamento de 10000 artículos a diversos almacenes, hallar la probabilidad de que...

- en el embarque vayan 10 artículos defectuosos.
- en el embarque vayan no más de siete artículos defectuosos.
- en el embarque vayan entre cinco y 10 artículos defectuosos.

2.36 Una fábrica de equipos de cómputo dio a conocer que de cada 20000 discos duros que produce, 19 880 discos cumplen con las especificaciones de control de calidad. Si los lotes de embarque están compuestos por 600 unidades, la gerencia de control de calidad quiere conocer las siguientes probabilidades:

- De que en un lote vayan no más de 6 discos duros defectuosos.
- De que en un lote vayan más de 6 discos duros defectuosos.

c) De que en un lote vayan más de 6, pero menos de 10 discos duros defectuosos.

2.37 Las estadísticas de salud más recientes de nuestro país, afirman que en la zona sur-oriente de la Ciudad de México se presenta una alta incidencia de malestares estomacales, provocados por la falta de higiene (esto es, 200 casos por cada 10000 habitantes). Supón que también se realizan exámenes a 1000 habitantes de un municipio conurbado y se asume que para estos la tasa de incidencia es la misma que para toda la región sur-oriente de la Ciudad de México. Determina:

- ¿Cuál es la probabilidad de que ninguna de las personas examinadas padezca alguna infección estomacal?
- ¿Cuál es la probabilidad de que al menos tres personas padezcan alguna infección estomacal?
- ¿Cuál es la probabilidad de que al menos ocho personas padezcan alguna infección estomacal?

2.38 Una empresa dedicada a la fabricación de focos incandescentes, reporta que 7% de la producción es defectuosa. Halla la probabilidad de que en una muestra de 270 bombillas suceda que...

- más de cinco focos sean defectuosos.
- entre uno y tres focos sean defectuosos.
- dos focos o menos sean defectuosos.

2.39 Después de un estudio realizado por una institución bancaria, se determinó que existe una probabilidad de 40% de que los clientes que cuentan con tarjeta de crédito paguen antes de su fecha de corte. Para revisar la vigencia de esta condición, el gerente de crédito selecciona cinco cuentas, con el fin de conocer: ¿cuál es la probabilidad de que se presenten cargos por intereses en 0, 1, 2, 3, 4 y 5 cuentas? Con base en ese objetivo desarrolla la curva de distribución discreta correspondiente.

2.40 En la inspección de hojalata producida por un proceso electrolítico continuo, se identifican 0.2 imperfecciones en promedio por minuto. Determina las probabilidades de identificar:

- Una imperfección en tres minutos.
- Dos o menos imperfecciones en cinco minutos.
- Una imperfección o menos en 15 minutos.

2.41 De acuerdo con el planteamiento del problema 2.36, resuelve este por aproximación de Poisson a la binomial.

2.42 Considerando el planteamiento del problema 2.37, resuelve por aproximación de Poisson a la binomial.

2.43 De acuerdo con la información expuesta en el problema 2.34, acerca del examen de manejo, se sabe que los examinadores aplican el examen a 2.5 candidatos por hora en promedio. Si la jornada de trabajo de cada empleado es de 8 horas, ¿cuál es la probabilidad de que durante una jornada de trabajo se examinen 20, 21, 22, 23, 24, 25 candidatos?

SR

2.44 Una empresa dedicada a la fabricación de placas para pisos de madera, encontró que después del proceso de barnizado existe 0.2% de probabilidad de hallar un defecto por cada metro cuadrado, extensión donde el límite de imperfecciones máximo es de tres por metro cuadrado. Ante esta situación, la gerencia de control de calidad desea conocer lo siguiente:

- La probabilidad de encontrar 3 defectos por metro cuadrado.
- La probabilidad de encontrar no más de 3 defectos por metro cuadrado.
- La probabilidad de encontrar más de 2 defectos por metro cuadrado.

2.45 La probabilidad de acertar a un blanco en movimiento con una nueva metralleta montada en un auto blindado maniobrando sobre terreno escabroso es de 0.004% en cada disparo. Si se realizan 3 000 disparos y se considera un análisis por distribución binomial, ¿cuál es la probabilidad de acertar al menos cuatro disparos?

2.46 La probabilidad de que un individuo sufra una reacción negativa como consecuencia de la aplicación de un suero es de 0.001%. A últimas fechas, el suero ha sido aplicado a 2 000 personas. Determinar mediante la distribución de Bernoulli lo siguiente:

- La probabilidad de que se presente la reacción exactamente en tres personas.
- La probabilidad de que se presente la reacción en al menos tres personas.
- La probabilidad de que se presente la reacción en más de tres personas.

2.47 Con el objetivo de solucionar en la medida de lo posible el problema de obesidad entre los estudiantes de una preparatoria, la dirección de servicios médicos del plantel

encontró que la media de los pesos de 500 estudiantes fue de 68.5 kilogramos, considerando una tolerancia de 7 kilogramos. La dirección de servicios médicos desea conocer lo siguiente:

- La probabilidad de que los estudiantes tengan un peso entre 54.5 y 70.0 kilogramos.
- La probabilidad de hallar estudiantes que tengan un peso por arriba de los 84 kilogramos.
- La probabilidad de hallar estudiantes que pesen menos de 58 kilogramos.
- La probabilidad de ubicar estudiantes que pesen entre 54.5 y 58 kilogramos, ya que es el rango que corresponde a su edad y desarrollo físico.

2.48 Un proyecto de ingeniería tiene un presupuesto promedio mensual de 8 300 dólares, con una variación de 764 dólares. De acuerdo con estos datos, establecer cuál es la probabilidad de que el presupuesto...

- Se encuentre entre 8 300 dólares, más-menos 780 dólares.
- Se encuentre entre 9 000 y 11 000 dólares.
- Sea menor a 6 873 dólares.

2.49 De acuerdo con un estudio que contempló la revisión de las cuentas de ahorro populares, se estableció que el monto promedio mensual de cada cuenta es de 830 pesos, presentando una variación de 76.40 pesos. Determina: ¿cuál es la probabilidad de que el monto de una cuenta de ahorro se ubique entre...

- 830 pesos más-menos 80 pesos?
- 900 y 1100 pesos?
- sea menor a 700 pesos?

2.50 Considerando la función de densidad de la distribución normal, grafica en Excel $\mu = 500$ y $\sigma = 460$ en un rango entre -500 y 2 500.

SR



PROBLEMAS RETO

Desarrolla el espacio muestral del lanzamiento de tres dados, de tal manera que se establezcan las siguientes probabilidades:

- Que la suma de puntos sea mayor a 9.
 - Que la suma de puntos sea entre 15 y 18.
 - Que la suma de puntos sea exactamente 12.

- Considerando la función de densidad de la distribución normal, realiza la tabla de áreas bajo la curva comprendida entre $Z = 0$ y $Z = 3$, con una precisión de 0.01 unidades.



REFERENCIAS

Levin, Richard I. y Charles A. Kirkpatrick (1989). *Enfoques cuantitativos a la administración*. México: Continental y McGraw-Hill.

Lind, Douglas A., William G. Marchal y Samuel A. Wathen (2005). *Estadística aplicada a los negocios y a la economía* (12a. ed.). México: McGraw-Hill.

Maris, Diez Stella (2005). *Estadística aplicada a los negocios utilizando Microsoft EXCEL*. Argentina: MP Ediciones.

Quevedo, Urias Héctor y Blanca Rosa Pérez Salvador (2008). *Estadística para Ingeniería y Ciencias*. México: Grupo Editorial Patria.



DIRECCIONES ELECTRÓNICAS

Problemas resueltos de probabilidad (nivel básico)

[http://www.vitutor.com/pro/2/a_g.html]

Problemas de probabilidad resueltos y propuestos

[<http://webdelprofesor.ula.ve/ciencias/jlchacon/materias/discreta/probpro.pdf>]

Simulador de la distribución binomial

[<http://virtual.uptc.edu.co/ova/estadistica/applets/simulacionbino/Simulacion.htm>]

Portal académico CCH-distribuciones de probabilidad

[<http://portalacademico.cch.unam.mx/alumno/sitiosdeinteres/matematicas/estadistica2>]

Distribución normal

[http://es.wikipedia.org/wiki/Distribuci%C3%B3n_normal]



Estadística inferencial

OBJETIVOS

- Comprender la importancia de la estadística inferencial.
- Entender los conceptos de población y muestra, así como establecer la diferencia entre ellos.
- Distinguir la diferencia entre muestras probabilísticas y no probabilísticas.
- Entender los conceptos de estadístico y parámetro, así como la diferencia entre ellos.
- Deducir las fórmulas para determinación del tamaño muestral.
- Comprender el concepto de error muestral.
- Comprender el concepto de error estándar.
- Entender el concepto de estimador.
- Distinguir la diferencia entre un estimador sesgado y uno insesgado.
- Entender el concepto de nivel de confianza.
- Comprender la trascendencia del tamaño muestral en relación con el análisis inferencial.
- Entender el concepto de grados de libertad.
- Comprender las características de la distribución t de Student.

¿QUÉ SABES?

- ¿Cuál es la trascendencia de la estadística inferencial?
- ¿Cuál es la diferencia entre una población y una muestra?
- ¿Cuál es la diferencia entre las muestras probabilísticas y no probabilísticas?
- ¿Cuál es la diferencia entre un estadístico y un parámetro?
- ¿Cómo se puede determinar el tamaño muestral?
- ¿Qué es el error muestral?
- ¿Qué es el error estándar?
- ¿Qué es un estimador?
- ¿Cuál es la diferencia entre un estimador sesgado y uno insesgado?
- ¿Qué es el nivel de confianza?
- ¿Qué importancia tiene el tamaño muestral para el desarrollo del análisis inferencial?
- ¿Qué son los grados de libertad?
- ¿Cuáles son las principales características de la distribución t de Student?

3.1 Introducción

Es común que en la vida diaria se hable de muestras y de resultados de estudios estadísticos relativos a opiniones sobre un tema o referentes a las características de un producto o servicio, por lo que es de interés exponer la teoría y los conceptos básicos que permiten estructurar los procesos de análisis correspondientes.

Alerta

La estadística inferencial expone el estudio del comportamiento de un fenómeno que ocurre en una población a través del estudio y análisis de muestras.

3.2 Concepto y propósito de la estadística inferencial

La estadística inferencial expone el estudio del comportamiento de un fenómeno que ocurre en una población a través del estudio y análisis de muestras; de manera práctica, es establecer las dimensiones estadísticas de una población (parámetros) a través de las dimensiones estadísticas de las muestras (estadísticos).

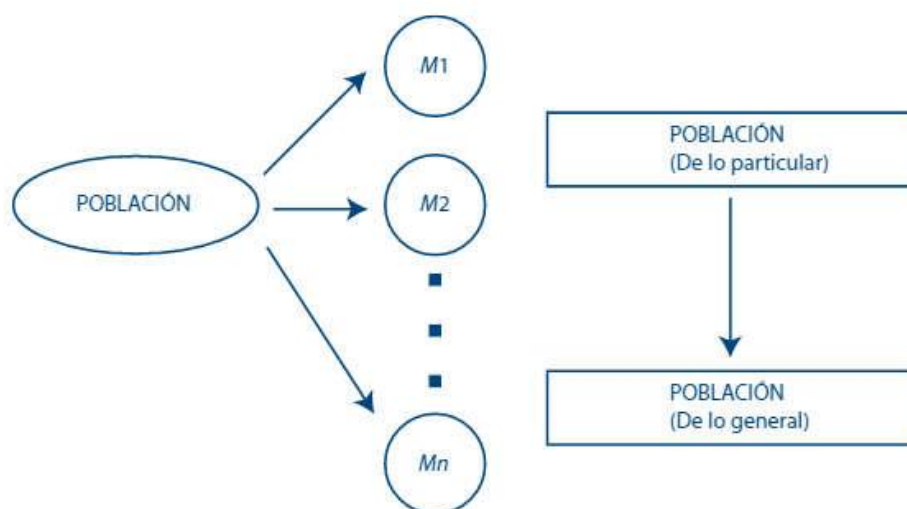


Figura 3.1

Alerta

Las dimensiones estadísticas de una muestra se denominan estadísticos, mientras que en la población se denominan parámetros.

Las dimensiones estadísticas son las mismas en cuanto a su concepto; sin embargo, operacional y simbólicamente se distinguen como se muestra en la tabla 3.1.

Tabla 3.1 Dimensiones estadísticas

Dimensión estadística	Población (parámetros)	Muestra (estadísticos)
Media	μ	\bar{x}
Desviación estándar	σ	S
Número de elementos	N	n

■ Concepto de población

Se define **población** como el conjunto de personas, objetos o entidades que cumplen con ciertas características definidas por el observador.

Las poblaciones se pueden clasificar como:

Finitas, cuando el número de elementos es contabilizable y ubicable.

Infinitas, cuando el número de elementos es tan grande que se dificulta su contabilización pese a ser ubicables.

Difusas, cuando el número de elementos es difícilmente contabilizable debido a que la ubicación de los elementos no se puede establecer.

Alerta

Una población es el conjunto de personas, objetos o entidades que cumplen con ciertas características definidas por el observador.

■ Concepto de muestra

Se define como muestra al **conjunto** de elementos representativo de una población.

Las muestras se pueden clasificar en:

- I. **Probabilísticas**, donde el proceso de selección de los elementos que la estructuran se basa en las reglas de la probabilidad. Entre las muestras probabilísticas se pueden citar las siguientes:
 - **Simples**. Son aquellas donde los elementos de una población son etiquetados numéricamente de manera que puedan ser seleccionados en forma aleatoria, y pueden ser con o sin reemplazo.
 - **Sistemáticas**. Son aquellas donde los elementos de la población son etiquetados numéricamente y catalogados por intervalos, seleccionándose en forma aleatoria de cada intervalo y repitiéndose el procedimiento hasta completar la muestra.
 - **Estratificadas**. Son aquellas donde los elementos de la población se clasifican en subgrupos o estratos de acuerdo con una característica definida, procediendo a realizar un muestreo simple de cada estrato, de manera que la muestra contenga elementos de todos los estratos.
 - **De conglomerado**. Son aquellas donde los elementos de la población se catalogan en subgrupos de acuerdo con una característica de orden general, como lo es una clasificación por género, colonia de residencia o lugar de nacimiento. Así, una vez formados los conglomerados, estos deben ser seleccionados en forma aleatoria, de manera que se elijan tantos elementos de cada uno hasta que se complete la muestra.
- II. **No probabilísticas**, donde los elementos de una población no cuentan con una probabilidad de ser seleccionados para estructurar una muestra. Dentro de las muestras no probabilísticas se pueden citar las siguientes:
 - **De conveniencia**. Son aquellas donde los elementos son seleccionados por cumplir con ciertas condiciones o características que faciliten el análisis del fenómeno; inclusive pueden estructurarse por voluntarios.
 - **De criterio**. Son aquellas donde los elementos son seleccionados con base en un criterio o juicio del observador.
 - **Por cuotas**. Son aquellas donde la población se cataloga en estratos, y se procede a seleccionar tantos elementos de cada uno hasta que se complete la muestra.
 - **De bola de nieve**. Son aquellas donde se selecciona un cierto número de elementos a los cuales se les solicita ofrezcan referencias sobre otros elementos de la población, y se repite el procedimiento hasta completar la muestra o satisfacer algún requisito en el número de elementos.

Alerta

Una muestra es el conjunto de elementos representativo de una población, y puede ser probabilística o no probabilística.

3.3 Estimadores

Como se ha expuesto, el propósito del análisis inferencial es determinar el valor de los parámetros de la población a través de los estadísticos de las muestras. Por consiguiente, se puede señalar que cuando el valor del estadístico coincide con el valor del parámetro se dice que el estadístico es un estimador insesgado, o sea que no existe diferencia entre el valor del parámetro y el estadístico; pero cuando no coincide, y existe diferencia, se dice que el estadístico es un estimador sesgado.

De hecho, se puede señalar que si se pudieran analizar todas las muestras, consistentes y válidas, de una población, se pueden estimar sus parámetros.

Para demostrar lo expuesto considérese el siguiente problema.

Problema resuelto

A partir de la población,

$$P = \{10, 17, 24, 32\}$$

Determina todas las muestras, consistentes y válidas, de tres elementos a efecto de calcular y comparar sus estadísticos con los parámetros de la población.

Alerta

Un estadístico es un estimador insesgado cuando no existe diferencia con el valor del parámetro y el estadístico; y cuando existe diferencia es un estimador sesgado.

Alerta

Si pudieran analizarse todas las muestras disponibles de una población, se estaría analizando a la población misma.

Solución

En primer lugar se determina el número de muestras, de manera que estas no se repitan (consistentes), por lo que se aplica la fórmula de las combinaciones:

$${}_4C_3 = \frac{4!}{3!(4-3)!} = 4$$

Por lo que, desarrollando los arreglos, las muestras resultantes son las siguientes:

$$m_1 = 10, 17, 24$$

$$m_2 = 10, 17, 32$$

$$m_3 = 10, 24, 32$$

$$m_4 = 17, 32, 24$$

Se procede a calcular el valor de la media y varianza de cada muestra.

$$m_1 \quad \bar{X} = 17.00$$

x	$(x - \bar{X})^2$
10	49.00
17	0.00
24	49.00
$\Sigma =$	98.00

$$S_1^2 = \frac{\sum_{i=1}^n (x_i - \bar{X}_1)^2}{n} = \frac{98.00}{3} = 32.66$$

$$m_2 \quad \bar{X}_2 = 19.67$$

x	$(x - \bar{X}_2)^2$
10	93.3156
17	7.0756
32	152.2756
$\Sigma =$	252.67

$$S_2^2 = \frac{252.67}{3} = 84.22$$

$$m_3 \quad \bar{X}_3 = 22.00$$

x	$(x - \bar{X}_3)^2$
10	144.00
24	4.00
32	100.00
$\Sigma =$	248.00

$$S_3^2 = \frac{248.00}{3} = 82.66$$

$$m_4 \quad \bar{X}_4 = 24.33$$

x	$(x - \bar{X}_4)^2$
17	53.72
32	58.82
24	0.108
$\Sigma =$	112.64

$$S_4^2 = \frac{112.64}{3} = 37.54$$

Se obtiene el valor de la media y la variación en la población.

$$\mu = 20.75$$

x	$(x - \mu)^2$
10	115.5625
17	14.0625
24	10.5625
32	126.5625
$\Sigma =$	266.75

$$\sigma^2 = \frac{266.75}{4} = 66.68$$

Solución (continuación)

El valor representativo de las medias muestrales es el promedio de las mismas, por lo que se calcula la media de medias y se compara con la media de la población.

$$\begin{aligned}\bar{\bar{X}} &= \frac{17 + 19.67 + 22 + 24.33}{4} = 20.75 \\ \mu &= 20.75 \\ \bar{\bar{X}} &= \mu\end{aligned}$$

Por tanto, se puede concluir que la media de medias es igual a la media poblacional, o sea, la media de las medias es el estimador insesgado de la media poblacional.

De igual forma se obtiene el promedio de las varianzas muestrales.

$$\bar{S}^2 = \frac{32.66 + 84.22 + 82.66 + 37.54}{4} = 59.27$$

Por lo que se puede apreciar que, a diferencia de los promedios de las medias, el valor promedio de las varianzas muestrales es diferente a la varianza poblacional.

$$\sigma^2 \neq \bar{S}^2$$

La pregunta lógica ante este resultado es: ¿por qué ocurre esto? Lo anterior se explica en razón de que tanto la población como las muestras como conjuntos pueden dar origen a otros tantos subconjuntos, pero de hecho no pueden variar en sí mismos, por lo que las varianzas deben calcularse considerando $N - 1$ y $n - 1$.

Por tanto, se calculan las varianzas muestrales y poblacional, y se procede a comparar el valor promedio de las varianzas muestrales con la varianza poblacional.

$$\begin{aligned}\hat{S}^2 &= \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n - 1} \\ \hat{S}_1^2 &= 49.00 \quad \hat{S}_2^2 = 126.34 \quad \hat{S}_3^2 = 124.00 \quad \hat{S}_4^2 = 56.32 \\ \hat{S}_{Prom}^2 &= \frac{49.00 + 126.34 + 124.00 + 56.32}{4} = 88.91 \\ \sigma^2 &= \frac{\sum_{i=1}^N (x_i - \mu)^2}{N - 1} = 88.91\end{aligned}$$

Por tanto se concluye que \hat{S}_{Prom}^2 es un estimador insesgado de σ^2 .

3.4 Distribución de las medias muestrales

Si se pudiera contar con todas las muestras, consistentes y válidas, de una población, se podría observar que las medias muestrales también cuentan con una distribución de frecuencias; esto es, se puede observar cuántas medias se repiten.

Para demostrarlo, considérese el siguiente problema.

Problema resuelto

A partir de la población que se expone construye la distribución de las medias muestrales a partir de muestras de 4 elementos:

$$P = \{3, 1, 2, 4, 3, 5\}$$

Solución

Con base en el número de elementos de la población ($n = 6$) se determina el número de muestras de 4 elementos:

$$P = \{3, 1, 2, 4, 3, 5\}$$

$${}_6C_4 = \frac{6!}{4!(6-4)!} = 15$$

Para determinar las muestras se procede a etiquetar cada uno de los elementos de la población:

3	1	2	4	3	5
a	b	c	d	e	f

Por lo que las muestras correspondientes son las siguientes:

						\bar{X}
m_1	a	b	c	d	3 1 2 4	2.50
m_2	a	b	c	e	3 1 2 3	2.25
m_3	a	b	c	f	3 1 2 5	2.75
m_4	a	b	d	e	3 1 4 3	2.75
m_5	a	b	d	f	3 1 4 5	3.25
m_6	a	b	e	f	3 1 3 5	3.00
m_7	a	c	d	e	3 2 4 3	3.00
m_8	a	c	d	f	3 2 4 5	3.50
m_9	a	c	e	f	3 2 3 5	3.25
m_{10}	a	d	e	f	3 4 3 5	3.75
m_{11}	b	c	d	e	1 2 4 3	2.50
m_{12}	b	c	d	f	1 2 4 5	3.00
m_{13}	b	c	e	f	1 2 3 5	2.75
m_{14}	b	d	e	f	1 4 3 5	3.25
m_{15}	c	d	e	f	2 4 3 5	3.50

Donde el valor de la media de medias es: $\bar{\bar{X}} = 3.00$.

Asimismo, el valor de la media poblacional es: $\mu = 3.00$.

Se procede a graficar las frecuencias de las medias de acuerdo con la siguiente tabla:

\bar{X}	f
2.25	1
2.50	2
2.75	3
3.00	3
3.25	3
3.50	2
3.75	1
4.00	0
4.25	0

Solución (continuación)

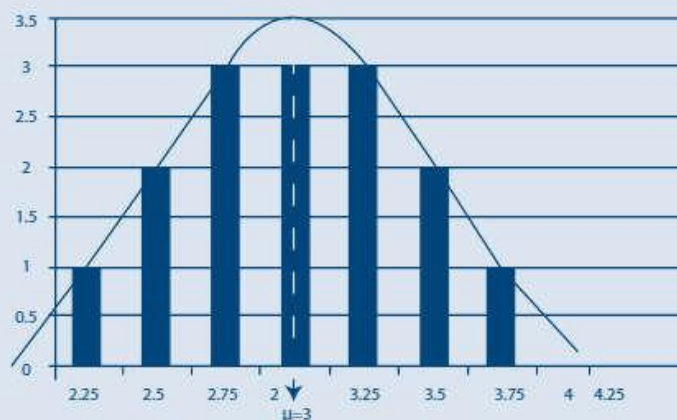


Figura 3.2

Observa la localización de la media poblacional, la cual se ubica en el centro de la distribución.

Para complementar lo expuesto se plantea resolver el siguiente problema con base en las condiciones que se exponen.

Problema resuelto

A partir de la población del problema anterior, construye la distribución de las medias muestrales de 5 elementos.

Solución

A partir del número de elementos de la población ($n = 6$) se determina el número de muestras de 5 elementos:

$${}_6C_5 = \frac{6!}{5!(6-5)!} = 6$$

Por lo que las muestras correspondientes son las siguientes:

											\bar{X}
m_1	a	b	c	d	e	3	1	2	4	3	2.60
m_2	a	b	c	d	f	3	1	2	4	5	3.00
m_3	a	b	c	e	f	3	1	2	3	5	2.80
m_4	a	b	d	e	f	3	1	4	3	5	3.20
m_5	a	c	d	e	f	3	2	4	3	5	3.40
m_6	b	c	d	e	f	1	2	4	3	5	3.00

Donde el valor de la media de medias es: $\bar{\bar{X}} = 3.00$.

Asimismo, el valor de la media poblacional es: $\mu = 3.00$.

Se procede a graficar las frecuencias de las medias de acuerdo con la siguiente tabla:



Alerta

Las medias muestrales tienden a distribuirse normalmente, donde la tipificación de la curva depende del número de muestras en razón de su tamaño.

Solución (continuación)

\bar{x}	f
2.60	1
2.80	1
3.00	2
3.20	1
3.40	1

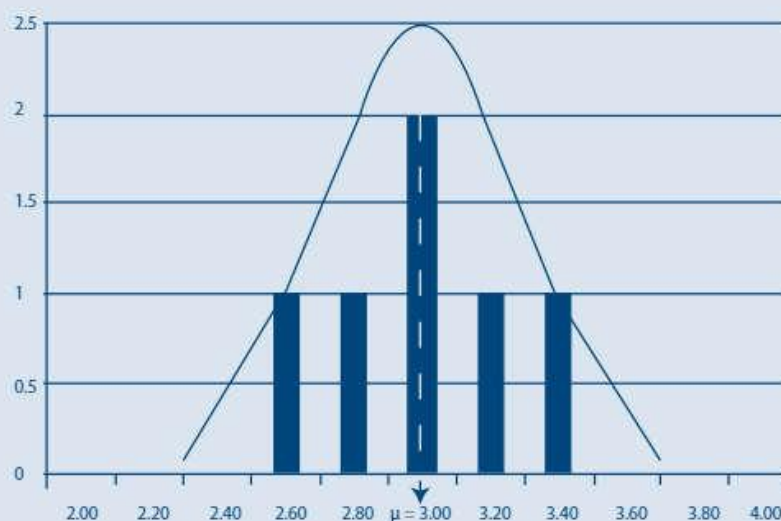


Figura 3.3

En la figura se observa la localización de la media poblacional, la cual se ubica en el centro de la distribución.

Por obvias razones la distribución de las medias dependerá de la estructura y el número de muestras a considerar, ya que si se cambia el número de elementos que estructuran la muestra, también cambiará la forma de la distribución, pero el valor de la media no cambiará.

Alerta

La diferencia entre el estadístico y el parámetro de la media se denomina error muestral de la media.

■ El error muestral

El concepto expone que puede existir diferencia entre la media muestral (estadístico) y la media poblacional (parámetro) debido a que por ciertas circunstancias no es posible tener acceso a los elementos de la población; en otras palabras, existe una deficiencia en la estructuración de la muestra que genera una diferencia entre el estadístico y el parámetro de la media, donde a la diferencia se le denomina error muestral de la media, el cual matemáticamente queda expresado como:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Sin embargo, para fines prácticos, si el tamaño de la población N se considera demasiado grande o infinito, el error muestral se puede definir como:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Para ejemplificar lo anterior, se calculan los errores muestrales de los problemas propuestos en la sección 3.3.

Problema resuelto

Considerando los cálculos de la distribución de las medias muestrales de la población:

$$P = \{3, 1, 2, 4, 3, 5\}$$

Para tamaños muestrales $n = 4$ y $n = 5$, calcula los valores de los errores muestrales para cada caso.

Solución

a) Para el caso de $n = 4$ se tiene lo siguiente:

$$N = 6 \quad n = 4 \quad \sigma = 1.2909$$

$$\sigma_{\bar{X}} = \frac{\sigma^2}{n} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{1.2909}{\sqrt{4}} \cdot \sqrt{\frac{6-4}{6-1}} = 0.4082$$

b) Para el caso de $n = 5$ se tiene lo siguiente:

$$N = 6 \quad n = 5 \quad \sigma = 1.2909$$

$$\sigma_{\bar{X}} = \frac{\sigma^2}{n} \cdot \sqrt{\frac{N-n}{N-1}} = \frac{1.2909}{\sqrt{5}} \cdot \sqrt{\frac{6-5}{6-1}} = 0.2582$$

La diferencia entre los valores de los errores muestrales se puede explicar de manera gráfica, ya que en las gráficas de las distribuciones de las medias se aprecia que cuando $n = 4$ la curva tiene mayor amplitud, condición que no se observa en la curva cuando $n = 5$.

■ Cálculo de probabilidades de ocurrencia sobre las medias muestrales

Pese a que la distribución de medias muestrales es una variable discreta, se puede observar que su gráfica corresponde a una variable aleatoria continua pudiéndose estimar la probabilidad de seleccionar una muestra con una media muestral con valor \bar{X} cualquiera, por lo que la fórmula de la z se ajusta como sigue:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

De manera que se pueden determinar las probabilidades utilizando las tablas de áreas bajo la curva normal. Para ejemplificar se muestra el siguiente problema.

Problema resuelto

Considerando los parámetros y estadísticos resultantes de la distribución de medias cuando $n = 4$, ¿cuál es la probabilidad de seleccionar una muestra con una media entre 3.0 y 3.4?

Solución

Los datos necesarios para el cálculo son:

$$\mu = 3 \quad \sigma_{\bar{X}} = 0.4082 \quad \bar{X} = 3.4$$

Solución (continuación)

por tanto,

$$z = \frac{|\bar{X} - \mu|}{\sigma_{\bar{X}}} = \frac{|3.4 - 3|}{0.4082} = 0.98$$

Para $P(z = 0.98) = 0.3365 = 33.65\%$

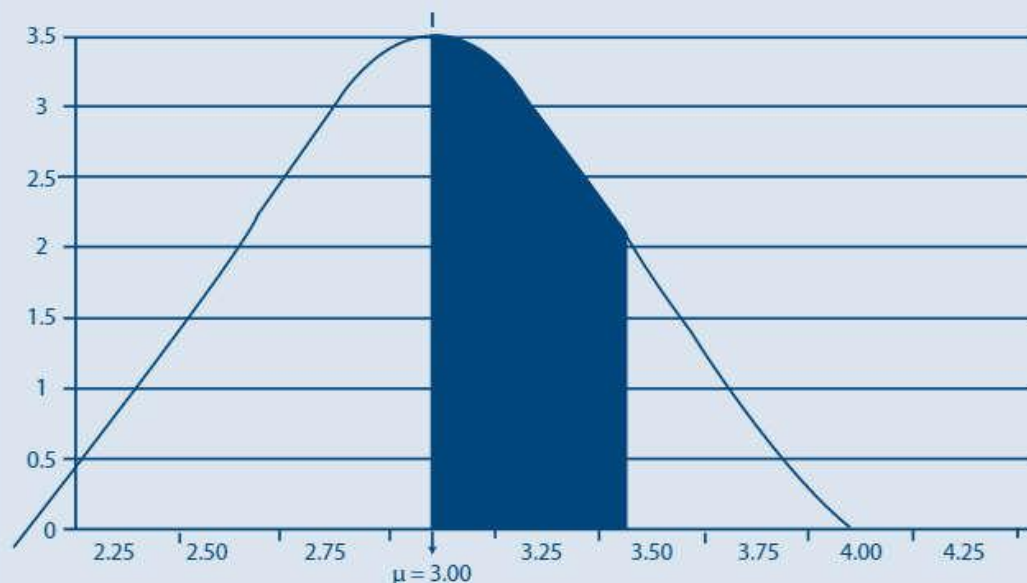


Figura 3.4

Alerta

La distribución normal de las medias muestrales permite calcular la probabilidad de ocurrencia de una muestra con una media que tenga un valor determinado.

3.5 Distribución muestral de las proporciones

De manera similar a la distribución de las medias, puede resultar de interés establecer la relación de la proporción en que ciertos elementos de una población cumplan con ciertas condiciones y la distribución de esta proporción dentro de las muestras; en otras palabras, la distribución muestral de la proporción refiere al conjunto de proporciones de todas las muestras del mismo tamaño, consistentes y válidas, de una población.

Por tanto se puede establecer que la proporción promedio se puede determinar como:

$$\bar{p} = \frac{\sum_{i=1}^{Nm} p_i}{Nm}$$

Donde:

\bar{p} , = Promedio muestral de las proporciones.

p_i , = Proporción de la muestra i ($i = 1, 2, 3, \dots, Nm$)

Nm = Número de muestras.

Asimismo, el error estándar ($\sigma_{\bar{p}}$) de las proporciones queda definido como:

$$\sigma_{\bar{p}} = \sqrt{\frac{PQ}{n} \cdot \frac{N-n}{N-1}}$$

Donde:

P = Valor de la proporción dentro de la población.

Q = Valor de la proporción de que no cumpla dentro de la población ($Q = 1 - P$).

N = Número de elementos de la población (tamaño de la población).

n = Número de elementos de la muestra (tamaño de la muestra).

Asimismo, de manera práctica el error estándar se puede definir como:

$$\sigma_{\bar{p}} = \sqrt{\frac{PQ}{n} \cdot \frac{N-n}{N-1}}$$

En consecuencia, se puede determinar el valor de la probabilidad de que se obtenga una muestra que cumpla con una proporción determinada (p) aplicando la curva de distribución normal, donde la fórmula de conversión a unidades estandarizadas se ajusta como sigue:

$$z = \frac{p - P}{\sigma_{\bar{p}}}$$

Para mostrar lo antes expuesto se plantea el siguiente problema.

Problema resuelto

Una oficina de seguridad y vialidad realiza un estudio sobre el número de conductores multados por infracciones al reglamento de tránsito considerando la siguiente población:

Multados	No multados
A	e
B	f
C	g
D	h

Los responsables del estudio están interesados en realizar un estudio sobre la distribución de la proporción de conductores multados considerando muestras de 5 conductores sobre la población.

Solución

Se observa que la proporción de conductores multados en la población es de 50% ($P = 50\%$).

Donde las proporciones dentro de las muestras de 5 elementos son las que se expresan y siendo la proporción promedio la siguiente:

$$P = 50\% \quad Q = 1 - P = 50\%$$

$$\bar{p} = \frac{\sum_{i=1}^{Nm} p_i}{Nm} = \frac{22}{44} = 50\%$$

Donde el valor del error muestral es:

$$\sigma_{\bar{p}} = \sqrt{\frac{PQ}{n} \cdot \frac{N-n}{N-1}} = \sqrt{\frac{0.5 \cdot 0.5}{5} \cdot \frac{6-5}{6-1}} = 0.1$$

Solución (continuación)

						P
m ₁	a	E	f	g	h	0.2
m ₂	b	e	f	g	h	0.2
m ₃	c	e	f	g	h	0.2
m ₄	d	e	f	g	h	0.2
m ₅	a	b	e	F	g	0.4
m ₆	a	b	e	F	h	0.4
m ₇	a	b	f	g	h	0.4
m ₈	a	c	e	F	g	0.4
m ₉	a	c	e	F	h	0.4
m ₁₀	a	c	f	g	h	0.4
m ₁₁	a	d	e	F	g	0.4
m ₁₂	a	d	e	F	h	0.4
m ₁₃	a	d	f	g	h	0.4
m ₁₄	b	c	e	F	g	0.4
m ₁₅	b	c	e	F	h	0.4
m ₁₆	b	c	f	g	h	0.4
m ₁₇	b	d	e	F	g	0.4
m ₁₈	b	d	e	F	h	0.4
m ₁₉	b	d	f	g	h	0.4
m ₂₀	c	d	e	F	g	0.4
m ₂₁	c	d	e	F	h	0.4
m ₂₂	c	d	f	g	h	0.4

						P
m ₂₃	a	b	c	e	f	0.6
m ₂₄	a	b	c	e	g	0.6
m ₂₅	a	b	c	e	h	0.6
m ₂₆	a	b	c	f	g	0.6
m ₂₇	a	b	c	f	h	0.6
m ₂₈	a	b	c	g	h	0.6
m ₂₉	a	c	d	e	f	0.6
m ₃₀	a	c	d	e	g	0.6
m ₃₁	a	c	d	e	h	0.6
m ₃₂	a	c	d	f	g	0.6
m ₃₃	a	c	d	f	h	0.6
m ₃₄	a	c	d	g	h	0.6
m ₃₅	b	c	d	e	f	0.6
m ₃₆	b	c	d	e	g	0.6
m ₃₇	b	c	d	e	h	0.6
m ₃₈	b	c	d	f	g	0.6
m ₃₉	b	c	d	f	h	0.6
m ₄₀	b	c	d	g	h	0.6
m ₄₁	a	b	c	d	e	0.8
m ₄₂	a	b	c	d	f	0.8
m ₄₃	a	b	c	d	g	0.8
m ₄₄	a	b	c	d	h	0.8

Donde el valor del error muestral es:

$$\sigma_{\bar{p}} = \sqrt{\frac{PQ}{n} \cdot \frac{N-n}{N-1}} = \sqrt{\frac{0.5 \cdot 0.5}{5} \cdot \frac{6-5}{6-1}} = 0.1$$

Por tanto, la distribución de las proporciones muestrales es la siguiente:

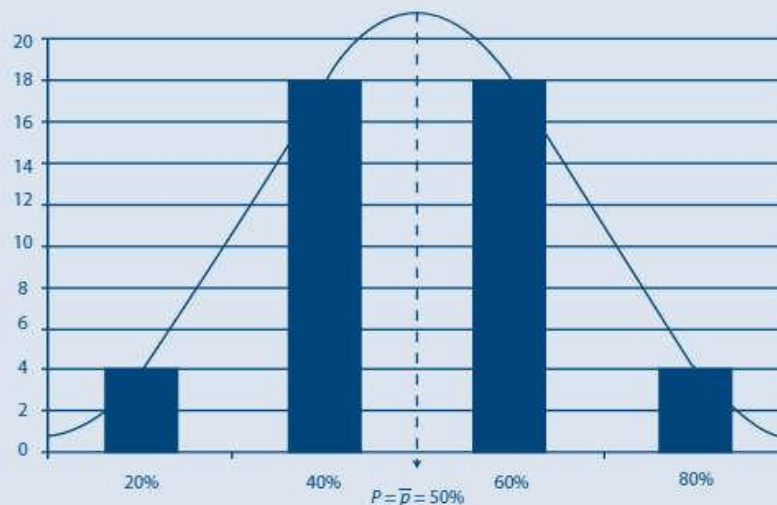


Figura 3.5

Al igual que en la distribución de las medias muestrales, es posible determinar la probabilidad de que se pueda obtener una muestra con una proporción determinada, tal como se ejemplifica a continuación.

Problema resuelto

Con base en los estadísticos y parámetros del problema anterior, ¿cuál es la probabilidad de seleccionar una muestra al azar con una proporción de conductores multados entre 50% y 70%?

Solución

Considerando que:

$$P = 50\% \quad \sigma_{\bar{p}} = 0.1 \quad p = 70\%$$

entonces:

$$z = \frac{|0.7 - 0.5|}{0.1} = 2$$

Por lo que:

$$P(z = 2) = 0.4772 = 47.72\%$$

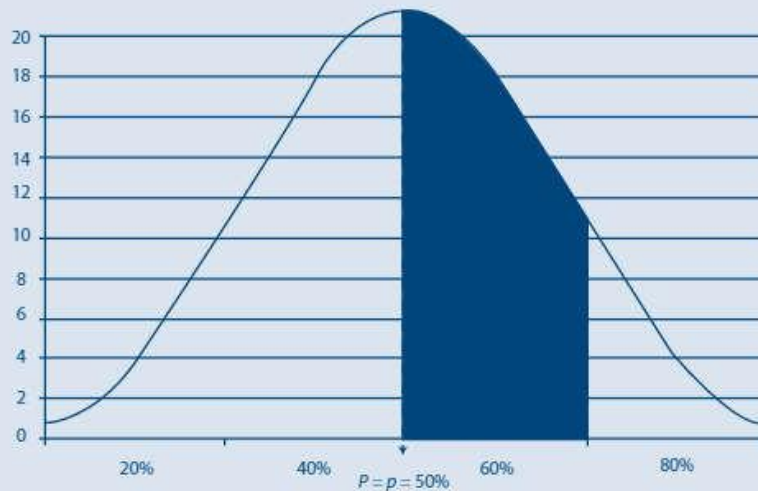


Figura 3.6

Alerta

Al igual que en la distribución de las medias muestrales, es posible determinar la probabilidad de que se pueda obtener una muestra con una proporción determinada.

3.6 Intervalos de confianza

Se puede señalar que es posible determinar los parámetros de la población siempre y cuando se conozcan todas las muestras representativas de la población; o sea que se estaría trabajando con la población misma. Sin embargo, en la práctica en muy pocas ocasiones es posible tener acceso a todas las muestras posibles de la población, por lo que las muestras presentarán ciertas limitaciones de diversos órdenes, por lo que se reconoce que existe error muestral.

Por lo anterior, no es posible conocer el valor del parámetro de manera sesgada; sin embargo, reconociendo que las medias muestrales se distribuyen normalmente, es posible determinar un rango dentro del cual se ubique el valor del mismo bajo ciertos niveles de probabilidad de ocurrencia, los cuales son mejor conocidos como niveles de confianza.

Los niveles de confianza se encuentran definidos en razón de la media y la desviación estándar.

En la siguiente gráfica se muestran los intervalos de confianza, así como las áreas bajo la curva (las probabilidades) que cada uno cubre.

Alerta

La distribución normal de las medias muestrales facilita determinar un rango dentro del cual se puede ubicar el valor de la media poblacional bajo niveles de probabilidad de ocurrencia, los cuales son mejor conocidos como niveles de confianza.

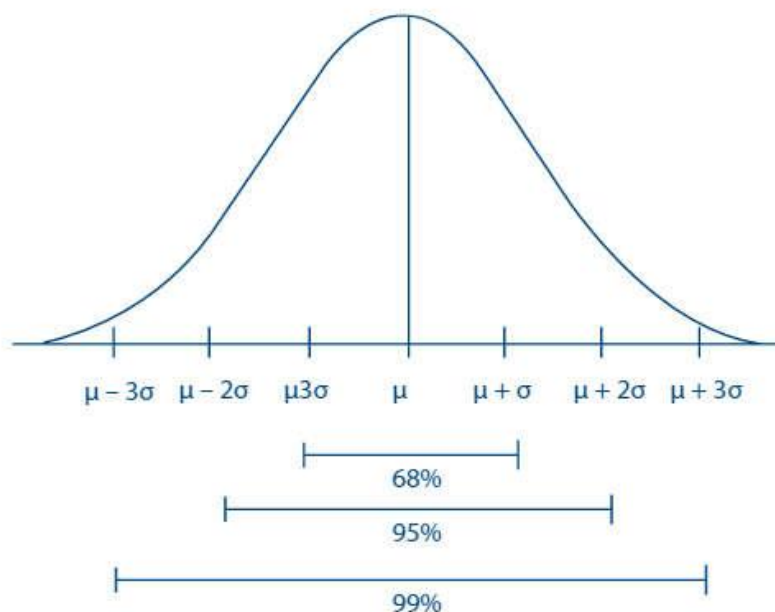


Figura 3.7

De manera práctica se puede señalar que bajo unidades estandarizadas se observa lo siguiente:

- ± 1 Cuando $Z = 1$ El intervalo de confianza es de 68.27%
- ± 2 Cuando $Z = 2$ El intervalo de confianza es de 95.45%
- ± 3 Cuando $Z = 3$ El intervalo de confianza es de 99.73%

De manera complementaria los valores de Z para los diferentes niveles de confianza más comunes son los siguientes:

Nivel de confianza	Z
50.00%	0.6745
68.10%	1.0000
90.00%	1.6545
95.00%	1.9600
95.45%	2.0000
99.00%	2.5800
99.73%	3.0000

■ Estimación del intervalo de confianza para la media poblacional

Considerando el valor de una media muestral, así como el valor del error muestral, es posible determinar los valores de los límites del intervalo de confianza donde:

El valor del límite inferior está determinado como: $\bar{X} - z \sigma_{\bar{X}}$.

El valor del límite superior está determinado como: $\bar{X} + z \sigma_{\bar{X}}$.

Para ejemplificar, considérese el siguiente problema.

Problema resuelto

Un estudio sobre la obesidad juvenil consideró tomar el peso a una muestra de 80 estudiantes, donde el peso promedio fue de 66 kilos, estimando una desviación estándar de la población estudiantil de 3 kilos.

Los responsables del estudio están interesados en conocer los límites de confianza de la media poblacional considerando niveles de confianza de 68.27%, 95.45% y 99.73%.

Solución

Los datos básicos para el cálculo son:

$$\bar{X} = 66 \quad \sigma = 3 \quad n = 80$$

Por lo que se procede a calcular el error muestral:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{80}} = 0.3354$$

Por tanto, los valores de los límites de los intervalos de confianza son los siguientes:

Nivel de confianza (NC)	Z	Límite inferior ($\bar{X} - z \sigma_{\bar{X}}$)	Límite superior ($\bar{X} + z \sigma_{\bar{X}}$)
68.27%	1.0000	65.6646	66.3354
95.45%	2.0000	65.3292	66.6708
99.73%	3.0000	64.9938	67.0062

Sin embargo, cuando no es posible conocer o estimar el valor de la desviación estándar poblacional, los límites de confianza se pueden determinar calculando el error muestral considerando la desviación estándar muestral, por tanto:

$$S_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Considerando que

$$\hat{S} = S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}}$$

Por tanto:

El valor del límite inferior está determinado como: $\bar{X} - z S_{\bar{X}}$.

El valor del límite superior está determinado como: $\bar{X} + z S_{\bar{X}}$.

Sin embargo, debe señalarse que $S_{\bar{X}}$ es buen estimador de $\sigma_{\bar{X}}$ siempre y cuando el valor de n sea grande, ya que el tamaño muestral permite que \hat{S} se aproxime al valor de σ , pero n debe ser mayor que 30.

■ Estimación del intervalo de confianza para la proporción poblacional con base en la proporción muestral

El intervalo de confianza para la proporción de la población queda definido como:

El valor del límite inferior está determinado como: $\bar{p} - z \sigma_{\bar{p}}$.

El valor del límite superior está determinado como: $\bar{p} + z \sigma_{\bar{p}}$.

Recordando que el error estándar está definido como: $\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}}$.



Alerta

Cuando no se conoce la desviación estándar poblacional es posible estimar los valores del intervalo de confianza utilizando el estimador de la desviación estándar muestral.

Pero si de alguna forma no se puede estimar o determinar el valor del error estándar poblacional, se pueden determinar los valores de los límites de confianza del intervalo considerando el error estándar muestral:

$$S_p = \sqrt{\frac{pq}{n}}$$

Sin embargo, debe señalarse que S_p es un buen estimador de $\sigma_{\bar{p}}$ siempre y cuando el valor de n sea grande, pero mayor que 30.

Asimismo, debe acotarse que el valor máximo del producto (pq) se alcanza cuando ambos valen 50%.

Para ejemplificar lo anterior, considérese el siguiente problema.

Problema resuelto

Los coordinadores del proceso de elección de los representantes estudiantiles de una facultad tienen registradas tres planillas: "Evolución", "Pro alumno" y "Mujeres en acción". Se sabe que de una muestra de 160 alumnos registrados en la facultad, 56 tienen la intención de votar por "Mujeres en acción".

¿Cuáles son los límites de confianza al 95% del porcentaje total de los alumnos de la facultad que votarán por la planilla en cuestión?

Solución

Se determina la proporción de la muestra:

$$p = \frac{56}{160} = 0.35 = 35\%$$

entonces $q = 1 - p = 1 - 0.35 = 0.65$

Por lo que, calculando el error estándar:

$$S_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.35)(0.65)}{160}} = 0.03771$$

Si el nivel de confianza es de 95% entonces $Z = \pm 1.96$, por lo que:

Nivel de confianza (NC)	Z	Límite inferior ($p - zS_p$)	Límite superior ($p + zS_p$)
95.00%	1.9600	27.61%	42.39%

3.7 Determinación del tamaño muestral para estimar una media poblacional

Considerando que la distribución de las medias muestrales corresponde a una distribución normal, se puede señalar que el error muestral (E) refiere la diferencia entre una media muestral (\bar{X}) y la media poblacional (μ); por tanto, si se considera a partir de la media el error muestral define un intervalo de confianza:

Alerta

Cuando no se puede estimar o determinar el valor del error estándar poblacional, se pueden determinar los valores de los límites de confianza del intervalo considerando el error estándar muestral.

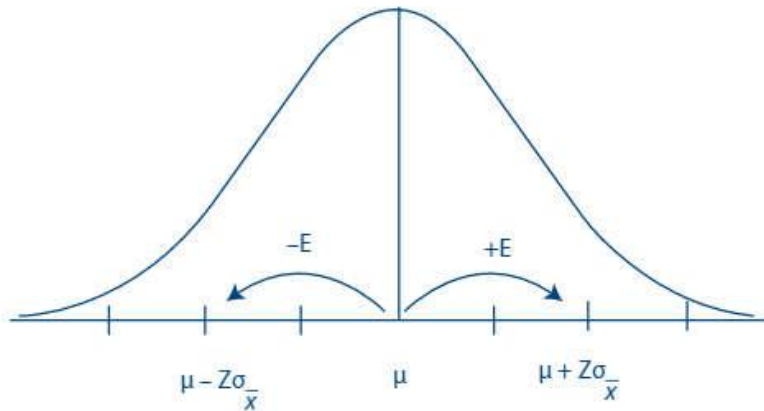


Figura 3.8

Por tanto:

$$e = \bar{X} - \mu = z \sigma_{\bar{x}}$$

Se obtiene que el intervalo de confianza queda definido como:

$$E = z \sigma_{\bar{x}} = z \frac{\sigma}{\sqrt{n}}$$

$$\mu \pm E = \mu \pm z \sigma_{\bar{x}} = \mu \pm z \frac{\sigma}{\sqrt{n}}$$

Por lo que, despejando n , se obtiene la fórmula para determinar el tamaño muestral:

$$E = z \frac{\sigma}{\sqrt{n}}$$

$$\sqrt{n} = \frac{z \cdot \sigma}{E}$$

$$n = \left(\frac{z \cdot \sigma}{E} \right)^2$$

En la fórmula anterior, si aplica, se puede estimar el valor de σ , así como el valor del error. Es importante indicar que el tamaño de la muestra también dependerá del nivel de confianza con el que se desee desarrollar la propuesta.

Considérese el siguiente problema.

Problema resuelto

Una tienda departamental considera realizar un estudio sobre el monto de ventas durante un periodo de crisis económica. Para ello desea determinar el tamaño de la muestra de las notas de ventas más conveniente si se estima una desviación estándar de \$255, un margen de error de \$50 y niveles de confianza de 95% y 99%.

Solución

$$\sigma = \$255 \quad E = \$50$$

Si el nivel de confianza es de 95% entonces $Z = 1.96$, por tanto:

$$n = \left(\frac{z \cdot \sigma}{E} \right)^2 = \left(\frac{(1.96) (\$255)}{\$50} \right)^2 = 99.92 \approx 100 \text{ notas}$$

Alerta

El tamaño muestral para media población depende del nivel de confianza que se defina, siempre y cuando las demás variables permanezcan constantes.

3.8 Determinación del tamaño muestral para estimar la proporción poblacional

De manera similar a lo expuesto en la sección anterior, también se puede determinar el tamaño de la muestra para proporciones partiendo de premisas similares como se expone a continuación. Se considera la siguiente gráfica.

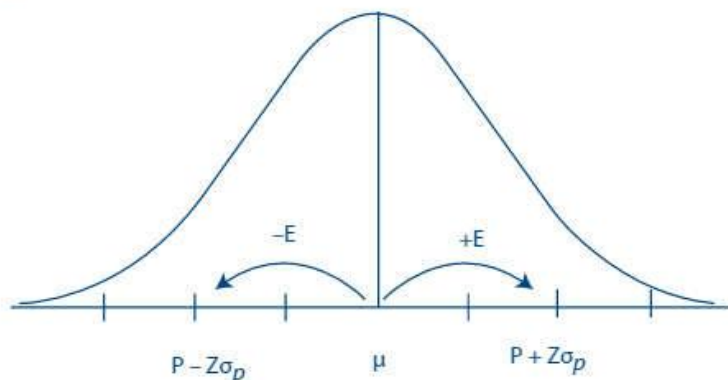


Figura 3.9

Por tanto:

$$E = p - P = z\sigma_p$$

Se obtiene que el intervalo de confianza queda definido como:

$$E = z\sigma_p = z\sqrt{\frac{P \cdot Q}{n}}$$

$$P \pm E = P \pm z\sigma_p = P \pm z\sqrt{\frac{P \cdot Q}{n}}$$

Por lo que despejando n se obtiene la fórmula para determinar el tamaño muestral:

$$E = z\sqrt{\frac{P \cdot Q}{n}}$$

$$E^2 = z^2 \frac{P \cdot Q}{n}$$

$$n = \frac{z^2 \cdot P \cdot Q}{E^2}$$

Debiéndose considerar que el tamaño muestral dependerá en mucho del nivel de confianza propuesto, así como del producto $(P \cdot Q)$ debido, tal como se explicó anteriormente, a que el valor máximo del producto máximo se logra cuando $P = Q = 50\%$.

Para ejemplificar, considérese el siguiente problema.

Problema resuelto

El estudio de mercado para el lanzamiento de un nuevo modelo de tableta estima que 35% del público aficionado a la tecnología estaría dispuesto a comprarla. Para consolidar el estudio se desea estimar el tamaño muestral más conveniente para el estudio de preferencia de compra, por lo que se proponen niveles de confianza de 68%, 95% y 99%, así como un error de 5%.



Alerta

El tamaño muestral por proporciones depende del nivel de confianza y de los valores de P y Q .

Solución

Si $P = 35\%$ entonces $Q = 1 - P = 65\%$, por tanto:

Nivel de confianza	Z	n (No. de personas)
68%	1.0000	$91.00 \approx 91.00$
95%	1.9600	$349.59 \approx 350.00$
99%	2.5800	$605.73 \approx 606.00$

Cabe señalar que entre más grande sea la muestra el error estándar se reduce.

3.9 Grados de libertad

Los grados de libertad pueden definirse como el número de elementos que pueden variar en su valor dentro de una colección.

Supóngase que se cuenta con tres variables: K, L y M. Si se dispone que el valor de K es 10, pero depende de los valores enteros que pueden tomar L y M, entonces los valores que pueden tomar son los siguientes:

$$K = 10 \text{ pero } K = L + M$$

A	10	9	8	7	6	5	4	3	2	1	0
B	0	1	2	3	4	5	6	7	8	9	10

O sea, de 3 variables pueden "variar 2"; entonces hay 2 grados de libertad.

**Alerta**

Los grados de libertad refieren el número de elementos que pueden variar en su valor dentro de una colección.

3.10 Intervalos de confianza para muestras pequeñas

Se considera como muestra pequeña aquella cuyo número de elementos es menor que 30.

Se ha determinado que la falta de elementos genera que el error muestral se incremente y por tanto la distribución de probabilidad no se ajuste a la normal. Por lo que, para compensar lo anterior, se utiliza la curva de distribución denominada t de Student, la cual es similar a la curva normal, pero un poco más robusta, y cuenta con las siguientes características:

- El valor del área bajo la curva es 1.
- Es simétrica con respecto a la media, pero es más extendida y chata en el centro en comparación con la curva de distribución normal.
- Al ser t una distribución más extendida, los valores de los límites de los intervalos de confianza son mayores en comparación con sus similares en la distribución normal.
- No hay una distribución única, sino un grupo de distribuciones, considerando que todas tienen como media 0 (cero), pero la desviación estándar dependerá del tamaño de la muestra.

La gráfica de la curva de distribución t es la siguiente:

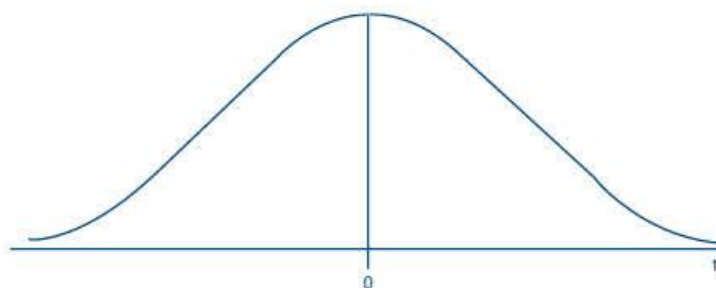


Figura 3.10

Una de las principales premisas es que se desconoce el valor de la desviación estándar poblacional, por lo que se deben tener como referentes los estadísticos de la muestra. En consecuencia, al igual que en la curva normal, para determinar el valor de las unidades de t se utiliza la fórmula:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

donde

$$\hat{S} = s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}}$$

Con base en el resultado obtenido de t se deberá considerar el nivel del intervalo de confianza.

Asimismo, para la determinación de los valores de los límites del intervalo de confianza para la estimación de la media poblacional con desviación estándar (σ) desconocida se aplica:

$$\bar{X} \pm t \frac{s}{\sqrt{n}}$$

Donde el valor de t depende del nivel de confianza propuesto y los grados de libertad, definidos como:

$$gl = n - 1$$

El nivel de confianza determina el área bajo la curva donde se puede ubicar el valor de la media poblacional, mientras que las áreas fuera de los límites del intervalo se denominan áreas de error (al error se le identifica con α). Para ubicar el valor de t se utilizan las tablas de distribución t , pudiendo ser de una cola (derecha o izquierda) o de dos colas. Por lo común se utiliza la tabla de distribución de dos colas, de manera que el error queda definido por $\frac{\alpha}{2}$.



Figura 3.11

La definición del nivel de confianza se puede establecer de manera directa, por ejemplo 95%, pero también se puede definir el mismo tan solo señalando el valor del error, o sea que un valor de $\alpha = 1\%$ indica un intervalo de confianza de 99%.

Para ejemplificar lo anterior, revítese el siguiente problema.

Problema resuelto

Una obra de construcción cuenta con un laboratorio de campo para realizar pruebas para la medición de la resistencia de cilindros de concreto. De un concreto diseñado para 250 kg/cm^2 se cuenta con una muestra de 10 cilindros cuyas resistencias se muestran en la siguiente tabla.

Problema resuelto (continuación)

CL #1	CL #2	CL #3	CL #4	CL #5
251	253	254	250	253
CL #6	CL #7	CL #8	CL #9	CL #10
252	250	253	251	253

Se desea estimar el valor de los límites de confianza de la media poblacional considerando un error de 5%.

Solución

Con base en los datos, el valor de la media y la desviación estándar muestrales son:

$$\bar{X} = 252.00$$

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}} = 1.4142$$

Grados de libertad:

$$gl = n - 1 = 10 - 1 = 9$$

Para $\alpha = 5\%$ (nivel de confianza de 95%) y $gl = 10$, el valor de t en dos colas es de 2.262, por lo que los valores de los límites del intervalo de confianza son:

$$\text{Límite inferior: } \bar{X} - t \frac{s}{\sqrt{n}} = 252 - (2.262) \frac{1.4142}{\sqrt{10}} = 250.99$$

$$\text{Límite superior: } \bar{X} + t \frac{s}{\sqrt{n}} = 252 + (2.262) \frac{1.4142}{\sqrt{10}} = 253.01$$

Alerta

Cuando el tamaño de una muestra es de 30 o más elementos, se considera que se puede distribuir con apego a la normal. Se considera como muestra pequeña aquella cuyo número de elementos es menor que 30 y su análisis debe apegarse a la distribución t de Student.

3.1 Con base en el conjunto:

$$U = \{18, 39, 66, 79, 90\}$$

determina las muestras de 3 elementos.

3.2 Con base en los resultados del problema anterior comprueba:

$$\bar{X} = \mu \text{ y } \hat{\sigma}^2 = S^2$$

3.3 Con base en el conjunto:

$$U = \{13, 21, 34, 45, 66\}$$

Determina las muestras consistentes y válidas de 4 elementos.

3.4 Con base en los resultados del problema anterior comprueba:

$$\bar{X} = \mu \text{ y } \hat{\sigma}^2 = S^2$$

3.5 Con base en el conjunto:

$$U = \{13, 24, 37, 56, 78, 98\}$$

determina muestras de 5 elementos.

3.6 Con base en los resultados del problema anterior comprueba:

$$\bar{X} = \mu \text{ y } \hat{\sigma}^2 = S^2$$

3.7 Considerando el planteamiento del problema 3.5, determina muestras de 4 elementos y comprueba.

3.8 Con base en los resultados del problema anterior comprueba:

$$\bar{X} = \mu \text{ y } \hat{\sigma}^2 = S^2$$

3.9 Considerando el conjunto:

$$P = \{200, 212, 234, 268, 294\}$$

determina las muestras, consistentes y válidas, de 3 elementos.

3.10 Con base en los resultados del problema anterior comprueba:

$$\bar{X} = \mu \text{ y } \hat{\sigma}^2 = S^2$$

3.11 Considerando el conjunto:

$$Q = \{10, 26, 84, 171, 192, 227, 258\}$$

determina las muestras, consistentes y válidas, de 5 elementos.

3.12 Con base en los resultados del problema anterior comprueba:

$$\bar{X} = \mu \text{ y } \hat{\sigma}^2 = S^2$$

3.13 Con base en los datos del problema 3.11 determina muestras de 3 elementos y comprueba que:

$$\bar{X} = \mu \text{ y } \hat{\sigma}^2 = S^2$$

3.14 Una fábrica de focos para el hogar está diseñando un foco de 60 watts con componentes de bajo costo. Las pruebas iniciales de ajuste de la vida útil se desarrollaron con una población de prototipos de 6 focos cuyas horas de servicio fueron las siguientes:

Foco	a	b	c	d	e	f
Horas	19	34	53	89	97	116

El jefe de proyecto te asigna el desarrollo del análisis estadístico considerando muestras de 4 elementos a efecto de determinar la media y la desviación estándar poblacional, así como la comprobación de estas por estimadores insesgados.

3.15 Un despacho jurídico cuenta con 4 secretarías (2 por turno matutino o vespertino rotando turnos), las cuales llenan un mismo formato para el control de documentos del despacho, y que cometen el siguiente número de errores.

Secretaría	No. de errores
A	3
B	2
C	1
D	4

Si se seleccionan muestras de dos secretarías a partir de la población de 4 secretarías, determina y comprueba los parámetros poblacionales (media y desviación estándar) a través de los estimadores insesgados de las muestras.

3.16 Una fábrica de rodamientos produce baleros de 2 pulgadas de diámetro. El departamento de control de calidad desea determinar el tamaño de la muestra de análisis al 90% de confianza, considerando una desviación estándar poblacional estimada de 0.04 pulgadas y un error de 0.06 pulgadas.

3.17 Una institución bancaria desarrolla un proceso de auditoría sobre los cortes de caja de sus diferentes sucursales. Si los responsables saben que la desviación estándar es de \$120 y un error de \$50, ¿cuál debe ser el tamaño de la muestra si considera un nivel de confianza de 99%?

3.18 Un proceso de remodelación urbana afecta la hora de ingreso de los empleados de una empresa. El departamento de Recursos Humanos desea realizar un estudio para determinar si se incrementa de manera temporal el tiempo de tolerancia que es de 15 minutos. Se requiere determinar el tamaño de la muestra de los registros del tiempo de ingreso al 95% de confianza, considerando un error de 8 minutos.

3.19 Una institución bancaria tiene por política que los ejecutivos de cuenta se conecten al sistema 5 minutos antes o después de la hora de apertura de las sucursales,

SR que es a las 8:30 a.m. Los responsables de la evaluación de calidad en el servicio (SERVQUAL) desean determinar el tamaño de la muestra al 95% y 99% de confianza de los registros de conexión al sistema, considerando que la desviación estándar es de 2 minutos 40 segundos y un error de 1 minuto 15 segundos.

SR **3.20** Una aerolínea regional desea realizar un estudio sobre el número de pasajeros en lista de espera por mes. Se sabe que el número de pasajeros en la lista fluctúa en más o menos 5 por vuelo. Se desea estructurar una muestra sobre las listas de espera al 99% de confianza, considerando un margen de 2 pasajeros.

SR **3.21** Un estudio sobre pacientes diabéticos demuestra que requieren aplicarse después de los alimentos más o menos 4 unidades de una nueva insulina. Si se considera un error de 1.5 unidades, se desea determinar el tamaño de la muestra de pacientes para desarrollar el estudio sobre el nuevo fármaco a un nivel de confianza de 90%.

SR **3.22** Con base en los datos del problema anterior, se considera oportuno desarrollar una investigación con una muestra al 99% de confianza.

3.23 La evaluación sobre el índice de reprobación de una materia universitaria indica que 45% de los alumnos reprueba. Se requiere conformar una muestra de los alumnos considerando 5% de error y un nivel de confianza de 95.45% con el propósito de desarrollar las medidas que permitan reducir el índice de reprobación.

SR **3.24** Una importante cadena de supermercados sabe que en 25% de las operaciones los cajeros no cuentan con el dinero fraccionario para dar el cambio a los clientes. Los responsables de la tesorería de la cadena desean establecer un estudio inferencial sobre esta situación, estructurando una muestra al 99% de confianza y considerando un error de 10%.

SR **3.25** Una cadena de tlapalerías sabe que en 78% de las ventas de thinner no se ofrece el litro exacto, o sea que se ofrece más o menos producto. Se desea desarrollar un proceso de análisis sobre los inventarios del solvente, considerando muestras al 90%, 95%, 95.45%, 99% y 99.73% y con un nivel de error de 10%.

SR **3.26** Una fábrica de acumuladores para automóviles encuentra que la vida media de una muestra de 26 acumuladores es de 39 meses con una desviación estándar de 5 meses. Si se consideran grados de confianza de 90%, 95% y 99%, ¿cuáles son los valores de los límites de confianza donde se puede encontrar la media poblacional?

3.27 Con base en los datos de la muestra, procede a ordenar y determinar los límites de confianza para la media poblacional al 95% y 99%.

142 161 150 145 130 184 176 154 149 178
112 90 162 77 198 181 116 100 165 199

3.28 Con base en los datos del problema anterior, determina los límites de los intervalos de confianza al 50% y 68%.

3.29 Con base en los datos de la muestra, procede a ordenar y determinar los límites de confianza para la media poblacional al 50% y 68%.

1234 1178 1340 1189 1567
1645 1934 1937 1756 1111
1023 1404 1023 987 1032

3.30 Con base en los datos del problema anterior, determina los límites de los intervalos de confianza al 95.45% y 99%.

3.31 Considerando datos de la muestra, procede a determinar los límites de confianza al 50%, 68% y 90%.

169 128 140 176 161 119 164 154
140 136 154 136 148 144 175 148
136 162 130 141 123 159 151 150
146 139 160 146 136 148 144 152
145 142 156 145 128 150 135 161

3.32 Con base en los datos del problema anterior, determina los límites de los intervalos de confianza al 95%, 95.45% y 99%.

3.33 A partir de la colección de datos, determina el valor de los límites de los intervalos de confianza al 95%, 95.45% y 99%.

169 128 140 176 161 119 164 154
140 136 154 136 148 144 175 148
136 162 130 141 123 159 151 150
146 139 160 146 136 148 144 152
145 142 156 145 128 150 135 161

3.34 Considerando datos de la muestra, procede a determinar los límites de confianza al 50%, 68% y 90%.

SR **3.35** Una empresa de elementos prefabricados de madera procedió a clasificar los excedentes de bastidores de una pulgada de espesor por su longitud en centímetros. Donde el detalle del inventario se expone a continuación.

25 37 47 60 74
34 38 52 63 66
27 38 49 61 72
46 64 70 44 45
41 53 62 67 42
45 59 71 72 60
50 58 51 56 52
45 49 52 53 57

Considerando esta información, procede a determinar el intervalo de confianza de la media poblacional, considerando el tamaño muestral en cuanto a la aplicación de la distribución normal o t de Student bajo niveles de 68% y 95.45%.

3.36 Al término de un proyecto de construcción, una empresa de proyectos electromecánicos procede a realizar el levantamiento del inventario de los tramos de cable de 500 MCM considerando su largo en pulgadas. Los responsables del departamento de Ingeniería de Procura (Compras) estructuran la muestra que se detalla a continuación.

37	38	38	64	53	59	58	49
74	66	72	45	42	60	52	57
60	63	61	44	67	72	56	53
25	34	27	46	41	45	50	45
65	73	39	70	54	73	71	61
47	52	49	70	62	71	51	52

Considerando esta información, procede a determinar el intervalo de confianza de la media poblacional, considerando el tamaño muestral en cuanto a la aplicación de la distribución normal o *t* de Student bajo niveles de error de 10% y 5%.

3.37 Considerando el planteamiento del problema anterior, determina el valor de los límites del intervalo de confianza al 99%.

3.38 Un fabricante de focos desarrolla un estudio sobre la vida útil de un foco ahorrador de 20 watts a través de una muestra de 15 unidades, donde las horas de servicio de cada foco son las siguientes:

520	521	511	513	510
513	522	500	521	495
496	488	500	502	512

Por lo que, en atención a la información mostrada, determina los intervalos de confianza sobre la media poblacional al 90%, 95% y 99%.

3.39 Determina los valores de los intervalos de confianza para la media poblacional al 50% y 68% a través de los estadísticos de la muestra que se expone a continuación:

11	27	34	42
22	56	35	56
45	67	37	67
78	73	66	98
65	90	78	95

3.40 Considerando los datos del problema anterior, determina los valores de los límites de los intervalos de confianza cuando $\alpha = 5\%$ y $\alpha = 1\%$.

3.41 Una empresa internacional de control de calidad en materiales de construcción obtiene una muestra de 10 cilindros de concreto, los cuales son ensayados a compresión obteniendo las siguientes resistencias en kg/cm^2 .

Cil.1	Cil.2	Cil.3	Cil.4	Cil.5	Cil.6	Cil.7	Cil.8	Cil.9	Cil.10
248	252	254	251	249	250	252	255	253	248

Determina el intervalo de confianza de la media poblacional al 90%, 95% y 99%.

3.42 El gerente de una tienda de refacciones para tornos de control numérico desea estimar la cantidad media de venta de refacciones. Una muestra de 20 facturas (en dólares) muestra las siguientes cantidades:

\$48.16	\$52.64	\$51.35	\$23.78	\$61.83
37.92	61.46	51.45	46.94	43.88
49.17	46.82	50.82	58.84	54.86
42.22	48.59	52.68	41.86	61.69

Si eres nombrado asesor de operaciones del gerente, procede a determinar lo siguiente:

¿Cuál es el mejor intervalo de confianza para la media poblacional, considerando niveles de error de 10%, 5% y 1% en prueba *t* de dos colas?

3.43 Una fábrica de grasa grafitada para maquinaria envasa la grasa en latas de 500 g. El área de control de calidad realiza una revisión sobre el proceso de envasado, para ello consideró una muestra de 10 paquetes con los siguientes pesos en gramos:

497	492	510	505	494
503	500	492	501	498

A las gerencias de producción y de control de calidad les interesa determinar el estimado de la media poblacional al 90%, 95% y 99%.

3.44 Una fábrica de bastidores de madera para muebles mide la resistencia a la compresión de 2 muestras de bastidores de $\frac{1}{3}$ pulg de espesor, donde las pruebas de laboratorio arrojaron los siguientes resultados en kg/cm^2 .

Muestra A		Muestra B
10.14		10.56
11.24		10.66
10.38		10.45
10.65		10.90
10.80		11.00

Como responsable del análisis de producción, procede a desarrollar los cálculos necesarios de manera que se determinen los intervalos de confianza al 90% para la media poblacional de cada muestra.

3.45 Con base en el planteamiento del problema anterior, determina el intervalo de confianza al 95% para la media poblacional considerando la muestra completa.

3.46 Con base en el planteamiento del problema 3.36, procede a desarrollar los cálculos necesarios de manera que se determinen los intervalos de confianza al 99% para la media poblacional de cada muestra.

3.47 Con base en el planteamiento del problema anterior, determina el intervalo de confianza al 90 y 95% para la media poblacional considerando la muestra completa.

3.48 Con base en el planteamiento del problema de la fábrica de focos para el hogar, determina el intervalo de confianza para la media poblacional al 95%.

Foco	a	b	c	d	e	f
Horas	19	34	53	89	97	116

3.49 Un análisis inferencial considera una media muestral de 8, una desviación estándar poblacional de 2, un tamaño muestral de 44, y límites de confianza de $LI = 7.41$ y $LS = 8.59$. Considerando una distribución normal, determina:

- El nivel de confianza del análisis.
- El valor del error muestral.



PROBLEMAS RETO

1

El Ministerio de Finanzas señala que el llenado de un formato vía Internet por parte de los ciudadanos promedia 120 minutos. Sin embargo, una muestra de 40 personas expuso que el tiempo promedio es de 80 minutos. Bajo condiciones de aproximación a la normal, al director de la oficina le interesaría conocer lo siguiente:

- ¿Cuál es el valor del error entre las medias?
- ¿Cuál es el valor de la desviación estándar poblacional al 68%, 95% y 99% de confianza?

2

Con base en el planteamiento del problema anterior, al director le interesa responder de manera adicional las siguientes interrogantes:

- Al 99% de confianza, ¿cuál es la probabilidad de que los contribuyentes tarden más de 120 minutos?
- Al 99% de confianza, ¿cuál es la probabilidad de que los contribuyentes tarden entre 120 y 150 minutos?
- Al 99% de confianza, ¿cuál es la probabilidad de que tarden más de 150 minutos?



REFERENCIAS

- Berenson, Mark L. y David M. Levine (1996). *Estadística básica en administración* (6a. ed.). México: Prentice Hall Hispanoamericana.
- Hines, William W., Douglas C. Montgomery, David M. Goldsman y Connie M. Borror (2009). *Probabilidad y estadística para ingenieros* (4a. ed.). México: Patria, 3a. reimpresión.
- Kohler, Heinz (1996). *Estadística para negocios y economía* (1a. ed.). México: Compañía Editorial Continental.
- Lind, Douglas, William G. Marchal y Samuel A. Wathen (2005). *Estadística aplicada a los negocios y la economía* (12a. ed.). México: McGraw-Hill.
- Mendenhall, William (1999). *Estadística para administradores* (2a. ed.). México: Grupo Editorial Iberoamérica.
- Quevedo, Urias Héctor y Blanca Rosa Pérez Salvador (2008). *Estadística para Ingeniería y Ciencias*. México: Grupo Editorial Patria.



DIRECCIONES ELECTRÓNICAS

Portal académico CCH-Distribuciones de Probabilidad

[<http://portalacademico.cch.unam.mx/alumno/sitiosdeinteres/matematicas/estadistica2>]

Distribución normal

[http://es.wikipedia.org/wiki/Distribuci%C3%B3n_normal]

t de Student

[<http://jacroman.blogspot.mx/2009/12/distribucion-t-student.html>]

Tabla t de Student

[<ftp://ftp.fcien.edu.uy/Programas/ftp/bioe2007/archivos/tablat.pdf>]



Análisis estadístico de experimentos

OBJETIVOS

- Entender la importancia de la aplicación de la estadística inferencial al análisis de los resultados derivados de un experimento.
- Comprender y distinguir los conceptos de experimento y prueba.
- Comprender el concepto de tratamiento.
- Distinguir la diferencia entre variable dependiente y variable independiente.
- Comprender el concepto de hipótesis estadística.
- Distinguir la diferencia entre hipótesis nula e hipótesis alternativa.
- Comprender la importancia de la aplicación de la distribución t de Student para el análisis de pruebas de hipótesis estadísticas.
- Entender el concepto de análisis de la varianza (ANOVA).
- Comprender la aplicación de la distribución de la F de Fisher para el análisis de las pruebas de hipótesis estadísticas.
- Entender el concepto de discriminantes para el análisis de diferenciación de resultados.

¿QUÉ SABES?

- ¿Cuál es la trascendencia de la aplicación de la estadística inferencial al análisis de los resultados derivados de un experimento?
- ¿Cuál es la diferencia entre un experimento y una prueba?
- ¿Qué es un tratamiento?
- ¿Cuál es la diferencia entre una variable dependiente y una independiente?
- ¿Qué es una hipótesis estadística?
- ¿Cuál es la diferencia entre una hipótesis nula y una hipótesis alternativa?
- ¿Cuál es la importancia de la aplicación de la distribución t de Student para el análisis de pruebas de hipótesis estadísticas?
- ¿Qué es análisis de la varianza (ANOVA)?
- ¿En qué forma la distribución de la F de Fisher apoya el análisis de las pruebas de hipótesis estadísticas?
- ¿Qué son los discriminantes para el análisis de diferenciación de resultados de un experimento?

i Alerta

Los elementos sujetos a observación y medición en un experimento son las variables.

i Alerta

Las variables en un experimento, de manera básica, se clasifican en dependientes e independientes.

i Alerta

La estructura básica de las hipótesis de investigación científica es la hipótesis estocástica: "si x entonces y ".

i Alerta

Una hipótesis estadística expone una conjetura o condición sobre un parámetro de la población, la cual se comprueba o refuta de acuerdo con el análisis de los estadísticos de la muestra.

i Alerta

El análisis de las hipótesis estadísticas puede generar que se cometan errores tipo 1 o tipo 2.

4.1 Introducción

En el transcurso de las investigaciones científicas e industriales se realizan procesos planificados de investigación y desarrollo en los cuales se llevan a cabo experimentos. Los resultados obtenidos en los procesos de experimentación deben ser organizados, computados y representados mediante gráficas para proceder a su interpretación y análisis mediante la estadística inferencial con el fin de comprobar, refutar o generar conocimientos.

4.2 Concepto de experimento

Experimento es el conjunto de actividades interrelacionadas, las cuales se realizan con el propósito de practicar o replicar un fenómeno de manera planificada y controlada para obtener un resultado esperado en razón al número y magnitud de las variables que lo estructuran.

De manera básica se puede señalar que las variables son los elementos sujetos a observación y medición durante un experimento, si consideramos que una o más variables son manipuladas por los investigadores para generar un resultado específico. Por consecuencia, a las variables se les puede clasificar en las siguientes.

- a) **Variables independientes.** Son aquellas manipuladas a discreción por los investigadores para generar un resultado específico y se les representa con la letra x .
- b) **Variables dependientes.** Son las que muestran el resultado o fenómeno en estudio y están representadas por la letra y .

Lo anterior dispone que las variables independientes causan o explican a las dependientes. Este argumento da origen a la estructura básica de las hipótesis de investigación científica, o sea las hipótesis estocásticas: "si x entonces y ".

En resumen, durante una investigación científica de orden aplicado a la industria, los investigadores manipularán las variables x buscando generar un resultado esperado referido por la variable y .

Ante lo expuesto, se puede señalar que en muy contadas ocasiones los resultados esperados de un experimento se obtienen en primera instancia. Ante ello los investigadores deben realizar ajustes durante el experimento, así como repetirlos bajo diferentes condiciones y estructura de las variables a efecto de observar el comportamiento del fenómeno en estudio. A este concepto se le conoce como tratamiento, es decir, los investigadores pueden repetir un experimento a partir de diferentes tratamientos.

Asimismo, cuando se realiza (ejecuta) un tratamiento se le denomina ensayo o prueba, por lo que un experimento puede constar de varios tratamientos y ensayos, de manera que los resultados de un tratamiento pueden ser contrastados con los de otro u otros tratamientos a efecto de obtener conclusiones referentes al fenómeno en estudio, incluso comparar uno o varios tratamientos contra otro considerado como testigo. A lo expuesto se le conoce como estudios comparativos.

4.3 Hipótesis estadísticas

Son argumentos que exponen una condición sobre un parámetro de la población, el cual está sujeto a verificación mediante la evidencia cuantitativa derivada del análisis estadístico sobre una o varias muestras.

Las hipótesis estadísticas de manera general quedan definidas como sigue:

- **Hipótesis nula (H_0).** Es una afirmación sobre un parámetro poblacional, el cual se trata de anular mediante la evidencia.
- **Hipótesis alternativa (H_1).** Se refiere a la negación del argumento expuesto en la hipótesis nula, la cual se valida en el momento que se refuta H_0 .

Sin embargo, puede llegar a suceder que existan condiciones que impidan el buen desarrollo de un experimento, así como el proceso de análisis estadístico que dé origen a los llamados errores estadísticos, los cuales pueden ser de dos tipos.

- **Tipo 1 o α .** Se rechaza H_0 cuando es verdadera.
- **Tipo 2 o β .** Se acepta H_0 cuando es falsa.

4.4 Aplicación de la distribución t de Student en el análisis de experimentos

Debe señalarse que la realización de experimentos tanto académicos como industriales depende en gran medida de la disponibilidad de recursos y el tiempo para llevarlos a cabo. Puede ocurrir que en los experimentos se utilicen muestras pequeñas o prototipos, es decir, que se cuente con un número limitado de elementos.

Lo anterior expone que se manejen tamaños muestrales pequeños, por lo que el análisis de experimentos se fundamenta en la distribución t de Student.

4.5 Estudios comparativos simples

La profundidad en cuanto al análisis estadístico de un experimento depende en gran medida de su propósito, pero en esencia se procura realizar varios tratamientos cuyos resultados se comparen a efecto de obtener conclusiones.



Figura 4.1

De manera general, los estudios comparativos simples exponen las siguientes hipótesis estadísticas:

- $H_0 = \mu_1 = \mu_2 = \dots = \mu_n$ No hay diferencia entre las medias.
- $H_1 = \mu_1 \neq \mu_2 \neq \dots \neq \mu_n$ Hay diferencia entre las medias.

El planteamiento de las hipótesis es simple de explicar, ya que en cada muestra o grupo se realiza un tratamiento distinto, de manera que si las medias no difieren en su valor, es posible que los tratamientos no generen el efecto o resultado esperado y por consecuencia se valida la hipótesis nula.

De manera práctica, la prueba de hipótesis consiste en contrastar los resultados de los tratamientos por pares de muestras. De modo que a partir de una varianza combinada,

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Donde

S_1^2 = varianza de la muestra uno

S_2^2 = varianza de la muestra dos

n_1 = tamaño de la muestra uno

n_2 = tamaño de la muestra dos

Se determina el valor de t calculada identificada por t_0 ,

$$t_0 = \frac{\bar{Y}_2 - \bar{Y}_1}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Alerta

En la teoría del análisis de experimentos el término grupo se utiliza como sinónimo de muestra.

Análisis estadístico de experimentos

Donde el valor de t_0 se deberá contrastar con una t_{tab} teórica o tabular definida como: $t_{\text{tab}}\left(\frac{\alpha}{2}, \text{gl}\right)$, donde los grados de libertad se definen como: $\text{gl} = (n_1 + n_2) - 2$. De manera gráfica la t_{tab} tabular define los límites de las áreas entre H_0 y H_1 .

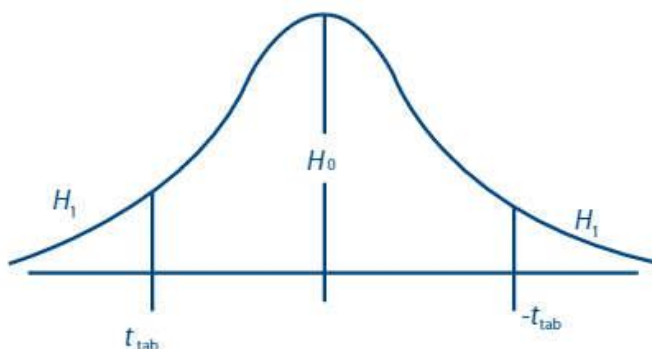


Figura 4.2

De manera que si,

$$t_0 < t_{\text{tab}} \text{ o } t_0 > t_{\text{tab}} \text{ se acepta } H_1.$$

Para ejemplificar lo anterior considérese el siguiente problema.

Problema resuelto

Una empresa que manufactura armas crea un nuevo diseño de cartuchos para escopeta. Con el fin de darle mayor estabilidad e impermeabilidad al cuerpo del cartucho, experimenta con dos tipos de resinas plásticas cuyo grosor origina que el peso del cartucho varíe.

Realiza un análisis comparativo simple para determinar si existe diferencia en los dos tipos de cartuchos, dependiendo del tipo de resina del cuerpo si se considera el peso, en gramos, de seis cartuchos por muestra.

Resina A (RA)	Resina A-1 (RA1)
32.30	31.89
31.90	32.10
31.45	32.34
32.80	32.40
33.01	32.30
32.78	31.99

Solución

Considérese que x = Resina, y = Peso y que:

$$H_0 = \mu_1 = \mu_2$$

$$H_1 = \mu_1 \neq \mu_2$$

Solución (continuación)

Por lo que, si seguimos el procedimiento de cálculo, se determinan los valores de la media y la varianza de cada muestra.

$$\begin{aligned}\bar{Y}_1 &= 32.37 & \bar{Y}_2 &= 32.18 \\ S^2 &= \frac{\sum (Y_i - \bar{Y})^2}{n-1} \\ S_1^2 &= 0.367 & S_2^2 &= 0.043\end{aligned}$$

El valor de la varianza combinada es:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(6 - 1)(0.367) + (6 - 1)(0.043)}{6 + 6 - 2} = 0.2025$$

Por lo que el valor de t_0 es:

$$t_0 = \frac{\bar{Y}_2 - \bar{Y}_1}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{32.37 - 32.17}{0.45274 \sqrt{\frac{1}{6} + \frac{1}{6}}} = 0.7651$$

Ya que el valor de t_{tab} :

$$\alpha = 5\% \quad \frac{\alpha}{2} = 0.025$$

$$gl = (n_1 + n_2) - 2 = 12 - 2 = 10$$

$$t_{\text{tab}}\left(\frac{0.05}{2}, 10\right) = 2.228$$

Al contrastar los valores de manera gráfica:

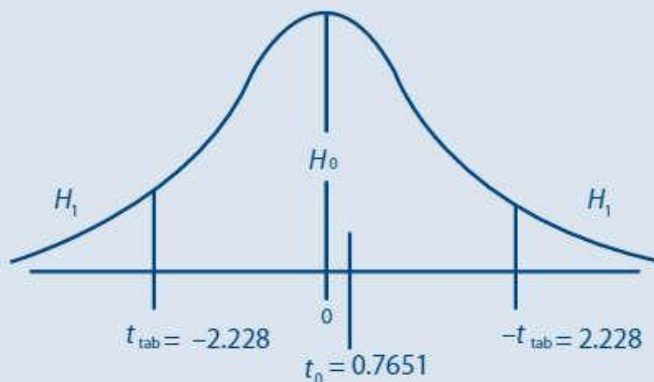


Figura 4.3

Los cálculos exponen que el valor de t_0 se ubica en el área correspondiente a la hipótesis nula. Es decir, las resinas no generan diferencias entre los pesos de los cartuchos.

4.6 Estudios comparativos basados en pruebas de hipótesis sustentados en el análisis de la varianza (ANOVA) de un factor con una muestra por grupo



Alerta

El término ANOVA es el acrónimo de *Analysis Over Variance*.

Este tipo de estudios se fundamenta en el análisis de las posibles fuentes de variación en los resultados de un experimento con respecto a la media de la variable dependiente, como son las siguientes.

- El error aleatorio con origen en el proceso de medición.
- Errores en los factores controlados, por ejemplo:
 - Desarrollo del método propuesto para la experiencia.
 - Fallas en los equipos e instrumentos.
 - Condiciones del medio ambiente.
 - Condiciones del analista.

El análisis de ANOVA permite realizar estudios comparativos entre más de dos muestras (grupos) que han sido sometidas a diferentes tratamientos con base en establecer qué tan grande es la variación de los datos de una muestra con respecto a su media (variación dentro de grupos), con respecto a la variación entre las medias muestrales y con respecto a la media general (variación entre grupos).

La premisa anterior permite definir las siguientes hipótesis estadísticas:

- $H_0 = \mu_1 = \mu_2 = \dots = \mu_n$ No hay diferencia entre las medias.
- $H_1 = \mu_1 \neq \mu_2 \neq \dots \neq \mu_n$ Hay diferencia entre las medias.

La prueba de hipótesis bajo ANOVA se fundamenta en la distribución F de Fisher, la cual es una distribución de probabilidad continua, también conocida como distribución F de Snedecor (por George Snedecor).

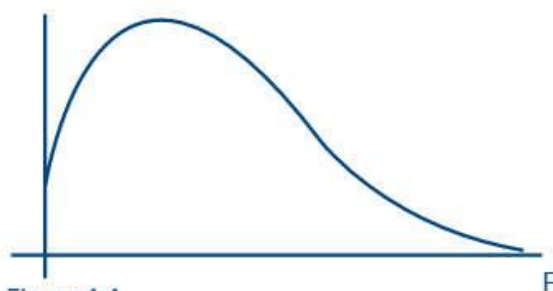


Figura 4.4



Alerta

La prueba de hipótesis por medio de ANOVA se fundamenta en la distribución F de Fisher.

La distribución F de Fisher es la más conocida por su aplicación a las pruebas de hipótesis estadísticas y el problema más simple del análisis de varianza, y se utiliza debido a que la hipótesis propone que las medias de múltiples poblaciones son iguales si están normalmente distribuidas y con la misma desviación estándar.

De manera concreta, la prueba de hipótesis propuesta consiste en contrastar el valor de una F calculada (F_c) con respecto al valor de una F tabular o teórica (F_t); de hecho, tal como se aprecia en la gráfica siguiente, cuando $F_t < F_c$ se cumple H_1 .

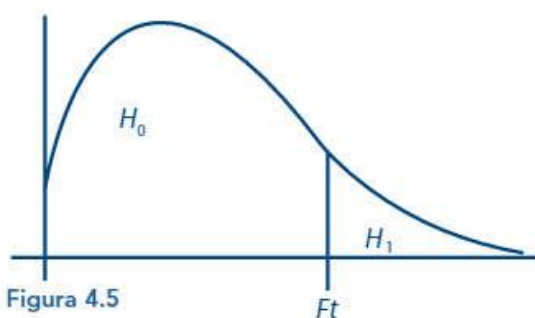


Figura 4.5

F_t

El procedimiento de cálculo de ANOVA parte de la determinación de valor de las variaciones. El valor de la variación dentro de grupos se calcula así:

$$Var_{DG} = \sum_{i=1}^n (X_i - \bar{X})^2$$

Mientras que el valor de la variación entre grupos se determina por la fórmula:

$$Var_{EG} = \sum_{j=1}^n (X_n - \bar{\bar{X}})^2$$

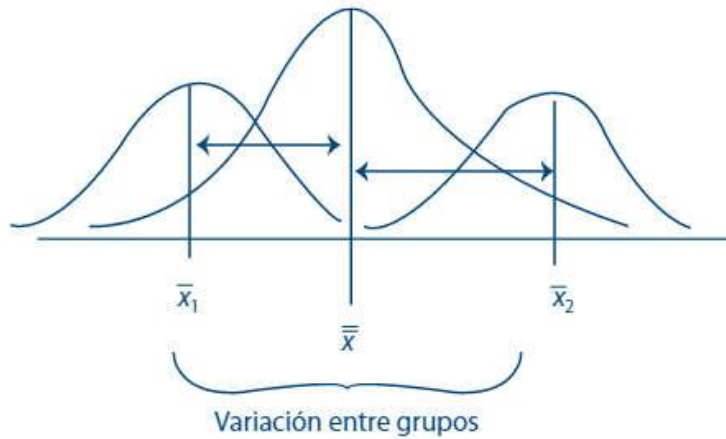


Figura 4.6

En consecuencia, el valor de la F calculada se determina así:

$$F_c = \frac{\frac{Var_{EG}}{N-1}}{\frac{Var_{DG}}{N-Nm}}$$

Donde

N = el número total de datos aportados por las muestras.

Nm = el número de muestras.

Var_{EG} = la variación entre grupos.

Var_{DG} = la variación dentro de grupos.

Para determinar el valor de la F tabular se requiere conocer el valor del error (α), así como los grados de libertad a efecto de ubicar el valor en la tabla correspondiente:

$$F_t [\alpha, (Nm - 1), (N - Nm)]$$

donde los valores de los grados de libertad están definidos así:

$Nm - 1$ = denominados grados de libertad del numerador.

$N - Nm$ = denominados grados de libertad del denominador.

Para ejemplificar lo anterior se presenta el siguiente problema.

Problema resuelto

Realiza una prueba de hipótesis por ANOVA del experimento de las resinas para cartuchos de escopeta a efecto de determinar si existe diferencia en el peso de los cartuchos con base en el tipo de resina utilizada. Considera un nivel de confianza del 95%.

Solución

Se procede a calcular las variaciones dentro de grupos y entre grupos.

$$\bar{X}_{RA} = 32.373$$

$$\bar{X}_{RA1} = 32.170$$

RA	RA1	$(X^i - \bar{X}_{RA})^2$	$(X^i - \bar{X}_{RA1})^2$
32.30	31.89	0.005329	0.0784
31.90	32.10	0.223729	0.0049
31.45	32.34	0.851929	0.0289
32.80	32.40	0.182329	0.0529
33.01	32.30	0.405769	0.0169
32.78	31.99	0.165649	0.0324
	Σ	1.834734	0.2144

De manera que

$$Var_{DG} = 1.8313 + 0.2144 = 2.0457$$

$$Var_{EG} = 6(32.373 - 32.2716)^2 + 6(32.17 - 32.2716)^2 = 0.1236$$

$$F_c = \frac{\frac{Var_{EG}}{Nm-1}}{\frac{Var_{DG}}{N-Nm}} = \frac{\frac{0.1236}{2-1}}{\frac{2.0457}{12-2}} = 0.6037$$

De la tabla de la distribución F de Fisher para $\alpha = 5\%$:

$$F_t[0.5, 1, 10] = 4.9650$$

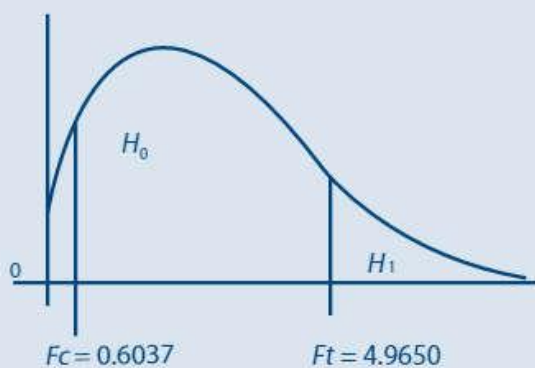


Figura 4.7

Como $F_t > F_c$ entonces se comprueba H_0 , lo que significa que no hay diferencia entre las medias, o sea que el tipo de resina del cuerpo de los cartuchos no provoca que haya variaciones en sus pesos.

■ Limitaciones de las pruebas de hipótesis fundamentadas en ANOVA

El resultado de la prueba tan solo señala si existe o no diferencia, pero en caso de que se cuente con más de dos muestras y la prueba señale que hay diferencia, la prueba no discrimina cuál muestra es, por lo que es necesario llevar a cabo otro tipo de procedimientos de cálculo para discriminar entre las muestras.

4.7 Estudios comparativos basados en ANOVA de dos factores

Este tipo de estudios sirven para evaluar el comportamiento individual y el conjunto de dos o más variables (factores) sobre una variable dependiente cuantitativa. El análisis permite observar la incidencia o efectos de cada variable en lo individual, así como la interacción entre ambas; es decir, se estudian tres efectos. Por señalar si se contará con el análisis de tres variables (factores), entonces los efectos serían siete: tres individuales, tres por pares de factores y uno de interacción entre ellos.

Para precisar considérese el caso que se cuente con m muestras con tamaño n dispuestas de forma matricial, donde uno de los factores en estudio, denotado por la letra a estará referido por los renglones, mientras que el segundo factor, denominado b , estará referido por las columnas.

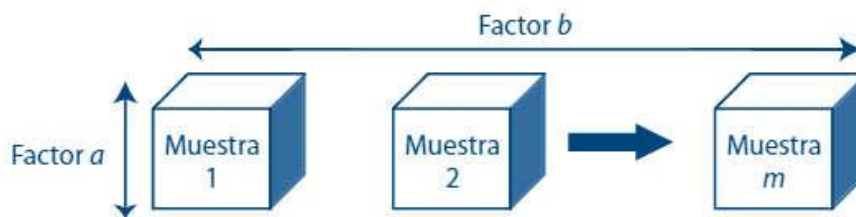


Figura 4.8

De manera práctica el procedimiento de cálculo se expone mediante el siguiente ejemplo.

Una empresa pretende innovar en el proceso de fabricación de uno de sus productos, por lo que prueba tres presentaciones diferentes del insumo principal: barra, granulado y en polvo. Ensayó cada presentación en las ocho máquinas que tiene. La gerencia de manufactura se interroga si existen diferencias significativas en el nivel de producción dependiendo del tipo de presentación del insumo y del tipo de máquina.

Tabla 4.1				
Factor <i>b</i> Tipo de insumo				
←				
Factor <i>a</i> Máquina		Barra	Granulado	En polvo
	Máquina 1	28	31	30
	Máquina 2	26	29	31
	Máquina 3	26	27	29
	Máquina 4	24	27	28
	Máquina 5	20	25	27
	Máquina 6	22	23	28
	Máquina 7	24	25	27
	Máquina 8	25	26	29

Con el enfoque de ANOVA de dos factores el tamaño muestral deberá entenderse por el número de elementos que se ubican bajo la relación renglón-columna, que en el caso del ejemplo es uno, por lo que el tamaño total de observaciones está definido como: $N = abn$.



Alerta

Al ANOVA de dos factores también se le conoce como prueba ANOVA de dos vías.



Alerta

La suma de cuadrados del factor a se basa en los cuadrados de las columnas, y la suma de cuadrados del factor b se basa en los cuadrados de los renglones.

Por lo que para calcular el ANOVA se deben obtener las sumas de los cuadrados de los efectos de cada factor.

- a) Para el factor a (renglones).

$$SC_a = \frac{\sum_{j=1}^b \text{Sum_Col}_j^2}{b \cdot n} - \frac{\text{Sum_Total}^2}{N}$$

- b) Para el factor b (columnas).

$$SC_b = \frac{\sum_{i=1}^a \text{Sum_Reng}_i^2}{a \cdot n} - \frac{\text{Sum_Total}^2}{N}$$

- c) Se calcula el valor de los cuadrados totales.

$$SC_{Tot} = \sum_{i=1}^a \sum_{j=1}^b Y_{ij}^2 - \frac{\text{Sum_Total}^2}{N}$$

- d) Se calcula el valor de los cuadrados del error o simplemente el error.

$$SC_{Error} = SC_{Tot} - [SC_a + SC_b]$$

Con base en los parámetros anteriores se estructura el cuadro de análisis.

Tabla 4.2 Análisis de varianza

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F_{calc}	F_{tab}	Prob.
Factor a	SC_a	$(a-1)$	$\frac{SC_a}{a-1} = CM_a$	$\frac{CM_a}{CM_e}$	F_{tab_a}	Pa
Factor b	SC_b	$(b-1)$	$\frac{SC_b}{b-1} = CM_b$	$\frac{CM_b}{CM_e}$	F_{tab_b}	Pb
Error	SC_{Error}	$(a-1)(b-1)$	$\frac{SC_{error}}{(a-1)(b-1)} = CM_e$			
Total	SC_{Tot}	$N-1$				

En la tabla 4.2 llama la atención la columna titulada Promedio de los cuadrados, también denominada Cuadrados medios y regularmente se simboliza por CM .

La información contenida en el cuadro propone que habrá dos pruebas de hipótesis, bajo el criterio expuesto anteriormente para la F de Fisher, una para determinar si el efecto del factor a genera o no diferencias y otra para determinar si el efecto del factor b genera o no diferencias.

Con el propósito de mostrar de manera aplicada el procedimiento anterior, considérese el planteamiento del problema propuesto como ejemplo.



Alerta

ANOVA para dos factores propone dos pruebas de hipótesis bajo la F de Fisher, una para el factor a y otra para el factor b .

Problema resuelto

Una empresa pretende innovar en el proceso de fabricación de uno de sus productos, por lo que prueba tres presentaciones diferentes del insumo principal: barra, granulado y en polvo. Ensayó cada presentación en las ocho máquinas de las que dispone, de manera que la gerencia de manufactura se interroga si existen diferencias significativas en el número de unidades producidas dependiendo del

Problema resuelto (continuación)

tipo de presentación del insumo y del tipo de máquina. Procede a realizar un análisis de prueba de hipótesis estadísticas por ANOVA de dos factores de una muestra con un nivel de significancia del 95%.

Tabla 4.3

		Factor b Tipo de insumo		
Factor a Máquina		Barra	Granulado	En polvo
	Máquina 1	28	31	30
	Máquina 2	26	29	31
	Máquina 3	26	27	29
	Máquina 4	24	27	28
	Máquina 5	20	25	27
	Máquina 6	22	23	28
	Máquina 7	24	25	27
	Máquina 8	25	26	29

Solución

Se realizan los cálculos correspondientes para la determinación de las variaciones. Los resultados se muestran en la siguiente tabla.

Tabla 4.4

	Barra	Granulado	En polvo	\bar{Y}_j	Sum_Reng	(Sum_Reng) ²
M1	28.00	31.00	30.00	29.667	89.00	7 921.00
M2	26.00	29.00	31.00	28.667	86.00	7 396.00
M3	26.00	27.00	29.00	27.333	82.00	6 724.00
M4	24.00	27.00	28.00	26.333	79.00	6 241.00
M5	20.00	25.00	27.00	24.000	72.00	5 184.00
M6	22.00	23.00	28.00	24.333	73.00	5 329.00
M7	24.00	25.00	27.00	25.333	76.00	5 776.00
M8	25.00	26.00	29.00	26.667	80.00	6 400.00
\bar{Y}_j	24.37	26.62	28.62	$\Sigma (Sum_Reng)^2 =$		50 971.00
Sum_Col	195.00	213.00	229.00			
(Sum_Col) ²	38 025.00	45 369.00	52 441.00			
$\Sigma (Sum_Col)^2 = 135 835.000$						

Por tanto,

a) Promedio de promedios

$$\bar{\bar{Y}} = 26.540$$

Solución (continuación)

b) Suma total y suma total al cuadrado

$$Sum_Total = \sum_{i=1}^a \sum_{j=1}^b Y_{ij} = (28 + 31 + 30 + \dots + 28 + 29) = 637$$

$$Sum_Total^2 = \left(\sum_{i=1}^a \sum_{j=1}^b Y_{ij} \right)^2 = (637)^2 = 405\,769$$

c) Suma de cuadrados

$$\sum_{i=1}^a \sum_{j=1}^b Y_{ij}^2 = (28^2 + 31^2 + 30^2 + \dots + 28^2 + 29^2) = 17\,081$$

d) Suma de cuadrados del factor a

$$SC_a = \frac{\sum_{j=1}^b Sum_Col_j^2}{b \cdot n} - \frac{Sum_Total^2}{N} = \frac{50\,971}{3 \cdot 1} - \frac{405\,769}{8 \cdot 3 \cdot 1} = 83.292$$

e) Suma de cuadrados del factor b

$$SC_b = \frac{\sum_{i=1}^a Sum_Reng_i^2}{a \cdot n} - \frac{Sum_Total^2}{N} = \frac{135\,835}{8 \cdot 1} - \frac{405\,769}{8 \cdot 3 \cdot 1} = 72.333$$

f) Suma de cuadrados totales

$$SC_{Tot} = \sum_{i=1}^a \sum_{j=1}^b Y_{ij}^2 - \frac{Sum_Total^2}{N} = 17\,081 - \frac{405\,769}{8 \cdot 3 \cdot 1} = 173.958$$

g) Se calcula el valor del cuadrado del error o simplemente el error.

$$SC_{Error} = SC_{Tot} - [SC_a + SC_b] = 173.958 - [83.292 + 72.333] = 18.333$$

Por lo que, con los parámetros calculados se forma el cuadro de análisis.

Tabla 4.5 Análisis de la varianza

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados
Factor a	83.291	$(8 - 1) = 7$	$CM_a = \frac{83.291}{7} = 11.898$
Factor b	72.333	$(3 - 1) = 2$	$CM_b = \frac{72.333}{2} = 36.167$
Error	18.333	$(8 - 1)(3 - 1) = 14$	$CM_e = \frac{18.333}{14} = 1.309$
Total	173.958	$24 - 1 = 23$	

Tabla 4.6 Origen de las variaciones

Origen de las variaciones	F_{calc}	F_{tab}
Factor a	$\frac{CM_a}{CM_e} = \frac{11.898}{1.309} = 9.089$	$F_{tab_a} = 2.764$
Factor b	$\frac{CM_b}{CM_e} = \frac{36.167}{1.309} = 27.695$	$F_{tab_b} = 3.783$

Solución (continuación)

A efecto de comprobar los cálculos anteriores, se presenta la tabla de resumen de resultados de una prueba de análisis de la varianza de dos factores con una sola muestra por grupo.

Tabla 4.7 Análisis de la varianza

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Prob.	Valor crítico para F
Filas	83.291	7	11.898	9.0863	0.000272	2.7641
Columnas	72.333	2	36.166	27.618	1.382E-05	3.7388
Error	18.333	14	1.309			
Total	173.958	23				

El análisis de los resultados, en relación con la prueba de hipótesis estadística, muestra que tanto en el factor a como en el factor b se cumplen las hipótesis H_1 (hay diferencia), lo que quiere decir que los factores generan un efecto sobre el número de unidades producidas.



Alerta

El ANOVA para dos factores se puede llevar a cabo a través de Excel mediante la prueba de análisis de la varianza de dos factores con una sola muestra por grupo.

4.8 Estudios comparativos basados en ANOVA de dos factores de varias muestras por grupo

Este tipo de estudios sirven para evaluar el comportamiento individual y en conjunto de dos factores sobre una variable dependiente cuantitativa, pero considerando que los tratamientos pueden ser estudiados a través de varias muestras.

Tal como se expuso, el análisis permite observar la incidencia o efectos de cada variable individual, así como la interacción entre ambas; es decir, se estudian tres efectos.

Para precisar considérese el caso que se cuente con m muestras con tamaño n dispuestas de forma matricial, donde uno de los factores en estudio, denotado por la letra a , estará definido por un número de ensayos que componen la muestra, mientras que el segundo factor, denominado b , estará referido por las columnas.

Tabla 4.8		Factor b		
Factor a	Cantidad X	Barra	Granulado	En polvo
		28.000	31.000	30.000
		26.000	29.000	31.000
		26.000	27.000	29.000
	Cantidad Y	24.000	27.000	28.000
		20.000	25.000	27.000
		22.000	23.000	28.000
		24.000	25.000	27.000
		25.000	26.000	29.000



Alerta

El ANOVA de dos factores con varias muestras por grupo refiere que el factor a cuenta con varios ensayos.

De manera práctica, el procedimiento de cálculo se expone al tener como base el ejemplo del procedimiento anterior y en el siguiente planteamiento: la empresa está interesada en analizar el impacto de la cantidad de insumo según la presentación en el número de piezas producidas, así como en saber si de manera combinada también inciden.

Obsérvese que se ensayan dos cantidades diferentes en las que la relación renglón-columna cuenta con un número de cuatro ensayos, es decir, el tamaño muestral es de cuatro elementos.

De manera que el procedimiento de cálculo bajo el enfoque de ANOVA de dos factores de varias muestras requiere del cálculo de las sumas de los cuadrados de los efectos de cada factor, así como en forma conjunta, tal como se expone a continuación.

- a) Para el factor a (renglones):

$$SC_a = \frac{\sum_{j=1}^b \text{Sum_Reng}^2}{b \cdot n} - \frac{\text{Sum_Total}^2}{N}$$

- b) Para el factor b (columnas):

$$SC_b = \frac{\sum \text{Sum_Col}^2}{a \cdot n} - \frac{\text{Sum_Total}^2}{N}$$

- c) Suma de cuadrados de ambos factores:

$$SC_{ab} = \frac{\sum Am^2}{n} - \left(\frac{\text{Sum_Total}^2}{N} + SC_a + SC_b \right)$$

Am es el acumulado por muestra.

- d) Se calcula el valor de los cuadrados totales:

$$SC_{Tot} = \sum_{i=1}^a \sum_{j=1}^b Y_{ij}^2 - \frac{\text{Sum_Total}^2}{N}$$

- e) Se calcula el valor de los cuadrados del error o simplemente el error:

$$SC_{Error} = SC_{Tot} - [SC_a + SC_b + SC_{ab}]$$

Con base en los parámetros anteriores se estructura el cuadro de análisis.

Tabla 4.9 Análisis de la varianza

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados
Factor a	SC_a	$(a-1)$	$\frac{SC_a}{a-1} = CM_a$
Factor b	SC_b	$(b-1)$	$\frac{SC_b}{b-1} = CM_b$
Interacción ab	SC_{ab}	$(a-1)(b-1)$	$\frac{SC_{ab}}{(a-1)(b-1)} = CM_{ab}$
Error	SC_{Error}	$N - (ab)$	$\frac{SC_{error}}{N-(ab)} = CM_e$
Total	SC_{Tot}	$N-1$	

Tabla 4.10			
Origen de las variaciones	F_{calc}	F_{tab}	Prob.
Factor a	$\frac{CM_a}{CM_e}$	F_{tab_a}	Pa
Factor b	$\frac{CM_b}{CM_e}$	F_{tab_b}	Pb
Interacción ab	$\frac{CM_{ab}}{CM_e}$	F_{tab_b}	Pab
Error			
Total			

La información contenida en el cuadro propone que habrá tres pruebas de hipótesis, bajo el criterio expuesto para la F de Fisher: una para determinar si el efecto del factor a genera o no diferencias; una segunda para determinar si el efecto del factor b origina o no diferencias; y una tercera para determinar si la interacción de los factores genera o no diferencias.

Con el propósito de mostrar de manera aplicada el procedimiento anterior, considérese el planteamiento del problema propuesto como ejemplo.

Alerta

El ANOVA para dos factores de varias muestras propone tres pruebas de hipótesis bajo la F de Fisher: una primera para el factor a ; una segunda para el factor b ; y una tercera para la interacción de los dos factores.

Problema resuelto

La empresa está interesada en analizar el impacto de la cantidad de insumo dependiendo de la presentación en el número de piezas producidas, así como en saber si de manera combinada también inciden. Para ello realiza un análisis de prueba de hipótesis estadísticas por ANOVA de dos factores de una muestra con un nivel de significancia del 95%.

Tabla 4.11				
		Factor b		
		Barra	Granulado	En polvo
Factor a	Cantidad X	28.000	31.000	30.000
		26.000	29.000	31.000
		26.000	27.000	29.000
		24.000	27.000	28.000
	Cantidad Y	20.000	25.000	27.000
		22.000	23.000	28.000
		24.000	25.000	27.000
		25.000	26.000	29.000

Solución

Se realizan los cálculos correspondientes para la determinación de las variaciones. Los resultados se fundamentan en los obtenidos del procedimiento anterior complementándolos como se muestra. Se calculan los valores que se muestran en la siguiente tabla.

Tabla 4.12							
	Barra		Granulado		En polvo		Suma de renglones
Cantidad X	28.000		31.000		30.000		
	26.000	Am_1	29.000	Am_2	31.000	Am_3	
	26.000	104.000	27.000	114.000	29.000	118.000	336.000
	24.000		27.000		28.000		
Cantidad Y	20.000		25.000		27.000		
	22.000	Am_4	23.000	Am_5	28.000	Am_6	
	24.000	91.000	25.000	99.000	27.000	111.000	301.000
	25.000		26.000		29.000		
Suma de columnas	195.000		213.000		229.000		Total 637
Cantidad X	\bar{X}_{CX}	26.000	28.500		29.500		
Cantidad Y	\bar{X}_{CY}	22.750	24.750		27.750		
	\bar{x}	26.542					

$$\Sigma(\text{Sum_Col})^2 = 195^2 + 213^2 + 229^2 = 135\,835$$

$$\Sigma(\text{Sum_Reng})^2 = 336^2 + 301^2 = 203\,497$$

$$\Sigma(Am)^2 = 104^2 + 114^2 + 118^2 + 91^2 + 99^2 + 111^2 = 68\,139$$

Considérese que el factor a está conformado por dos grupos, mientras que el factor b , por tres grupos, por tanto:

$$a = 2 \qquad b = 3$$

Adicionalmente, la relación renglón-columna expone que se compone de cuatro elementos, por lo que el tamaño muestral (n) es 4.

Por consiguiente,

a) Suma total y suma total al cuadrado:

$$\text{Sum_Total} = \sum_{i=1}^a \sum_{j=1}^b Y_{ij} = (28 + 31 + 30 + \dots + 28 + 29) = 637$$

$$\text{Sum_Total}^2 = \left(\sum_{i=1}^a \sum_{j=1}^b Y_{ij} \right)^2 = (637)^2 = 405\,769$$

Solución (continuación)

b) Suma de cuadrados del factor a:

$$SC_a = \frac{\sum Sum_Reng^2}{b \cdot n} - \frac{Sum_Total^2}{N} = \frac{203\,497}{3 \cdot 4} + \frac{405\,769}{2 \cdot 3 \cdot 4} = 51.0417$$

c) Suma de cuadrados del factor b:

$$SC_b = \frac{\sum Sum_Col^2}{a \cdot n} - \frac{Sum_Total^2}{N} = \frac{135\,835}{2 \cdot 4} + \frac{405\,769}{2 \cdot 3 \cdot 4} = 72.333$$

d) Suma de cuadrados de la interacción de factores:

$$SC_{ab} = \frac{\sum Am^2}{n} - \left(\frac{Sum_Total^2}{N} + SC_a + SC_b \right) =$$

$$= \frac{68\,139}{4} - \left(\frac{405\,769}{24} + 51.0417 + 72.333 \right) = 4.333$$

e) Suma de cuadrados totales:

$$SC_{Tot} = \sum_{i=1}^a \sum_{j=1}^b Y_{ij}^2 = \frac{Sum_Total^2}{N} = 17\,081 - \frac{405\,769}{8 \cdot 3 \cdot 1} = 173.958$$

f) Se calcula el valor del cuadrado del error o simplemente el error.

$$SC_{Error} = SC_{Tot} - [SC_a + SC_b + SC_{ab}]$$

$$SC_{Error} = SC_{Tot} - [SC_a + SC_b + SC_{ab}]$$

$$= 173.958 - [51.0417 + 72.333 + 4.333] = 46.275$$

Tabla 4.13 Análisis de la varianza

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados
Factor a	51.0417	$(a-1) = 2-1 = 1$	$CM_a = \frac{51.0417}{1} = 51.0417$
Factor b	72.333	$(b-1) = 3-1 = 2$	$CM_b = \frac{72.333}{2} = 36.1666$
Interacción ab	4.3333	$(a-1)(b-1) = (2-1)(3-1) = 2$	$CM_{ab} = \frac{4.333}{2} = 2.1666$
Error	46.275	$N - (ab) = 24 - (2 \cdot 3) = 18$	$CMe = \frac{46.275}{18} = 2.5708$
Total	173.958	$N - 1 = 24 - 1 = 23$	

Alerta

El ANOVA para dos factores se puede llevar a cabo a través de Excel mediante la prueba de análisis de la varianza de dos factores con varias muestras por grupo.

Solución (continuación)

Tabla 4.14

Origen de las variaciones	F_{calc}	F_{tab}
Factor a	$\frac{CM_a}{CM_e} = 19.8544$	4.4139
Factor b	$\frac{CM_b}{CM_e} = 14.0682$	3.5546
Interacción ab	$\frac{CM_{ab}}{CM_e} = 0.8427$	3.5546
Error		
Total		

De primera instancia se interpreta que la cantidad de material aplicado al proceso de fabricación sí genera diferencias en el número de unidades producidas.

En cuanto al tipo de presentación del insumo, se puede señalar que éste también genera diferencias en el número de unidades producidas. Sin embargo, la interacción de los factores expone que no deberían generar diferencias.

Asimismo, se muestra la tabla generada por Excel para el análisis de la varianza de dos factores con varias muestras por grupo a efecto de corroborar los resultados obtenidos.

Tabla 4.15 Análisis de la varianza

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Prob.	Valor crítico para F
Muestra	51.042	1	51.042	19.865	0.000	4.414
Columnas	72.333	2	36.167	14.076	0.000	3.555
Interacción	4.333	2	2.167	0.843	0.447	3.555
Dentro del grupo	46.250	18	2.569			
Total	173.958	23				

4.9 Discriminantes para pruebas de hipótesis basadas en ANOVA

Las pruebas de hipótesis por ANOVA pueden establecerse de orden general, ya que tan solo señalan si se aprueba o refuta la hipótesis nula. Sin embargo, cuando se cuenta con más de dos grupos y se aprueba la hipótesis alternativa surge la interrogante: ¿entre qué grupos existe la diferencia? Un proceso lógico sería realizar pruebas de ANOVA por pares de grupos a efecto de determinar los grupos que difieren entre sí, pero esta alternativa de cálculo puede resultar larga y tediosa, por lo que en su momento se crearon procesos de discriminación fundamentados en los cuadrados promedio (varianzas) y la distribución t .

Para propósitos de los contenidos de esta obra se hará referencia a dos discriminantes: el factor de Diferencia Mínima Significativa (LSD , *Least Significant Difference*), así como la prueba de rangos múltiples de Duncan.

Alerta

Utilizar muestras pequeñas expone que el valor de la varianza poblacional se considera desconocida, pero puede ser sustituida por el promedio de los cuadrados del error.

■ Discriminante LSD para prueba de hipótesis de un factor con más de dos grupos

El análisis por el discriminante LSD se fundamenta en el supuesto de que la desviación estándar poblacional (σ) es desconocida, pero puede sustituirse por $\sqrt{CM_e}$, al considerar los resultados obtenidos a

través del cálculo del análisis de la varianza correspondiente, de manera que se puede complementar con la aplicación de la distribución *t* de Student considerando un factor de escala $\sqrt{\frac{CM_e}{n}}$.

Se entiende que el valor del CM_e corresponde al promedio de los cuadrados dentro de grupos.

De manera que cuando el análisis contempla grupos con el mismo tamaño muestral n , el valor del discriminante LSD se determina mediante la fórmula

$$LSD = t_{\frac{\alpha}{2}, N-N_m} \sqrt{\frac{2CM_e}{n}}$$

Pero cuando los grupos cuentan con tamaños muestrales diferentes, el valor del discriminante se obtiene al aplicar la fórmula

$$LSD = t_{\frac{\alpha}{2}, N-N_m} \sqrt{CM_e \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Donde

α = valor del error

$N - N_m$ = los grados de libertad del error (variación dentro de grupos)

n_i = tamaño muestral del primer grupo

n_j = tamaño muestral del segundo grupo

El criterio de discriminación es el siguiente:

Si $|\bar{y}_i - \bar{y}_j| > LSD$ se concluye que existe diferencia entre las medias poblacionales μ_i y μ_j .

Para ejemplificar el proceso de cálculo, considérese el ejemplo sobre la resina de los cartuchos de escopeta, pero ténganse en cuenta las tres resinas.

Problema resuelto

Realiza un análisis comparativo simple de ANOVA para determinar si existe diferencia en los cartuchos dependiendo del tipo de resina del cuerpo si se considera el peso, en gramos, de seis cartuchos por muestra. Considera un nivel de confianza del 95%.

Tabla 4.16		
Resina A (RA)	Resina A-1 (RA1)	Resina A-2 (RA2)
32.30	31.89	32.10
31.90	32.10	32.00
31.45	32.34	31.90
32.80	32.40	32.60
33.01	32.30	32.66
32.78	31.99	32.39

Alerta

El análisis por *LSD* de un factor se puede realizar tanto en el sentido vertical como en el horizontal, dependiendo de las necesidades de análisis.

Solución

La solución se logra mediante Excel: análisis de la varianza de un factor.

Tabla 4.17 Análisis de la varianza de un factor

Resumen				
Grupos	Cuenta	Suma	Promedio	Varianza
RA	6	194.24	32.3733	0.3669
RA1	6	193.02	32.1700	0.0429
RA2	6	193.63	32.2717	0.1029

Análisis de la varianza						
Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Prob.	Valor crítico para F
Entre grupos	0.1240	2	0.0620	0.3629	0.7016	3.6823
Dentro de los grupos	2.5636	15	0.1709			
Total	2.6876	17				

Donde se identifica que el valor del $CM_e = 0.1709$.

$$t_{\frac{\alpha}{2}, N-N_m} = t_{\frac{0.05}{2}, (18-3)} = t_{0.025, 15} = 2.131$$

Por lo que el valor del discriminante es:

$$LSD = t_{\frac{\alpha}{2}, N-N_m} \sqrt{\frac{2CM_e}{n}} = (2.131) \sqrt{\frac{(2)(0.1709)}{6}} = 0.2387$$

Se realiza el proceso de discriminación como sigue:

$$\bar{Y}_{RA} = 32.3733$$

$$\bar{Y}_{RA1} = 32.1700$$

$$\bar{Y}_{RA2} = 32.2717$$

$$|\bar{Y}_{RA} - \bar{Y}_{RA1}| = |32.3733 - 32.1700| = 0.2033 < LSD \quad \text{no hay diferencia}$$

$$|\bar{Y}_{RA} - \bar{Y}_{RA2}| = |32.3733 - 32.2717| = 0.1016 < LSD \quad \text{no hay diferencia}$$

$$|\bar{Y}_{RA1} - \bar{Y}_{RA2}| = |32.1700 - 32.2717| = 0.1017 < LSD \quad \text{no hay diferencia}$$

Se concluye que los tipos de resina no ofrecen diferencia entre los grupos.

■ Discriminante *LSD* para prueba de hipótesis de dos factores con más de dos grupos y varias muestras por grupo

Bajo este criterio se puede señalar que tanto un factor *a* puede generar diferencias entre las medias poblacionales como las puede generar un factor *b*, por lo que se tendrá que calcular un determinante *LSD* para *a* y un discriminante para el factor *b*.

De manera concreta, el procedimiento expone comparar las medias muestrales de los grupos por pares a efecto de precisar entre qué grupos existen las diferencias.

Para el factor a el valor del discriminante se calcula mediante la fórmula

$$LSD_a = t_{\frac{\alpha}{2}, ab(n-1)} \sqrt{CM_e \left(\frac{1}{n_{ai}} + \frac{1}{n_{aj}} \right)}$$

Asimismo, para el factor b el valor del discriminante se calcula mediante la fórmula

$$LSD_b = t_{\frac{\alpha}{2}, ab(n-1)} \sqrt{CM_e \left(\frac{1}{n_{bi}} + \frac{1}{n_{bj}} \right)}$$

Donde

α = valor del error

$ab(n-1)$ = grados de libertad del error (variación dentro de grupos)

n_{ai} o n_{bi} = tamaño muestral del primer grupo

n_{aj} o n_{bj} = tamaño muestral del segundo grupo

Se puede generar el discriminante para la interacción de los factores:

$$LSD_{ab} = t_{\frac{\alpha}{2}, ab(n-1)} \sqrt{CM_e \left(\frac{1}{n} + \frac{1}{n} \right)}$$

donde n es el número de observaciones en la relación factor a -factor b .

El criterio de discriminación para cualquiera de los factores es el siguiente.

Si $|\bar{y}_i - \bar{y}_j| > LSD$ se concluye que existe diferencia entre las medias poblacionales μ_i y μ_j . Debe apuntarse que en el análisis de la interacción las medias corresponden al promedio basado en los AG.

Nótese que en las fórmulas se consideran los tamaños muestrales de los grupos en análisis de manera que puedan compararse estos en caso de que el tamaño muestral no sea homogéneo.

Para ejemplificar lo anterior, considérese de nuevo el problema de los cartuchos del punto anterior pero bajo las condiciones que se exponen.

Problema resuelto

Realiza un análisis comparativo simple para determinar si existe diferencia en el peso de los cartuchos dependiendo del tipo de resina del cuerpo y de pólvora. Se consideran muestras de tres elementos y un nivel de confianza del 95%.

Tabla 4.18

	Resina A (RA)	Resina A-1 (RA1)	Resina A-2 (RA2)
Pólvora 1 (P1)	32.30	31.89	32.10
	31.90	32.10	32.00
	31.45	32.34	31.90
Pólvora 2 (P2)	32.80	32.40	32.60
	33.01	32.30	32.66
	32.78	31.99	32.39

Solución

La solución se logra mediante Excel: análisis de la varianza de dos factores de varias muestras por grupo.

Tabla 4.19

	Resina RA		Resina RA1		Resina RA2	
Pólvera P1	32.30	Amp1-RA	31.89	Amp1-RA1	32.10	Amp1-RA2
	31.90	31.88	32.10	32.11	31.96	31.98
	31.45		32.34		31.90	
Pólvera P2	32.80	Amp2-RA	32.40	Amp2-RA1	32.60	Amp2-RA2
	33.01	32.86	32.30	32.23	32.72	32.57
	32.78		31.99		32.39	

Tabla 4.20 Análisis de la varianza de dos factores con varias muestras por grupo

Resumen	RA	RA1	RA2	Total
Cuenta	3	3	3	9
Suma	95.6500	96.3300	95.9900	287.9700
Promedio	31.8833	32.1100	31.9967	31.9967
Varianza	0.1808	0.0507	0.0100	0.0700
Cuenta	3	3	3	9
Suma	98.5900	96.6900	97.6400	292.9200
Promedio	32.8633	32.2300	32.5467	32.5467
Varianza	0.0162	0.0457	0.0204	0.0958
Total				
Cuenta	6	6	6	
Suma	194.2400	193.0200	193.6300	
Promedio	32.3733	32.1700	32.2717	
Varianza	0.3669	0.0429	0.1029	

Tabla 4.21 Análisis de la varianza

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Prob.	Valor crítico para F
Muestra	1.3612	1	1.3612	25.2213	0.0003	4.7472
Columnas	0.1240	2	0.0620	1.1490	0.3495	3.8853
Interacción	0.5547	2	0.2774	5.1388	0.0244	3.8853
Dentro del grupo	0.6477	12	0.0540			

Solución (continuación)

Por tanto, el valor de los discriminantes para los dos factores es el siguiente

$$t_{\frac{\alpha}{2}, ab(n-1)} = t_{\frac{0.05}{2}, (2)(3)(3-1)} = t_{0.025, 12} = 2.179$$

$$LSD_a = t_{\frac{\alpha}{2}, ab(n-1)} \sqrt{CM_e \left(\frac{1}{n_{ai}} + \frac{1}{n_{aj}} \right)} = (2.179) \sqrt{(0.0540) \left(\frac{1}{9} + \frac{1}{9} \right)} = 0.2387$$

$$|\bar{y}_{P1} - \bar{y}_{P2}| = |31.9967 - 32.5467| = 0.55 > LSD \quad \text{existe diferencia}$$

Asimismo, para el factor b el valor del discriminante se calcula mediante la fórmula

$$LSD_b = t_{\frac{\alpha}{2}, ab(n-1)} \sqrt{CM_e \left(\frac{1}{n_{bi}} + \frac{1}{n_{bj}} \right)} = (2.179) \sqrt{(0.0540) \left(\frac{1}{6} + \frac{1}{6} \right)} = 0.2923$$

$$\bar{y}_{RA} = 32.3733$$

$$\bar{y}_{RA1} = 32.1700$$

$$\bar{y}_{RA2} = 32.2717$$

$$|\bar{y}_{RA} - \bar{y}_{RA1}| = |32.3733 - 32.1700| = 0.2033 < LSD \quad \text{no hay diferencia}$$

$$|\bar{y}_{RA} - \bar{y}_{RA2}| = |32.3733 - 32.2717| = 0.1016 < LSD \quad \text{no hay diferencia}$$

$$|\bar{y}_{RA1} - \bar{y}_{RA2}| = |32.1700 - 32.2717| = 0.1017 < LSD \quad \text{no hay diferencia}$$

Para el análisis LSD de la interacción se determinan los valores de los promedios por cada conjunto de observaciones ($n = 3$).

$$\bar{y}_{P1-RA} = \frac{31.88}{3} = 10.63$$

$$\bar{y}_{P2-RA} = \frac{32.86}{3} = 10.95$$

$$\bar{y}_{P1-RA1} = \frac{32.11}{3} = 10.70$$

$$\bar{y}_{P2-RA1} = \frac{32.23}{3} = 10.74$$

$$\bar{y}_{P1-RA2} = \frac{31.98}{3} = 10.66$$

$$\bar{y}_{P2-RA2} = \frac{32.57}{3} = 10.86$$

El valor del discriminante de la interacción es

$$LSD_{ab} = t_{\frac{\alpha}{2}, ab(n-1)} \sqrt{CM_e \left(\frac{1}{n} + \frac{1}{n} \right)} = (2.179) \sqrt{(0.0540) \left(\frac{1}{3} + \frac{1}{3} \right)} = 0.4134$$

Al realizar el proceso de discriminación por relación de factores, de manera básica son los siguientes:

$$|\bar{y}_{P1-RA} - \bar{y}_{P2-RA}| = |10.63 - 10.95| = 0.32 < LSD \quad \text{no hay diferencia}$$

$$|\bar{y}_{P1-RA1} - \bar{y}_{P2-RA1}| = |10.66 - 10.74| = 0.08 < LSD \quad \text{no hay diferencia}$$

$$|\bar{y}_{P1-RA2} - \bar{y}_{P2-RA2}| = |10.66 - 10.86| = 0.2 < LSD \quad \text{no hay diferencia}$$

Solución (continuación)

Pueden llevarse a cabo otras tantas discriminaciones dependiendo de los requerimientos de análisis que se consideren oportunos.

La conclusión es que los factores considerados para el experimento no generan una diferencia significativa en el diseño del cartucho; en otras palabras, se pueden fabricar indistintamente con cualquiera de ellos.



Alerta

El análisis por R de Duncan aplica únicamente para análisis de un solo factor con muestras del mismo tamaño.

4.10 Método de discriminación de la R de Duncan

El método creado en 1955 permite comparar pares de medias poblacionales y se conoce como la Prueba de Rangos Múltiples de Duncan para análisis de un solo factor con muestras del mismo tamaño.

El método se fundamenta en el promedio del error estándar definido por:

$$S\bar{y}_i = \sqrt{\frac{CME}{n}}$$

De manera que se procede a encontrar el valor de la R en las tablas de los rangos significativos de Duncan:

$$Rp = r_{\alpha}(p, f)S\bar{y}_i$$

Donde

r_{α} = rangos significativos de Duncan

f = grados de libertad del error

P = valores (2, 3, 4, ... a)

Para realizar el análisis se debe hacer lo siguiente.

1. Calcular el valor de $S\bar{y}_i$.
2. Calcular los valores de $Rp(R_2, R_3, \dots, R_a)$.
3. Ordenar las medias muestrales de menor a mayor.
4. Encontrar las diferencias entre las medias empezando por la de valor más alto contra las restantes en orden descendente.
5. Comparar las diferencias contra los valores de Rp de manera descendente. El criterio de evaluación es:

$$\bar{y}_i - \bar{y}_j > Rp \quad \text{existe diferencia}$$

Para mostrar el procedimiento expuesto considérese el problema de los cartuchos.

Problema resuelto

La fábrica de armas experimenta con otros tipos de resina modificados para generar estabilidad e impermeabilidad a los cartuchos de uno de sus tipos de escopeta. Realiza un análisis comparativo simple, basado en el método de la R de Duncan, para determinar si existe diferencia entre los cartuchos dependiendo del tipo de resina del cuerpo. El nivel de confianza es de 95%.

Problema resuelto (continuación)

Tabla 4.22

Resina A (RA)	Resina A-1 (RA1)	Resina A-2 (RA2)	Resina A-3 (RA3)	Resina A-4 (RA4)
32.400	33.450	32.260	32.703	32.925
32.500	32.980	32.340	32.607	32.740
31.950	32.340	32.200	32.163	32.145
32.800	33.150	31.980	32.643	32.975

Solución

Con base en los datos se determinan las medias de cada grupo.

Tabla 4.23

Resina A (RA)	Resina A-1 (RA1)	Resina A-2 (RA2)	Resina A-3 (RA3)	Resina A-4 (RA4)
32.400	33.450	33.440	33.097	32.925
32.500	32.980	32.420	32.633	32.740
31.950	32.340	32.960	32.417	32.145
32.800	33.150	33.110	33.020	32.975
32.413	32.980	32.983	32.792	32.696

Se procede a determinar el valor del CM_e por medio de la rutina de Excel de análisis de la varianza para un solo factor.

Tabla 4.24 Análisis de la varianza

Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Prob.	Valor crítico para F
Entre grupos	0.8918	4	0.2229	1.4411	0.2689	3.0556
Dentro de los grupos	2.3204	15	0.1547			
Total	3.2122	19				

Calculamos el valor promedio del error estándar

$$S\bar{y}_i = \sqrt{\frac{CM_e}{n}} = \sqrt{\frac{0.1547}{4}} = 0.1967$$

Determinamos los valores de los rangos significativos de Duncan utilizando las tablas

$$r_{0.05}(2,15) = 3.01$$

$$r_{0.05}(3,15) = 3.16$$

$$r_{0.05}(4,15) = 3.25$$

$$r_{0.05}(5,15) = 3.31$$

Solución (continuación)

Por lo que

$$R_2 = r_{0.05}(2, 15) S\bar{y}_i = (3.01)(0.1967) = 0.5961$$

$$R_3 = r_{0.05}(3, 15) S\bar{y}_i = (3.16)(0.1967) = 0.6216$$

$$R_4 = r_{0.05}(4, 15) S\bar{y}_i = (3.25)(0.1967) = 0.6393$$

$$R_5 = r_{0.05}(5, 15) S\bar{y}_i = (3.31)(0.1967) = 0.6511$$

$$R_p = r_{\alpha}(p, f) S\bar{y}_i$$

Al ordenar las medias de menor a mayor

$$\bar{Y}_{RA} = 32.413$$

$$\bar{Y}_{RA4} = 32.696$$

$$\bar{Y}_{RA3} = 32.786$$

$$\bar{Y}_{RA2} = 32.965$$

$$\bar{Y}_{RA1} = 32.980$$

$$RA1 \text{ v.s. } RA: 32.980 - 32.413 = 0.5670 < 0.6511 (R_5) \quad \text{no existe diferencia}$$

$$RA1 \text{ v.s. } RA4: 32.980 - 32.696 = 0.284 < 0.6393 (R_4) \quad \text{no existe diferencia}$$

$$RA1 \text{ v.s. } RA3: 32.980 - 32.786 = 0.194 < 0.6216 (R_3) \quad \text{no existe diferencia}$$

$$RA1 \text{ v.s. } RA2: 32.980 - 32.965 = 0.015 < 0.5961 (R_2) \quad \text{no existe diferencia}$$

$$RA2 \text{ v.s. } RA: 32.965 - 32.413 = 0.5520 < 0.6393 (R_4) \quad \text{no existe diferencia}$$

$$RA2 \text{ v.s. } RA4: 32.965 - 32.696 = 0.2690 < 0.6216 (R_3) \quad \text{no existe diferencia}$$

$$RA2 \text{ v.s. } RA3: 32.965 - 32.786 = 0.1790 < 0.5961 (R_2) \quad \text{no existe diferencia}$$

$$RA3 \text{ v.s. } RA4: 32.796 - 32.696 = 0.1000 < 0.6216 (R_3) \quad \text{no existe diferencia}$$

$$RA2 \text{ v.s. } RA: 32.965 - 32.413 = 0.5520 < 0.5961 (R_2) \quad \text{no existe diferencia}$$

$$RA4 \text{ v.s. } RA: 32.696 - 32.413 = 0.2830 < 0.5961 (R_2) \quad \text{no existe diferencia}$$

La conclusión es que las resinas para el cuerpo del cartucho no proponen ninguna diferencia en su peso.

4.1 Una empresa fabricante de llantas especiales para vehículos de la industria de la construcción cuenta con cuatro plantas donde se fabrica el mismo modelo de llanta con los niveles de producción que se muestran.

Tabla 4.25

Planta A	Planta B	Planta C	Planta D
26	37	41	36
38	34	38	32
32	38	30	30
36	38	26	34
34	44	22	36

En cada planta se reproduce de igual manera el proceso de manufactura y la gerencia de producción desea saber si existen diferencias en la producción de cada planta a través de un estudio comparativo simple si se considera un error del 1%.

4.2 Resuelve el problema anterior con un error al 5%.

4.3 Una planta fabricante de un filamento especial para lámparas incandescentes labora dos turnos de ocho horas (matutino y vespertino) con los niveles de producción en metros que se citan en la siguiente tabla.

Tabla 4.26

Matutino	Vespertino
22.420	23.510
18.360	20.620
21.460	26.470
19.200	19.750
23.400	20.300
27.380	17.840
23.460	26.340

La gerencia de control de calidad está interesada en conocer si hay diferencia en la producción entre los turnos, por lo que propone efectuar un estudio comparativo simple con un nivel de confianza de 90%.

4.4 Resuelve el problema anterior considerando un error del 5%.

4.5 Una empresa empaadora de camarón café para exportación cuenta con tres líneas de empackado con los siguientes volúmenes de producción de cajas por hora.

Tabla 4.27

Línea 1	Línea 2	Línea 3
51	49	54
50	53	54
55	48	53
48	50	52

Eres analista de una empresa y el dueño te encomienda realizar un análisis por estudios comparativos simples para establecer si existe o no diferencia en la producción de cada línea al 95% de confianza.

4.6 Una empresa fabricante de bolsas industriales para agua pone a prueba cuatro de sus diseños inyectando aire a presión (KPa) hasta que estallen.

Tabla 4.28

Tipo 1	Tipo 2	Tipo 3	Tipo 4
2.4	2.5	2.5	2.7
2.3	2.5	2.5	2.6
2.6	2.4	2.6	2.8
2.5	2.3	2.7	2.6
2.7	2.6	2.7	2.6
2.3	2.5	2.5	2.5

El fabricante desea saber si existe diferencia en la presión máxima de las bolsas dependiendo de su tipo. Se propone realizar un análisis comparativo simple con un nivel de confianza del 90%.

4.7 Analiza el problema anterior, considerando un nivel de confianza del 95%.

4.8 Un laboratorio realiza pruebas sobre un nuevo tipo de insulina inyectable para pacientes diabéticos insulino dependientes. Se seleccionaron 21 pacientes voluntarios con base en sus características físicas y expediente clínico. Se repartieron al azar en tres grupos; a cada grupo se le fijó una dieta igual pero con diferentes dosis del medicamento. Los investigadores realizan una prueba de hipótesis estadísticas por estudios comparativos simples para determinar si existe diferencia entre las dosis aplicadas a cada grupo al considerar un error del 5%.

Tabla 4.29

Dosis 1	Dosis 2	Dosis 3
115	102	98
108	105	98
120	100	97
110	98	100
109	95	95
108	100	102
112	104	105

4.9 Una fábrica de refacciones para autos prueba tres nuevas aleaciones para las veteas de dirección de un auto económico popular. Los resultados de las pruebas de dureza en unidades Rockwell son las que se muestran en la tabla. La gerencia de control de calidad está interesada en saber si existen diferencias en los materia-

les propuestos para la refacción, por lo que realiza una prueba de hipótesis estadísticas para estudios comparativos simples al 99% de confianza.

Tabla 4.30

Alcación 1	Alcación 2	Alcación 3
26.0	29.0	29.4
26.5	28.9	29.3
27.5	28.5	29.0
27.0	28.6	29.3

4.10 Para optimizar el tiempo de acceso a la información, una fábrica de discos duros portátiles de 500 GB realiza pruebas en los tres modelos que se encuentran en desarrollo en computadoras con el mismo procesador y obtiene los siguientes tiempos en milisegundos.

Tabla 4.31

Disco 1	Disco 2	Disco 3
8.2	8.4	8.8
8.0	8.3	8.6
8.2	8.5	8.9
8.3	8.4	8.7
8.1	8.6	8.6
8.1	8.4	8.6
8.2	8.5	8.5
8.2	8.5	8.6

Para determinar si existen diferencias entre los modelos, se pone en práctica un proceso de análisis de pruebas estadísticas simples por estudios comparativos simples al 99% de confianza.

4.11 Con base en el planteamiento del problema 4.1, establece si existen diferencias entre los grupos propuestos a través de un análisis de la varianza de un solo factor y corrobora los resultados de los cálculos mediante la rutina de Excel.

4.12 Argumenta, ¿coinciden los resultados obtenidos por el estudio comparativo simple con el de ANOVA de un solo factor en el planteamiento del problema 4.1?

4.13 Con fundamento en el planteamiento del problema 4.3 establece si existen diferencias entre los grupos propuestos a través de un análisis de la varianza de un solo factor. Corrobora los resultados de los cálculos mediante la rutina de Excel; considera un nivel de confianza del 99%.

4.14 Con base en el planteamiento del problema 4.5, establece si existen diferencias entre los grupos propuestos a través de un análisis de la varianza de un solo factor. Corrobora los resultados de los cálculos mediante Excel; considera niveles de confianza del 90% y 95%.

4.15 Con fundamento en el planteamiento del problema 4.6, establece si hay diferencias entre los grupos propuestos a través de un análisis de la varianza de un solo factor. Corrobora los resultados de los cálculos mediante la rutina de Excel; considera un nivel de confianza del 99%.

4.16 Una empresa de concretos experimenta con un concreto basado en tres nuevos tipos de cemento: puzolánico I, puzolánico II y puzolánico III. Se llevan a cabo pruebas de resistencia en una mezcla de concreto de 250 kg/cm². Realiza una prueba de ANOVA de un factor de una muestra para comparar la resistencia entre las mezclas de concreto y establece si hay diferencia en la resistencia del concreto de acuerdo con el tipo de cemento utilizado. Considera un nivel de confianza del 95%.

Tabla 4.32

Resistencia en concreto con cemento Puzolánico I	Resistencia en concreto con cemento Puzolánico II	Resistencia en concreto con cemento Puzolánico III
258	248	254
263	252	256
256	246	257
256	248	255
262	247	251
262	246	253
256	250	253
258	250	254
256	251	257
254	250	256
255	252	255
256	255	254

4.17 Con base en los resultados del problema anterior, realiza un proceso de discriminación por LSD de un factor.

4.18 Una empresa dedicada a la fabricación de vidrio desea extender la variedad de sus productos de seguridad. Quiere lanzar al mercado un nuevo producto al cual le han añadido resinas para variar la resistencia del vidrio dependiendo de la cantidad que se añada. La empresa desea saber si existe variación en la resistencia del vidrio (kg/cm²) entre una nueva resina y la actual al realizar un análisis por discriminante LSD. En la tabla que sigue se muestra la comparación de las resistencias del vidrio con la resina actual y la nueva.

Tabla 4.33

Resistencia del vidrio Resina actual	Resistencia del vidrio Resina nueva
85	96
80	87
75	84
76	72
50	60

4.19 Con base en el planteamiento del problema 4.8, haz una prueba de discriminante *LSD* de un factor con el fin de establecer si existen diferencias entre los grupos propuestos.

4.20 Con fundamento en el planteamiento del problema 4.9, realiza una prueba de discriminante *LSD* de un factor con el propósito de establecer si existen diferencias entre los grupos propuestos.

4.21 Con base en el planteamiento del problema 4.10, realiza una prueba de discriminante *LSD* de un factor con el fin de establecer si hay diferencias entre los grupos propuestos.

4.22 Una cooperativa agroindustrial ensaya con cinco tipos diferentes de aditivos proteínicos para la fermentación de sus quesos y procede a medir el pH. La empresa le encarga a un laboratorio en alimentos que les indique si existen diferencias entre los aditivos, por lo que los asesores llevan a cabo una prueba de ANOVA de un solo factor al 95% de confianza. En la tabla que sigue aparecen los resultados de las pruebas.

Tabla 4.34

Aditivo 1	Aditivo 2	Aditivo 3	Aditivo 4	Aditivo 5
4.01	4.07	4.20	4.20	4.14
4.04	4.12	4.17	4.15	4.12
4.02	4.16	4.21	4.19	4.14
4.04	4.15	4.19	4.17	4.14
4.03	4.08	4.20	4.14	4.11

4.23 Con base en el planteamiento del problema anterior, realiza un proceso de discriminación basado en la *R* de Duncan para establecer las diferencias entre los grupos.

4.24 Una empresa fabricante de focos de vapor ensaya los nuevos modelos de focos de baja presión para uso comercial tratando de decidir entre dos modelos. De momento se realizan las pruebas de vida útil tal como se muestra en la siguiente tabla. Realiza una prueba de ANOVA de un factor de una muestra para comparar la vida útil entre los modelos de foco con un nivel de confianza del 95%.

Tabla 4.35

Modelo I Vida útil en horas	Modelo II Vida útil en horas
8 090	7 994
8 090	8 115
8 120	7 810
8 111	8 110
8 112	7 999
8 115	7 811
8 126	8 050
8 116	8 068
8 108	7 956
8 112	8 106

4.25 Con fundamento en los resultados del problema anterior, realiza un proceso de discriminación por *LSD*.

4.26 Un proceso de fabricación se reproduce en una planta por medio de tres máquinas al mismo tiempo, las 24 horas del día durante los siete días de la semana. El nivel de piezas producidas por día se muestra a continuación:

Tabla 4.36

	Máquina A	Máquina B	Máquina C
Lunes	400	395	398
Martes	385	393	397
Miércoles	410	390	399
Jueves	405	397	401
Viernes	406	399	401
Sábado	393	394	396
Domingo	400	395	398

Realiza una prueba de hipótesis estadística basada en ANOVA de dos factores con un error del 5%.

4.27 Con base en los resultados del problema anterior, realiza un proceso de discriminación por *LSD* para los factores *a* y *b*.

4.28 Con fundamento en los datos de la prueba de concreto, la empresa productora realiza pruebas sobre un nuevo aditivo acelerador para concreto con el cual se pretende estandarizar la resistencia en los concretos de 250 kg/cm² desarrollados con los tres tipos de cemento: puzolánico I, puzolánico II y puzolánico III. Lleva a cabo un análisis de hipótesis estadísticas basado en ANOVA de dos factores con varias muestras al 95% para determinar si existe o no diferencia en la resistencia de los concretos.

Tabla 4.37

	Resistencia en concreto con cemento Puzolánico I	Resistencia en concreto con cemento Puzolánico II	Resistencia en concreto con cemento Puzolánico III
Aditivo I	258	248	254
	263	252	256
	256	246	257
	256	248	255
Aditivo II	262	247	251
	262	246	253
	256	250	253
	258	250	254
Aditivo III	256	251	257
	254	250	256
	255	252	255
	256	255	254

4.29 Con base en los resultados del problema anterior, haz un análisis por discriminantes *LSD* de dos factores incluyendo el análisis por interacción.

4.30 La empresa fabricante de focos no solo ensaya con el diseño de los nuevos modelos de focos de vapor de baja presión para uso comercial en cuanto a su vida útil, sino también con el uso del vapor de sodio y del de mercurio. Realiza una prueba de ANOVA de dos factores con varias muestras para determinar si los tipos de vapor y modelos generan diferencias en la vida útil de los focos. Considera un nivel de confianza del 95%.

Tabla 4.38

	Modelo I Vida útil en horas	Modelo II Vida útil en horas
Vapor de sodio	8 090	7 994
	8 090	8 115
	8 120	7 810
	8 111	8 110
	8 112	7 999
Vapor de mercurio	8 115	7 811
	8 126	8 050
	8 116	8 068
	8 108	7 956
	8 112	8 106

4.31 Con base en los resultados del problema anterior, lleva a cabo un análisis por discriminantes *LSD* de dos factores, incluye el análisis por interacción.

4.32 Una empresa fabricante de autos eléctricos utilitarios para usos especiales realiza pruebas sobre el rendimiento del banco de baterías de su nuevo modelo de auto ultraligero, el cual cuenta con tres versiones. Para el banco de baterías propone seis baterías de plomo ácido o seis de níquel-cadmio. La eficiencia eléctrica del banco en unidades porcentuales se muestra en la siguiente tabla. Realiza una prueba de ANOVA de dos factores al 95% de confianza para establecer si los tipos de banco y de auto generan diferencias en los rendimientos.

Tabla 4.39

	Auto tipo I	Auto tipo II	Auto tipo III
Plomo-ácido	82%	84%	78%
	78%	86%	79%
	79%	82%	80%
	74%	83%	82%
	80%	85%	77%
	75%	84%	80%
Níquel-cadmio	86%	92%	90%
	82%	88%	86%
	84%	85%	92%
	87%	90%	88%
	84%	86%	84%
	86%	83%	90%

4.33 Con base en los resultados del problema anterior, realiza un análisis por discriminantes *LSD* de dos factores incluyendo el análisis por interacción.

4.34 Un experimento en cuanto a la mejora del diseño de un helicóptero de aeromodelismo se fundamenta en el rendimiento de la distancia recorrida en razón del tipo de motor y del tipo de combustible empleado. Las distancias recorridas se muestran en la siguiente tabla.

Tabla 4.40

	Motor 1	Motor 2	Motor 3	Motor 4
Tipo 1	630	590	525	604
Tipo 2	680	5701	540	603
Tipo 3	670	610	534	595
Tipo 4	650	585	548	579
Tipo 5	640		554	594
Tipo 6	668			600

Realiza un análisis comparativo simple a efecto de determinar si existen o no diferencias entre los tipos de motor con respecto al tipo de combustible empleado.

4.35 Un experimento en cuanto a la mejora de un combustible para autos de tipo ecológico se estructuró con base en probar diferentes tipos de autos compactos. El experimento midió entre otros parámetros el rendimiento en kilómetros por litro. Estos se muestran en la siguiente tabla.

Tabla 4.41

Combustible	Auto 1	Auto 2	Auto 3	Auto 4	Auto 5	Auto 6
Tipo 1	12.70	13.10	14.10	12.40	14.20	13.14
Tipo 2	13.20	13.00	13.92	12.78	14.25	13.15
Tipo 3	12.70	13.05	13.89	12.80	14.30	13.20
Tipo 4	12.90	12.95	14.00	12.75	14.20	13.25
Tipo 5	12.82	12.92	13.90	12.82	14.16	13.17
Tipo 6	12.87	12.94	13.95	12.78	14.22	13.16
Tipo 7	12.92	12.97	13.98	12.80	14.22	13.17

Realiza una prueba de hipótesis basada en el ANOVA de dos factores de una muestra, al 95% de confianza, a efecto de determinar si existen o no diferencias entre los tipos de autos con respecto al tipo de combustible empleado.

4.36 Con fundamento en los resultados del problema anterior, realiza un análisis por discriminantes *LSD*. De manera adicional desarrolla un proceso de discriminación por coeficiente *LSD* para cada uno de los factores, así como para la interacción de los mismos.

4.37 Con base en el planteamiento del problema 4.35, realiza una prueba de hipótesis de un factor por tipo de auto de

manera que se realice un análisis por discriminantes basado en la *R* de Duncan.

T 4.38 Con fundamento en el planteamiento del problema 35, procede a realizar una prueba de hipótesis

de un factor por tipo de combustible de manera que se desarrolle un análisis por discriminantes basado en la *R* de Duncan.



PROBLEMA RETO

Un experimento en agronomía se estructuró con base en diferentes formas de tratar el suelo y sembrar una semilla mejorada de cierto tipo de oleaginosa. La cantidad sembrada dependió del clima y de la calidad del agua; las producciones en kilos se muestran en la tabla.

Tabla 4.42

	Trat 1	Trat 2	Trat 3	Trat 4	Trat 5
Cantidad 1	425	380	393	370	400
Cantidad 2	385	378	397	382	405
Cantidad 3	378	394	394	378	397
Cantidad 4	410	372	392	380	399
Cantidad 5	420	383	390	381	402
Cantidad 6	390	398	388	384	401
Cantidad 7	382	374	390	386	397
Cantidad 8	398	368	389	388	395
Cantidad 9	379	397	396	390	400
Cantidad 10	396	392	394	383	398

1

- Haz un análisis comparativo simple para los tratamientos a efecto de determinar si existen diferencias entre los mismos.
- Realiza un análisis comparativo simple para las cantidades de semilla a efecto de determinar si existen diferencias entre las mismas.
- Desarrolla un proceso de discriminación basado en la *R* de Duncan para determinar si existen diferencias entre los tratamientos.
- Realiza un proceso de discriminación basado en la *R* de Duncan para determinar si existen diferencias entre las cantidades.
- Haz una prueba de hipótesis estadística por ANOVA de un factor para los tratamientos y compara estos resultados con los del inciso a). Argumenta, ¿existen diferencias entre los resultados?
- Realiza una prueba de hipótesis estadística por ANOVA de un factor para las cantidades de semilla. Compara estos resultados con los del inciso b). Argumenta, ¿existen diferencias entre los resultados?
- Desarrolla un análisis por discriminantes *LSD* para los tratamientos con el propósito de determinar si existen diferencias entre los mismos.
- Con base en los resultados de los incisos c) y g), ¿existen diferencias entre los resultados de los discriminantes?

- i) Realiza un análisis por discriminantes *LSD* para las cantidades de semilla con el propósito de determinar si existen diferencias entre ellos.
- j) Con base en los resultados de los incisos d) e i), ¿existen diferencias entre los resultados de los discriminantes?
- k) Con fundamento en el planteamiento del problema reto, pon en práctica una prueba de hipótesis estadística basada en el ANOVA de dos factores con el propósito de determinar si existen diferencias tanto en los tratamientos, como en cantidades y la interacción entre ellos.
- l) Realiza un análisis por discriminantes *LSD* de dos factores tanto para los tratamientos, como para las cantidades y la interacción entre ellos con el propósito de determinar si existen diferencias.



REFERENCIAS

Box, George E. P., William G. Hunter y J. Stuart Hunter (1993). *Estadística para investigadores-introducción al diseño de experimentos, análisis de datos y construcción de modelos*. España: Editorial Reverté, S.A.

Griesbrecht, F. G. y Marcia L. Gumperts (2004). *Planning, Construction and Statistical Analysis*. Estados Unidos: Wiley and Sons.

Kruehl, Robert O. (2001). *Diseño de experimentos* (2a. ed.). Estados Unidos: Thomson.

Montgomery, Douglas C. (2001). *Design and Analysis of Experiments* (5a. ed.). Estados Unidos: Wiley and Sons.

Vicente, Ma. Lina, Pedro Girón, Carmen Nieto y Teresa Pérez (2005). *Diseño de Experimentos-Soluciones SAS y SPSS*. México: Pearson-Prentice Hall.



DIRECCIONES ELECTRÓNICAS

Práctica 3. Diseño de experimentos con dos o más factores

[http://dm.udc.es/assignaturas/estadistica2/secprac_3.html]

Bases del análisis de la varianza

[http://www.hrc.es/bioest/Anova_2.html]

El análisis de la varianza (ANOVA)- 2. Estimación de componentes de varianza,

[<http://argo.urv.es/quimio/general/anova2cast.pdf>]

Tabla t de Student

[www.emp.uva.es/inf_acad/hermer/.../e2t_tabla_t_de_student.pdf]

Valores F de la distribución F de Fisher

[<http://dcb.fi-c.unam.mx/profesores/irene/Notas/tablas/Fisher.pdf>]

Tablas de rangos significativos de Duncan

[<http://costaricalinda.com/Estadistica/duncan1.htm>]

Fisher's *LSD* (Least Significant Difference)

[<http://www.mzandee.net/~zandee/statistiek/syllabus/LSD.pdf>]

An example of analysis of variance with two factors, compare means

[www.seedtest.org/upload/cms/.user/ANOVA_2.pdf]