PROBABILIDAD Y ESTADÍSTICA MATEMÁTICA

Código 3002

GUÍA DE ESTUDIO

Preparada por
Luis Armando Hernández Solís

UNIVERSIDAD ESTATAL A DISTANCIA
VICERRECTORÍA ACADÉMICA
ESCUELA DE CIENCIAS EXACTAS Y NATURALES

Edición académica
Virginia Ramírez

Encargado de cátedra
Cristian Quesada

Revisión filológica
María Benavides

Esta guía de estudio ha sido confeccionada para ser utilizada en la carrera de profesorado en la Enseñanza de la Matemática que imparte la UNED.

PRESENTACIÓN

"[…] el sujeto es el propio constructor de su conocimiento, así el intercambio que establezca con el medio y las acciones ejercidas sobre los objetos es fundamental para el desarrollo del conocimiento" (Pereira, 1990, p. 10).

Esta guía de estudio tiene como propósito acompañar al estudiante, en su proceso de aprendizaje a distancia, con la creación de ambientes basados en la exploración, la conjetura y el descubrimiento.

Se plantean puntos de atención, ejemplos, ejercicios resueltos y glosario; además, introduce la visualización y manipulación de los datos, varias actividades de mediación pedagógica utilizando el *software* libre *winstats.exe*, el cual le permitirá individualizar, de alguna manera, su aprendizaje, presentándole un ambiente autodidacta.

Kaput (1994) hace las siguientes distinciones entre los métodos tradicionales y los métodos tecnológicos:

- medios estáticos vs. medios dinámicos,
- medios inertes vs. medios interactivos y
- manipulaciones físicas vs. manipulaciones basadas en el computador.

Por eso, insto al educando a mantener un ritmo adecuado con respecto a los avances tecnológicos actuales, así como establecer mecanismos en la incorporación de *software* en su proceso de aprendizaje.

Asimismo, el estudiante debe tener claro que, al utilizar la tecnología y un *software* apropiado, se tienen las siguientes ventajas:

- Realizar diferentes experimentos, variando los parámetros, para poder relacionarlos activamente unos con otros.

- Lograr un buen manejo de objetos matemáticos y sus respectivas relaciones.
- Conectar experimentos reales con formalismos matemáticos usando una mezcla de datos reales y simulaciones.

El *software* libre *winstats.exe* se utiliza para la obtención de medidas de tendencia central y de dispersión para diferentes tipos de datos, graficación de datos unidimensionales y bidimensionales, resolución de problemas de probabilidad y estadística, comparación de resultados mediante varias simulaciones realizadas por el *software* libre *winstats.exe* en las cuales se manipularán parámetros; descubrimiento de propiedades de diferentes distribuciones de probabilidad conocidas, entre otras actividades.

Por consiguiente, este trabajo pretende ser una herramienta más en el curso de Probabilidad y Estadística Matemática, la cual le permita comprender, de manera más profunda, los temas de probabilidad y estadística, convirtiéndola en un facilitador de los aprendizajes.

CONTENIDO

GENERALIDADES

El estudiante de Enseñanza de la Matemática debe manejar conceptos tanto de estadística como de probabilidad, puesto que los métodos estadísticos contribuyen al proceso de realizar juicios científicos frente a la incertidumbre y a la variación; y la probabilidad, por su parte, apoya el estudio de fenómenos puramente aleatorios.

La Cátedra de Matemáticas Intermedias eligió como texto básico el libro:

> Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers y Keying Ye. *PROBABILIDAD Y ESTADÍSTICA PARA INGENIERÍAS Y CIENCIAS*. Octava edición. Editorial Pearson Educación, México, 2007.

Aunque no es un libro escrito, precisamente, para la educación a distancia, se escogió porque presenta una serie de cualidades que le ayudarán en la comprensión de los contenidos, como las siguientes:

- En cada sección se mezcla la teoría con los ejemplos, de una forma muy detallada y explícita.
- Al final de las secciones hay gran cantidad de problemas, con aplicaciones en ingeniería, biología, física y computación, que facilitan la comprensión de los métodos estadísticos y probabilísticos en otras áreas.
- Hay material nuevo y de repaso al final de los capítulos, el cual destaca las ideas clave, así como los riesgos y peligros de los que debe estar consciente el estudiante.
- En cada capítulo se muestra la relación con los otros capítulos.

La presente guía tiene como finalidad acompañar y *"llevar de la mano"* al alumno en su proceso de aprendizaje a distancia; parte de temas básicos, como son la descripción e interpretación de datos estadísticos y el análisis teórico de los modelos de distribución de probabilidades, para poder efectuar inferencias estadísticas.

Esta guía consta de cinco secciones, se mencionan a continuación:

- Sección 1. Estadística descriptiva
- Sección 2. Conceptos básicos de probabilidad
- Sección 3. Variables aleatorias y distribuciones de probabilidad
- Sección 4. Elementos de inferencia estadística
- Sección 5. Regresión lineal y correlación

Cada sección consta de las siguientes partes:

- **Objetivos generales.** Son los que se establecen en la descripción curricular de la asignatura.
- **Objetivos específicos.** Son los que se quieren que el alumno logre al final del estudio de cada sección de la guía.
- **Puntos de atención.** Es una breve sinopsis de los principales tópicos de las secciones desarrolladas en el libro de texto con ejemplos ilustrativos. Además, cuenta con recomendaciones y observaciones importantes para la resolución de diferentes ejercicios.
- **Laboratorios.** En los procesos de enseñanza-aprendizaje de la probabilidad y estadística, es importante basarse en la exploración, la conjetura y el descubrimiento. Para esto, la tecnología introduce una nueva era en la enseñanza de la matemática fundamentada en la visualización y manipulación de los datos. Existen varios *software* para la estadística y probabilidad; en este caso, se eligió el *winstats.exe* que podrá encontrar gratuitamente en <http://math.exeter.edu/rparris/>. El cual permite la instalación tantas veces y en tantas máquinas como el usuario lo desee, es un *software* fácil de usar y se puede accesar fácilmente a Internet.

  En los laboratorios se mostrarán varias actividades que se pueden formular con *winstats.exe*, tales como la obtención de medidas de tendencia central y de dispersión para diferentes tipos de datos; la graficación de datos unidimensionales y bidimensionales; la resolución de problemas de probabilidad y estadística de una forma heurística e inductiva, comparando resultados mediante varias simulaciones realizadas por el *software,* en las cuales se manipularán parámetros; el descubrimiento de propiedades de diferentes distribuciones de probabilidad conocidas.

- **Ejercicios de autoevaluación.** Su finalidad es que el estudiante posea una base adicional de ejercicios, donde pueda probar los conocimientos adquiridos. Estos no sustituyen a los del libro de texto, por lo cual el alumno debe resolver la mayoría de los ejercicios planteados en cada sección y, luego, los de la guía.

- **Soluciones de los ejercicios de autoevaluación.** Es el desarrollo de la solución de cada uno de los ejercicios de autoevaluación, de una forma explícita y clara.

- **Glosario.** Es la definición de las palabras clave de cada sección.

Por consiguiente, esta guía pretende ser una herramienta más en el curso de Probabilidad y Estadística Matemática, la cual permita al estudiante individualizar, de alguna manera, su aprendizaje, presentándole un ambiente autodidacta, de exploración y descubrimiento. Es un facilitador de los aprendizajes.

SECCIÓN 1
ESTADÍSTICA DESCRIPTIVA

En la sección se desarrollan conceptos básicos de la estadística descriptiva, por ejemplo medidas de tendencia central, de dispersión, de posición para datos unidimensionales; así como su representación en tablas e histogramas.

Asimismo, se presentan dos actividades dinámicas utilizando el *software* gratuito *winstats.exe*, con el fin de buscar una mayor comprensión de los conceptos desarrollados, e incentivar el uso de nuevas tecnologías para el proceso de enseñanza-aprendizaje de la matemática.

OBJETIVOS GENERALES

1. Conocer y aplicar conceptos básicos de la estadística que permitan aclarar los alcances de esta disciplina, así como la presentación e interpretación de datos mediante tablas y gráficas.
2. Describir características de un conjunto de datos utilizando las principales medidas de tendencia central, posición y variabilidad usadas en la estadística.

OBJETIVOS ESPECÍFICOS

1. Comprender el significado de la estadística descriptiva.
2. Calcular medidas de tendencia central tales como media, moda y mediana.
3. Calcular medidas de dispersión como la varianza y desviación estándar, para datos unidimensionales.
4. Calcular medidas de posición tales como cuartiles y percentiles.

5. Representar datos unidimensionales mediante diagramas tallo-hoja y tablas de frecuencia.

6. Representar gráficamente datos unidimensionales mediante histogramas.

PUNTOS DE ATENCIÓN DE LAS SECCIONES 1.1 A LA 1.5 Y LA 1.8

Para las secciones 1.1, 1.2 y 1.3 del libro de texto se le recomienda realizar una lectura, con el propósito de introducir el tema de la estadística descriptiva, además de conceptos básicos tales como población, muestreo, diseño experimental, entre otros. Esto le ayudará a cumplir con el primer objetivo específico, ya que se le presenta un panorama general de la Estadística, el empleo de datos científicos, la variabilidad en los datos científicos, el papel de la probabilidad y los diferentes procedimientos de muestreo. Aunque estos temas no serán evaluados específicamente, es significativo que se forme una perspectiva general del área por tratar.

En la sección 1.4 se presentan los conceptos básicos de dos medidas de posición: la media y la mediana de datos unidimensionales de una muestra. Se menciona un concepto adicional, el de la moda de una serie de datos importante para describirlos, por esta razón, se definirá como el valor que se presenta con mayor frecuencia, es decir, es el más común. La moda puede no existir; incluso, si existe, puede no ser única.

El apartado *"Otras medidas de posición"*, que se encuentra en las páginas 12 y 13 del libro de texto, no es de interés evaluativo del curso.

Observación. Para calcular la media poblacional (μ) se utiliza la misma fórmula que para la media muestral $(\overline{X})$, siendo $\{x_1, x_2, ..., x_n\}$ el conjunto de todos los datos y n el tamaño de la población.

EJEMPLO 1

Sea el conjunto $\{3,4,4,5,6,8,8,8,10\}$ una muestra de 9 datos, se tiene que:

- la media es $\overline{X}=\dfrac{3+4+4+...+10}{9}=6,\overline{2}$;

- la mediana es $\tilde{X}=6$, ya que al haber 9 datos (cantidad impar), la mediana será el valor medio, o sea el valor numérico que está en la posición 5;

- la moda es 8, ya que es el valor que se presenta con mayor frecuencia. La muestra es unimodal, puesto que la moda es única.

En la sección 1.5 se presentan los conceptos básicos de las medidas de variabilidad o dispersión: la varianza y la desviación estándar de una muestra, las cuales hacen referencia al grado en que los datos numéricos tienden a extenderse alrededor de un valor medio.

Se le recomienda utilizar la siguiente fórmula para la varianza, equivalente a la planteada en la definición 1.2 del libro de texto, puesto que es más fácil de calcular para valores grandes de n:

$$s^2=\frac{\sum x_i^2-\frac{\left(\sum x_i\right)^2}{n}}{n-1}.$$

Observación. La fórmula para la varianza poblacional $\left(\sigma^2\right)$ no es la misma que la de la varianza muestral $\left(s^2\right)$, ya que cuando se calcula σ^2 se divide por n y no por $n-1$. Es decir, se supone que $\{x_1,x_2,...,x_n\}$ son los n valores numéricos de una población de tamaño n, con una media μ.

$$\sigma^2=\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{n}.$$

EJEMPLO 2

Calcular la varianza y la desviación estándar de los datos del ejemplo 1.

Para realizar el cálculo de forma ordenada y clara es recomendable construir la tabla 1.

Tabla 1

x_i	x_i^2
3	9
4	16
4	16
5	25
6	36
8	64
8	64
8	64
10	100
$\sum x = 56$	$\sum x^2 = 394$

Ahora, al utilizar la fórmula alternativa para la varianza muestral se obtiene:

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{394 - \frac{(56)^2}{9}}{8} = 5{,}69\overline{4}.$$

Para calcular la desviación estándar s, se tiene que:

$$s = \sqrt{s^2} = 2{,}3863\ldots$$

En la sección 1.8 se explican diferentes formas de ordenar y representar los datos, una manera es mediante un diagrama de tallo y hojas.

EJEMPLO 3

Si se toman en cuenta los datos del ejercicio 1.1 de la página 13 del libro de texto, se pueden ordenar mediante un diagrama de tallo y hojas, como el que se muestra en el cuadro 1.

El ejercicio dice así: "se registran las siguientes mediciones para el tiempo de secado (en horas) de cierta marca de pintura esmaltada."

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Cuadro 1. Diagrama tallo y hojas del tiempo de secado

Tallo	Hoja	Frecuencia
2	5 8 8 9	4
3	0 3 4 6 7	5
4	0 4 8 8	4
5	2 6	2

Otra forma de ordenar los datos es por medio de una tabla de distribución de frecuencias, como se muestra en el ejemplo 4.

EJEMPLO 4

Si se toman los datos del ejemplo 3 y se elige $k=4$, la amplitud de cada clase será de $0,9$; así, se obtiene el cuadro 2 de distribución de frecuencias.

Cuadro 2. Distribución de frecuencias

Intervalo de clases	**Frecuencia absoluta (f_i)**	**Frecuencia relativa (f_r)**
2,0 – 2,9	4	0,266…
3,0 – 3,9	5	0,333…
4,0 – 4,9	4	0,266…
5,0 – 5,9	2	0,133…
Totales	15	1

Observación. En este caso el intervalo de clase es un intervalo cerrado, por ejemplo $2{,}0-2{,}9$ representa $[2{,}0;2{,}9]$, pero también se pueden tomar intervalos semiabiertos, como por ejemplo $[2{,}0;3{,}0[$,$[3{,}0;4{,}0[$, $[4{,}0;5{,}0[$, $[5{,}0;6{,}0[$; porque lo importante es que no hayan valores que estén en dos o más clases.

Un histograma es simplemente un diagrama de barras, donde la altura de cada barra indica el número de veces que el dato aparece en la serie.

EJEMPLO 5

Si se representa gráficamente la distribución de frecuencias del ejemplo 4, mediante un histograma, se tiene el gráfico 1.

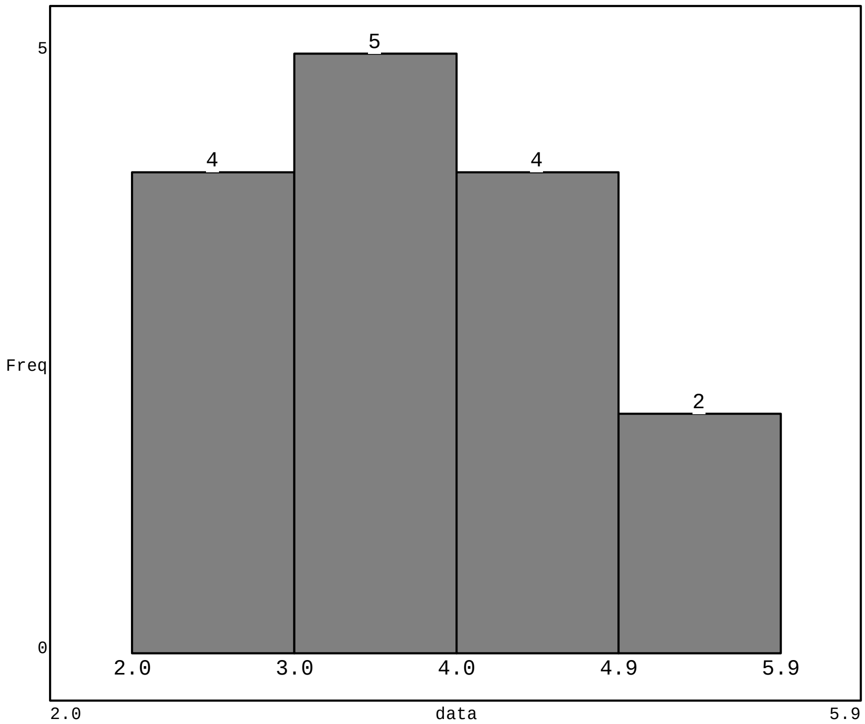


Gráfico 1

MEDIDAS DE POSICIÓN. CUARTILES Y PERCENTILES

En el libro de texto no se desarrolla el tema; por su importancia, se explica en esta guía.

Cuartiles. Si una serie de datos se colocan en orden de magnitud, el valor medio (o la media aritmética de los dos valores medios) que divide al conjunto de datos en dos partes iguales es la mediana, como se explicó anteriormente. Los valores representados por Q_1, Q_2, Q_3 se llaman primero, segundo y tercer cuartil, respectivamente. El valor Q_2 es igual a la mediana.

Percentiles. Suponga n valores de datos colocados en orden creciente. El percentil k, que se llama P_k, es el número para el cual el k por ciento de los valores son menores que P_k y el $(100-k)$ por ciento son superiores. P_k se define como sigue:

1. se realiza la operación $\dfrac{k \cdot n}{100}$;
2. se le asigna a E la parte entera y a D la parte decimal de la operación anterior (es decir $\dfrac{k \cdot n}{100} = E + D$);
3. Luego,

$$P_k = \begin{cases} \text{Valor } (E+1) \text{ cuando } D \neq 0, \\ \dfrac{\textit{Valor } E + \text{Valor } (E+1)}{2} \text{ cuando } D = 0. \end{cases}$$

EJEMPLO 6

Los siguientes ejercicios se realizarán con base en los datos del *ejemplo 3*:

a) determinar Q_1 y Q_3;

b) determinar P_{35} y P_{60}.

Solución

Para la resolución de los ejercicios utilice el diagrama de tallo y hoja del *ejemplo 4*.

a) Tome en cuenta que hay 15 valores, se tiene que la mediana es el valor de la posición 8 (o sea 3,6), esto quiere decir que Q_1 será el valor de la posición 4 (o sea 2,9); ya que de 7 valores anteriores a la mediana, el punto medio estará en la posición 4; de igual forma Q_3 será el valor de la posición 12 (o sea 4,8).

Observación. Los percentiles P_{25} y P_{75} corresponden al primer y tercer cuartil, respectivamente. Por lo que el primer, segundo y tercer cuartiles se pueden averiguar mediante los percentiles 25, 50 y 75, respectivamente.

b) Para calcular los percentiles se toma en cuenta que $n=15$.

Para P_{35} se tiene que $\frac{k \cdot n}{100} = \frac{35 \cdot 15}{100} = 5{,}25 \Rightarrow E = 5$ y $D = 0{,}25$.

Como $D \neq 0$ entonces $P_{35} = valor(5+1) = valor 6 = 3{,}3$.

Para P_{60} se tiene que. $\frac{k \cdot n}{100} = \frac{60 \cdot 15}{100} = 9 \Rightarrow E = 9$ y $D = 0$.

Como $D=0$ entonces $P_{60} = \frac{valor(9)+valor(10)}{2} = \frac{3{,}7+4{,}0}{2} = 3{,}85$.

Laboratorios

Si se tuviera que calcular la media y varianza de muestras de 100, 500 o 1000 datos, la labor sería demasiado complicada y engorrosa. Para poder enfrentarse a este tipo de ejercicios de una forma rápida y segura, es necesario aprender a utilizar un *software* que facilite esta labor. Se ha escogido el *software winstats.exe,* puesto que es gratuito, accesible y muy fácil de usar.

A continuación se le presentan los laboratorios 1 y 2 del apartado 1.4, que tienen como fin aprender a utilizar la aplicación para datos univariables.

Laboratorio 1 (obtención de información)

Para realizar este laboratorio se utilizan los datos del ejercicio de autoevaluación 7.

Instrucciones

1. Abra el *software winstats.exe,* haga doble clic en el ícono correspondiente. Se abrirá una pequeña pantalla verde con dos opciones en el menú: $\boxed{Window}$, $\boxed{Help}$

2. Ubíquese en $\boxed{Window}$ y escoja la opción $\boxed{1 - var\ data\ \ F1}$. Esta es la opción para trabajar con datos unidimensionales. Se abrirá una pantalla blanca donde se ubican los datos de la muestra.

3. Para poner los datos en 0, se debe ubicar en la primera opción del menú principal de la ventana $\boxed{File}$, escoger la opción $\boxed{New}$ y luego $\boxed{Zeros}$.

4. Para digitar los datos debe hacer clic en la casilla correspondiente para entrar, digitar el dato, luego oprimir la tecla "*enter*" para salir de la casilla. Repetir el proceso hasta que esté digitada toda la muestra. No importa el orden de los datos, tampoco si se digitan en filas o columnas, pues corresponden a datos en una variable.

5. Para obtener una serie de información como medidas de posición, de variabilidad, cuartiles y amplitud, basta con ubicarse en la cuarta opción del menú principal $\boxed{Stats}$ y escoger la opción $\boxed{Overall\ ...}$. Se abre una ventana como la siguiente:

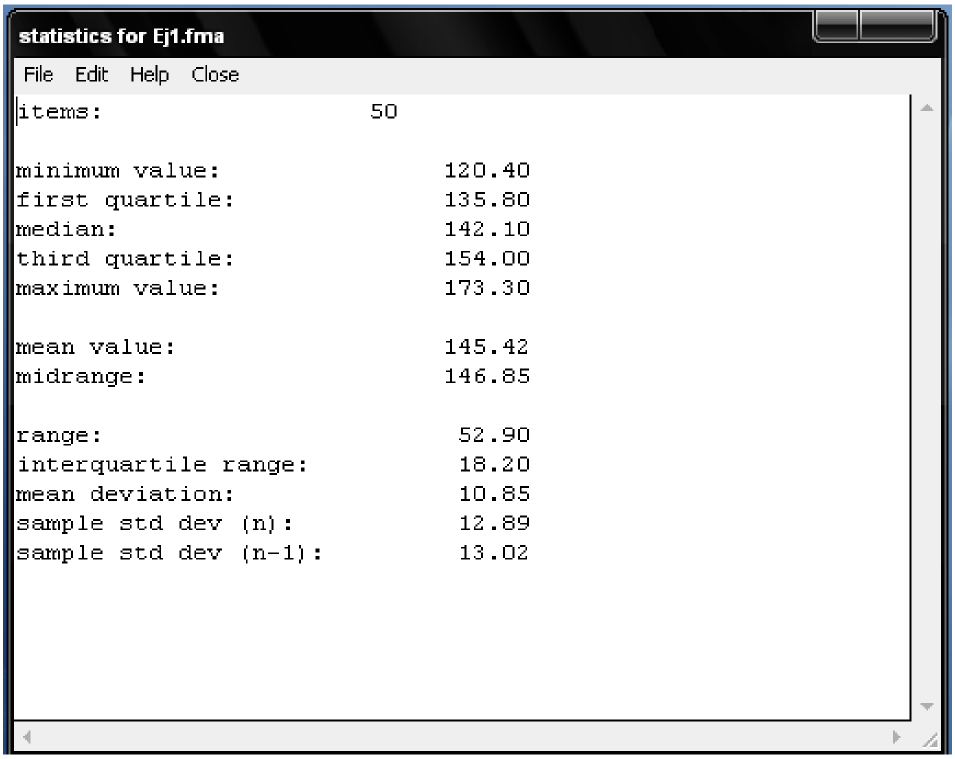


Imagen 1

Cuadro 3. Significados

Información	**Significado**
Ítems	Cantidad de datos
Mínimum value	Valor mínimo de la muestra
First quartile	Primer cuartil
Median	Mediana
Third quartile	Tercer cuartil
Maximum value	Valor máximo de la muestra
Mean value	Media aritmética
Midrange	Punto medio del rango
Range	Rango
Interquartile range	Rango intercuartílico (IRQ)
Mean deviation	Desviación media
Sample std dev (n)	Desviación típica
Sample std dev (n-1)	Cuasi-desviación típica

Laboratorio 2 (construcción del histograma)

Para este laboratorio utilice los datos ya digitados en el laboratorio 1.

Instrucciones

1. Para indicar al *software* el intervalo de trabajo, se ubica en $\boxed{Stats}\Rightarrow\boxed{Histogram}\Rightarrow\boxed{Interval\ ...}$. En ese momento se abre la siguiente pantalla, en la cual se digitan el dato mínimo y el máximo:

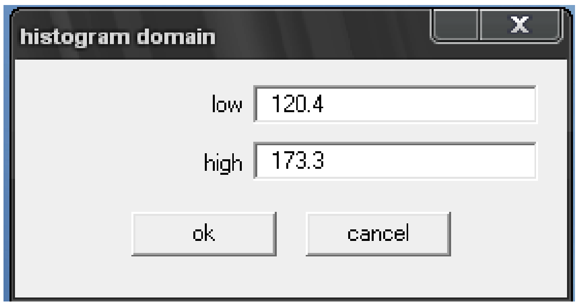


Imagen 2

2. Seguidamente, se le indica a la aplicación de cuántas clases desea el histograma. Para ello, realice la siguiente secuencia: $\boxed{Stats}\Rightarrow\boxed{Histogram}\Rightarrow\boxed{Number\ of\ groups}$; en ese momento, se abre una ventana, en este caso, indica 8 clases, de la siguiente manera:

Imagen 3

3. Ahora tiene todo listo para ver el histograma, siguiendo la secuencia: $\boxed{Stats}\Rightarrow\boxed{Histogram}\Rightarrow\boxed{Histogram}$. Se abre la siguiente ventana:

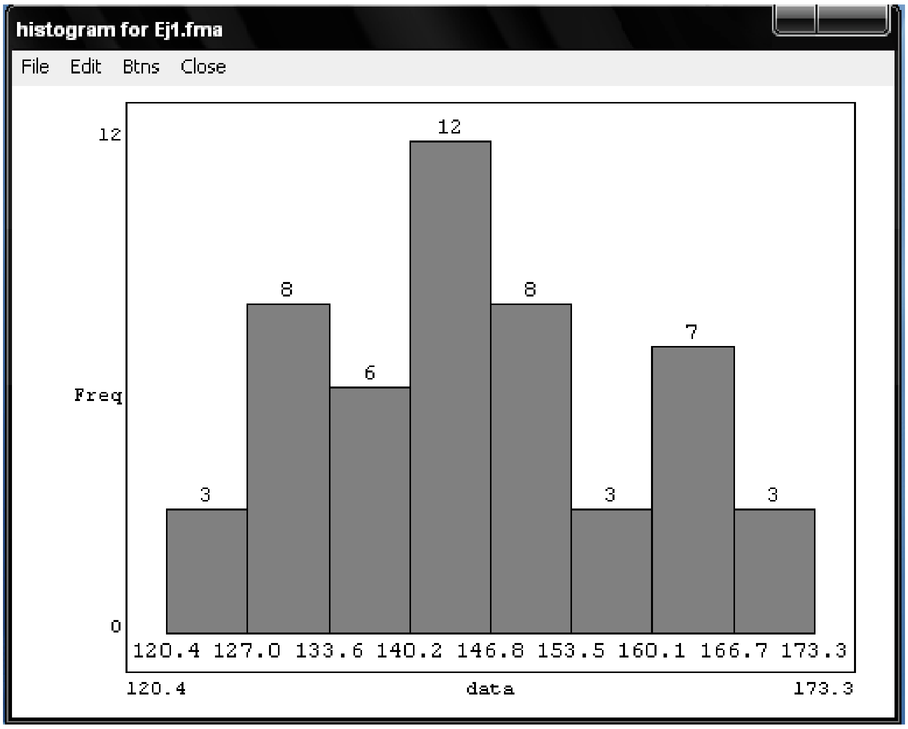


Imagen 4

4. Adicionalmente, puede observar la distribución de frecuencia de los datos y el diagrama tallo y hoja realizando las siguientes secuencias, respectivamente: $\boxed{Stats} \Rightarrow \boxed{Histogram} \Rightarrow \boxed{Frequencies}$

frequencies for Ej1.fma

File Edit Help Close

120.4	...	127.0	3	6.00%
127.0	...	133.6	8	16.00%
133.6	...	140.2	6	12.00%
140.2	...	146.8	12	24.00%
146.8	...	153.5	8	16.00%
153.5	...	160.1	3	6.00%
160.1	...	166.7	7	14.00%
166.7	...	173.3	3	6.00%

Imagen 5

Stats → Stemplots → Stemplots ...

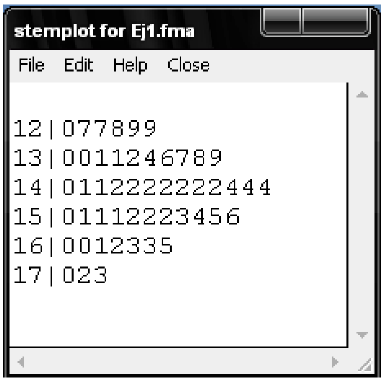


Imagen 6

EJERCICIOS DE AUTOEVALUACIÓN

1. Sea el conjunto $\{1,5,6,7,9,10\}$, calcule la media, la mediana y la moda (si existe).

2. Hallar la media y la varianza muestrales $\overline{X}$ y s^2 para el conjunto de datos $\{2,3,5,7,11,13,17,19\}$.

3. Muestre que las n piezas de información en $\sum_{i=1}^{n}(x_1 - x_2)^2$ no son independientes; es decir, muestre que:

   $\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = 0$ (Ejercicio 1.16, página 28 del libro de texto).

4. Si se tienen 46 datos ordenados de menor a mayor, ¿en cuál posición se encuentra el dato que corresponde a P_{12}?

5. Hallar los cuartiles Q_1, Q_2 y Q_3 para los siguientes datos:

5,7	8,9	10,1	13,0	16,2	20,8	20,9	22,7

6. Considere la tabla 2 de distribuciones, donde se agrupa la cantidad de goles anotados por 6 jugadores (x) del futbol nacional.

Tabla 2

x	***f***	***F***	***fr***
1	4		0,100
2	8		
3		15	0,075
4	8	23	
5		28	0,125
6			

   a) Complete los datos que faltan en la tabla 2, donde *f, F, fr* representan, respectivamente, la frecuencia absoluta, la frecuencia acumulada y la frecuencia relativa.

   b) Calcule la media, la mediana y la moda de esta distribución.

7. La siguiente es una muestra de 50 pesos (en libras) de jóvenes varones de 18 a 25 años de edad:

120,4	126,8	127,0	128,4	129,1	129,4	129,8	130,3	131,0	131,1
132,3	134,3	135,8	136,7	137,8	139,3	140,1	141,2	141,3	141,5
141,7	141,8	141,9	141,9	141,9	142,3	143,5	143,5	143,8	149,8
150,9	151,0	151,3	151,5	152,0	152,2	153,0	154,0	154,8	155,6
160,1	160,4	160,8	162,0	162,8	163,0	164,5	170,0	172,0	173,3

   Considere la muestra anterior y realice lo siguiente:

   a) determine las medidas de tendencia central: media, mediana y moda;

   b) determine las medidas de variabilidad: varianza y desviación estándar;

   c) realice una tabla de distribución de frecuencias que posea 5 clases o categorías;

d) represente los datos de la tabla de distribución de frecuencias mediante un histograma;

e) determine Q_1 y Q_3;

f) determine P_{23} y P_{54}.

8. Los datos que se presentan a continuación corresponden a las notas de aprovechamiento del curso de Ecuaciones diferenciales, en la Universidad Estatal a Distancia.

6,3	8,1	9,4	6,9	4,9
5,2	3,4	5,2	5,5	6,1
8,0	6,6	6,1	7,2	5,9
8,7	7,7	8,5	4,4	7,2
7,5	6,4	8,9	6,1	6,9

De acuerdo con los datos anteriores, realice lo que se le solicita a continuación:

a) represente los datos mediante un diagrama tallo-hoja;

b) calcule la media, mediana y moda de la población;

c) clasifique los datos en 5 clases y haga una tabla de distribución de frecuencias;

d) calcule la varianza poblacional de los datos;

e) calcule Q_1 y Q_3;

f) calcule P_{70}.

9. Una revista de modas hizo un estudio acerca de la estatura (en cm) de la mujer joven costarricense. Para realizarlo tomó como muestra la estatura de 30 mujeres

con edades que varían entre los 20 y los 40 años. Los resultados del estudio fueron los siguientes:

62	53	51	63	61	67	72	53	55	58
61	52	68	62	56	61	62	70	65	60
57	55	61	75	62	63	66	58	57	61

De acuerdo con los datos anteriores, realice lo que se le solicita a continuación:

a) calcule las medidas de tendencia central: media, mediana y moda;

b) clasifique los datos en 8 clases y haga una tabla de distribución de frecuencias;

c) calcule P_{31} y P_{80}.

10. Una empresa de venta de artículos para automóvil realiza un estudio acerca de la *vida útil* de las baterías de una marca X. Para ello recoge los datos de 40 baterías, con respecto al tiempo (en años) de su *vida útil*. Las baterías para automóvil se garantizan por 3 años.

2,2	4,1	3,5	4,5	3,2	3,7	3,0	2,6
3,4	1,6	3,1	3,3	3,8	3,1	4,7	3,7
2,5	4,3	3,4	3,6	2,9	3,3	3,9	3,1
3,3	3,1	3,7	4,4	3,2	4,1	1,9	3,4
4,7	3,8	3,2	2,6	3,9	3,0	4,2	3,5

a) Elabore un diagrama tallo-hojas.

b) Elabore un histograma de frecuencias; para ello, construya una tabla de frecuencias, cuyo número de clases sea 10. Tome como amplitud el intervalo $[1{,}50;4{,}90]$.

SOLUCIÓN A LOS EJERCICIOS DE AUTOEVALUACIÓN

1. Sea el conjunto $\{1,5,6,7,9,10\}$, con una muestra de 6 datos se tiene que:

- la media es $\overline{X} = \dfrac{1+5+6+...+10}{6} = 6,\bar{3}$;

- la mediana es $\overline{X} = \dfrac{6+7}{2} = 6,5$, ya que al haber 6 datos (cantidad par), la mediana será la media aritmética de los 2 valores medios, o sea la media de los valores numéricos de las posiciones 3 y 4;

- para esta muestra no existe la moda.

2. Hay que calcular la media y varianza muestrales $\overline{X}$ y s^2, para el conjunto de datos $\{2,3,5,7,11,13,17,19\}$. Entonces se sigue el siguiente procedimiento:

$$\overline{X} = \frac{2+3+5+7+11+13+17+19}{8} = 9,625.$$

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{1027 - \frac{(77)^2}{8}}{7} = 40,8393.$$

3. Desarrolle la expresión $\sum_{i=1}^{n}(X_i - \overline{X})$

$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\bar{x} = \sum_{i=1}^{n} x_i - n\bar{x} = \sum_{i=1}^{n} x_i - n\left(\frac{x_1 + x_2 + ... + x_n}{n}\right)$$

$$\sum_{i=1}^{n} x_i - \left(x_1 + x_2 + ... + x_n\right) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0.$$

Por lo tanto, se demuestra que n piezas de información en $\sum_{i=1}^{n}(X_i - \overline{X})^2$ no son independientes.

4. Para P_{12} se tiene que:

$$\frac{k \cdot n}{100} = \frac{12 \cdot 46}{100} = 5{,}51 \Rightarrow E = 5 \text{ y } D = 0{,}51.$$

Como $D \neq 0$ entonces $P_{12} = valor\left(5+1\right) = valor 6$.

Por lo tanto, P_{12} representa la posición 6 de la muestra.

5. Tome en cuenta que hay 8 valores y la mediana es la media aritmética de los dos valores medios, o sea la media de los valores numéricos de las posiciones 4 y 5; esto quiere decir que Q_1 será la media de los valores de las posiciones 2 y 3 (o sea $\frac{8{,}9+10{,}1}{2} = 9{,}5$), ya que de cuatro valores anteriores a la mediana el punto medio estará entre la posición 2 y 3; asimismo, Q_3 será el valor entre la posición 6 y la posición 7 (o sea $\frac{20{,}8+20{,}9}{2} = 20{,}85$).

6. La tabla 2 se completa con los siguientes valores:

x	***f***	***F***	***fr***
1	4	**4**	0,100
2	8	**12**	**0,200**
3	**3**	15	0,075
4	8	23	**0,200**
5	**5**	28	0,125
6	**12**	**40**	**0,300**

7. En este ejercicio, los datos se presentan ordenados, sin embargo, en la mayoría de los casos no se encuentran así, por lo tanto, se recomienda organizarlos en forma ascendente para así identificar, más claramente, los valores máximo y mínimo, y para facilitar su posterior clasificación.

a) Cálculo de las medidas de posición.

La media es $\overline{X} = \dfrac{120,4 + 126,8 + 127,0 + ... + 173,3}{50} = 145,42$.

La mediana es $\overline{X} = \dfrac{141,9 + 142,3}{2} = 142,1$, ya que al haber 50 datos (cantidad par), la mediana será la media aritmética de los 2 valores medios, o sea la media de los valores numéricos de las posiciones 25 y 26.

La moda es 141,9, pues es el valor que se presenta con mayor frecuencia. La muestra es *unimodal*, puesto que la moda es única.

b) Para realizar el ejercicio de forma ordenada y clara utilice la tabla 3.

Tabla 3

Dato #	x_i	x_i^2
1	120,4	14496,16
2	126,8	16078,24
3	127	16129
4	128,4	16486,56
5	129,1	16666,81
.	.	.
.	.	.
49	172	29584
50	173,3	30032,89
	∑x=7270,9	$\sum x^2$ =1065630,51

Ahora, al utilizar la fórmula alternativa para la varianza muestral se obtiene:

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{1065630,51 - \frac{(7270,9)^2}{50}}{49} = 169,6076...$$

Al calcular la desviación estándar s se tiene que: $s = \sqrt{s^2} = 13,0233...$

c) Se quiere construir una distribución de frecuencias de los datos anteriores en 5 clases $(k = 5)$, como la que se muestra en la tabla 4. Los datos máximo y mínimo son 173,3 y 120,4, respectivamente, por lo que la amplitud es:

$$c = \frac{173{,}3 - 120{,}4}{5} = 10{,}64.$$

Como los datos se presentan en décimas, la amplitud debe redondearse a décimos, entonces $c = 10{,}6$.

Tabla 4

Intervalo de clases	**Frecuencia absoluta(f_i)**	**Frecuencia relativa(f_r)**
120,4 – 130,9	8	0,16
131,0 – 141,5	12	0,24
141,6 – 152,1	15	0,30
152,2 – 162,7	9	0,18
162,8 – 173,3	6	0,12
Totales	50	1

Observación. Note que el valor de la amplitud de clase no se le suma al límite inferior de cada una, sino que se suma el valor unitario menor a la amplitud de clase, en este caso ese valor es 10,5.

d) Al representar mediante un histograma la anterior distribución de frecuencias, se obtiene el gráfico 2.

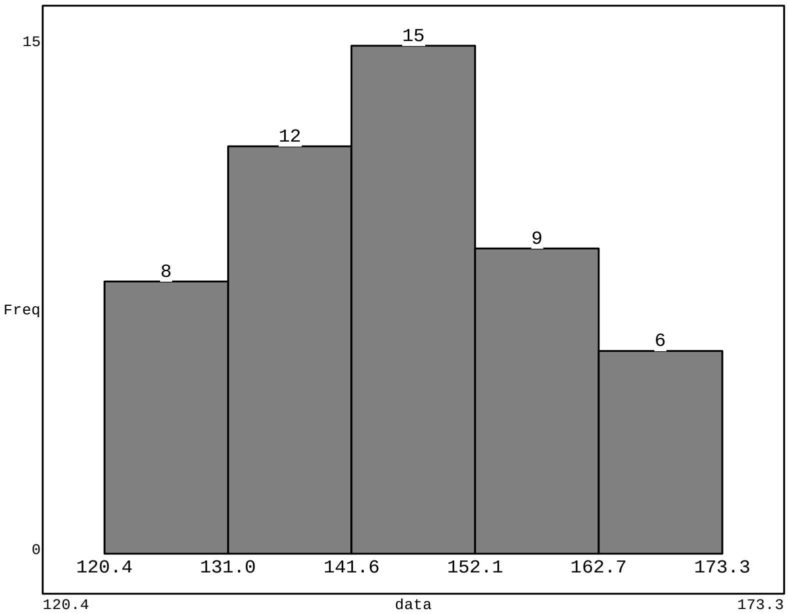


Gráfico 2

Al considerar que hay 50 valores y la mediana es la media aritmética de los 2 valores medios, o sea la media de las posiciones 25 y 26, Q_1 es el valor de la posición 13 (o sea 135,8), ya que de 25 valores anteriores a la mediana el punto medio está en la posición 13; y Q_3 es el valor de la posición 38 (o sea 154,0).

e) Para calcular los percentiles se toma $n = 50$.

Para P_{23} se tiene que:

$$\frac{k \cdot n}{100} = \frac{23 \cdot 50}{100} = 11{,}5 \Rightarrow E = 11 \text{ y } D = 0{,}5.$$

Como $D \neq 0$, $P_{23} = valor(11+1) = valor12 = 134{,}3$.

Para P_{54} se tiene que:

$$\frac{k \cdot n}{100} = \frac{54 \cdot 50}{100} = 27 \Rightarrow E = 27 \text{ y } D = 0.$$

Como $D=0$, $P_{54}=\dfrac{valor(27)+valor(28)}{2}=\dfrac{143{,}5+143{,}8}{2}=143{,}65$.

8. Primero es conveniente ordenar los datos de menor a mayor.

3,4	5,5	6,3	7,2	8,1
4,4	5,9	6,4	7,2	8,5
4,9	6,1	6,6	7,5	8,7
5,2	6,1	6,9	7,7	8,9
5,2	6,1	6,9	8	9,4

a) Ordenados los datos de menor a mayor proceda a realizar el diagrama tallo-hoja, con 7 clases.

Tallo	Hoja	Frecuencia
3	4	1
4	4 9	2
5	2 2 5 9	4
6	1 1 1 3 4 6 9 9	8
7	2 2 5 7	4
8	0 1 5 7 9	5
9	4	1

b) Se procederá a calcular las medidas de tendencia central.

La media poblacional es $\mu=\dfrac{3{,}4+4{,}4+4{,}9+\ldots+9{,}2}{25}=6{,}684$.

La mediana poblacional es $\overline{X}=6{,}6$, ya que,. al haber 25 datos (cantidad impar), es el valor medio, o sea el valor numérico que está en la posición 13. La moda de la población es 6,1, dado que se presenta con mayor frecuencia. La población es unimodal, puesto que la moda es única.

c) Se quiere hacer una tabla de distribución de frecuencias de 5 clases (o sea $k=5$) y se tiene que la amplitud o rango es $M-m=9{,}4-3{,}4=6$. Por lo tanto el tamaño de la clase será $c=\frac{6}{5}=1{,}2$.

Intervalo de clases	Frecuencia absoluta(f_i)	Frecuencia relativa(f_r)
[3,4; 4,6[	2	0,08
[4,6; 5,8[	4	0,16
[5,8; 7,0[	9	0,36
[7,0; 8,2[	6	0,24
[8,2; 9,4]	4	0,16
Totales	25	1

d) Para calcular la varianza poblacional se debe utilizar la siguiente fórmula:

$$\sigma^2=\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n}=\frac{(3{,}4-6{,}684)^2+(4{,}4-6{,}684)^2+\ldots(9{,}4-6{,}684)^2}{25}=2{,}0609.$$

e) Tome en cuenta que hay 25 valores, entonces la mediana es el valor que esté en la posición 13 (o sea 6,6), por lo tanto antes y después de la mediana quedan 12 valores. Por lo tanto, Q_1 es la media de los valores de las posiciones 6 y 7 (o sea $\frac{5{,}5+5{,}9}{2}=5{,}7$), ya que de 12 valores anteriores a la mediana, el punto medio estará entre la posición 6 y 7; de igual manera, Q_3 será el valor entre la posición 19 y la posición 20 (o sea $\frac{7{,}7+8}{2}=7{,}85$).

f) Para P_{70} se tiene que: $\frac{k\cdot n}{100}=\frac{70\cdot 25}{100}=17{,}5\Rightarrow E=17$ y $D=0{,}5$.

Como $D \neq 0$, $P_{12} = valor(17+1) = valor\ 18 = 7{,}5$.

9. Primero se ordenan los datos de la siguiente manera:

51	52	53	53	55	55	56	57	57	58
58	60	61	61	61	61	61	62	62	62
62	63	63	65	66	67	68	70	72	75

a) Se procederá a calcular las medidas de tendencia central.

La media muestral es $\overline{X} = \dfrac{51+52+53+53+...+70+72+75}{30} = 60{,}9$.

La mediana es $\overline{X} = \dfrac{61+61}{2} = 61$, ya que al haber 30 datos (cantidad par), la mediana es la media de los valores de las posiciones 15 y 16.

La moda de la población es 61, ya que es el valor que se presenta con mayor frecuencia. La población es unimodal, puesto que la moda es única.

b) Se quiere hacer una tabla de distribución de frecuencias de 8 clases (o sea $k = 8$) y se tiene que la amplitud o rango es $M - m = 75 - 51 = 24$. Por lo tanto, el tamaño de la clase será $c = \dfrac{24}{8} = 3$.

Intervalo de clases	**Frecuencia absoluta *(f_i)***
$[51,54[$	4
$[54,57[$	3
$[57,60[$	4
$[60,63[$	10
$[63,66[$	3
$[66,69[$	3
$[69,72[$	1
$[72,75[$	2
Totales	30

c) Para calcular los percentiles se toma en cuenta que $n = 30$.

Para P_{31} se tiene que:

$$\frac{k \cdot n}{100} = \frac{31 \cdot 30}{100} = 9{,}3 \Rightarrow E = 9 \text{ y } D = 0{,}3.$$

Como $D \neq 0$, $P_{35} = valor(9+1) = valor 10 = 58$.

Para P_{85} se tiene que:

$$\frac{k \cdot n}{100} = \frac{80 \cdot 30}{100} = 24 \Rightarrow E = 24 \text{ y } D = 0.$$

Como $D = 0$, $P_{80} = \dfrac{valor(24) + valor(25)}{2} = \dfrac{65+66}{2} = 65{,}5$.

10. Primero ordene los datos.

1,6	1,9	2,2	2,5	2,6	2,6	2,9	3,0
3,0	3,1	3,1	3,1	3,1	3,2	3,2	3,2
3,3	3,3	3,3	3,4	3,4	3,4	3,5	3,5
3,6	3,7	3,7	3,7	3,8	3,8	3,9	3,9
4,1	4,1	4,2	4,3	4,4	4,5	4,7	4,7

(a) Ordenados de menor a mayor, proceda a realizar el diagrama tallo-hoja, con 7 clases.

Tallo	Hoja	Frecuencia
1	6 9	2
2	2 5 6 6 9	5
3	0 0 1 1 1 1 2 2 2 3 3 3 4 4 4 5 5 6 7 7 7 7 8 8 9 9	25
4	1 1 2 3 4 5 7 7	8

(b) Se quiere hacer una tabla de distribución de frecuencias de 10 clases (o sea $k = 10$) y se tiene que la amplitud o rango es $M - m = 4{,}90 - 1{,}50 = 3{,}40$.

Por lo tanto, el tamaño de la clase será $c = \dfrac{3{,}40}{10} = 0{,}34$.

Intervalo de clases	Frecuencia absoluta(f_i)
1,50 – 1,84	1
1,84 – 2,18	1
2,18 – 2,52	2
2,52 – 2,86	2
2,86 – 3,20	10
3,20 – 3,54	8
3,54 – 3,88	6
3,88 – 4,22	5
4,22 – 4,56	3
4,56 – 4,90	2
Totales	40

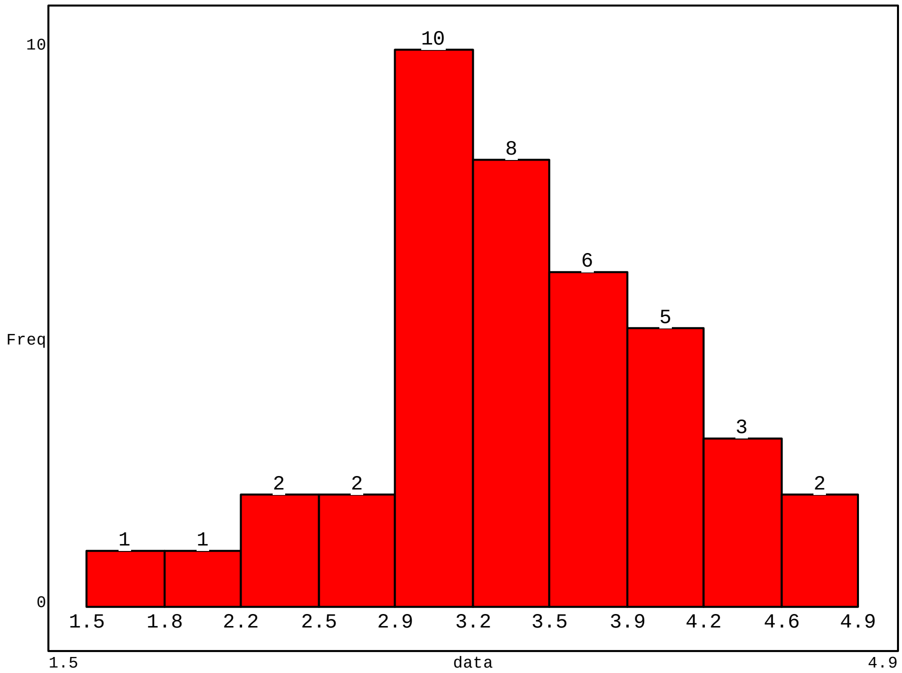


Gráfico 3

Glosario

clase. Subconjunto de un conjunto de datos que se distinguen de otros por algún rasgo peculiar.

cuartiles. Son aquellos valores que dividen al conjunto de datos en cuatro partes iguales.

datos agrupados. Datos ordenados y resumidos.

desviación estándar. Es el promedio de desviación o diferencia de los datos con respecto a la media aritmética.

medidas de dispersión. Son valores que miden cuán concentrados están los datos de una muestra o población alrededor de la medida de posición.

estadística descriptiva. La parte de la estadística que trata solamente de describir y analizar un grupo dado sin sacar conclusiones o inferencias de un grupo mayor.

estadística inferencial. La parte de la estadística que trata de las condiciones bajo las cuales tales inferencias son válidas.

estimación. Valoración numérica total de una unidad a partir de datos incompletos.

frecuencia absoluta. Es el número de unidades estadísticas que pertenecen a la clase o categoría.

frecuencia acumulada. Es la suma de la frecuencia de la clase, más las frecuencias de todas las categorías o clases anteriores.

frecuencia relativa. Es la frecuencia absoluta de una clase dividida por el total de frecuencias absolutas de todas las clases, y se expresa generalmente como porcentaje.

histograma. Gráfico utilizado para representaciones estadísticas, formado por rectángulos de igual anchura y altura proporcional a las cantidades que representan.

media aritmética. Es la suma de todos los valores dividida entre la cantidad de valores sumados. Es una medida de tendencia central que es sensible a la magnitud de los valores de cada uno de sus lados.

mediana. Es el valor medio o media aritmética de los dos valores medios. Es una medida de tendencia central que es sensible al número de valores de dichos lados.

medidas de tendencia central. Es un valor, que es típico o representativo de un conjunto de datos.

moda. La moda de una serie de valores es aquel valor que se presenta con la mayor frecuencia, es decir, es el valor más común. La moda puede no existir, incluso si existe puede no ser única.

muestra. Subconjunto de una población.

percentil. Son aquellos valores que dividen a los datos en cien partes iguales.

población. Colección de datos que atañen a las características de un grupo de individuos u objetos. La población puede ser un conjunto finito o infinito de elementos.

rango. El rango de un conjunto de números es la diferencia entre el mayor y el menor de todos ellos. También se le conoce como amplitud.

varianza. Cantidad que mide la dispersión de los valores que recorre una variable aleatoria.

SECCIÓN 2
CONCEPTOS BÁSICOS DE PROBABILIDAD

En esta sección se desarrollan varios conceptos básicos referidos a probabilidad, tales como: espacio muestral y evento, axiomas de probabilidades, probabilidad condicional, reglas de multiplicación y el teorema de Bayes. Asimismo, se expone una introducción a la teoría combinatoria, con temas como: conteo, permutaciones y combinaciones.

Adicionalmente, se le presenta una actividad dinámica con el *software* gratuito *winstats.exe*, para una mayor comprensión de los conceptos desarrollados en este capítulo, así como incentivar el uso de nuevas tecnologías para el proceso de enseñanza-aprendizaje de la matemática.

OBJETIVO GENERAL

Estudiar los conceptos básicos y las leyes que rigen las probabilidades de eventos simples y compuestos.

OBJETIVOS ESPECÍFICOS

1. Utilizar las técnicas de conteo.
2. Calcular las permutaciones y combinaciones de un conjunto finito.
3. Construir diagramas de árbol.
4. Reconocer los axiomas de la probabilidad.
5. Calcular la probabilidad de los elementos de un espacio probabilístico.
6. Diferenciar los espacios muestrales finitos e infinitos.
7. Aplicar los teoremas relativos a la probabilidad condicionada.

8. Reconocer y aplicar la fórmula de Bayes.
9. Determinar cuándo dos eventos son independientes y calcular la probabilidad de que sucedan simultáneamente.

PUNTOS DE ATENCIÓN DE LAS SECCIONES 2.1 Y 2.2

En la sección 2.1 se da una explicación de lo que es un experimento estadístico, para luego definir el espacio muestral como el conjunto de todos los resultados (puntos muestrales) de un experimento estadístico.

Observación: se recomienda describir, mediante un enunciado o regla, el espacio muestral, cuando este tenga una cantidad grande o infinita de elementos.

EJEMPLO 1

Realice el siguiente experimento: lance al aire una moneda de 20 colones, una de 10 y otra de 5 colones al mismo tiempo; luego observe la cara superior de cada moneda, la cual puede ser escudo (E) o corona (C). Con este experimento se pueden obtener los siguientes espacios muestrales:

a) Registrar el número de escudos (E) posible en un lanzamiento. Por lo tanto, el espacio muestral sería $S_1 = \{0,1,2,3\}$.

b) Registrar la secuencia de coronas (C) y escudos (E), anotando como primer resultado la moneda de 20 colones; como segundo la de 10 y como tercer resultado la de 5 colones. Por lo tanto, el espacio muestral obtenido es: $S_2 = \{EEE, CEE, CCE, CCC, ECE, EEC, CEC\}$.

c) Registrar el hecho de que las tres monedas coincidan en la cara (*s*) y que no coincidan en la cara (*n*). Por lo tanto, el espacio muestral es: $S_3 = \{s, n\}$.

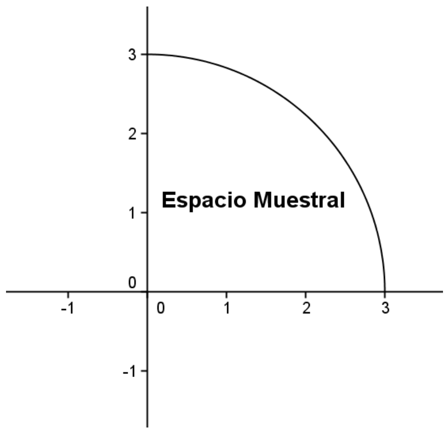
En los tres casos anteriores los espacios muestrales son finitos.

EJEMPLO 2

Tome como ejemplo al ejercicio 2.2 de la página 38, el que dice así:

"Utilice el método de la regla para describir el espacio muestral S, que consiste en todos los puntos del primer cuadrante dentro de un círculo de radio 3 con centro en el origen".

Como se ve en la gráfica 1, este es un espacio muestral infinito.



Gráfica 1

Ahora bien, si se conoce la ecuación del círculo se tiene que el espacio muestral es:

$$S = \left\{ (x, y) / x^2 + y^2 < 9; x \geq 0 \wedge y \geq 0 \right\}.$$

En la sección 2.2 del libro de texto, se desarrolla el concepto de evento (*E*) como un subconjunto del espacio muestral.

Así por ejemplo, si toma en cuenta el ejemplo 1 se tiene que un evento del espacio muestral S_1 es que en un lanzamiento no salgan escudos por lo que $E = \{0\}$. En el espacio muestral S_2, de ese mismo ejemplo, un evento sería que en el experimento al menos 2 caras sean corona (C), entonces $E = \{CCE, CEC, ECC, CCC\}$.

Igualmente, un evento del espacio muestral del ejemplo 2 es $E=\left\{(1,2),\left(\frac{5}{2},\frac{1}{2}\right)\right\}$ puesto que: $1^2+2^2<9 \quad \wedge \quad \left(\frac{5}{2}\right)^2+\left(\frac{1}{2}\right)^2<9$.

PUNTOS DE ATENCIÓN DE LA SECCIÓN 2.3

En la sección 2.3 se introduce el tema de análisis combinatorio de una forma básica pues es de gran importancia para la obtención de la cardinalidad de un evento o espacio muestral complejo, para facilitar esta tarea, es preciso conocer los siguientes principios básicos:

Un *diagrama de árbol* es un esquema para enumerar todas las apariciones posibles de una secuencia de experimentos o eventos, donde cada uno puede ocurrir de un número infinito de maneras. La construcción de los diagramas se ilustra en el ejemplo 3.

EJEMPLO 3

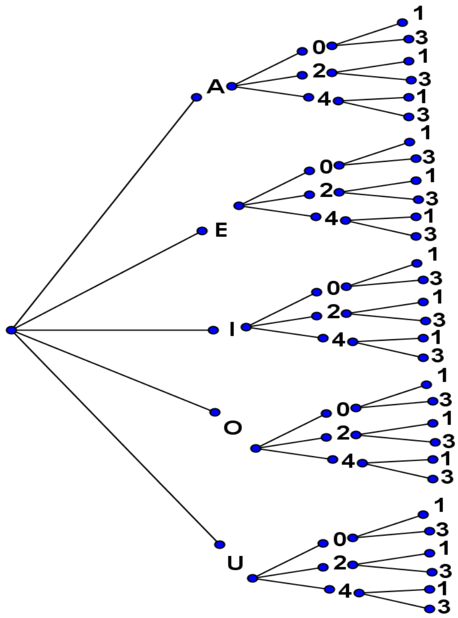
Para elegir una clave de acceso a un juego de video, el cual se lleva a cabo en línea, se tiene que hacer lo siguiente: la clave consta tres 3 elementos, el primero es escoger una con las cinco vocales, luego un número par de 0 a 5, y por último un número impar de 0 a 5. Realice lo siguiente:

a) un diagrama de árbol para representar la información;

b) ¿cuántos puntos muestrales hay en el espacio muestral?

SOLUCIÓN

a) En la gráfica 2 se muestra el diagrama de árbol representativo.



Gráfica 2

b) Contando el último nivel de las ramas se observa que hay 30 claves de acceso posibles.

Un concepto de mucha importancia es el concepto de factorial de n, el cual se denota $n!$ y se define de la siguiente manera: $n!=1\cdot 2\cdot 3\cdots(n-2)\cdot(n-1)\cdot n$. Donde $n\in IN \quad \wedge \quad 0!=1$. Por ejemplo, $5!=1\cdot 2\cdot 3\cdot 4\cdot 5=120$ y $4!\cdot 3!=(1\cdot 2\cdot 3\cdot 4)(1\cdot 2\cdot 3)=144$.

Para calcular de forma rápida, en la calculadora se le presenta la opción de cálculo de factorial con las teclas $\boxed{n!}$ o $\boxed{x!}$. Pero, a veces, n es muy grande, por lo que conseguir un resultado exacto de $n!$ es imposible hasta para los modernos computadores. Por tanto, frecuentemente, se utiliza la fórmula de aproximación de Stirling $n!\sim\sqrt{2\pi n}\cdot n^{n}\cdot e^{-n}$.

El símbolo $\sim$ significa que, cuanto más grande sea n, es decir cuando $n\rightarrow+\infty$, el valor de los dos términos tiende a 1.

En las páginas de la 43 a la 47 del texto se explican los distintos tipos de permutaciones, así como el concepto de combinaciones.

Se utiliza el símbolo $_nP_r$ para representar el número de permutaciones de n objetos distintos tomados de r a la vez; pero, en otros libros, se utilizan los símbolos $P(n,r)$, $P_{n,r}$, P_r^n.

De igual forma, se utiliza $_nC_r$ para representar el número de combinaciones de n objetos distintos tomados de r a la vez; en otros textos, se emplean $C(n,r)$, $C_{n,r}$, C_r^n. A continuación se le presenta la tabla 1, que es un resumen de los conceptos de permutaciones y combinaciones.

Tabla 1

Tipo	Fórmula	**Ejemplo.** Sea el conjunto de las 5 vocales del abecederario
Permutaciones de n objetos	Sin repetición de objetos $n!$	Las permutaciones de las 5 vocales, sin repetición de vocales $5!=120$
	Con repetición de objetos n^n	Las permutaciones de las 5 vocales, con repetición de vocales. Por ejemplo: aiaao, eiaie, uuueu, $5^5=3125$
Permutaciones de n objetos tomados de r en r	Sin repetición de objetos $_nP_r=\dfrac{n!}{(n-r)!}$	Las permutaciones de las 5 vocales, tomando parejas $_5P_2=\dfrac{5!}{(5-2)!}=\dfrac{5!}{3!}=\dfrac{120}{6}=20$
	Con repetición de objetos $_nP_r=n^r$	Las permutaciones de las 5 vocales, tomando parejas, permitiendo repetición de vocales. Por ejemplo,: aa, ee, ii, oo, uu $_5P_2=5^2=25$
Permutaciones de n objetos arreglados en círculo	$(n-1)!$	Las diferentes maneras de acomodar las 5 vocales en forma de un círculo $(5-1)!=4!=24$
Permutaciones distintas de n objetos de los que n_1 son de una clase, n_2 de una segunda clase, …, n_4 de una k-ésima clase	$\dfrac{n!}{n_1\cdot n_2\cdots n_k}$	Diferentes maneras de acomodar las 5 vocales si solo se distingue que hay 3 vocales graves y 2 agudas. $\dfrac{5!}{3!\cdot 2!}=\dfrac{120}{6\cdot 2}=\dfrac{120}{12}=10$
Combinaciones de n objetos tomados de r en r	$_nC_r=\dfrac{n!}{r!(n-r)!}$	Las combinaciones de las 5 vocales, tomando parejas $_5C_2=\dfrac{5!}{2!(5-2)!}=\dfrac{5!}{2!\cdot 3!}=\dfrac{120}{12}=10$ Observación*:* el orden de las vocales no importa. O sea, es el mismo elemento ae que ea

EJEMPLO 4

Se explicará el ejercicio 2.38 de la página 48, el cual dice así:

"Cuatro matrimonios compran 8 lugares en la misma fila para un concierto. ¿De cuántas maneras diferentes se pueden sentar

a) sin restricciones?

b) si cada pareja se sienta junta?

c) si todos los hombres se sientan juntos a la derecha de todas las mujeres?"

SOLUCIÓN

En la parte (a), como no hay restricciones, se tiene que es un caso de permutaciones sin repetición de 8 personas. Por lo tanto, hay $8!$ (o sea 40320) maneras diferentes como pueden sentarse. En la parte (b), como cada pareja se sienta junta, entonces hay $4!$ formas de sentarse y como el esposo se puede sentar a la izquierda o a la derecha de su esposa, por cada pareja existen 2 posibilidades de sentarse (o sea 2^4), por lo tanto, con la regla de la multiplicación se tiene que existen $4! \cdot 2^4$ (o sea 384) formas diferentes de sentarse dado que cada pareja se sienta junta. En la (c), como ya se sabe que los hombres estarán juntos y las mujeres también, entonces los hombres permutarán entre sí, al igual que las mujeres. Por tanto, por la regla de la multiplicación (teorema 2.1), se tiene que hay $4! \cdot 4!$ (o sea 576) formas diferentes de sentarse, de tal manera que los hombres se sienten juntos a la derecha de todas las mujeres.

PUNTOS DE ATENCIÓN DE LAS SECCIONES 2.4 Y 2.5

En las secciones 2.4 y 2.5, se desarrolla la introducción al tema de probabilidad básica, que consiste en el estudio de fenómenos puramente aleatorios; y se usa para indicar la posibilidad de que ocurra un evento o resultado.

En el teorema 2.9, de la página 50, se le presenta la fórmula para la probabilidad de eventos equiprobables (igual probabilidad). En el ejemplo 5 se ilustra el teorema.

Ejemplo 5

Sea un mazo ordinario de naipes con 52 cartas. Averigüe las siguientes probabilidades:

a) sacar una carta de corazones,

b) sacar una reina o un rey.

Solución

Como el espacio muestral consta de 52 posibilidades, entonces $N=52$.

a) Al haber 13 cartas de corazones, entonces $n=13$. Por lo tanto, se tiene que la probabilidad de sacar una de un mazo ordinario de naipes y que esta sea de corazones es de $\frac{13}{52}=\frac{1}{4}$.

b) Al haber en total 4 cartas de reinas y 4 de rey, entonces $n=4+4=8$. Por lo tanto, se tiene que la probabilidad de sacar una de un mazo ordinario de naipes y que esta sea un rey o una reina es de $\frac{8}{52}=\frac{2}{13}$.

En el ejemplo 6, se ilustra el uso de las propiedades aditivas de la probabilidad que se desarrolla en la sección 2.5 del texto.

Ejemplo 6

Se desarrolla el ejercicio 2.54 del libro, que dice así:

“Suponga que en un grupo de último año de facultad de 500 estudiantes se encuentra que 210 fuman, 258 consumen bebidas alcohólicas, 216 comen entre comidas; 122 fuman y consumen bebidas alcohólicas, 83 comen entre comidas y consumen bebidas alcohólicas, 97 fuman y comen entre comidas, y 52 tienen esos tres hábitos nocivos para la salud. Si se selecciona al azar a un miembro de este grupo, encuentre la probabilidad de que el estudiante

(a) fume, pero no consuma bebidas alcohólica;

(b) coma entre comidas y consuma bebidas alcohólicas, pero no fume;

(c) ni fume ni coma entre comidas."

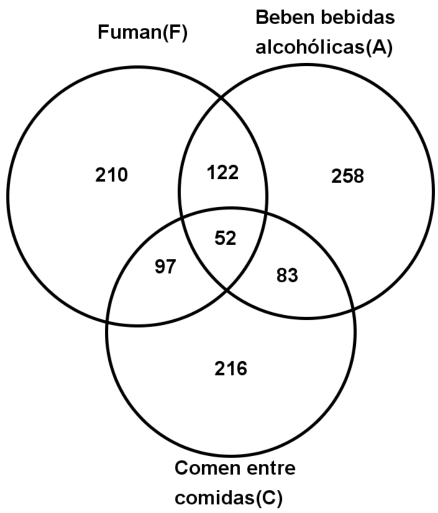
SOLUCIÓN

Para este tipo de ejercicios es conveniente hacer un diagrama de Venn.

(a) $P(F \cap A') = \frac{88}{500} = \frac{22}{125}$

(b) $P(C \cap A \cap F') = \frac{31}{500}$

(c) $P(F' \cap C') = \frac{171}{500}$



PUNTOS DE ATENCIÓN DE LAS SECCIONES 2.6 Y 2.7

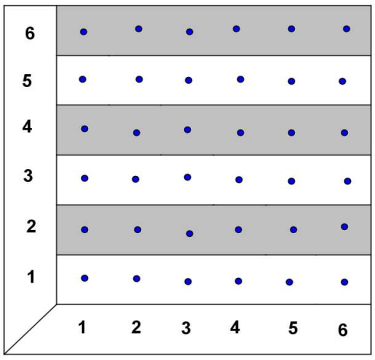
En la sección 2.6 se desarrolla la probabilidad de que ocurra un evento, siempre que haya sucedido otro antes. Esto recibe el nombre de probabilidad condicional, porque se conoce la condición de que haya pasado uno anteriormente. En el ejemplo 7 se ilustra el concepto.

EJEMPLO 7

Si se lanzan dos dados, y se sabe que uno, de antemano, muestra un número par. ¿Cuál es la probabilidad de que la suma de puntos de ambos dados sea mayor a 6?

SOLUCIÓN

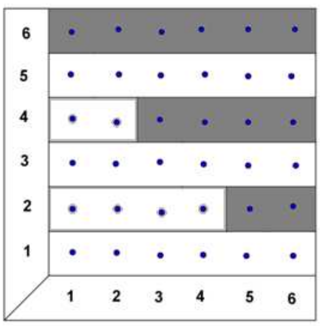
La condición de que el número mostrado, por el primer dado, sea par cambia el espacio muestral. En particular, el nuevo contiene solamente 18 puntos, que se ven en las filas sombreadas 2, 4 y 6 representadas en la gráfica 3; donde seis filas representan los resultados de un dado y seis columnas los del otro.



Gráfica 3

¿Para cuántos de los puntos del otro dado la suma es mayor que 6?

Como se muestra en la gráfica 4, solamente 12 puntos sombreados cumplen las dos condiciones.



Gráfica 4

Por lo tanto, la probabilidad de que los dados sumen más que 6, si uno mostró un número par, es de $\frac{12}{18}$, o sea 67% aproximadamente.

PUNTOS DE ATENCIÓN DE LA SECCIÓN 2.8

En la sección 2.8 se desarrolla el teorema de la probabilidad total y el teorema de Bayes. En el ejemplo 8 se ilustra el de la probabilidad total.

EJEMPLO 8

Tome como ejemplo el ejercicio 2.102 de la página 72, el cual se enuncia:

> "La policía planea hacer cumplir los límites de velocidad usando un sistema de radar en cuatro diferentes puntos dentro de la ciudad. Las trampas de radar en cada uno de los sitios L_1, L_2, L_3 y L_4 operan 40, 30, 20 y 30% del tiempo, y si una persona maneja a gran velocidad cuando va a su trabajo tiene las probabilidades de 0.2, 0.1, 0.5 y 0.2, respectivamente, de pasar por esos lugares. ¿Cuál es la probabilidad de que reciba una multa por conducir con exceso de velocidad?"

SOLUCIÓN

Sea E_1, E_2, E_3 y E_4 los eventos en los que una persona haya sido detectada con alta velocidad por los radares en los sitios L_1, L_2, L_3 y L_4, respectivamente. Sea M el evento de recibir una multa por conducir con exceso de velocidad. Al utilizar la ley de probabilidad total se tiene que:

$$P(M) = \sum_{i=1}^{4} P(M|E_i)P(E_i) = (0,4)(0,2) + (0,3)(0,1) + (0,2)(0,5) + (0,3)(0,2)$$

$$\therefore P(M) = 0,27.$$

Por lo tanto, hay 27% de probabilidad de que reciba una multa por conducir con exceso de velocidad.

El teorema 2.17 (página 71), conocido como Regla de Bayes, debido al matemático inglés Thomas Bayes (1702-1761), se apoya en el proceso inverso de la ley de la probabilidad total. En el ejemplo 9 se expone.

EJEMPLO 9

En un aeropuerto utilizan 3 perros policías (*A*, *B* y *C*) para detectar droga en las maletas de los pasajeros. El control dura 12 horas, de las cuales el perro A trabaja 5 horas, el perro B 4 y el perro C 3 horas. El perro A tiene 90% de confiabilidad en su detección de droga, el B tiene 92% y el C tiene 95%. Si un día se encuentra droga en las maletas de un pasajero, ¿cuál es la probabilidad de que lo haya detectado el perro A?

SOLUCIÓN

Sea *D* el evento de haber encontrado droga en la maleta de un pasajero; y $P(A)$ la probabilidad de que el perro *A* esté trabajando en ese momento, $P(B)$ la de que sea el perro *B* y $P(C)$ el perro *C*. Sean $P(D|A)$, $P(D|B)$ y $P(D|C)$ las probabilidades de confiabilidad que hay de los perros *A*, *B* y *C*, respectivamente.

$P(A|D)$ es la probabilidad de que el perro *A* sea el que ha detectado la droga, dado que se encontró en la maleta de un pasajero. Al utilizar la Regla de Bayes se tiene que:

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

$$P(A|D) = \frac{\left(\frac{5}{12}\right)(0{,}90)}{\left(\frac{5}{12}\right)(0{,}90) + \left(\frac{4}{12}\right)(0{,}92) + \left(\frac{3}{12}\right)(0{,}95)} \approx 0{,}41.$$

Por lo tanto, hay aproximadamente 41% de probabilidad de que si se descubre droga en una maleta de un pasajero haya sido el perro *A* el que la detectó.

EJERCICIO ADICIONAL. Compruebe que se cumple $P(A|D)+P(B|D)+P(C|D)=1$.

LABORATORIO

Sea el experimento de lanzar un dardo hacia un tablero cuadrado de lado 4 cm, con un círculo concéntrico de radio 2 cm. ¿Cuál es la probabilidad de que al tirar el dardo este "caiga" en el círculo?

Este problema se puede resolver como el ejemplo 11. El espacio muestral (*S*) es el área total del tablero cuadrado de lado 4 cm y el evento (*A*) es el área del círculo concéntrico de radio 2 cm. Por lo tanto, $N=(4)^2=16cm^2$ y $n=\pi(2)^2=4\pi cm^2$. Entonces,

$$P(A)=\frac{n}{N}=\frac{4\pi}{16}=\frac{\pi}{4}\approx 0,785.$$

La probabilidad de que, al lanzar un dardo, este "caiga" en el círculo es de 78,5%, aproximadamente.

Ahora bien, al utilizar el programa *winstats.exe*, se puede realizar este experimento y comparar los resultados.

Instrucciones

1. Abra el *software winstats.exe,* haga doble clic en el ícono correspondiente. Se abrirá una pequeña pantalla verde con dos opciones en el menú: $\boxed{Window}$, $\boxed{Help}$.

2. Ubíquese en $\boxed{Window}$ y realice la siguiente secuencia: $\boxed{Simulations}\rightarrow\boxed{throw\ darts}$. Al hacerlo, le aparecerá una ventana con un tablero cuadrado con un círculo concéntrico, el cual le ayudará a simular el problema. En adelante, se trabaja solo con el menú principal de esta ventana.

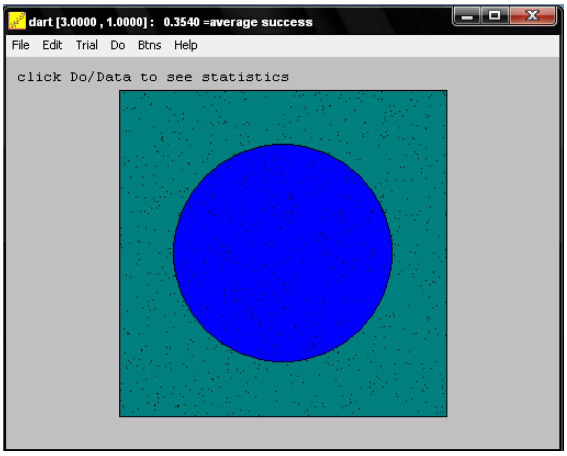


Imagen 1

3. Por defecto, el cuadrado tiene 3 unidades de lado y el círculo 1 unidad de radio; para cambiar estas dimensiones, en el menú principal, se realiza la secuencia: $\boxed{edit} \rightarrow \boxed{Parameters\ldots}$. En ese momento, emergerá una ventana con los parámetros: *square side* y *target radius*. En *square side* digite 4 y en *target radius,* 2, oprima el botón $\boxed{ok}$. De momento no cambiará la imagen del experimento, pero en el paso siguiente lo hará.

4. Con la tecla $\boxed{F1}$ de su computadora, se ejecutará el evento de tirar un dardo, esto lo puede efectuar cuantas veces quiera oprimiendo la tecla. Realícelo para observar su comportamiento. Por defecto, el ancho del "dardo" es pequeño, si quiere una mejor visualización siga la secuencia s: $\boxed{edit} \rightarrow \boxed{Dartsize\ldots}$ y digite 2 o 3 según su gusto.

5. Puede también tirar varias veces el dardo de una forma simultánea. Para elegir la cantidad de dardos por lanzar siga la secuencia $\boxed{Do} \rightarrow \boxed{Number\ldots}$ y en la casilla de texto escriba 1000. Luego, oprima la tecla $\boxed{F12}$ y podrá observar el evento.

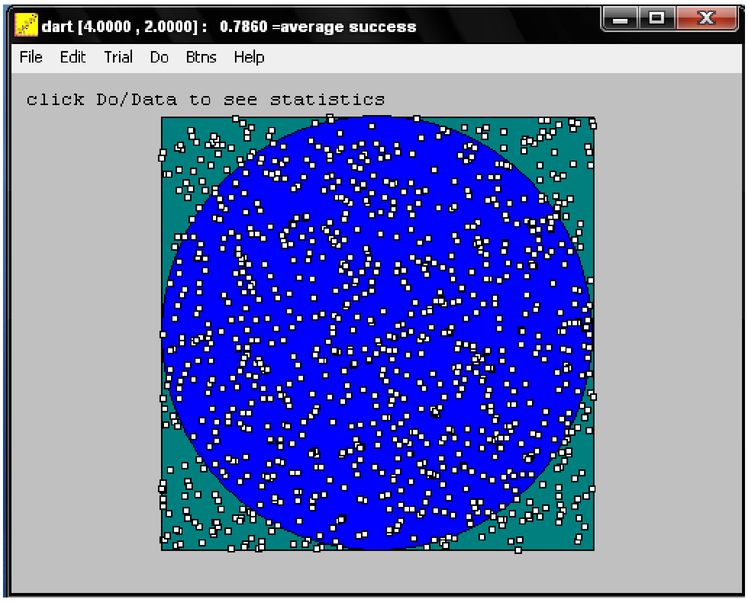


Imagen 2

6. El objetivo de este laboratorio es comparar el resultado obtenido, con la definición de probabilidad, con el conseguido por la simulación. Para esto, puede visualizar la tabla de frecuencias del evento, realice la secuencia: Do, Data, Freqencies ..., entonces aparecerá una ventana donde 1 es el evento de que el dardo "caiga" en el círculo y 0 el caso contrario.

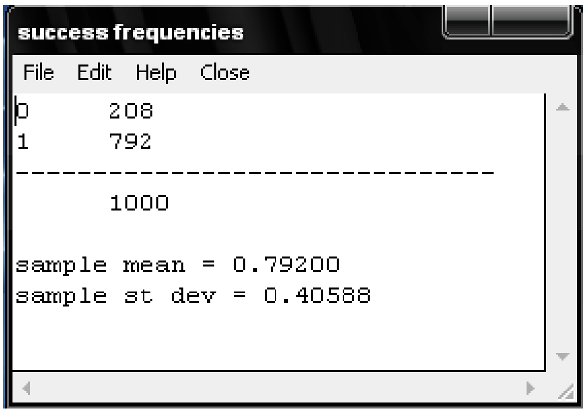


Imagen 3

7. Compare los resultados de la simulación con los conseguidos con la definición básica de la probabilidad de un evento.

EJERCICIOS DE AUTOEVALUACIÓN

1. Se realiza el experimento de lanzar 2 dados y se registran. Halle los elementos de los siguientes eventos:

    a) $A = \{\text{los dos números son iguales}\}$

    b) $B = \{\text{los dos números sumen más de 10}\}$

    c) $C = \{\text{sale un 3 en el segundo dado}\}$

    d) $D = \{\text{la suma de los dos números sea impar y mayor que 5}\}$

    e) $A \cap B$

    f) $C \cup D$

    g) $(A')'$

    h) $(B \cup C)'$

2. Calcule $\dfrac{2010!}{2009!}$.

3. Una clave de acceso de un banco consta de 4 dígitos seguidos de 3 letras minúsculas del alfabeto.

    a) ¿Cuántas claves diferentes se forman si los dígitos y las letras pueden repetirse?

    b) ¿Cuántas claves diferentes existen si los dígitos y las letras no pueden repetirse?

4. Calcule ${}_{n+1}P_{n-1}$.

5. En un grupo de 10 amigos, ¿cuántas distribuciones de sus fechas de cumpleaños se dan al año?

6. Para una fiesta se escogen 4 de 10 discos de salsa, 8 de 12 de merengue y 3 de 7 de cumbia. ¿Cuántas maneras de seleccionar la música existen?

7. En una clase de 10 alumnos van a distribuirse 3 premios. Averiguar de cuántos modos puede hacerse si

   a) los premios son diferentes y una persona no puede recibir más de uno;

   b) los premios son iguales y una persona no puede recibir más de un uno;

   c) los premios son diferentes y una persona puede recibir más de uno.

8. Hay que colocar a 5 hombres y 4 mujeres en una fila, de modo que las mujeres ocupen los lugares pares. ¿De cuántas maneras puede hacerse?

9. Considere un espacio muestral con 2 eventos A y B. ¿Es posible que $P(A)$ sea igual a $\frac{a}{a+b}$ y $P(B)$ sea igual a $\frac{b}{a+b}$?

10. Supóngase que, en un experimento, se tiene el espacio muestral $\{2,4,5,7,8\}$ y se han determinado las siguientes probabilidades elementales:

$$P(2)=\frac{1}{8} \quad P(4)=\frac{1}{12} \quad P(5)=\frac{3}{16} \quad P(7)=\frac{5}{12} \quad P(8)=\frac{3}{16}.$$

Sea A el evento "El resultado sea un número par". Calcule $P(A)$.

11. Considerar el experimento de escoger un número real del intervalo de 0 a 1, inclusive.

   a) ¿Cuál es el espacio muestral de este experimento?

   b) ¿Cuál es la probabilidad de escoger el número $\frac{1}{3}$?

   c) ¿Es posible escoger $\frac{1}{3}$?

12. ¿Cuál es la probabilidad de que si se escoge un punto al azar dentro del triángulo equilátero de lado a, como el de la figura 1, no esté dentro del círculo inscrito en el triángulo?

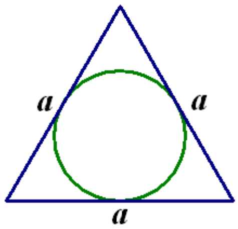


Figura 1

13. Demuestre que $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

14. En una escuela del centro de San José, algunos estudiantes tienen problemas dentales, 37% de ellos tiene caries, el 48% sarro, y el 12% caries y sarro a la vez.

    a) ¿Cuál es la probabilidad de que un estudiante tenga caries, si se sabe que tiene sarro?

    b) Si se escoge un niño al azar, ¿cuál es la probabilidad de que no tenga caries, dado que tiene sarro?

15. Una muestra aleatoria de 80 personas adultas se clasifica, en la tabla 2, por sexo y nivel de escolaridad.

Tabla 2

Nivel de escolaridad	Hombre	Mujer
Primaria	12	19
Secundaria	10	23
Universidad	9	7

Si se elige una persona al azar de este grupo, encuentre la probabilidad de que:

a) sea hombre, dado que tiene un nivel de escolaridad de secundaria;

b) sea mujer dado que no tiene un nivel de escolaridad universitario.

16. En la final del campeonato mundial de tiro al blanco, se enfrentan el competidor de Alemania (*A*) y el de Francia (*F*). Por resultados anteriores, se tiene que la probabilidad de que *A* dé en el blanco es de 92% y la de *F* es de 88%. Si ambos disparan, hallar la probabilidad de que:

    a) ambos den en el blanco;

    b) uno de ellos dé en el blanco;

    c) ninguno dé en el blanco.

17. En el equipo de Barcelona F. C., se tienen 3 delanteros titulares: Eto'o, Henry y Messi. Tienen la obligación de hacer goles, los últimos resultados indican que:

    - Eto'o realiza el 35% de los tiros a marco, y convierte en gol un 52% de los que ejecuta.
    - Henry realiza el 27% de los tiros a marco, y convierte en gol un 40%.
    - Messi realiza el 38% de los tiros a marco, y convierte en gol un 67%

    Realice lo que se le solicita:

    a) ¿qué probabilidad existe de que si alguno de los delanteros tira a marco, se convierta en gol?

    b) si en un partido hay un gol, ¿qué probabilidad hay de que lo haya hecho Lionel Messi?

SOLUCIÓN A LOS EJERCICIOS DE AUTOEVALUACIÓN

1. A continuación se le presentan los elementos de cada uno de los anteriores eventos.

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

   b) $B=\{(6,4),(5,5),(4,6),(6,5),(5,6),(6,6)\}$

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

   d) $D=\left\{\begin{matrix}(1,6),(3,4),(3,6),(5,2),(5,4),(5,6)\\(6,1),(4,3),(6,3),(2,5),(4,5),(6,5)\end{matrix}\right\}$

   e) $A\cap B=\{(5,5),(6,6)\}$

   f) $C\cup D=\left\{\begin{matrix}(1,6),(3,4),(3,6),(5,2),(5,4),(5,6),(6,1),(4,3),\\(6,3),(2,5),(4,5),(6,5),(6,4),(4,6),(5,5),(6,6)\end{matrix}\right\}$

   g) $(A')'=A$

   h) $(B\cup C)'=\left\{\begin{matrix}\text{la suma de los dos números sea menor o igual a 10,}\\\text{dado que el segundo dado no es 3}\end{matrix}\right\}$

2. Si intenta calcular $\frac{2010!}{2009!}$, con una calculadora científica, se dará cuenta de que no lo podrá hacer por falta de memoria. Mejor realice el siguiente procedimiento:

$$\frac{2010!}{2009!}=\frac{2009!\ \cdot\ 2010}{2009}=2010.$$

3. Asuma que hay 27 letras del alfabeto castellano y 10 dígitos tomados de 0 a 9.

   a) La cantidad de claves que se pueden formar permitiendo la repetición de dígitos y letras es $10^4\cdot 27^3$, o sea 196 830 000.

b) La cantidad de claves que se pueden formar sin permitir que se repitan dígitos ni letras es ${}_{10}P_4 \cdot {}_{27}P_3$, o sea (5040)(17550) = 88 452 000.

4. Según la fórmula de permutación se tiene que:

$$ {}_{n+1}P_{n-1} = \frac{(n+1)!}{\left[(n+1)-(n-1)\right]!} = \frac{(n+1)!}{(n+1-n+1)!} = \frac{(n+1)!}{2!} = \frac{(n+1)!}{2}. $$

5. Considere que el año tiene 365 días y puede darse el caso de que varias personas cumplan años en la misma fecha, hay 365^{10} distribuciones distintas.

6. Utilice la fórmula de las combinaciones y la ley de la multiplicación, existe ${}_{10}C_4 \cdot {}_{12}C_8 \cdot {}_7C_3 = 210 \cdot 495 \cdot 35 = 3638250$ formas de seleccionar la música.

7.

   a) Hay ${}_{10}P_3 = 720$ maneras de distribuir los premios si estos son diferentes.

   b) En caso de que los premios sean iguales, pueden repartirse de ${}_{10}C_3 = 120$ maneras.

   c) Si estos son diferentes y una persona puede recibir más de un premio, entonces existen $10^3 = 1000$ formas diferentes.

8. El cuadro 1 representa el problema:

Cuadro 1

Posibilidades	5	4	4	3	3	2	2	1	1
Género	H	M	H	M	H	M	H	M	H
Lugares	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

Aplicando la ley de la multiplicación, hay $5 \cdot 4 \cdot 4 \cdot 3 \cdot 3 \cdot 2 \cdot 2 \cdot 1 \cdot 1 = 2880$ posibilidades de ubicarse en la fila.

9. Sea S el espacio muestral tal que $S = A \cup B$ entonces,

$$P(A)+P(B)=\frac{a}{a+b}+\frac{b}{a+b}=\frac{a+b}{a+b}=1.$$

Por lo tanto, es posible que $P(A)$ sea igual a $\frac{a}{a+b}$ y $P(B)$ sea igual a $\frac{b}{a+b}$.

10. Como en el espacio muestral existen tres números pares, entonces se tiene que:

$$P(A)=\frac{1}{8}+\frac{1}{12}+\frac{3}{16}=\frac{19}{48}\approx 0,40.$$

Por lo tanto, hay aproximadamente un 40% de probabilidad de que el resultado sea un número par.

11.

a) El espacio muestral es el conjunto infinito de los números reales en el intervalo de 0 a 1.

b) La probabilidad de escoger el número $\frac{1}{3}$ es aproximadamente 0.

c) Aunque la probabilidad es una fracción muy pequeña, aproximadamente es 0, no es imposible escoger a $\frac{1}{3}$.

12. El área del círculo es πa^2 y el área del triángulo equilátero es igual a $\frac{\left(2a\sqrt{3}\right)^2 \cdot \sqrt{3}}{4}=3a^2\sqrt{3}$. Entonces, el evento de escoger un punto al azar dentro del triángulo equilátero y que no esté dentro del círculo inscrito en el triángulo es:

$$A_E = 3a^2\sqrt{3}-\pi a^2 = a^2\left(3\sqrt{3}-\pi\right).$$

El espacio muestral es: $A_S = 3a^2\sqrt{3}$.

Por lo tanto, la probabilidad buscada es:

$$P(E) = \frac{A_E}{A_S} = \frac{a^2\left(3\sqrt{3}-\pi\right)}{3\sqrt{3}} \approx 0,40.$$

Hay aproximadamente 40% de probabilidad de que si se escoge un punto al azar dentro del triángulo equilátero no esté dentro del círculo inscrito.

13. Se sabe que $P(A \cup B) = P(A) + P(B|A)$ y como $P(B|A) = P(B) - P(A \cap B)$ entonces $P(A \cup B) = P(A) + P(B) - (A \cap B)$.

14. Sea S el evento de tener sarro y C el de tener caries, se tiene que:

 a) La probabilidad de que un estudiante tenga caries, si se sabe que tiene sarro es:

$$P(C|S) = \frac{P(C \cap S)}{P(S)} = \frac{0,12}{0,48} = \frac{1}{4}.$$

 Por lo tanto, hay un 25% de probabilidad de que un estudiante tenga caries dado que tiene sarro.

 b) Si se escoge un niño al azar, la probabilidad de que no tenga caries, dado que tenga sarro, es $P(C'|S)$ pero no se sabe $P(C' \cap S)$ entonces se aplica la siguiente propiedad:

$$P(C|S) + P(C'|S) = 1$$

$$\Rightarrow P(C'|S) = 1 - P(C|S) = 1 - \frac{1}{4} = \frac{3}{4}.$$

 Por lo tanto, hay 75% de probabilidad de que el niño escogido al azar no tenga caries, dado que tiene sarro.

15. Sea H el evento de que sea hombre y M mujer. Sea P, S y U el nivel de escolaridad del adulto (primaria, secundaria y universidad, respectivamente).

a) La probabilidad de que la persona sea hombre, dado que tiene grado de escolaridad de secundaria, se puede calcular con $P(H|S)$.

$$P(H|S)=\frac{P(H\cap S)}{P(S)}=\frac{\frac{10}{80}}{\frac{33}{80}}=\frac{10}{33}\approx 0{,}30.$$

Por lo tanto, hay aproximadamente 30% de probabilidad de que si se escoge a un individuo de educación secundaria este sea un hombre.

b) Si no tiene un grado académico universitario, dado que es mujer, se puede calcular con $P((P\cup S)|M)$.

$$P((P\cup S)|M)=\frac{(P\cup S)\cap M}{P(M)}=\frac{\frac{42}{80}}{\frac{49}{80}}\approx 0{,}86.$$

Por lo tanto, hay aproximadamente 86% de probabilidad de escoger una persona que no tiene grado académico universitario, dado que es mujer.

16. Se tiene que $P(A)=0{,}92$ y $P(F)=0{,}88$ y se asume que los eventos son independientes.

a) Como los eventos son independientes,

$$P(A \text{ y } F)=P(A\cap F)=P(A)\cdot P(F)=0{,}92\cdot 0{,}88=0{,}8096\approx 0{,}81.$$

Por lo tanto, hay aproximadamente 81% de probabilidad de que ambos den en el blanco.

b) Por la regla de la adición, se tiene que:

$$P(A \text{ o } F) = P(A \cup F) = P(A) + P(F) - P(A \cap F) = 0{,}92 + 0{,}88 - 0{,}81 \approx 0{,}99.$$

Por lo tanto, hay aproximadamente 99% de probabilidad de que uno de los dos dé en el blanco.

c) Utilizando la ley de De Morgan,

$$P(A' \cap F') = P((A \cup F)') = 1 - P(A \cup F) = 1 - 0{,}99 \approx 0{,}01.$$

O también, como A' y F' son eventos independientes entonces,

$$P(A' \cap F') = P(A') \cdot P(F') = 0{,}08 \cdot 0{,}12 = 0{,}0096 \approx 0{,}01.$$

Por lo tanto, hay 1% de que ninguno dé en el blanco.

17.

a) Sea G el evento de que al tirar a marco, alguno de estos delanteros, se convierta en gol; E el de Eto'o, H el de Henry y M el de Messi.

Al utilizar la ley de probabilidad total se tiene que:

$$P(G) = P(E) \cdot P(G|E) + P(H) \cdot P(G|H) + P(M) \cdot P(G|M),$$

$$P(G) = 0{,}35 \cdot 0{,}52 + 0{,}27 \cdot 0{,}40 + 0{,}38 \cdot 0{,}67 \approx 0{,}54.$$

Por lo tanto, hay aproximadamente 54% de probabilidad de que, si alguno de los 3 delanteros tira a marco, se concrete en gol.

b) Por el teorema de Bayes, se tiene que:

$$P(M|G)=\frac{P(M)\cdot P(G|M)}{P(G)}=\frac{0{,}38\cdot 0{,}67}{0{,}54}\approx 0{,}47.$$

Por lo tanto, hay aproximadamente 47% de probabilidad de que si el Barcelona F. C. hizo un gol, sea de Lionel Messi.

Glosario

combinación. Cada uno de los conjuntos de m elementos que pueden formarse con los n elementos de un conjunto dado.

evento. Conjunto de resultados o subconjunto del espacio muestral.

eventos equiprobable. Son eventos que tienen la misma probabilidad de ocurrir.

eventos mutuamente excluyentes o disjuntos. Son eventos que no tienen puntos muestrales en común.

eventos independientes. Dos eventos son independientes si la ocurrencia de uno de ellos no influye en la del otro.

espacio muestral. Conjunto de todos los resultados posibles de un experimento estadístico.

experimento estadístico. Proceso que genera un conjunto de datos estadísticos.

permutación. Una permutación de un número de objetos es cualquiera de los diferentes arreglos de esos objetos en un orden definido.

probabilidad. Medida del grado de ocurrencia de un evento.

probabilidad condicional. Es la probabilidad de que ocurra un evento, dado que ya ha ocurrido otro.

punto muestral. Es un resultado particular o elemento simple del espacio muestral.

teoría combinatoria. Parte de las matemáticas que estudia las propiedades de los elementos en cuanto a su posición y a los grupos que pueden formarse con dichos elementos.

SECCIÓN 3
VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD

En la sección se desarrollan los conceptos de variable aleatoria discreta y continua; así como los de media y varianza. Además, se exponen las distribuciones de probabilidad discreta, en las cuales se enfatizan la uniforme discreta, la binomial y la multinomial, se explican las distribuciones continuas de probabilidad en las que destacan la uniforme continua y la normal. Por último, se presenta el método para aproximar la distribución binomial por medio de la distribución normal.

Adicionalmente, se muestran dos actividades dinámicas con el uso del *software* gratuito *winstats.exe*, para buscar una mayor comprensión de los conceptos desarrollados en esta sección, así como incentivar el empleo de nuevas tecnologías para el proceso de enseñanza-aprendizaje de la matemática.

OBJETIVO GENERAL

Aplicar el concepto de variable aleatoria en problemas que involucren distribuciones de probabilidad discreta y continua.

OBJETIVOS ESPECÍFICOS

1. Aplicar el concepto de variable aleatoria discreta y continua.
2. Encontrar la distribución de probabilidad de una variable aleatoria discreta.
3. Encontrar la distribución de probabilidad acumulada de una variable aleatoria discreta.
4. Representar la función de densidad de una variable aleatoria continua.

5. Representar la función de probabilidad acumulada de una variable aleatoria continua.
6. Calcular la media y desviación típica de las distribuciones de probabilidad de variables aleatorias discretas y continuas.
7. Comprender y aplicar el concepto de distribución uniforme discreta.
8. Comprender y aplicar el concepto de distribución binomial.
9. Calcular la media y la varianza de un experimento binomial.
10. Aplicar la distribución multinomial a problemas probabilísticos.
11. Reconocer las propiedades de la curva de distribución normal.
12. Calcular áreas bajo la curva de la distribución normal.
13. Aproximar la probabilidad de una variable binomial por medio de la curva de distribución normal.

PUNTOS DE ATENCIÓN DE LAS SECCIONES 3.1 A LA 3.3

En las secciones 3.1, 3.2 y 3.3 del libro de texto se desarrolla un concepto fundamental para la probabilidad y la estadística, el de *variable aleatoria*, se usa cuando se desea asignar un número específico a cada resultado de un experimento.

EJEMPLO 1

Sea el experimento de lanzar 2 dados y observar la cantidad de puntos en sus caras superiores. El espacio muestral S se compone de 36 resultados posibles, es decir

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

Utilizando a S como conjunto de partida, se pueden definir 2 variables aleatorias, X e Y.

(a) Sea X el valor absoluto de la diferencia del número de puntos de los dados, es decir, $X(a,b) = |a-b|$, $\forall (a,b) \in S$.

(b) Sea Y el producto del número de puntos de los dados, es decir, $Y(a,b) = a \cdot b$, $\forall (a,b) \in S$.

Como S contiene un número finito de posibilidades, es un espacio muestral discreto, por ende, X e Y son *variables aleatorias discretas*. Si una variable puede tomar valores en una escala continua, se le denomina *variable aleatoria continua*. En el ejemplo 2, se ilustra este concepto.

EJEMPLO 2

Sea X la variable aleatoria definida por la proporción de contaminación con materia fecal que hay, por metro cúbico de agua, en el río Virilla, X toma todos los valores de x para los cuales $0 \leq x \leq 1$.

Observación. Las notaciones abreviadas $P(X=a)$ y $P(a \leq X \leq b)$ se usan, respectivamente, para la probabilidad de que exista un $s \in S$ tal que $X(s)=a$ y que exista un $s \in S$ tal que $a \leq X(s) \leq b$.

Las variables aleatorias tienen propiedades importantes, las cuales no se indican en el libro de texto.

Definición 1. Sean X e Y variables aleatorias con el mismo espacio muestral S y *k,* un número real. Entonces se cumple que:

1) $(X+Y)(s) = X(s) + Y(s)$.

2) $(X+k)(s) = X(s) + k$.

3) $(k \cdot X)(s) = k \cdot X(s)$.

4) $(X \cdot Y)(s) = X(s) \cdot Y(s)$.

EJEMPLO 3

Sean X e Y variables aleatorias definidas respectivamente así: $X(a,b)=2ab$ e $Y(a,b)=a^2+b^2$, $\forall(a,b)\in IR^2$. Calcule:

a) $(X+Y)(-3,4)$.

b) $(X+2)(-3,4)$.

Solución

Con la definición 1 se tiene que:

a) $(X+Y)(-3,4)=X(-3,4)+Y(-3,4)=\left[2\cdot(-3)\cdot 4\right]+\left[(-3)^2+4^2\right]=-24+25=1.$

b) $(X+2)(-3,4)=X(-3,4)+2=-24+2=-22.$

EJERCICIO PROPUESTO. Utilice las propiedades 3) y 4) para el punto (-3,4) siendo *k = -2*.

PUNTOS DE ATENCIÓN DE LAS SECCIONES 4.1 Y 4.2

En el teorema 4.1 (página 110) se plantean las fórmulas para calcular el valor esperado de $g(X)$, el cual depende de la variable aleatoria X; es decir, cada $g(X)$ está determinado al conocer X; en el ejemplo 4, se ilustra este teorema, en el caso de que X sea discreta.

EJEMPLO 4

El ejercicio 4.17 (página 114) dice así:

"Sea X; una variable aleatoria con la siguiente distribución de probabilidad:

x	-3	6	9
$f(x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

Encuentre $\mu_{g(X)}$, donde $g(X)=(2X+1)^2$."

SOLUCIÓN

Al utilizar el teorema 4.1, en el caso de que X; sea discreta, se obtiene el siguiente resultado:

$$\mu_{g(X)} = E(g(X)) = \sum_x g(x) \cdot f(x).$$

$$\mu_{g(X)} = E(g(X)) = (2 \cdot -3 + 1)^2 \cdot \frac{1}{6} + (2 \cdot 6 + 1)^2 \cdot \frac{1}{2} + (2 \cdot 9 + 1)^2 \cdot \frac{1}{3} = 209.$$

Por lo tanto, el valor esperado de $g(X)$ es 209.

La definición 4.2 (página 111) no es tema del curso, por lo tanto no es prioritario su estudio.

En la definición 4.3 (página 116), se plantean las fórmulas para calcular la varianza (σ^2) de una variable aleatoria, ya sea discreta o continua. Estas se tornan un poco engorrosas al aplicarlas, se recomienda utilizar el teorema 4.2 (página 117), el cual muestra una para el cálculo de la varianza de una variable aleatoria, de forma más sencilla.

En el teorema 4.3 (página 118) se plantea la fórmula para calcular la varianza ($\sigma^2_{g(X)}$) de una variable aleatoria de la forma $g(X)$, donde $f(X)$ es la distribución de probabilidad de la variable aleatoria X. En el ejemplo 5, se ilustra la aplicación de este teorema.

EJEMPLO 5

Calcule la varianza ($\sigma^2_{g(X)}$) de la variable aleatoria $g(X)$ del ejercicio 4.17 (página 114). Recuerde que en el ejemplo 4 se calculó la media ($\mu_{g(X)}$) de la variable aleatoria $g(X)$.

SOLUCIÓN

Por el teorema 4.3 (página 118), para el caso donde X es una variable aleatoria discreta, se tiene que la fórmula para calcular la varianza ($\sigma^2_{g(X)}$) es la siguiente:

$$\sigma_{g(X)}^2 = E\left\{\left[g(X) - \mu_{g(X)}\right]^2\right\} = \sum_x \left[g(x) - \mu_{g(X)}\right]^2 \cdot f(x).$$

En el ejemplo 4 se obtuvo como resultado $\mu_{g(X)} = 209,$ se tiene que:

$$\sigma_{g(X)}^2 = \sum_x \left[\left(2x+1\right)^2 - 209\right]^2 \cdot f(x) = \sum_x \left[4x^2 + 4x - 208\right]^2 \cdot f(x).$$

$$\sigma_{g(X)}^2 = \left[4\cdot(-3)^2 + 4\cdot(-3) - 208\right]^2 \cdot \frac{1}{6} + \left[4\cdot(6)^2 + 4\cdot(6) - 208\right]^2 \cdot \frac{1}{2} + \left[4\cdot(9)^2 + 4\cdot(9) - 208\right]^2 \cdot \frac{1}{3}.$$

$$\sigma_{g(X)}^2 = 33856 \cdot \frac{1}{6} + 1600 \cdot \frac{1}{2} + 23104 \cdot \frac{1}{3} = 14144.$$

Por lo tanto, la varianza de $g(X)$ es $\sigma_{g(X)}^2 = 14144$.

La definición 4.4 (página 119) y el teorema 4.4 (página 120) no son temas de este curso, por lo tanto no son prioritarios para su estudio.

PUNTOS DE ATENCIÓN DE LAS SECCIONES 5.1 A LA 5.3

En la sección 5.1, se presenta una introducción y motivación al tema de distribuciones de probabilidad para variables aleatorias discretas.

En la página 149, se desarrolla la forma general para la distribución de probabilidad de un *experimento multinomial*. En el ejemplo 6, se plantea uno similar.

EJEMPLO 6

Un equipo de futbol, con base en resultados anteriores al jugar de local, se dice que posee 50% de probabilidad de ganarle a otro equipo, 30% de empatar y 20% de perder. En los siguientes 5 partidos de local, hallar la probabilidad P de que el equipo gane 3 partidos, pierda 1 y empate 1.

Solución

Como los partidos son independientes uno del otro, el orden de los resultados no importa. Según la fórmula para la distribución multinomial, que está en la página 149, se obtiene el siguiente resultado:

$$P=\binom{n}{x_1,x_2,\ldots,x_k}p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}=\frac{5!}{3!\cdot 1!\cdot 1!}\left(0{,}5\right)^3\cdot(0{,}3)^1\cdot\left(0{,}2\right)^1=0{,}15.$$

Por lo tanto, existe 15% de probabilidad de que, en los siguientes 5 partidos, el equipo gane 3, empate 1 y pierda 1.

Puntos de atención de las secciones 6.1 a la 6.4

En la sección 6.1 se plantea el tema de *distribución de probabilidad uniforme continua*, la cual se caracteriza por tener una función de densidad con una monotonía constante. Esta no es muy aplicada en probabilidad, pero debe considerarla como una introducción.

En la página 171 se da la definición de la distribución uniforme continua y en la 172 está el teorema 6.1, que plantea las fórmulas para el cálculo de la media (μ) y la varianza (σ^2) de una distribución uniforme.

En la sección 6.2, se desarrolla el concepto de una de las distribuciones de probabilidad más significativas en el campo de la estadística inductiva, la *distribución normal*. Se le recomienda al estudiante estudiar las cinco propiedades desarrolladas en las páginas 173, 174 y 175.

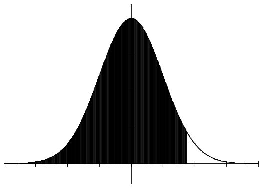
Observación. La distribución normal es continua con una gráfica muy famosa en forma de campana, a veces llamada *Campana de Gauss,* en honor a Karl Friendrich Gauss $(1777-1855)$.

En la sección 6.3, se desarrolla el tema del cálculo del área bajo la curva normal. Al depender una distribución normal de su media μ y de su varianza σ^2 existe infinita cantidad de curvas normales diferentes, por lo tanto, se trabajará con una *distribución normal estándar*, la cual tiene una media μ=0 y una varianza σ^2=1**.** Todos los valores de esta distribución se encuentran en la tabla A.3, páginas 751 y 752.

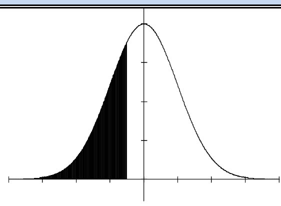
En la tabla A.3 se muestra el área bajo la curva normal estándar, donde $-3{,}49 \le z \le 3{,}49$ y vienen dados en $0{,}01$. Esta se representa por $P(Z < z_1)$, en la tabla 1 se expone un resumen para el cálculo bajo la curva.

Tabla 1

(Caso 1) ÁREA BAJO LA CURVA = $P(Z < z_1)$

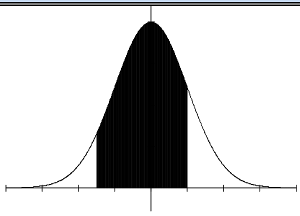


$0 \ge z_1$

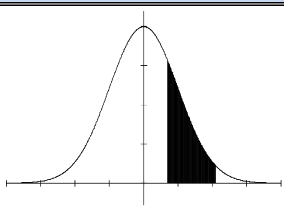


$0 \le z_1$

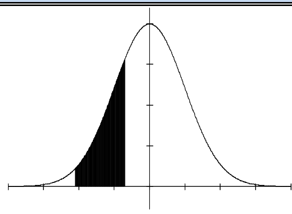
(Caso 2) ÁREA BAJO LA CURVA = $P(Z < z_2) - P(Z < z_1)$



$z_1 \le 0 \le z_2$

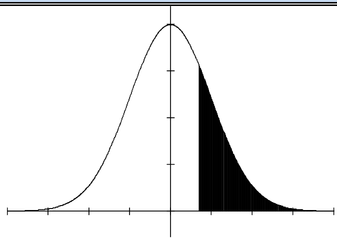


$0 \le z_1 \le z_2$

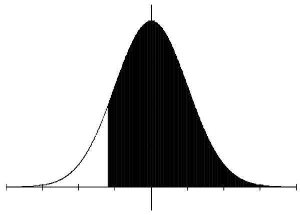


$z_1 \le z_2 \le 0$

(Caso 3) ÁREA BAJO LA CURVA = $P(Z > z_1) = 1 - P(Z < z_1)$



$0 \ge z_1$



$0 \le z_1$

En el ejemplo 7, se muestra cómo calcular el área bajo la curva de una distribución normal estándar.

EJEMPLO 7

Sea Z una variable aleatoria continua con distribución normal estándar, calcule las siguientes áreas bajo la curva:

a) $P(Z < -0{,}59)$.

b) $P(1{,}76 < Z < 2{,}30)$.

c) $P(Z > 1{,}07)$.

SOLUCIÓN

Con las tablas 1 de esta sección de la guía y la A.3, se pueden calcular las áreas propuestas.

a) Para calcular el área menor a -0,59 se debe utilizar el **caso 1** de la tabla 3, donde el área bajo la curva es $P(Z < z_1)$. Ahora, en la tabla A.3 se busca, en la primera columna, el valor **-0,5** y en la primera fila **0,09**. Siga la fila donde está **-0,5** y siguiendo la columna donde está **0,09**, la intersección entre la fila y la columna es el valor del área bajo la curva de $]-\infty, -0{,}59[$, el cual es 0,2776. Por lo tanto, $P(Z < -0{,}59) = 0{,}2776$.

b) De acuerdo con la tabla 1, este ejercicio representa el **caso 2**, por lo tanto, se encuentran las áreas bajo la curva al punto 1,76 y al punto 2,30, con ayuda de la tabla A.3, como se hizo en el punto anterior, y se obtiene $P(Z < 1{,}76) = 0{,}9608$ y $P(Z < 2{,}30) = 0{,}9893$ se tiene entonces que:

$$P(1{,}76 < Z < 2{,}30) = 0{,}9893 - 0{,}9608 = 0{,}0285.$$

c) De acuerdo con la tabla 1, este es el **caso 3**, se encuentra el área bajo la curva al punto 1,07, con ayuda de la tabla A.3 como se hizo en los puntos (a) y (b) de este ejemplo. Se obtiene $P(Z < 1{,}07) = 0{,}8577$ se tiene entonces que:

$$P(Z > 1{,}76) = 1 - P(Z < 1{,}76) = 1 - 0{,}9608 = 0{,}1423\,.$$

En el ejemplo 8, se muestra cómo calcular el área bajo la curva de una distribución normal no estandarizada.

EJEMPLO 8

En un estudio acerca del peso en niños recién nacidos en determinada población, se encontró que este se distribuía aproximadamente según una normal de media (μ) de 6,70 libras y desviación estándar (σ) de 0,25 libras. Hallar el porcentaje de niños con peso

a) menor a 6,00 libras;

b) entre 6,30 y 7,00 libras;

c) mayor que 7,20 libras.

SOLUCIÓN

Sea X la variable aleatoria del peso de niños recién nacidos, la cual tiene una distribución aproximadamente normal pero no estandarizada, ya que $\mu \neq 0$ y $\sigma^2 \neq 1$; por lo tanto, en todos los casos hay que estandarizar los valores de X con la fórmula que está en la página 177.

a) Hay que encontrar $P(X < 6{,}00)$, para ello estandarice el valor 6,00.

$$z_1 = \frac{6{,}00 - 6{,}7}{0{,}25} = -2{,}80.$$

Como $P(X < 6{,}00) = P(Z < -2{,}80)$ se busca en la tabla A.3 el valor del área bajo la curva al punto -2,8 y así se obtiene que $P(Z < -2{,}80) = P(X < 6{,}00) = 0{,}0026$. Esto quiere decir que la probabilidad de que un niño pese menos de 6,00 libras es casi nula (0,26%).

b) De manera similar, estandarice los valores 6,30 y 7,00 para calcular $P(6,30 < X < 7,00)$.

$$z_1 = \frac{6,30 - 6,7}{0,25} = -1,60, \qquad z_2 = \frac{7,00 - 6,7}{0,25} = 1,20.$$

Como $P(6,30 < X < 7,00) = P(-1,6 < Z < 1,2)$, de acuerdo con la tabla 3, este es el **caso 2**, se encuentran las áreas bajo la curva al punto 1,60 y al punto 1,20 utilizando la tabla A.3, como se hizo en el punto anterior de este ejemplo, se obtiene que $P(Z < -1,60) = 0,0548$ y $P(Z < 1,20) = 0,8849$ entonces

$$P(-1,60 < Z < 1,20) = 0,8849 - 0,0548 = 0,8301.$$

Por lo tanto, hay aproximadamente un 83% de niños recién nacidos que tiene un peso entre 6,30 y 7,00 libras.

c) Análogamente a los puntos (a) y (b) del ejemplo, se necesita estandarizar el valor 7,20 para encontrar $P(X > 7,20)$.

$$z_1 = \frac{7,20 - 6,7}{0,25} = 2,00.$$

Como $P(X > 7,20) = P(Z > 2,00)$ se encuentra mediante la tabla A.3 del área bajo la curva al valor de 2,00 y, de acuerdo con la tabla 3, este es un **caso 3**, entonces se tiene que:

$$P(X > 7,20) = P(Z > 2,00) = 1 - P(Z < 2,00) = 1 - 0,9777 = 0,0228.$$

Por lo tanto, hay aproximadamente 2,3% de niños recién nacidos que tiene un peso mayor a 7,20 libras.

En la sección 6.4, se explica cómo en muchos problemas es aplicable la distribución normal, y se exponen varios ejemplos.

PUNTOS DE ATENCIÓN DE LA SECCIÓN 6.5

En la sección 6.5 se explica cómo utilizar la distribución normal para aproximar la binomial.

La probabilidad binomial $b(x;n,p)$ resulta cada vez más difícil de calcular a medida que n aumenta, pero se puede aproximar $b(x;n,p)$ por medio de la distribución normal con el teorema 6.2 (página 188), el cual plantea una fórmula para pasar un dato binomial a uno normal estandarizado.

LABORATORIOS

Laboratorio 1 (distribución de probabilidad de una variable aleatoria discreta)

Sea el experimento de lanzar 2 dados y observar la cantidad de puntos de sus caras superiores. El espacio muestral S se compone de 36 resultados posibles, es decir

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

Si usa *S* como conjunto de partida, puede definir la variable aleatoria *X* como la suma de la cantidad de puntos de las caras superiores, es decir, $X(a,b) = a + b \ \forall (a,b) \in S.$ En la tabla 2 se muestra la *distribución de probabilidades* de *X*.

Tabla 2

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Con la distribución de probabilidad de *X*, se obtiene:

$$\mu = E(X) = \sum_{2}^{12} x \cdot f(x) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + \ldots + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7.$$

Y como $\sigma^2 = E(X^2) - \mu^2$ sea

$$\mu = E(X^2) = \sum_{2}^{12} x^2 \cdot f(x) = 4 \cdot \frac{1}{36} + 9 \cdot \frac{2}{36} + \ldots + 121 \cdot \frac{2}{36} + 144 \cdot \frac{1}{36} = 54{,}8\overline{3}.$$

Al aplicar la fórmula del teorema 4.2 se obtiene:

$$\sigma^2 = E(X^2) - \mu^2 = 54{,}8\overline{3} - (7)^2 = \frac{35}{6} = 5{,}8\overline{3}$$

$$\Rightarrow \sigma \approx 2{,}42.$$

Ahora bien, con el programa *winstats.exe* se puede realizar este experimento y comparar los resultados.

Instrucciones

1. Abra el *software winstats.exe,* haga doble clic en el ícono correspondiente. Se desplegará una pequeña pantalla verde con dos opciones en el menú: $\boxed{Window}$, $\boxed{Help}$.

2. Ubíquese en $\boxed{Window}$ y realice la secuencia $\boxed{Simulations} \rightarrow \boxed{Roll\ dice}$. Al hacerlo, aparecerá una ventana con la cara superior de dos dados, la cual le ayudará a simular el problema. En adelante, se trabaja solo con el menú principal de esta ventana.

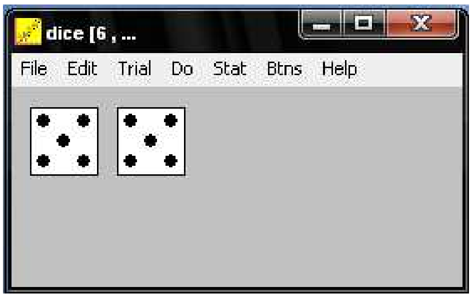


Imagen 1

3. Por defecto aparecen dos dados, pero se podría cambiar la cantidad. Para hacer esto, en el menú principal se sigue la secuencia: $\boxed{edit} \rightarrow \boxed{Parameters\ldots}$. No es necesario variar los parámetros.

4. Para asignarle a la variable aleatoria *X* la suma de los puntos de los dados, se realiza la siguiente secuencia $\boxed{Stat} \rightarrow \boxed{Sum}$.

5. Con la tecla $\boxed{F1}$ podrá tirar los dos dados. Realícelo varias veces para poder observar su comportamiento con respecto a la suma de los puntos.

6. Puede también efectuar este evento de forma simultánea. Para elegir la cantidad de veces que tirará los dados siga la secuencia $\boxed{Do} \rightarrow \boxed{Number\ldots}$ y digite en la casilla de texto 1000. Luego, al oprimir la tecla $\boxed{F12}$, se ejecutará el evento 1000 veces.

7. El objetivo de este laboratorio es comparar la distribución de probabilidad, la media y la desviación estándar de la variable aleatoria *X*, con los resultados obtenidos en la simulación. Para esto, puede visualizar la tabla de frecuencias del evento, realice la secuencia $\boxed{Do} \rightarrow \boxed{Data} \rightarrow \boxed{Freqencies\ldots}$ entonces aparecerá una ventana donde la primera columna representa los diferentes valores de *X* (de 2 a 12) y en la segunda, la frecuencia con la cual la suma de los puntos de los dados dio el valor de *X*. Adicionalmente, al final de la pantalla,

aparece la media (*sample mean*) y la desviación estándar (*sample st dev*) de los datos obtenidos en 1000 lanzamientos.

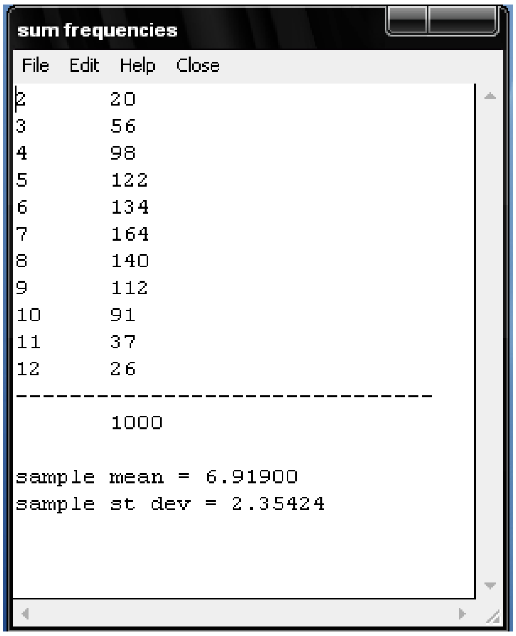


Imagen 2

8. Compare los resultados obtenidos en la simulación con los conseguidos al utilizar las definiciones de distribución de probabilidad, media y varianza de una variable aleatoria discreta.

Laboratorio 2 (distribución binomial y aproximación de la normal a la binomial)

Un francotirador del ejército tiene una probabilidad de dar en el blanco de $\frac{3}{5}$. Si en un entrenamiento el francotirador dispara hacia un objetivo 50 veces. Realice lo siguiente:

a) ¿Cuál es la probabilidad de que dé en el blanco 35 veces?

b) Hallar el número esperado (μ) de veces que dará en el blanco.

c) Hallar la desviación estándar (σ) del experimento binomial.

d) Hallar la probabilidad de que dé en el blanco al menos 35 veces.

Solución

Se conocen los siguientes datos: $n=50$, $x=35$, $p=\dfrac{3}{5}$, $q=1-\dfrac{3}{5}=\dfrac{2}{5}$. Entonces se obtiene:

a) $b\left(35;50,\dfrac{3}{5}\right)=\dbinom{50}{35}\cdot\left(\dfrac{3}{5}\right)^{35}\cdot\left(\dfrac{2}{5}\right)^{15}\approx 0,04.$

b) $\mu=np=50\cdot\dfrac{3}{5}=30.$

c) $\sigma=\sqrt{npq}=\sqrt{50\cdot\dfrac{3}{5}\cdot\dfrac{2}{5}}=2\sqrt{3}\approx 3,46.$

d) Con el teorema 6.2 y la distribución normal estandarizada.

$$Z=\frac{X-np}{\sqrt{npq}}=\frac{35-50\cdot\frac{3}{5}}{\sqrt{50\cdot\frac{3}{5}\cdot\frac{2}{5}}}\approx 1,44.$$

Ahora, al usar la tabla A.3 para calcular $P(Z<1,44)$ se tiene:

$$P(X\geq 35)\approx P(Z\geq 1,44)=1-P(Z<1,44)=1-0,9251=0,0749.$$

Ahora bien, con el programa *winstats.exe*, se puede realizar este experimento y comparar los resultados.

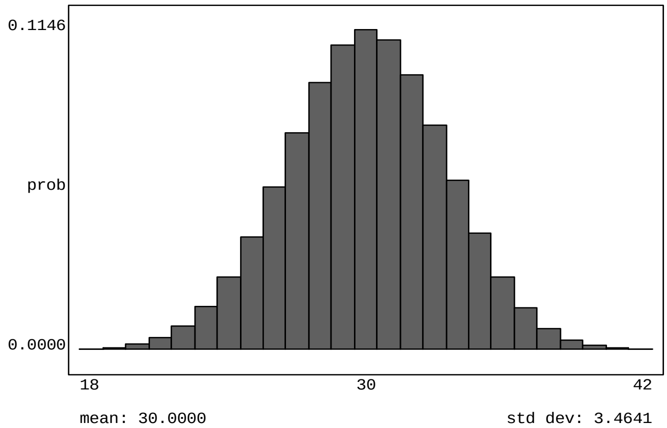
Instrucciones

1. Abra el *software winstats.exe,* haga doble clic en el ícono correspondiente. Se abrirá una pequeña pantalla verde con dos opciones en el menú: $\boxed{Window}$, $\boxed{Help}$.

2. Ubíquese en $\boxed{Window}$ y siga la secuencia $\boxed{Probability} \to \boxed{Binomial}$. Al efectuarla, se abrirá una ventana con un histograma de una distribución binomial con parámetros ya establecidos. Para cambiarlos por los del ejemplo, se sigue la secuencia $\boxed{Edit} \to \boxed{Parameters...}$ en el menú de la ventana nueva. Aparecerá una ventana con dos casillas de texto. En "*trials*" que representará *n* (cantidad de experimentos binomiales) digite 50 y en la otra casilla "*Success prob*" que representará *p* (la probabilidad de tener éxito), digite 0,6 (que es igual a $\frac{3}{5}$) y oprima el botón *OK*.

Imagen 3

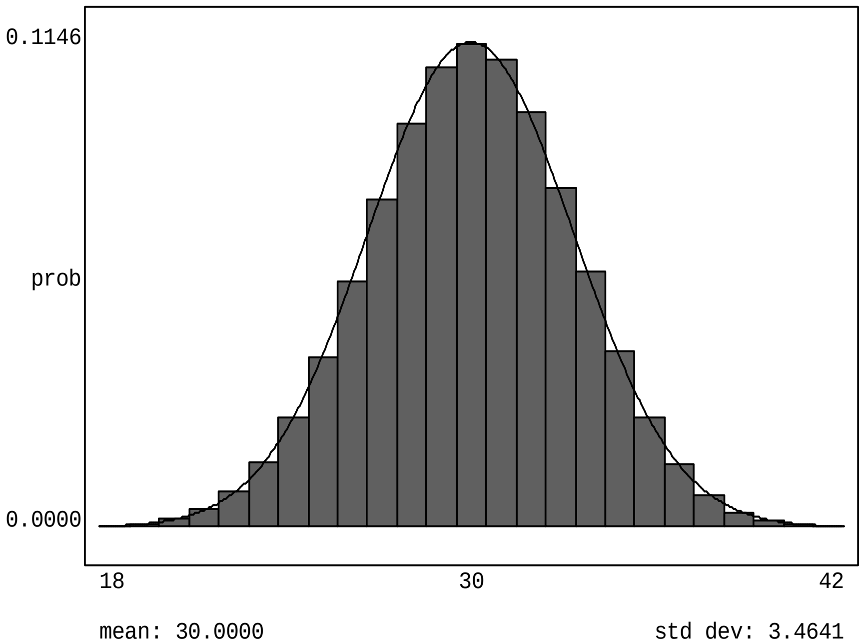
3. Al realizar el punto anterior, aparece el histograma de la distribución de probabilidad con los valores de la media (*mean*) y de la desviación estándar (*std dev*) en la parte inferior, como se muestra en la gráfica 1.



Gráfica 1

4. Con la secuencia $\boxed{Calc} \rightarrow \boxed{Interval...}$, se calcula la probabilidad de un intervalo, digite en la pantalla emergente el extremo inferior (*low x*) y el extremo superior (*high x*). Así, por ejemplo, para calcular $P(X \geq 35)$ en la casilla de *low x* se digita 0 y en la de *high x* se digita 35, y se oprime el botón *probability*.

5. Al ser una distribución binomial su comportamiento es aproximadamente normal, así que con la secuencia $\boxed{Edit} \rightarrow \boxed{Normal\ overlay}$ se puede ver en la gráfica 2 el histograma y la curva normal que aproxima la binomial.



Gráfica 2

6. Compare los resultados obtenidos en el laboratorio con los conseguidos al resolver los ejercicios por medios teóricos deductivos.

EJERCICIOS DE AUTOEVALUACIÓN

1. Sea el experimento de lanzar dos dados y observar el número de puntos que sale en cada uno. El espacio muestral S se compone de 36 resultados posibles, es decir

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

Use *S* como conjunto de partida y sea la variable aleatoria *X* la mayor cantidad de puntos que aparece en las caras superiores de los dados, es decir, $Y(a,b) = máximo(a,b)\ \forall (a,b) \in S$.

De acuerdo con la variable aleatoria *X*, realice lo siguiente:

a) represente la tabla de distribución de probabilidad;

b) averigüe la función de distribución de probabilidad acumulada;

c) calcule la media (μ) o valor esperado (E(x));

d) calcule la varianza (σ^2).

2. Averigüe el valor del parámetro k para que $f(x)$ sea una función de densidad probabilística.

$$f(x) = \begin{cases} \dfrac{x-k}{2} & \text{para} \qquad k \le x \le 2x \\ 0 & \text{en otro caso.} \end{cases}$$

3. De acuerdo con la función de densidad $f(x)$ obtenida en el ejercicio 2, averigüe:

a) la función de probabilidad acumulada;

b) la media (μ) o valor esperado *(E(x)).*

4. Un jugador tira un dado, si sale un número mayor a 4, él gana. Si *X* es la variable aleatoria que asigna a cada cara su probabilidad de salir, realice lo siguiente:

   a) encuentre la fórmula para la distribución de probabilidad de *X*;

   b) ¿cuál es la probabilidad de que el jugador gane?

   c) calcule la media (μ) y la varianza (σ^2) de *X*.

5. En VICESA (Vidriera Centroamericana, S.A) fabrican un tipo de botella. El 15% sale defectuoso. Si escoge al azar 10 botellas, averigüe

   a) la probabilidad de que una salga defectuosa;

   b) la media y la desviación estándar de este experimento binomial;

   c) la probabilidad de que, máximo, salgan tres defectuosas.

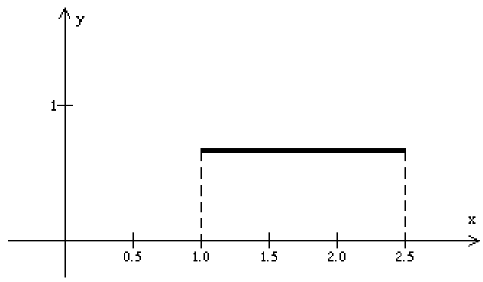
6. Si la función de densidad de la variable aleatoria uniforme continua *X* en el intervalo $[A,B]$ es:

$$f(x;A,B)=\begin{cases} \dfrac{1}{B-A} & \text{para} \qquad A \leq x \leq B \\ 0 & \text{en otro caso.} \end{cases}$$

   Demuestre que la media y la varianza de la distribución uniforme son:

$$\mu=\frac{A+B}{2}, \qquad \sigma^2=\frac{(B-A)^2}{12}.$$

7. Considere la gráfica 3, la cual representa la función de densidad para una variable aleatoria continua *X* en el intervalo $\left[1,\frac{5}{2}\right]$.



Gráfica 3

Realice lo que se le solicita a continuación:

a) ¿cuál es el criterio de la función de densidad de *X*?

b) calcule $P[X \geq 2]$;

c) calcule la media (μ) y la varianza (σ^2).

8. La temperatura en la provincia de Cartago, en el mes de diciembre, se distribuye normalmente con media $\mu = 21^{\circ}$ C y desviación estándar $\sigma = 1,6^{\circ}$ C. Hallar la probabilidad de que la temperatura durante dicho mes sea

   a) menor a 18° C;

   b) entre 20° C y 23° C;

   c) mayor a 22° C.

9. En un estudio realizado a una muestra de varones, se encontró que la media de edad en la que se casaban era de 28 años, con una desviación estándar de 2 años, estas edades (*X*) se distribuyen aproximadamente normal.

   a) Si el 5% de la muestra de menor edad se consideran muy jóvenes para casarse, ¿cuál fue la máxima edad con la que uno o varios varones de este grupo se casó?

b) ¿Cuál fue aproximadamente el varón de menos edad que se casó?

c) ¿Cuál fue aproximadamente el varón de mayor edad que se casó?

SOLUCIÓN A LOS EJERCICIOS DE AUTOEVALUACIÓN

1. De acuerdo con la variable aleatoria *X*, se tienen los siguientes resultados:

   a) En la tabla 3 se representa la tabla de distribución de probabilidad de *X*:

Tabla 3

x	1	2	3	4	5	6
$f(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

   b) Utilice la tabla 3 y aplique la definición 3.5 (página 81), se tiene que la función de probabilidad acumulada de *X* se representa mediante el siguiente criterio:

$$F(x)=\begin{cases} 0, & para \quad x<1 \\ \frac{1}{36} & para \quad 1\le x<2 \\ \frac{4}{36} & para \quad 2\le x<3 \\ \frac{9}{36} & para \quad 3\le x<4 \\ \frac{16}{36} & para \quad 4\le x<5 \\ \frac{25}{36} & para \quad 5\le x<6 \\ 1 & para \quad x\ge 6 \end{cases}$$

   c) Use la distribución de frecuencias de *X*, proceda a aplicar la definición 4.1 (página 108) en el caso de que la variable aleatoria sea discreta;

$$\mu = E(x) = \sum_{x} x \cdot f(x),$$

$$\mu = E(X) = \sum_{1}^{6} x \cdot f(x) = 1 \cdot \frac{1}{36} + 2 \cdot \frac{3}{36} + 3 \cdot \frac{5}{36} + 4 \cdot \frac{7}{36} + 5 \cdot \frac{9}{36} + 6 \cdot \frac{11}{36} = 4{,}47\bar{2}.$$

d) Utilice la distribución de frecuencias de la variable aleatoria *X*, aplique el teorema 4.2 (página 117).

$$\sigma^2 = E(X^2) - \mu^2.$$

Como ya se calculó μ, falta obtener $E(X^2)$,

$$E(X^2) = \sum_{1}^{6} x^2 \cdot f(x) = 1^2 \cdot \frac{1}{36} + 2^2 \cdot \frac{3}{36} + 3^2 \cdot \frac{5}{36} + 4^2 \cdot \frac{7}{36} + 5^2 \cdot \frac{9}{36} + 6^2 \cdot \frac{11}{36} = 21{,}97\bar{2}.$$

Entonces, al aplicar la fórmula del teorema 4.2 se obtiene:

$$\sigma^2 = E(X^2) - \mu^2 = 21{,}97\bar{2} - \left(4.47\bar{2}\right)^2 \approx 1{,}97.$$

2. Para que $f(x)$ sea una función de densidad debe cumplir la siguiente característica:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$\Rightarrow \int_{k}^{2k} \frac{x-k}{2} dx = 1$$

$$\Rightarrow \frac{1}{2}\left[\frac{x^2}{2} - kx\right]_{k}^{2k} = 1$$

$$\Rightarrow \frac{1}{2}\left[\frac{k^2}{2}\right] = 1$$

$$\Rightarrow \frac{k^2}{4} = 1 \Rightarrow k^2 = 4 \Rightarrow k = 2 \vee k = -2.$$

Por lo tanto, tome $k=2$ pues la función debe ser positiva en el intervalo, entonces $f(x)$ cumple con la segunda característica de la definición 3.6. Así se tiene que:

$$f(x) = \begin{cases} \frac{x}{2} - 1 & \text{para} \qquad 2 \le x \le 4 \\ 0 & \text{en otro caso.} \end{cases}$$

Como puede observar, se cumplen las otras dos características de la definición 3.6.

- $f(x) \ge 0$ para $2 \le x \le 4$.
- $P(a < x < b) = \int_a^b f(x)dx = \left[\frac{x^2 - 4x}{4} \right]_a^b = \frac{(b-a)^2 - 4(b-a)}{4}$ para $2 \le a < b \le 4$.

3. De acuerdo con la función de densidad $f(x)$ del ejercicio 2, se obtiene lo siguiente:

a) Al usar la definición 3.7 (página 86) se tiene que:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt \text{ para } -\infty < x < \infty.$$

$$F(x) = P(X \le x) = \int_2^x \frac{t}{2} - 1 dt = \left[\frac{t^2 - 4t}{4} \right]_2^x = \left(\frac{x^2 - 4x}{4} - \frac{2^2 - 4 \cdot 2}{4} \right) = \frac{x^2 - 4x + 4}{4} = \frac{(x-2)^2}{4}.$$

Por lo tanto, la función de distribución acumulada de la variable aleatoria X es:

$$F(x) = \begin{cases} 0 & \text{para} \quad x < 2 \\ \frac{(x-2)^2}{4} & \text{para} \quad 2 \le x < 4 \\ 1 & \text{para} \quad x \ge 4 \end{cases}$$

b) Con la definición 4.1 (página 108) del libro de texto, para el caso de que la variable aleatoria sea continua, se obtiene:

$$\mu = E(x) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

$$\mu = E(X) = \int_{2}^{4} x \cdot \left(\frac{x}{2} - 1 \right) dx = \int_{2}^{4} \frac{x^2}{2} - xdx = \left[\frac{x^3}{6} - \frac{x^2}{2} \right]_2^4$$

$$\mu = E(X) = \int_{2}^{4} x \cdot \left(\frac{x}{2} - 1 \right) dx = \left(\frac{32}{3} - 8 - \frac{4}{3} + 2 \right) = \frac{10}{3} = 3,\bar{3}.$$

4. Al tirar un dado, cada elemento del espacio muestral $S = \{1,2,3,4,5,6\}$ ocurre con una probabilidad de $\frac{1}{6}$.

   a) Por lo tanto, se tiene una distribución uniforme discreta, con

$$f(x;6) = \frac{1}{6} \quad \forall x \in S.$$

   b) Como es una distribución uniforme discreta, la probabilidad de que el jugador gane, o sea, obtenga un número mayor a 4 al tirar el dado, es:

$$P(X > 4) = \sum_{k=5}^{6} f(k,6) = \frac{1}{6} + \frac{1}{6} = 2 \cdot \frac{1}{6} = \frac{1}{3}.$$

   Por lo tanto, hay aproximadamente 33% de probabilidad de que el jugador gane.

   c) Al aplicar las fórmulas del teorema 5.1 (página 142) se obtiene:

$$\mu = \frac{1}{k} \sum_{i=1}^{k} x_i = \frac{1}{6} \cdot (1+2+3+4+5+6) = \frac{1}{6} \cdot 21 = 3,5.$$

$$\sigma^2=\frac{1}{k}\sum_{i=1}^{k}(x_i-\mu)^2=\frac{1}{6}\cdot\left[(1-3{,}5)^2+(2-3{,}5)^2+\ldots+(5-3{,}5)^2+(6-3{,}5)^2\right]=\frac{1}{6}\cdot 17{,}5\approx 2{,}92.$$

Por lo tanto, la media (μ) y la varianza (σ^2) de *X* son, respectivamente, 3,5 y 2,92.

5. Se tienen los siguientes datos: $n=10,\ \ p=0{,}15,\ \ q=1-0{,}15=0{,}85$.

   a) Use la fórmula de distribución de probabilidad binomial, donde $x=1$, se obtiene:

   $$b(1;10{,}0{,}15)=\binom{10}{1}\cdot(0{,}15)^1\cdot(0{,}85)^9\approx 0{,}35.$$

   Por lo tanto, hay aproximadamente 35% de probabilidad de que en 10 botellas escogidas al azar se encuentre una defectuosa.

   b) Con el teorema 5.2 se obtienen los siguientes resultados:

   $$\mu=np=10\cdot 0{,}05=1{,}5.$$

   $$\sigma=\sqrt{npq}=\sqrt{10\cdot 0{,}15\cdot 0{,}85}\approx 1{,}13.$$

   c) Hay que calcular $P(X\leq 3)$, con la fórmula de distribución de probabilidad binomial se tiene que

   $$P(X\leq 3)=\sum_{k=1}^{3}b(k;10{,}0{,}15)$$

   $$=\binom{10}{1}\cdot(0{,}15)^1\cdot(0{,}85)^9+\binom{10}{2}\cdot(0{,}15)^2\cdot(0{,}85)^8+\binom{10}{3}\cdot(0{,}15)^3\cdot(0{,}85)^7\approx 0{,}75.$$

6. Como sabe $f(x)=\dfrac{1}{B-A}$ para $A\leq x\leq B$ se tiene que:

$$\mu=\int_A^B \frac{x}{B-A}dx=\frac{1}{B-A}\left[\frac{x^2}{2}\right]_A^B=\frac{B^2-A^2}{2(B-A)}=\frac{(A+B)(B-A)}{2(B-A)}=\frac{A+B}{2}.$$

Con el teorema 4.2 (página 117), donde $\sigma^2=E(X^2)-\mu^2$.

Calcule $E(X^2)$

$$E(X^2)=\int_A^B \frac{x^2}{B-A}dx=\frac{1}{B-A}\left[\frac{x^3}{3}\right]_A^B=\frac{B^3-A^3}{3(B-A)}=\frac{(B-A)\left(B^2+AB+A^2\right)}{3(B-A)}=\frac{B^2+AB+A^2}{3}$$

$$\sigma^2=E(X^2)-\mu^2=\frac{B^2+AB+A^2}{3}-\left(\frac{A+B}{2}\right)^2=\frac{4\left(B^2+AB+A^2\right)-3(B^2+2AB+A^2)}{12}$$

$$\Rightarrow \sigma^2=\frac{B^2-2AB+A^2}{12}=\frac{(B-A)^2}{12}.$$

7. Según la definición de distribución uniforme de la página 171, donde $A=1$ y $B=\frac{5}{2}$, se obtiene la función de densidad para *X*:

$$f(x)=\begin{cases}\frac{2}{3}, & 1\le x\le \frac{5}{2},\\ 0, & \text{en otro caso.}\end{cases}$$

a) $P[X\ge 2]=\int_2^{\frac{5}{2}}\frac{2}{3}dx=\left[\frac{2}{3}x\right]_2^{\frac{5}{2}}=\frac{1}{3}.$

b) Al aplicar el teorema 6.1, de la página 172, se obtiene:

$$\mu=\frac{A+B}{2}=\frac{1+\frac{5}{2}}{2}=\frac{7}{4}=1{,}75.$$

$$\sigma^2 = \frac{(B-A)^2}{12} = \frac{\left(\frac{5}{2}-1\right)^2}{12} = \frac{3}{16} = 0{,}1875.$$

Por lo tanto, la media (μ) y la varianza (σ^2) de la variable aleatoria *X* son, respectivamente, 1,75 y 0,1875.

8. Sea *X* la variable aleatoria de la temperatura en la provincia de Cartago en el mes de diciembre, esta variable tiene una distribución aproximadamente normal, pero no estandarizada, pues $\mu \neq 0$ y $\sigma^2 \neq 1$, por lo tanto se utiliza la fórmula que está en la página 177.

   a) Hay que encontrar $P(X<18)$, para ello estandarice el valor 18.

   $$z_1 = \frac{18-21}{1{,}6} = -1{,}875 \approx -1{,}88.$$

   Como $P(X<18) = P(Z<-1{,}88)$, se busca en la tabla A.3 el valor del área bajo la curva al punto -1,88; así, obtiene $P(Z<-1{,}88) = P(X<18) = 0{,}0301$. Esto quiere decir que hay aproximadamente 3% de probabilidad de que la temperatura sea menor a 18° C.

   b) Igualmente, estandarice los valores 20 y 23 para calcular $P(20<X<23)$.

   $$z_1 = \frac{20-21}{1{,}6} = -0{,}625 \approx -0{,}63, \qquad z_2 = \frac{23-21}{1{,}6} = 1{,}25.$$

   Como $P(20<X<23) = P(-0{,}63<Z<1{,}25)$ entonces, de acuerdo con la tabla A.3 del libro de texto, $P(Z<-0{,}63) = 0{,}2643$ y $P(Z<1{,}25) = 0{,}8944$. Luego, de acuerdo con el **caso 2** de la tabla 3, se tiene que:

   $$P(-0{,}63<Z<1{,}25) = 0{,}8944 - 0{,}2643 = 0{,}6301.$$

Por lo tanto, hay aproximadamente un 63% de probabilidad de que la temperatura en Cartago, en el mes de diciembre, esté entre 20 y 23 grados centígrados.

c) Análogamente a los puntos (a) y (b) del ejemplo, estandarice el valor 22 para encontrar $P(X>22)$.

$$z_1 = \frac{22-21}{1,6} = 0,625 \approx 0,63.$$

Como $P(X>22)=P(Z>0,63)$ entonces, mediante la tabla A.3 del libro de texto, el área bajo la curva al valor de 0,63, y utilizando el **caso 3** de la tabla 1 de la guía, se tiene que:

$$P(X>22)=P(Z>0,63)=1-P(Z<0,63)=1-0,7357=0,2643.$$

Por lo tanto, hay aproximadamente 26% de probabilidad de que la temperatura en Cartago en el mes de diciembre sea mayor a 22 grados centígrados.

9. Para ello, invierta el proceso y comience con el área o probabilidad conocida, encuentre el valor *z* y después determine la *x* que busca.

a) Con ayuda de la tabla A.3, a la inversa, se obtiene $z=-1,64$ ya que en la fila de -1.6 y en la columna .04 está el área 0.0505, equivalente a un 5%. Ahora, al usar la fórmula de estandarización reacomodada se tiene:

$$x=\sigma z+\mu = 2\cdot -1,64+28=24,72.$$

Por lo tanto, la edad máxima de esa parte de la muestra fue aproximadamente de 24 años cumplidos.

b) El menor valor de la tabla A.3 es el de -3.49 con un área de 0.0002, al utilizarlo para saber cuál fue la menor edad de los varones que se casaron, se tiene que:

$$x=\sigma z+\mu = 2\cdot -3,49+28=21,02.$$

Por lo tanto, la edad mínima fue aproximadamente de 21 años cumplidos.

c) Análogamente, el mayor valor de la tabla A.3 es el de 3,49 con un área de 0,9998, al utilizarlo para saber cuál fue la mayor edad en que se casaron de los varones, se tiene:

$$x = \sigma z + \mu = 2 \cdot 3{,}49 + 28 = 34{,}98.$$

Por lo tanto, la edad máxima fue aproximadamente de 34 años cumplidos.

Glosario

distribución binomial. Es la función que establece la relación de cada número posible de éxitos con su probabilidad.

distribución de probabilidad. Subconjunto de un conjunto de datos que se distinguen de otros por algún rasgo peculiar.

distribución uniforme continua. Es la distribución continua más simple que se caracteriza por una función de densidad "plana".

distribución uniforme discreta. Es aquella distribución de probabilidad discreta más simple, donde la variable aleatoria toma cada uno de sus valores con una probabilidad idéntica.

experimento binomial. Es un experimento que a menudo consiste en pruebas repetidas, donde solo hay dos resultados posibles, los cuales se pueden marcar como **éxito** o **fracaso**.

experimento multinomial. El experimento binomial se convierte en multinomial si cada prueba tiene más de dos resultados posibles.

espacio muestral discreto. Es un espacio muestral que contiene un número finito de posibilidades o una serie interminable con tantos elementos como números enteros existen.

espacio muestral continuo. Es un espacio muestral que contiene un número infinito de posibilidades igual al número de puntos en un segmento de línea.

función de densidad. Es una función de los valores numéricos de la variable aleatoria continua, donde el área bajo la curva es igual a 1 en el intervalo definido por la variable aleatoria.

función de distribución acumulada. Es la función F de X tal que $F(a) = P(X \leq a)$.

variable aleatoria. Es una función que asocia un número real con cada elemento del espacio muestral.

variable aleatoria discreta. Es una variable aleatoria en la cual se puede contar su conjunto de resultados posibles.

variable aleatoria continua. Es una variable aleatoria que toma valores en una escala continua.

variable aleatoria de Bernoulli. Es una variable aleatoria en la que se eligen 0 y 1 para describir dos posibles valores.

variable aleatoria normal. Es una variable aleatoria continua que tiene la distribución de probabilidad con forma de campana.

variable aleatoria normal estándar. Es una variable aleatoria normal con media 0 y varianza 1.

SECCIÓN 4
ELEMENTOS DE INFERENCIA ESTADÍSTICA

En la sección se desarrollan nociones básicas de inferencia estadística, tales como la estimación de parámetros poblacionales, utilizando tanto estimadores puntuales como intervalos de confianza, la prueba de hipótesis estadística de una muestra cuando se conoce o no la varianza, con valores P para la toma de decisiones.

Asimismo, se le presenta un laboratorio utilizando el *software* gratuito *winstats.exe*, con el fin de buscar una mayor comprensión del concepto de *intervalo de confianza* de un parámetro poblacional.

OBJETIVO GENERAL

Estudiar algunos conceptos básicos de la inferencia estadística tales como el contrastar hipótesis y estimar parámetros poblacionales.

OBJETIVOS ESPECÍFICOS

1. Comprender el significado de inferencia estadística.
2. Estimar una media a través de un intervalo de confianza, cuando se conoce o no σ.
3. Encontrar el error máximo de una estimación de μ.
4. Hallar intervalos de predicción y de tolerancia para μ.
5. Diferenciar entre intervalos de confianza, de predicción y de tolerancia.
6. Calcular el error tipo I y II de una prueba de hipótesis estadística.

7. Elegir correctamente la hipótesis nula y alternativa para una prueba de hipótesis estadística.

8. Utilizar correctamente valores P para la toma de decisiones en una prueba de hipótesis.

9. Realizar correctamente el procedimiento formal para la prueba de hipótesis de medias, cuando se tiene una sola muestra y varianza conocida.

10. Realizar correctamente el procedimiento formal para la prueba de hipótesis de medias, cuando se tiene una sola muestra y varianza desconocida.

PUNTOS DE ATENCIÓN DE LAS SECCIONES 9.1 A LA 9.3

De las secciones de la 9.1 a la 9.3 se realiza una introducción a la inferencia estadística, se explican nociones básicas como estimar parámetros de una población mediante una muestra aleatoria.

Se le recomienda realizar una lectura comprensiva de estas secciones e intentar responderse, al final, las siguientes preguntas: ¿qué es la *inferencia estadística*?, ¿para qué se utiliza la *inferencia estadística*?, ¿qué es un *estimador puntual*?, ¿qué significa que el estimador sea *insesgado*?, ¿qué es y cómo se interpreta un *estimador por intervalo*?, ¿qué es un *intervalo de confianza*?, ¿a qué se le llama *coeficiente de confianza* o *grado de confianza*? El saber responder correctamente las anteriores interrogantes le ayudará a comprender, de una mejor manera, los conceptos posteriores.

Observación: cualquier procedimiento de muestreo que produzca inferencias las cuales sobreestimen, o subestimen, de forma consistente alguna característica de la población se dice que está *sesgado*. Para eliminar cualquier posibilidad de sesgo, es deseable elegir una *muestra aleatoria*, en el sentido de que las observaciones se realicen de forma independiente y al azar.

PUNTOS DE ATENCIÓN DE LAS SECCIONES 9.4 A LA 9.7

En estas secciones se desarrollan temas relacionados con la estimación de una media poblacional μ, mediante el estimador puntual $\bar{x}$ (media muestral), intervalo de confianza de μ cuando se conoce σ y cuando se desconoce, estimación del error máximo, tamaño de la muestra para un determinado grado de confianza; límites de confianza unilaterales, intervalos de predicción, límites de tolerancia, entre otras nociones.

Es importante distinguir claramente entre intervalos de confianza, de predicción y de tolerancia (comúnmente el estudiante los confunde). Por esto, se le recomienda realizar una lectura comprensiva de la página 285 del libro de texto. En el ejemplo 1 se intentará diferenciar estos conceptos

EJEMPLO 1

Use el enunciado del ejercicio 9.7 (página 286) para comparar los diferentes intervalos de estudio.

"Una muestra aleatoria de 100 propietarios de automóviles muestra que, en el estado de Virginia, un automóvil se maneja, en promedio, 23 500 kilómetros por año con una desviación estándar de 3900 kilómetros. Suponga que la distribución de las mediciones es aproximadamente normal."

a) Encuentre un *intervalo de confianza* de 99% para el promedio de kilómetros que se maneja un automóvil, anualmente, en Virginia.

b) Encuentre un *intervalo de predicción* de 99% para la cantidad de kilómetros de un automóvil por año.

c) Encuentre un *intervalo de tolerancia* de 99% que contenga 99% de los kilómetros que recorren los automóviles.

Solución

Se tiene que $n=100$, $\bar{x}=23500$ y $\sigma=3900$, entonces

a) Para encontrar un intervalo de confianza de 99% (o sea α = 0,01), se halla el valor z, que deja un área de 0,005 a la derecha y de 0,995 a la izquierda. Por lo tanto, al usar la tabla A.3, se tiene que $z_{0,005}=2,575$. Con la definición de la página 275, se tiene que el intervalo de confianza de 99% es:

$$23500-(2,575)\frac{3900}{10}<\mu<23500+(2,575)\frac{3900}{10}$$

$$22496<\mu<24504.$$

b) La predicción puntual para la cantidad de kilómetros de un automóvil por año es $x_i y_i$. El valor z es $z_{0,005}=2,575$. Por lo tanto, con la definición de la página 282, cuando σ es conocida, se tiene que el intervalo de predicción de 99% es:

$$23500-(2,575)(3900)\sqrt{1+\frac{1}{100}}<x_0<23500+(2,575)(3900)\sqrt{1+\frac{1}{100}}$$

$$13407<x_0<33593.$$

c) Utilizando la definición de *límites de tolerancia* de la página 284, se tiene que $s=3900$ y de la tabla A.7 para $n=100$, $1-\alpha=0,99$ y $\gamma=0,01$, además, $k=3,096$ para los límites de los dos lados. Por lo tanto, los límites de tolerancia de 99% son:

$$23500\pm(3,096)(3900).$$

Por ende, el intervalo de tolerancia $11425<\mu<35574$.

PUNTOS DE ATENCIÓN DE LAS SECCIONES 10.1 A LA 10.4

En la sección 10.1, se exponen varias nociones básicas del tema de prueba de hipótesis estadística. Se le recomienda realizar una lectura minuciosa, ya que se desarrollan varios argumentos interesantes para la comprensión integral de los conceptos.

En la sección 10.2 es importante comprender los conceptos estadístico de prueba y región crítica, para luego relacionarlos con la tabla 10.1 (página 325). Asimismo, la probabilidad de cometer el error tipo I y el error tipo II se puede reducir al aumentar el tamaño de muestra. Además, es importante que el estudiante domine las propiedades de una prueba de hipótesis.

En la sección 10.3 se estudia cómo elegir la hipótesis nula y alternativa para ponerlo en práctica en la 10.4, en la que se utilizan los valores P para la toma de decisiones. En el ejemplo 2 se ilustran, de forma conjunta, estos conocimientos.

EJEMPLO 2

Considere el enunciado del ejercicio 10.15 de la página 337.

"En un restaurante de carnes asadas una máquina de bebidas gaseosas se ajusta de manera que la cantidad de bebida que sirva esté distribuida de forma aproximadamente normal, con una media de 200 mililitros y una desviación estándar de 15 mililitros. La máquina se verifica periódicamente tomando una muestra de 9 bebidas y calculando el contenido promedio. Si $\bar{x}$ cae en el intervalo $191 < \bar{x} < 209$, se considera que la máquina opera de forma satisfactoria; de otro modo, concluimos que $\mu \neq 200$ mililitros."

a) Establezca la hipótesis nula (H_0) y la alternativa (H_1) del problema.

b) Encuentre la probabilidad de cometer el error tipo I cuando $\mu = 200$ mililitros.

c) Encuentre la probabilidad de cometer el error tipo II cuando $\mu = 215$ mililitros y la potencia de la prueba.

SOLUCIÓN

Se tiene que $n = 9$, $\sigma = 15$ y $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{9}} = 5$, entonces:

a) La hipótesis nula es $H_0 : \mu = 200$ y la alternativa es $H_1 : \mu \neq 200$. Por lo tanto, es una prueba de dos colas.

b) Como $\mu = 200$, se tiene que $z_1 = \frac{191 - 200}{5} = -1{,}8$ y $z_2 = \frac{209 - 200}{5} = 1{,}8$ por lo tanto, $\alpha = 2P(Z < -1{,}8) = 2 \cdot (0{,}0359) = 0{,}0718$. Existe aproximadamente un 7,2% de probabilidad de cometer un error tipo I.

c) Como $\mu = 215$ se tiene que $z_1 = \frac{191 - 215}{5} = -4{,}8$ y $z_2 = \frac{209 - 215}{5} = -1{,}2$, por lo tanto, $\beta = P(-4{,}8 < Z < -1{,}2) = 0{,}1151 - 0 = 0{,}1151$.

   Existe aproximadamente 11,5% de probabilidad de cometer un error tipo II. La potencia de la prueba es $1 - \beta = 0{,}8849$; lo cual indica que hay aproximadamente un 88% de probabilidad de rechazar H_0 dado que H_1 es verdadera.

PUNTOS DE ATENCIÓN DE LAS SECCIONES 10.5 A LA 10.7

En la sección 10.5 se expone el procedimiento formal para probar hipótesis cuando existe una sola media poblacional y se conoce la varianza. A manera de resumen, se presenta la tabla 1.

Tabla 1

H_0	Valor del estadístico de prueba	H_1	Región crítica
$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ (σ conocida)	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\frac{\alpha}{2}} \ \vee \ z > z_{\frac{\alpha}{2}}$

En la sección 10.6 se explica la analogía que existe entre el enfoque de la prueba de hipótesis con el del intervalo de confianza.

En la sección 10.7 se expone el procedimiento formal para probar hipótesis cuando existe una sola media poblacional y no se conoce la varianza. En este caso, se utiliza la variable aleatoria $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$, con S (desviación estándar muestral), pues se desconoce σ (desviación estándar poblacional). La variable aleatoria T tiene una distribución t con $v = n - 1$ grados de libertad. El área bajo la curva se encuentra en la tabla A.4.

A manera de resumen, se presenta la tabla 2.

Tabla 2

H_0	Valor del estadístico de prueba	$\sigma = 20$	Región crítica
$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}; \ v = n - 1$ (σ desconocida)	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\frac{\alpha}{2}} \ \vee \ t > t_{\frac{\alpha}{2}}$

En el ejemplo 3 se ilustra este procedimiento.

Ejemplo 3

Considere el ejercicio 10.25 de la página 357.

"Pruebe la hipótesis de que el contenido promedio de los envases de un lubricante específico es de 10 litros, si los contenidos de una muestra aleatoria de 10 envases son 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3 y 9.8 litros. Utilice un nivel de significancia de 0.01 y suponga que la distribución del contenido es normal."

SOLUCIÓN

Se tienen las siguientes hipótesis:

$$H_0 : \mu = 10,$$

$$H_1 : \mu \neq 10.$$

Además, se tiene la siguiente información:

a) $n = 10 \Rightarrow v = n - 1 = 9$ (9 grados de libertad);

b) se utiliza la tabla A.4 (página 753) con $\alpha = 0{,}01$ (nivel de significancia) y 9 grados de libertad, y se obtiene que la región crítica es $t_{\frac{\alpha}{2}} < -3{,}25 \vee t_{\frac{\alpha}{2}} > 3{,}25$;

c) con los datos de la muestra se obtiene que $\overline{x} = 10{,}06$ y $S = 0{,}246$.

Se calcula $t = \dfrac{10{,}06 - 10}{0{,}246 / \sqrt{10}} = 0{,}77$.

Como t no está en la región crítica, se decide no rechazar la hipótesis nula $H_0 : \mu = 10$.

LABORATORIO

EJEMPLO 4

Sea una variable aleatoria poblacional X tiene una media desconocida μ y desviación estándar $\sigma = 20$. Una muestra aleatoria de tamaño 100 da una media muestral $\overline{x} = 250$. Un intervalo de confianza del 90% para μ es $[246{,}7;253{,}3]$.

EJERCICIO ADICIONAL. Corrobore el intervalo de confianza del ejemplo 4.

El objetivo de este laboratorio es estudiar el comportamiento de los intervalos de confianza del 90% para μ. Esto significa que hay una probabilidad de 0,90 de que una muestra aleatoria de ellos, de igual tamaño, resulte estar en un valor $\overline{x}$ de $\overline{X}$, para que contenga a μ.

Instrucciones

9. Abra el *software* *winstats.exe,* haga doble clic en el ícono correspondiente. Se abrirá una pequeña pantalla verde con dos opciones en el menú: $\boxed{Window}$, $\boxed{Help}$.

10. Ubíquese en $\boxed{Window}$ y siga la secuencia $\boxed{Demos} \rightarrow \boxed{Confidence\ intervals}$. Al hacerlo, se mostrarán dos ventanas: una, el cuadro de diálogo donde se introduce la información y otra donde se visualiza la demostración. La imagen 1 representa el cuadro de diálogo.

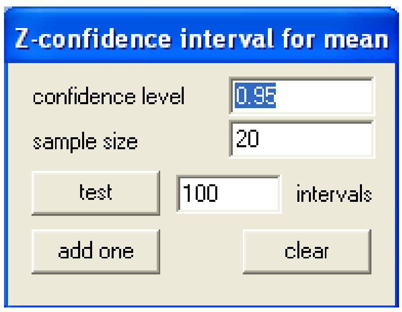


Imagen 1

11. Los tres cuadros de edición son para el nivel de confianza (*confidence level*), el tamaño de los intervalos (*simple size*) y la cantidad de la muestra.

12. Digite en el primer cuadro 0,90, en el segundo 6,6 (ya que ese es el tamaño del intervalo de confianza encontrado $253{,}3-246{,}7=6{,}6$) y en el último digite 20, como se muestra en la imagen 2.

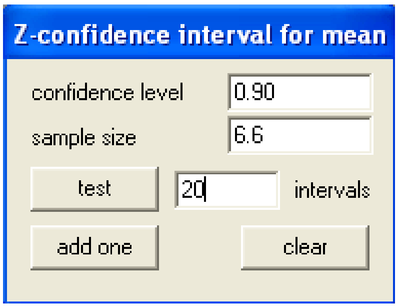


Imagen 2

13. Luego, haga clic en "*test*" (prueba) para ver los resultados. Cada intervalo representa el resultado de un sondeo de la misma población normal. La línea vertical es la media de la población, y los intervalos de confianza se muestran como segmentos horizontales gruesos. Aquellos que contienen la media son "éxito", los otros (mostrado en color de contraste) son "fallas", en que no puedan colocar la media dentro de él. Presione "*add one*" (agregar uno) para anexar un intervalo a la pantalla. Con "*test*", la visualización se actualiza automáticamente. También puede hacer "*clear*". En la imagen 3, se muestra lo anterior con 100 intervalos.

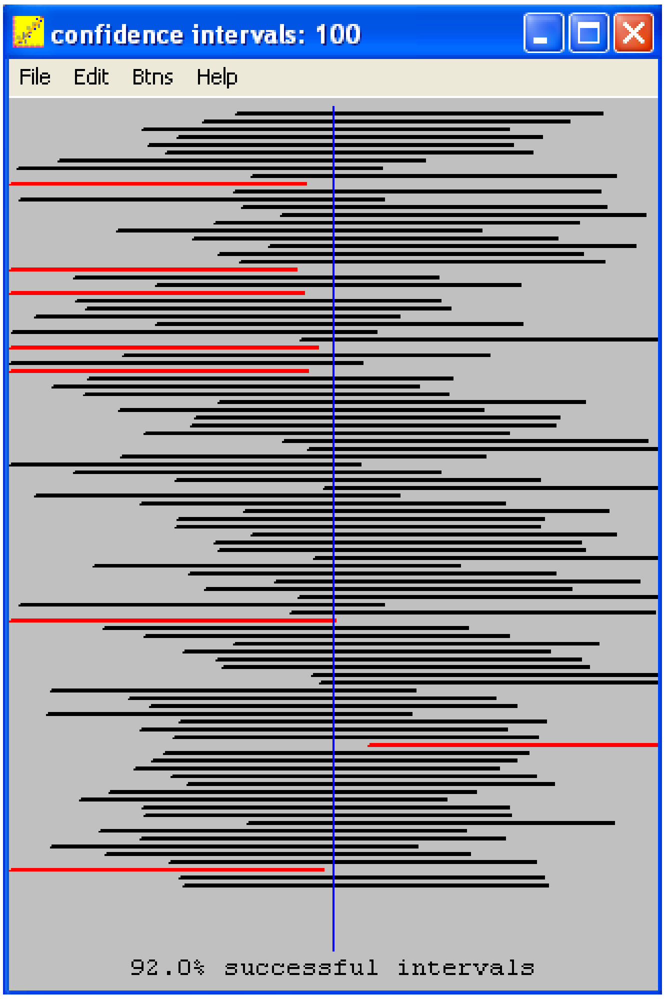


Imagen 3

14. Analice el comportamiento de este experimento, con el porcentaje de intervalos exitosos (*successful intervals*) que se muestra al final de la pantalla y compárelos con el grado de confianza establecido (90%).

EJERCICIOS DE AUTOEVALUACIÓN

1. Sea X una variable aleatoria normal de media 4,5 y varianza 0,3. Halle las estimaciones puntuales insesgadas de la media μ y de la varianza σ^2 de X, basadas en los valores de la muestra aleatoria $x_1 = 5{,}4$, $x_2 = 4{,}2$, $x_3 = 4{,}6$ y $x_4 = 5{,}1$.

2. Sea X una variable aleatoria normal de media μ, cuyo valor se desconoce, y desviación estándar σ igual a 1,5. De una muestra aleatoria de 36 valores de X se obtiene una media muestral $\overline{x} = 8$. Determinar el margen de error E para un intervalo de confianza del 90% para μ y hallar el correspondiente intervalo de confianza. Dar una interpretación del resultado.

3. De acuerdo con los datos del ejercicio 2, ¿cuál debe ser el tamaño de la muestra para obtener 99% de confianza de que la estimación de μ difiera por menos de 0,5?

4. De acuerdo con los datos del ejercicio 2,

   a) encuentre un *intervalo de predicción* de 99%,

   b) encuentre un *intervalo de tolerancia* de 99% que contenga 99% de los datos.

5. Si X es una variable aleatoria binomial, demuestre que

   a) $P = \dfrac{X}{n}$ es un estimador insesgado de p;

   b) $P' = \dfrac{X + \frac{\sqrt{n}}{2}}{n + \sqrt{n}}$ es un estimador sesgado de p;

   c) El estimador P' se vuelve insesgado conforme $n \to \infty$.

(Tomado de los ejercicios 9.2 y 9.3 del libro de texto).

6. En una población de tortugas adultas, se toma la siguiente muestra aleatoria de 8 tortugas para medir su largo (X) en centímetros.

14,6	21,7	15,2	19,1	20,5	12,8	18,9	24,4

Sea X una variable normalmente distribuida. Hallar el intervalo de confianza del 98% para la media μ de X.

7. Un fabricante de aceite de motor sintético, afirma que este tiene una duración media de 15 000 kilómetros con una desviación estándar de 350 kilómetros. Para probar la hipótesis de que $\mu = 12500$ contra la alternativa de que $\mu < 12500$ kilómetros, se prueba el aceite en 50 automóviles con las mismas características. La región crítica se define como $\bar{x} < 12350$.

    a) Encuentre la probabilidad de cometer un error tipo I cuando H_0 es verdadera.

    b) Evalúe β para las alternativas $\mu = 12200$ y $\mu = 12350$ kilómetros.

8. Se supone que los parachoques de una nueva línea de automóviles disminuirán los daños en colisiones a velocidades mayores a 5 millas por hora. En un contraste de 5 coches, la velocidad media para dicha disminución fue de 4,8 millas por hora con una desviación estándar muestral de 0,3 millas por hora.

    a) ¿Son estos resultados estadísticamente significativos al nivel 0,05?

    b) ¿A qué niveles serán estadísticamente significativos los resultados del contraste?

SOLUCIÓN A LOS EJERCICIOS DE AUTOEVALUACIÓN

1. Se utilizan los estimadores puntuales $\bar{x}$ y s^2.

- La media muestral es: $\bar{x} = \frac{1}{4}\sum_{i=1}^{4} x_i = \frac{5,4+4,2+4,6+5,1}{4} = 4,825 \approx 4,8.$

- La varianza muestral es:

$$s^2 = \frac{1}{4-1}\sum_{i=1}^{4}\left(x_i - \bar{x}\right)^2 = \frac{(5,4-4,8)^2 + (4,2-4,8)^2 + (4,6-4,8)^2 + (5,1-4,8)^2}{3} \approx 0,28.$$

2. Para determinar E, utilice el teorema 9.1 de la página 278. Se tiene que $\frac{\sigma}{\sqrt{n}}=\frac{1,5}{\sqrt{36}}=\frac{1}{4}=0,25$, y $(1-\alpha)100\%=90\% \Rightarrow \alpha=0,10$. Hay que obtener $z_{\frac{\alpha}{2}}=z_{0,05}$ (o sea, el valor z que deja un área de 0,05 a la derecha y 0,95 a la izquierda) entonces, según la tabla A.3 del libro, se tiene que $z_{0,05}=1,65$. Por lo tanto, el error máximo que se obtiene con una confianza del 90% es de:

$$E=z_{0,05}\cdot\frac{\sigma}{\sqrt{n}}=1,65\cdot 0,25=0,4125.$$

Entonces el intervalo de confianza es:

$$\left[\overline{x}-E,\overline{x}+E\right]=[8-0,4125,8+0,4125]=[7,5875;8,4125].$$

Esto quiere decir que hay 90% de probabilidad de que la media μ se encuentre en el intervalo $[7,5875;8,4125]$.

3. Por el teorema 9.2 de la página 277, se tiene que $(1-\alpha)100\%=99\% \Rightarrow \alpha=0,01$. Hay que obtener $z_{\frac{\alpha}{2}}=z_{0,005}$ (o sea, el valor z que deja un área de 0,005 a la derecha y 0,995 a la izquierda), según la tabla A.3, se tiene que $z_{0,005}=2,57$. Luego,

$$n=\left(\frac{2,575\cdot 1,5}{0,5}\right)^2=59,675625.$$

Por lo tanto, la muestra debe ser de tamaño 60.

4. Se tiene que $n=36$, $\overline{x}=8$ y $\sigma=1,5$, entonces:

   a) La predicción puntual es $\overline{x}=23500$, de acuerdo con la tabla A.3 se tiene que el valor z es $z_{0,005}=2,575$. Por lo tanto, utilizando la definición de la página 282 cuando σ es conocida, se tiene que el intervalo de predicción de 99% es:

$$8-(2{,}575)(1{,}5)\sqrt{1+\frac{1}{36}}<x_0<8+(2{,}575)(1{,}5)\sqrt{1+\frac{1}{36}}$$

$$4{,}08<x_0<11{,}92.$$

b) Por la definición de *límites de tolerancia* de la página 284, se tiene que $s=1{,}5$ y de la tabla A.7 para $n=36$, $1-\alpha=0{,}99$ y $\gamma=0{,}01$, entonces $k\approx 3{,}5924$ para los límites de los dos lados. Por lo tanto, los límites de tolerancia de 99% son:

$$8\pm(3{,}5924)(1{,}5).$$

Por lo tanto, el intervalo de tolerancia es $2{,}6114<\mu<13{,}39$.

5. Se tiene que $E(X)=n\cdot p$, entonces

a) $E\left(\hat{P}\right)=E\left(\frac{X}{n}\right)=\frac{E(X)}{n}=\frac{n\cdot p}{n}=p;$

b) $E(P')=\frac{E(X)+\frac{\sqrt{n}}{2}}{n+\sqrt{n}}=\frac{n\cdot p+\frac{\sqrt{n}}{2}}{n+\sqrt{n}}\neq p;$

c) $\lim\limits_{n\to\infty}\frac{np+\frac{\sqrt{n}}{2}}{n+\sqrt{n}}=\lim\limits_{n\to\infty}\frac{n\left(p+\frac{1}{2\sqrt{n}}\right)}{n\left(1+\frac{1}{\sqrt{n}}\right)}=\lim\limits_{n\to\infty}\frac{p+\frac{1}{2\sqrt{n}}}{1+\frac{1}{\sqrt{n}}}=p.$

6. Se utilizan los estimadores puntuales $\overline{x}$ y s^2.

- La media muestral es: $\overline{x}=\frac{1}{8}\sum_{i=1}^{8}x_i=\frac{14{,}6+21{,}7+...+24{,}4}{8}=18{,}4.$

- La varianza muestral es:

$$s^2 = \frac{1}{8-1}\sum_{i=1}^{8}\left(x_i - \bar{x}\right)^2 = \frac{\left(14,6-18,4\right)^2 + \left(21,7-18,4\right)^2 + \ldots + \left(24,4-18,4\right)^2}{7} = 15,44$$

$$\Rightarrow s = 3,93.$$

Por la tabla A.4, $t_{0,01} = 2,998$ para $v = 8-1 = 7$ grados de libertad. De aquí, el intervalo de confianza de 98% para µ es $18,4 - \left(2,998\right)\left(\frac{3,93}{\sqrt{8}}\right) < \mu < 18,4 + \left(2,998\right)\left(\frac{3,93}{\sqrt{8}}\right)$.que se reduce a $14,23 < \mu < 22,57$.

7. Se conoce $n = 50$, $H_0: \mu = 12500$, $H_1: \mu < 12500$, $\sigma = 350 \Rightarrow \sigma_{\bar{x}} = \frac{350}{\sqrt{50}} = 49,5$.

   a) Como $z = \frac{12350-12500}{49,5} = -3,03$ entonces $\alpha = P\left(Z < -3,03\right) = 0,0012$. Por lo tanto, existe 0,12% de probabilidad de cometer un error tipo I cuando H_0 es verdadera.

   b) Si $\mu = 12200$ entonces $z = \frac{12350-12200}{49,5} = 3,03$. Se tiene entonces que $\beta = P\left(Z > 3,03\right) = 0,9988$. Por lo tanto, existe 99,88% de probabilidad de cometer un error tipo II cuando H_0 es falsa.

   Ahora, si $\mu = 12350$, entonces $z = \frac{12350-12350}{49,5} = 0$. Se tiene que $\beta = P\left(Z > 0\right) = 0,5$. Por lo tanto, existe 50% de probabilidad de cometer un error tipo II cuando H_0 es falsa.

8. Asuma que la velocidad máxima para disminuir los daños está normalmente distribuida.

a) El estadístico de contraste es el valor t de la media muestral. Los resultados son estadísticamente significativos si, t, está en la región crítica de la variable aleatoria T con 4 grados de libertad. La hipótesis nula es $H_0 : \mu = 5$ y la alternativa es $H_0 : \mu < 5$. Según la tabla A.4, el valor crítico de t al nivel 0,05 con 4 grados de libertad es -2,13. Por lo tanto, la región crítica, que está en dirección de la hipótesis alternativa, se compone de todos los valores de t menores o iguales a -2,13. El valor t para la media muestral es $\hat{t} = \dfrac{4,8-5}{0,3/\sqrt{5}} = -1,49$. Como -1,49 no es menor o igual a -2,13, entonces los resultados del contraste no son estadísticamente significativos al nivel 0,05. No hay evidencia suficiente para rechazar la hipótesis nula a ese nivel.

b) Los resultados del contraste son estadísticamente significativos a cualquier nivel de α, para el que el valor de P del contraste sea menor o igual a α. El valor P del contraste es la probabilidad de una media muestral de 4,8 o menor si la actual es igual a 5. Es decir, el valor P es $P\left(\overline{X} \leq 4,8\right) = P\left(\dfrac{\overline{X}-5}{0,3/\sqrt{5}} \leq \dfrac{4,8-5}{0,3/\sqrt{5}}\right) = P\left(t \leq -1,49\right)$, donde t es una variable aleatoria T con 4 grados de libertad. El valor más cercano a -1,49 obtenido en la tabla A.4 es -1,533 que corresponde con $\alpha = 0,1$.

Glosario

error tipo I. Es el error que puede ocurrir al rechazar la hipótesis nula cuando es verdadera.

error tipo II. Es el error de no rechazar la hipótesis nula cuando es falsa.

estimador puntual. Es un valor de un estadístico que se usa para estimar un parámetro poblacional.

estimador insesgado. Es cuando la media de la distribución muestral de un estadístico es igual al correspondiente parámetro poblacional.

estimador más eficaz. De todos los posibles estimadores insesgados de algún parámetro, es el que tiene menor varianza.

hipótesis alternativa. Es una hipótesis opcional por si se rechaza H_0, se denota con H_1.

hipótesis estadística. Es una aseveración o conjetura con respecto a una o más poblaciones.

hipótesis nula. Se refiere a cualquier hipótesis que se desea probar y se denota con H_0.

inferencia estadística. Consiste en aquellos métodos por los cuales se realizan inferencias o generalizaciones acerca de una población.

intervalo de confianza. Intervalo donde se espera encontrar el valor de un parámetro de la población.

intervalo de predicción. Intervalo que intenta predecir los posibles valores de una observación futura. Brinda una buena estimación de la ubicación de una observación futura.

intervalo de tolerancia. Intervalo donde “cae” la mayoría de la población. Mide la dimensión de la población.

nivel de significancia. Probabilidad de riesgo máximo de cometer un error tipo I.

potencia de una prueba estadística. Es la probabilidad de rechazar la hipótesis nula dado que una alternativa específica es verdadera.

región crítica. Conjunto de todos los valores del estadístico de contraste en la dirección de la hipótesis alternativa con un valor P menor o igual que el nivel de significación.

valor P del contraste. Es la probabilidad de que un valor estadístico de contraste, en el sentido de la hipótesis alternativa y con el valor extremo obtenido, podría haber ocurrido si H_0 fuera cierta.

SECCIÓN 5

REGRESIÓN LINEAL Y CORRELACIÓN

En la sección se desarrollan varios conceptos referidos a datos bidimensionales, tales como: diagramas de dispersión, coeficientes de correlación y regresión lineal simple por medio del método de mínimos cuadrados.

Adicionalmente, se le presenta una actividad dinámica utilizando el *software* gratuito *winstats.exe*, con el fin de buscar una mayor comprensión de los conceptos que se desarrollan en las secciones 11.1, 11.2 y 11.3 del libro de texto; así como incentivar el uso de nuevas tecnologías para el proceso de enseñanza-aprendizaje de la matemática.

OBJETIVO GENERAL

Estudiar la relación entre variables mediante la correlación y la regresión lineal.

OBJETIVOS ESPECÍFICOS

1. Representar datos bivariados en un diagrama de dispersión.
2. Determinar, en un conjunto de datos bivariados, el coeficiente de correlación lineal.
3. Encontrar la recta de mejor ajuste mediante el método de mínimos cuadrados.

PUNTOS DE ATENCIÓN DE LAS SECCIONES 11.1 A LA 11.3

En la sección se estudia la relación que se puede hacer entre dos variables (datos bivariados), como por ejemplo edad y peso, peso y altura, edad y escolaridad, entre otros.

En la sección 11.1 del libro se plantea una introducción general a la regresión lineal, por lo cual se recomienda que esta sección sea destinada solamente para lectura introductoria al tema.

Antes de revisar los puntos de atención de las secciones 11.2 y 11.3, se debe tener claro los diagramas de dispersión, *¿qué es un diagrama de dispersión?*

Sea una serie de pares ordenados $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, que representan a las variables x e y, respectivamente. El diagrama de dispersión de los datos es simplemente la representación gráfica, por medio de puntos, de los pares ordenados (x_i, y_i) en un sistema de coordenadas rectangulares.

EJEMPLO 1

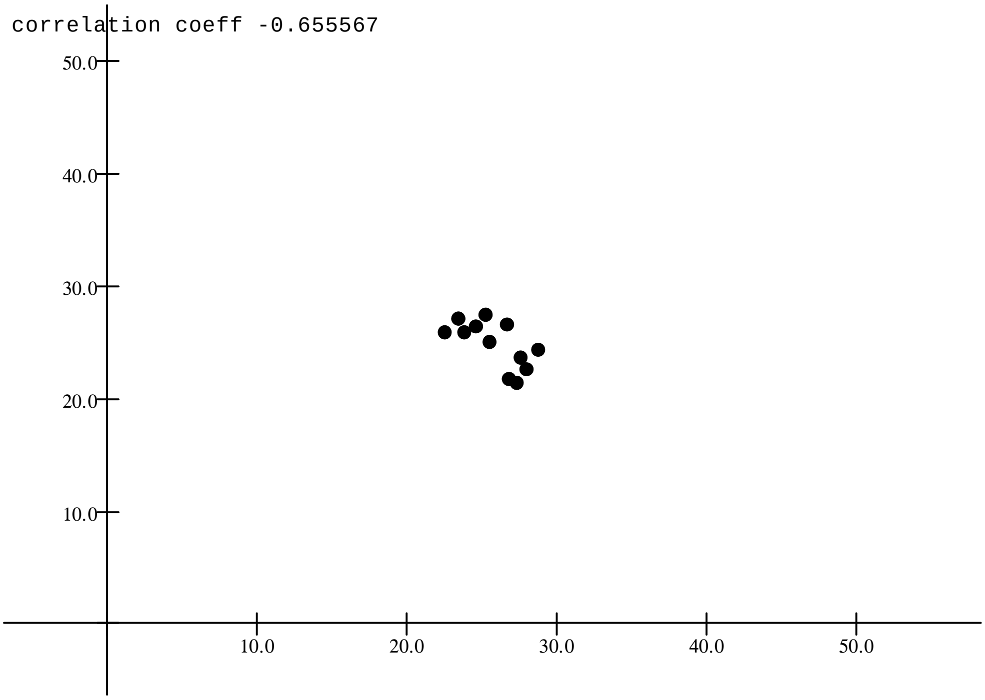
En el ejercicio 11.4, de la página 398, se dan dos variables: la tensión normal (*x*) y la resistencia al corte (*y*); por ello al ser datos bivariados, se pueden representar en un diagrama de dispersión.

"En cierto tipo de espécimen de prueba metálico, se sabe que la tensión normal sobre este se relaciona de manera funcional con la resistencia al corte. Los siguientes son un conjunto de datos experimentales obtenidos para las dos variables:"

Tensión normal (x)	**Resistencia al corte** (y)
26,8	26,5
25,4	27,3
28,9	24,2
23,6	27,1
27,7	23,6
23,9	25,9
24,7	26,3
28,1	22,5
26,9	21,7
27,4	21,4
22,6	25,8
25,6	24,9

SOLUCIÓN

Tomando en cuenta los datos, el diagrama de dispersión se representa en la gráfica 1.



Gráfica 1

Al tener dos variables, por lo general, interesa saber qué tipo de relación algebraica tienen (lineal, polinomial, exponencial, etc.) y cuál es su grado de relación. Por ejemplo, si X e Y representan la longitud y la circunferencia de una clase particular de hueso, en el cuerpo de un adulto, es importante mediante una muestra constatar si existe alguna relación algebraica entre esas medidas.

El análisis de correlación intenta medir la intensidad de tales relaciones entre dos variables por medio de un solo número, denominado coeficiente de correlación.

Solo interesa desarrollar este (r), el cual sirve como indicador numérico del grado de relación lineal que exista entre la variables x e y en una muestra.

Para calcular el coeficiente de correlación lineal se tiene la siguiente fórmula:

$$r = \frac{\sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum\left(x_i - \overline{x}\right)^2 \sum\left(y_i - \overline{y}\right)^2}}.$$

Ahora bien, se vuelve muy engorrosa y complicada, por lo que se utiliza una equivalente y mucho más sencilla.

$$r=\frac{\sum x_i y_i-\frac{\sum x_i\cdot\sum y_i}{n}}{\sqrt{\sum x_i^2-\frac{\left(\sum x_i\right)^2}{n}}\cdot\sqrt{\sum y_i^2-\frac{\left(\sum y_i\right)^2}{n}}}.$$

Observación: $r>0$ si y tiende a incrementar cuando se incrementa x y $r<0$ si y tiende a disminuir cuando disminuye x. Además $-1\leq r\leq 1$.

EJEMPLO 2

Con los datos del ejemplo 1 (*ejercicio 11.4 de la página 398*), calcule el coeficiente de correlación de los dados.

SOLUCIÓN

Para poder calcular el coeficiente de correlación lineal, de una forma ordenada y segura, se sugiere hacer la tabla 1.

Tabla 1

x_i	y_i	x_i^2	x_i^2	$x_i y_i$
26,8	26,5	718,24	702,25	710,2
25,4	27,3	645,16	745,29	693,42
28,9	24,2	835,21	585,64	699,38
23,6	27,1	556,96	734,41	639,56
27,7	23,6	767,29	556,96	653,72
23,9	25,9	571,21	670,81	619,01
24,7	26,3	610,09	691,69	649,61
28,1	22,5	789,61	506,25	632,25
26,9	21,7	723,61	470,89	583,73
27,4	21,4	750,76	457,96	586,36
22,6	25,8	510,76	665,64	583,08
25,6	24,9	655,36	620,01	637,44
∑x= 311,6	∑y= 297,2	$\sum x^2$ =8134,26	$\sum y^2$ =7407,8	∑xy=7687,76

Seguidamente, con la fórmula recomendada para el cálculo del coeficiente de correlación lineal.

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \cdot \sum y_i}{n}}{\sqrt{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} \cdot \sqrt{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}}}$$

$$r=\frac{7687{,}76-\frac{311{,}6\cdot 297{,}2}{12}}{\sqrt{8134{,}26-\frac{\left(311{,}6\right)^2}{12}}\cdot\sqrt{7407{,}8-\frac{\left(297{,}2\right)^2}{12}}}$$

$$r\approx\frac{7687{,}76-7717{,}29}{\sqrt{43{,}05}\cdot\sqrt{47{,}15}}$$

$$r\approx -0{,}6554.$$

Se obtiene así, que el coeficiente de correlación lineal es $r\approx -0{,}6554$.

Ahora bien, ¿cómo interpretar este resultado? Cuanto más fuerte sea la relación entre x e y, r estará más cerca de -1 ó de 1; cuanto más débil sea esa relación, r estará más cerca de 0.

En las secciones 11.1, 11.2 y 11.3 se hace una introducción de la regresión lineal, se explica el modelo de regresión lineal simple y con el método de mínimos cuadrados se determina la recta de mejor ajuste.

Otra manera de calcular los coeficientes a y b de la recta de mejor ajuste es con la tabla del ejemplo 2 y las siguientes fórmulas:

$$S_x=\sqrt{\frac{\sum x_i^2-\frac{\left(\sum x_i\right)^2}{n}}{n-1}},\qquad S_y=\sqrt{\frac{\sum y_i^2-\frac{\left(\sum y_i\right)^2}{n}}{n-1}}.$$

Entonces se tiene que:

$$b=\frac{rS_y}{S_x}\text{ y } a=\overline{y}-b\overline{x}.$$

Así, la recta de regresión ajustada es $\hat{y}=a+bx$.

Observación: si se tiene que $S_{xy}=\sqrt{\dfrac{\sum x_i y_i-\dfrac{\sum x_i \sum y_i}{n}}{n-1}}$ entonces $r=\dfrac{S_{xy}}{S_x S_y}$.

EJEMPLO 3

De acuerdo con los datos del ejemplo 1 (ejercicio 11.4 de la página 398) realice lo que se le pide a continuación:

a) determine la recta de mejor ajuste $\hat{y}=a+bx$;

b) estime la resistencia al corte para una tensión normal de 24,5 kilogramos por centímetro cuadrado.

SOLUCIÓN

Se tienen los siguientes datos que serán útiles:

$$r=-0{,}6554,\ \bar{x}=\frac{311{,}6}{12}\approx 25{,}97,\ \bar{y}=\frac{297{,}2}{12}\approx 24{,}77.$$

a) Como $b=\dfrac{rS_y}{S_x}$ y $a=\bar{y}-b\bar{x}$ entonces se calcula lo siguiente:

$$S_x=\sqrt{\frac{\sum x_i^2-\dfrac{\left(\sum x_i\right)^2}{n}}{n-1}}=\sqrt{\frac{8134{,}26-\dfrac{(311{,}6)^2}{12}}{11}}\approx\sqrt{\frac{43{,}05}{11}}\approx 1{,}9783.$$

$$S_y=\sqrt{\frac{\sum y_i^2-\dfrac{\left(\sum y_i\right)^2}{n}}{n-1}}=\sqrt{\frac{7407{,}8-\dfrac{(297{,}2)^2}{12}}{11}}\approx\sqrt{\frac{47{,}15}{11}}\approx 2{,}0703.$$

Por lo tanto,

$$b=\frac{rS_y}{S_x}=\frac{-0{,}6554\cdot 2{,}0703}{1{,}9783}=-0{,}6859 \text{ y } a=\bar{y}-b\bar{x}=24{,}77-0{,}6859(25{,}97)=42{,}5828.$$

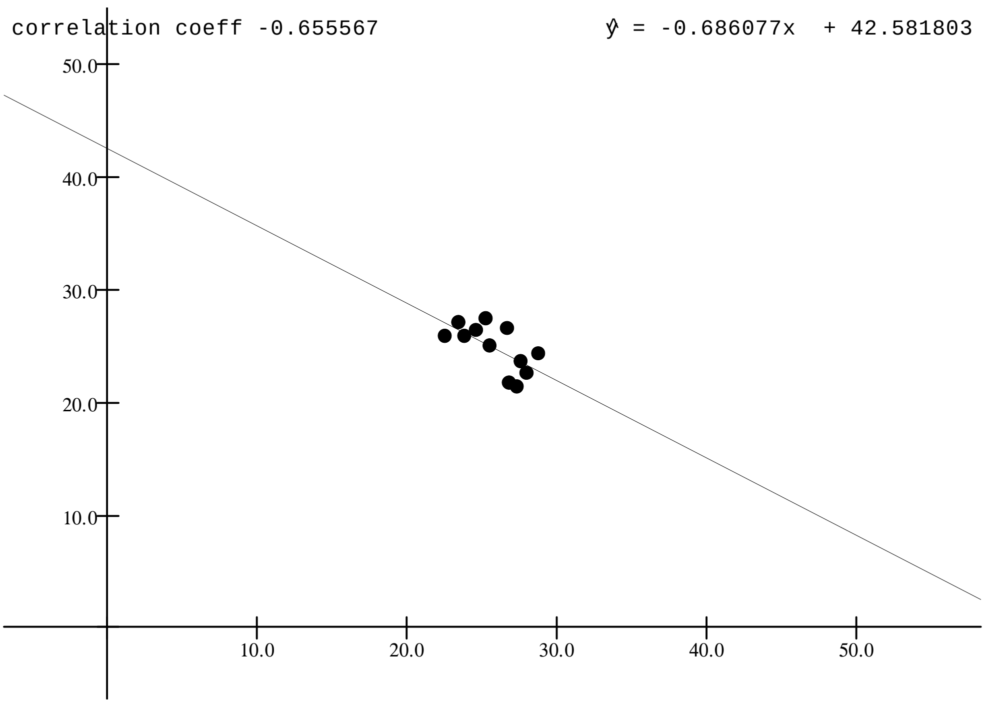
Entonces la recta de mejor ajuste, para la muestra del ejercicio 11.4, es:

$$\hat{y} = 42,5828 - 0,6859x.$$

b) Para estimar la resistencia al corte, para una tensión normal de 24,5 kilogramos por centímetro cuadrado, simplemente se sustituye, en la ecuación encontrada, la variable x por el valor 24,5, se desarrolla la operación y se obtiene el valor $\hat{y}$.

$$\hat{y} = 42,5828 - 0,6859 \cdot 24,5 = 25,8.$$

La gráfica 2 representa la recta en el diagrama de dispersión, con el objetivo de visualizar y comparar los datos de la muestra con el comportamiento de la recta de mejor ajuste.



Gráfica 2

LABORATORIOS (datos bidimensionales)

Como se observó en los ejemplos 1, 2 y 3, el manejo de datos bivariables es un poco engorroso y lento. Por eso, para hacerle frente a este tipo de ejercicios de una forma rápida y segura, es importante aprender a utilizar un *software* que facilite esta labor.

A continuación, se le presenta el siguiente laboratorio cuyo fin es aprender a utilizar la aplicación para datos bivariables.

Para realizarlo se usan los datos del ejemplo 1.

Instrucciones

1. Abra el *software winstats.exe* haciendo doble clic en el ícono correspondiente. Se despliega una pequeña pantalla verde con dos opciones en el menú: $\boxed{Window}$, $\boxed{Help}$.
2. Ubíquese en $\boxed{Window}$ y escoja la opción $\boxed{Multi - var\ data\ \ \mathbf{F1}}$. Esta es la opción para trabajar con datos unidimensionales. Se le abrirá una pantalla blanca donde irán los datos de la muestra.
3. Para poner los datos en 0, se debe ubicar en la primera opción del menú principal de la ventana $\boxed{File}$, escoja las opciones $\boxed{New}$ y luego $\boxed{Zeros}$.
4. Debe indicarle a la aplicación de cuántas variables es la muestra. Para ello, se debe ubicar en la segunda opción del menú principal de la ventana $\boxed{Edit}$ y escoger la opción $\boxed{Dimensions}$. Se abrirá una ventana donde se debe indicar la cantidad de filas (*rows*) y columnas (*columns*). Las columnas indican las variables de la muestra (en este caso 2) y las filas indican la cantidad de pares ordenados (en este caso 12).

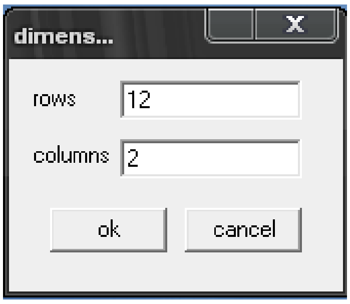


Imagen 1

5. Para poder digitar los datos, haga clic en la casilla correspondiente, luego debe oprimir la tecla "*enter*" para salir. En la columna vbl 1 se introducen los datos de la variable x y en la vb 2, los de la y, como se muestra en la siguiente imagen:

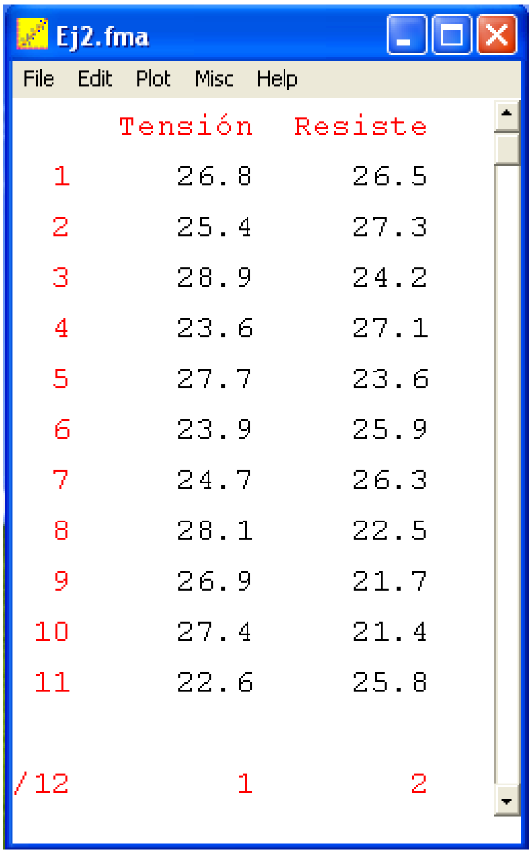


Imagen 2

6. Para observar el diagrama de dispersión debe ubicarse en la opción del menú principal $\boxed{Misc}$ y escoger la opción $\boxed{Correlation\ demo\ ...}$. En ese momento, aparecerá una ventana que indica la variable dependiente y la independiente; por defecto, la primera columna será independiente y la segunda dependiente, oprima el botón *OK*. Con las teclas $\boxed{PgDn}$ y $\boxed{PgUp}$ podrá acercar o distanciar el diagrama de dispersión.

7. Para observar, en la ventana del diagrama de dispersión, el coeficiente de correlación lineal, se debe realizar la secuencia en el menú de la ventana: $\boxed{Misc} \Rightarrow \boxed{Show} \Rightarrow \boxed{Correlation\ coefficient}$. En esa misma ventana puede ver la ecuación y la representación gráfica de la recta de regresión ajustada, siguiendo estas secuencias en el menú del diagrama de dispersión:

$$\boxed{Misc} \Rightarrow \boxed{Show} \Rightarrow \boxed{Regression\ Equation\ ...}$$

$$\boxed{Misc} \Rightarrow \boxed{Show} \Rightarrow \boxed{Regression\ line\ ...}$$

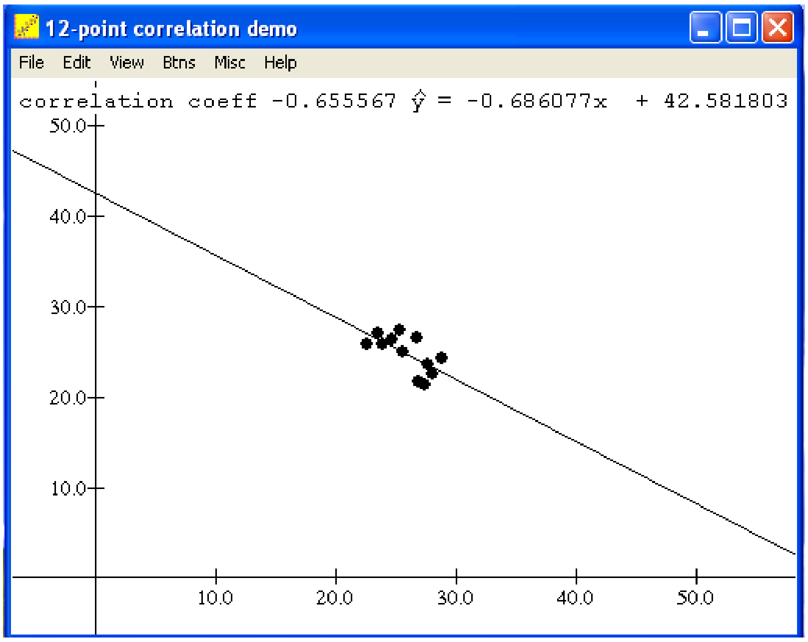


Imagen 3

EJERCICIOS DE AUTOEVALUACIÓN

1. Una cadena de restaurantes le pide a una empresa estadística que lleve a cabo un estudio, para determinar la relación entre los gastos de publicidad semanal x y las ventas del restaurante y. Se recolectaron los valores de estas variables en millones de colones, durante 8 semanas, obteniendo los siguientes datos:

$$\sum x = 310 \qquad \sum y = 4235$$

$$\sum x^2 = 13935 \qquad \sum y^2 = 2308421$$

$$\sum xy = 173524$$

   a) Calcular el coeficiente de correlación r.

   b) Hallar el criterio de la recta mínimo cuadrática $\hat{y} = a + bx$.

2. Considere la tabla 2 para realizar los ejercicios que se le presentan a continuación.

Tabla 2

x	2	4	5	8	10
y	12	8	10	2	4

Calcular el coeficiente de correlación r. Hallar la ecuación de la recta de regresión ajustada $\hat{y} = a + bx$.

3. Una marca nueva de autos deportivos realiza un estudio de costos de operación; para ello, primero se hace un muestreo de tamaño 30. Este consiste en que, cada día, se anota la cantidad de autos fabricados y el costo total de operación (en dólares). Al final de los 30 días se obtuvo la tabla 3.

Tabla 3

Cantidad	Costo	Cantidad	Costo	Cantidad	Costo
39	135500	26	67780	18	50160
31	92350	43	144900	52	153640
46	166500	55	203550	49	144500
31	99600	7	18300	35	188300
36	104560	31	84170	28	78450
63	243400	14	34170	18	49800
12	34100	25	69350	35	99600
19	53300	20	53150	39	112490
43	156400	9	49550	22	58250
25	65300	27	185450	13	34420

De acuerdo con la tabla 3, realice lo que se le pide a continuación.

a) Represente los datos mediante un diagrama de dispersión.

b) Calcular el coeficiente de correlación r.

c) Hallar la ecuación de la recta de regresión ajustada $\hat{y} = a + bx$.

d) Utilizando la ecuación de la recta de regresión lineal obtenida en el punto (c), cuál se esperaría que fuera el costo de fabricar 10 autos.

4. En la tabla 4, se muestran la edad (en meses) X y el peso (en libras) Y de 10 niños menores a un año de edad.

Tabla 4

x	2	5	3	8	6	11	1	5	8	3
y	9,8	15,3	12,0	19,4	17,2	22,7	7,4	12,8	22,3	13,4

a) Hallar el coeficiente de correlación entre X e Y.

b) Determinar la ecuación de la recta de mejor ajuste de Y sobre X.

c) Estimar el peso de un niño de 7 meses.

d) ¿Es confiable esta estimación, con respecto a los datos?

SOLUCIÓN A LOS EJERCICIOS DE AUTOEVALUACIÓN

1. Para encontrar lo indicado, se realizan los siguientes procedimientos:

a) Con la fórmula recomendada para el cálculo del coeficiente de correlación lineal.

$$r = \frac{\sum x_i y_i - \dfrac{\sum x_i \cdot \sum y_i}{n}}{\sqrt{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}} \cdot \sqrt{\sum y_i^2 - \dfrac{\left(\sum y_i\right)^2}{n}}}$$

$$r = \frac{173524 - \dfrac{(310)(4235)}{8}}{\sqrt{13935 - \dfrac{(310)^2}{8}}\sqrt{2308421 - \dfrac{(4235)^2}{8}}} = \frac{9417,75}{\sqrt{1922,5}\sqrt{66517,88}} \approx 0,8328.$$

Por lo tanto, el coeficiente r es $0,8328$.

b) Se tienen los siguientes datos que serán útiles:

$$r = 0,8328,\ \bar{x} = \frac{310}{8} = 38,75 \text{ e } \bar{y} = \frac{4235}{8} = 529,375.$$

Ahora bien, como $b = \dfrac{r \cdot S_y}{S_x}$ y $a = \overline{y} - b \cdot \overline{x}$ se tiene que calcular:

$$S_x = \sqrt{\frac{13935 - \frac{(310)^2}{8}}{7}} \approx 16{,}57 \text{ y } S_y = \sqrt{\frac{2308421 - \frac{(4235)^2}{8}}{7}} \approx 97{,}48.$$

Sustituyendo r , S_x y S_y en $b = \dfrac{r \cdot S_y}{S_x}$ y $a = \overline{y} - b \cdot \overline{x}$ se obtiene:

$$b = \frac{r \cdot S_y}{S_x} = \frac{0{,}8328 \cdot 97{,}48}{16{,}57} \approx 4{,}90.$$

$$a = \overline{y} - b \cdot \overline{x} = 529{,}375 - (4{,}90)(38{,}75) \approx 339{,}5.$$

Por lo tanto, la ecuación de la recta de mejor ajuste es $\hat{y} = 339{,}5 + 4{,}9x$.

2. Para poder calcular el coeficiente de correlación lineal, de una forma ordenada y segura, construya la tabla 5.

Tabla 5

x_i	y_i	x_i^2	x_i^2	$x_i y_i$
2	12	4	144	24
4	8	16	64	32
5	10	25	100	50
8	2	64	4	16
10	4	100	16	40
$\sum x = 29$	$\sum y = 36$	$\sum x^2 = 209$	$\sum y^2 = 328$	$\sum xy = 162$

a) Con la fórmula recomendada para el cálculo del coeficiente de correlación lineal.

$$r=\frac{\sum x_i y_i-\frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sqrt{\sum x_i^2-\frac{\left(\sum x_i\right)^2}{n}}\sqrt{\sum y_i^2-\frac{\left(\sum y_i\right)^2}{n}}}.$$

$$r=\frac{162-\frac{(29)(36)}{5}}{\sqrt{209-\frac{(29)^2}{5}}\sqrt{328-\frac{(36)^2}{5}}}=\frac{-46,8}{\sqrt{40,8}\sqrt{68,8}}\approx -0,8833.$$

Por lo tanto, el coeficiente r es $-0,8833$.

b) Los siguientes datos serán útiles:

$$r=-0,8833,\ \bar{x}=\frac{29}{5}=5,8 \text{ e } \bar{y}=\frac{36}{5}=7,2.$$

Ahora bien, como $b=\frac{r\cdot S_y}{S_x}$ y $a=\bar{y}-b\cdot\bar{x}$ se calcula:

$$S_x=\sqrt{\frac{209-\frac{(29)^2}{5}}{4}}\approx 3,1937 \text{ y } S_y=\sqrt{\frac{328-\frac{(36)^2}{5}}{4}}\approx 4,1473.$$

Sustituyendo r, S_x y S_y en $b=\frac{r\cdot S_y}{S_x}$ y $a=\bar{y}-b\cdot\bar{x}$ se obtiene:

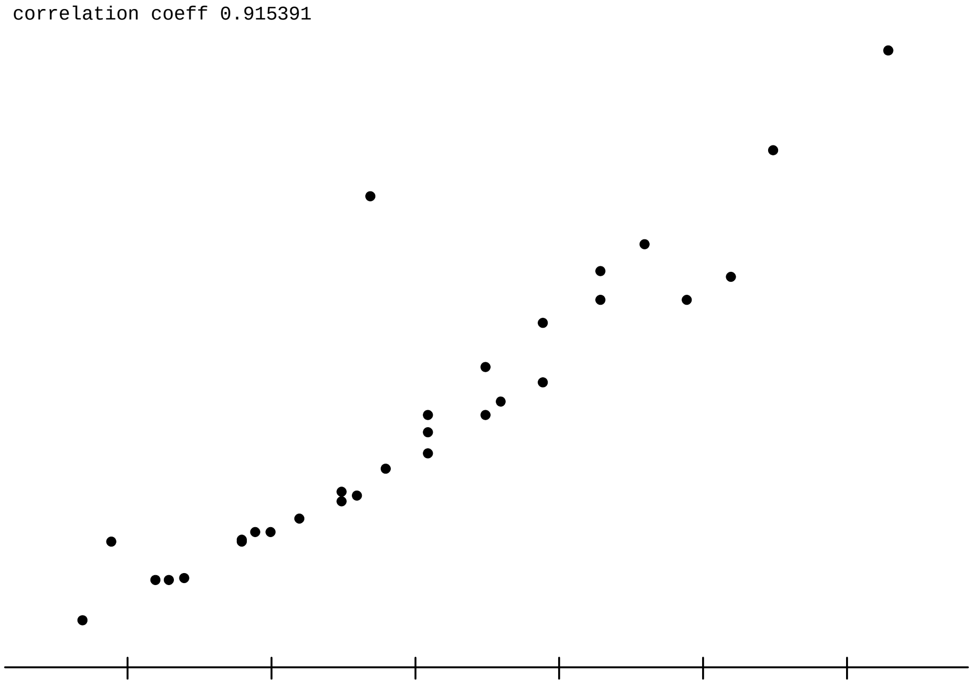
$$b=\frac{r\cdot S_y}{S_x}=\frac{-0,8833\cdot 4,1473}{3,1937}=-1,1470.$$

$$a=\bar{y}-b\cdot\bar{x}=7,2-(-1,470)(5,8)=13,8526.$$

Por lo tanto, la ecuación de la recta de regresión ajustada es $\hat{y}=13,8526-1,1470x$.

3. Por la cantidad de los datos, los siguientes ejercicios se realizarán con una calculadora científica en el modo estadístico.

   a) En la gráfica 3 se representa el diagrama de dispersión de los datos de la tabla 3.



Gráfica 3

   b) Con la fórmula recomendada para el cálculo del coeficiente de correlación lineal.

$$r=\frac{\sum x_i y_i-\frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sqrt{\sum x_i^2-\frac{\left(\sum x_i\right)^2}{n}}\sqrt{\sum y_i^2-\frac{\left(\sum y_i\right)^2}{n}}},$$

$$r=\frac{111242790-\frac{(911)(2961620)}{30}}{\sqrt{33555-\frac{(911)^2}{30}}\sqrt{384353612600-\frac{(2961620)^2}{30}}}=\frac{21308262{,}67}{\sqrt{5890{,}97}\sqrt{91980511790}}\approx 0{,}9154.$$

   Por lo tanto, el coeficiente r es $0{,}9154$.

c) Los siguientes datos que serán útiles:

$$r = 0{,}9154\,,\ \bar{x} = 30{,}37 \text{ e } \bar{y} = 98720{,}67.$$

Ahora se calcula:

$$S_x = \sqrt{\frac{33555 - \frac{(911)^2}{30}}{29}} \approx 14{,}253,$$

$$S_y = \sqrt{\frac{384353612600 - \frac{(2961620)^2}{30}}{29}} \approx 56318{,}219.$$

Sustituyendo r, S_x y S_y en $b = \frac{r \cdot S_y}{S_x}$ y $a = \bar{y} - b \cdot \bar{x}$ se obtiene:

$$b = \frac{r \cdot S_y}{S_x} = \frac{0{,}9154 \cdot 56318{,}219}{14{,}253} \approx 3617.$$

$$a = \bar{y} - b \cdot \bar{x} = 98720{,}67 - (3617)(30{,}37) \approx -11128.$$

Por lo tanto, la ecuación de la recta de mejor ajuste es $\hat{y} = -11128 + 3617x$.

d) Con la ecuación de la recta de regresión encontrada en el punto (c), se sustituye x por 10 para encontrar el costo aproximado $\hat{y}$, de la siguiente manera:

$$\hat{y} = 3617x - 11128 = 3617 \cdot (10) - 11128 = 25042.$$

Por lo tanto, el costo aproximado de construir 10 autos es de 25 042 dólares.

4. Para poder calcular el coeficiente de correlación lineal, de una forma ordenada y segura, construya la tabla 6.

Tabla 6

x_i	y_i	x_i^2	x_i^2	$x_i y_i$
2	9,8	4	96,04	19,6
5	15,3	25	234,09	76,5
3	12,0	9	144,00	36,0
8	19,4	64	376,36	155,2
6	17,2	36	295,84	103,2
11	22,7	121	515,29	249,7
1	7,4	1	54,76	7,4
5	12,8	25	163,84	64,0
8	22,3	64	497,29	178,4
3	13,4	9	179,56	40,2
$\sum x = 52$	$\sum y = 152,3$	$\sum x^2 = 358$	$\sum y^2 = 2557$	$\sum xy = 930,2$

a) Según la fórmula recomendada para el cálculo del coeficiente de correlación lineal.

$$r = \frac{930,2 - \frac{(52)(152,3)}{10}}{\sqrt{358 - \frac{(52)^2}{10}}\sqrt{2557 - \frac{(152,3)^2}{10}}} = \frac{138,24}{\sqrt{87,6}\sqrt{237,5}} \approx 0,9583.$$

Por lo tanto, el coeficiente r es $0,9583$.

Los siguientes datos que serán útiles: $r = 0,9584$, $\bar{x} = \frac{52}{10} = 5,2$ e $\bar{y} = \frac{152,3}{10} = 15,23.$

Ahora se calcula:

$$S_x = \sqrt{\frac{358 - \frac{(52)^2}{10}}{9}} \approx 3{,}1198 \text{ y } S_y = \sqrt{\frac{2557 - \frac{(152{,}3)^2}{10}}{9}} \approx 5{,}1367.$$

Sustituyendo r, S_x y S_y en $b = \frac{r \cdot S_y}{S_x}$ y $a = \overline{y} - b \cdot \overline{x}$ se obtiene:

$$b = \frac{r \cdot S_y}{S_x} = \frac{0{,}9583 \cdot 5{,}1367}{3{,}1198} = 1{,}5778,$$

$$a = \overline{y} - b \cdot \overline{x} = 15{,}23 - (1{,}5778)(5{,}2) = 7{,}0254.$$

Por lo tanto, la ecuación de la recta de mejor ajuste es $\hat{y} = 7{,}0254 + 1{,}5778x$.

b) Utilizando la ecuación de la recta de regresión, encontrada en el punto (b), se sustituye x por 7 para encontrar el peso aproximado $\hat{y}$ de un niño de 7 meses, de la siguiente manera:

$$\hat{y} = 1{,}5778x + 7{,}0254 = 1{,}5778 \cdot (7) + 7{,}0254 = 18{,}07.$$

Por lo tanto, el peso aproximado de un niño de 10 meses es de 18,07 libras.

c) Sí, ya que el coeficiente de correlación está muy cercano a 1.

Glosario

análisis de regresión. Método que se utiliza para encontrar la mejor relación entre dos o más variables que, al cuantificar la intensidad de dicha relación, se permitan predecir los valores de la respuesta ante valores dados del regresor.

análisis de correlación. Procedimiento que intenta medir la intensidad de la relación entre dos variables, por medio de un solo número denominado *coeficiente de correlación*.

correlación lineal. Es cuando la mayoría de puntos, en un diagrama de dispersión, parecen encontrarse cerca de una recta.

diagrama de dispersión. Es, simplemente, la representación gráfica, por medio de puntos, de los pares ordenados (x_i, y_i) en un sistema de coordenadas rectangulares.

método de mínimos cuadrados. Es un procedimiento de minimización para estimar los parámetros de la ecuación de regresión, en el que la suma de los cuadrados de los residuos sea mínima.

regresión. Estimación de una variable (la variable dependiente) a partir de una o más relacionadas entre sí (las variables independientes).

regresión lineal simple. Es el método de regresión que utiliza solamente una variable regresora.

recta de regresión ajustada. Es una estimación de la verdadera recta de regresión. Se espera que, cuando se disponga de una gran cantidad de datos, la recta ajustada esté más cerca de la verdadera línea de regresión