Business School

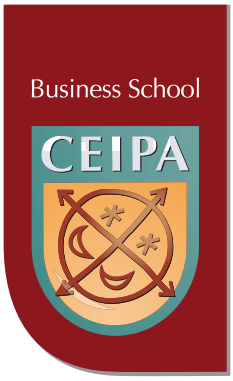
CEIPA

Notas de Estadística

Francisco Javier Jaramillo Álvarez
Docente tiempo completo CEIPA

Notas de Estadística

Francisco Javier Jaramillo Álvarez
Especialista en Estadística de la Universidad Nacional
Ingeniero de alimentos de la Corporación Universitaria Lasallista



Versión 4
Institución universitaria CEIPA
2011

Notas de Estadística
2011

Francisco Javier Jaramillo Álvarez
Docente tiempo completo CEIPA

ISBN 978-958-99767-3-9

CEIPA. Calle 77 sur Nº 40 - 165. Sabaneta, Colombia
Tel. (57-4) 3056100
www.ceipa.edu.co

Diagramación y diseño
Eliana Jaramillo Gaviria elianajga@gmail.com

Hecho en Colombia

SOBRE EL AUTOR

Francisco Javier Jaramillo Álvarez. Especialista en Estadística; Ingeniero de Alimentos. Estudiante de Maestría en Educación y nuevas Tecnologías. Docente Escuela de Administración, Institución Universitaria CEIPA. e-mail: francisco.jaramillo@ceipa.edu.co

Presentación

Para la Escuela de Administración de la Fundación Universitaria Ceipa, es muy grato e importante presentar hoy, lo que conversacionalmente, y muy modestamente digo yo, hemos denominado "Notas de".

Estos "cuadernos de administración", adquieren su importancia por ser parte muy relevante en nuestro modelo educativo; todo su contenido responde a la arquitectura del núcleo temático para el cual fue elaborado; no significa lo anterior, que no sean de gran utilidad también para el estudiante de administración de empresas de cualquier otra institución, y para los profesores de administración, ya que su contenido recoge, en forma muy clara y breve, los aspectos fundamentales del tema propuesto; a los profesionales de la administración y directivos en general, les facilita la recordación de importantes conceptos administrativos.

Son varios los aspectos que debemos destacar en estas "Notas de"; son elaboradas por nuestros profesores y ello garantiza su pertinencia dentro de nuestro diseño curricular por núcleos problémicos; si bien retoman lo fundamental de la teoría de los temas tratados, ya en los casos que plantean y en los ejercicios que proponen, éstos se elaboran a partir del conocimiento de nuestra realidad, de nuestro entorno, y en mi modesto sentir, esa pertinencia las diferencia de la gran mayoría de los tratados de administración que se manejan en nuestras instituciones, y que contemplan realidades culturalmente muy diferentes a la nuestra; en este sentido, estamos seguros de que estamos haciendo un aporte muy importante, a la ciencia de la administración de empresas en Colombia.

Otro aspecto que consideramos debe destacarse, es que su elaboración, apunta más a facilitar el desarrollo de competencias, que a la adquisición de una erudición en la ciencia administrativa; diríamos en un lenguaje muy sencillo: apuntan más a lo urgente que a lo eminente, sin que lo eminente esté ausente de ellas.

Celebramos la iniciativa de la Escuela de Administración, quienes idearon estas publicaciones; felicitamos al Decano, a su equipo de directores de programas y a todos y cada uno de los profesores autores de ellas; todos deben sentir hoy la gran satisfacción de estar construyendo "administración de empresas colombiana" y de estar haciendo un gran aporte a la calidad de la formación de los profesionales colombianos.

Antonio Mazo Mejía
Rector Fundador

Tabla de Contenido

INTRODUCCIÓN

Estas notas presentan los fundamentos generales de la estadística, de forma tal que puedan ser de utilidad para cualquier estudiante de administración, independientemente de su campo de especialización.

Se pretende que el estudiante obtenga una apreciación sobre la utilidad del método estadístico en su campo profesional, que adquiera una buena comprensión de los parámetros o estadísticos básicos y la lógica que refuerza la aplicación de las herramientas estadísticas, que sea capaz de seleccionar la técnica estadística apropiada y de realizar los cálculos necesarios, así como interpretar y comprender los resultados de su estudio.

La mayoría de textos de estadística utilizados en nuestro medio son traducciones de libros de Estados Unidos. Por eso, los ejemplos que traen son descontextualizados y desactualizados; uno de los objetivos de este texto es adaptar las herramientas estadísticas a casos de nuestro medio y, para ello, utiliza datos reales y actuales.

La estadística es una herramienta esencial en la formación profesional de un administrador. Toda persona responsable de la toma de decisiones debe tener una formación muy sólida en fundamentos de estadística, ya que ella interactúa directamente con la investigación, el mercadeo, el control de calidad, área de ventas, área de producción, entre otras.

La labor a desarrollar en una organización requiere de personas con una alta dosis de creatividad, asertividad, sentido común y con un conocimiento profundo de técnicas administrativas y de análisis, entre otras. Cotidianamente el administrador enfrenta situaciones, especialmente en las áreas financiera, contable, de mercadeo, producción e investigación, donde se requiere la utilización de métodos y procedimientos que faciliten la organización, análisis y representación de las informaciones. Este texto aspira a aportar luces que logren ese fin.

Los objetos de aprendizaje, que harán las veces de capítulos, son:

1. Técnicas descriptivas
2. Regresión y correlación
3. Probabilidades
4. Distribuciones de probabilidades
5. Estadística inferencial

Los objetivos de aprendizaje pueden resumirse como:

- Procesar en forma clara y precisa la información objeto de estudio.
- Adquirir la capacidad de interpretar adecuadamente los resultados de un reporte estadístico.
- Modelar y predecir el comportamiento de variables para apoyar la toma de decisiones administrativas.
- Realizar proyecciones por medio de ecuaciones de ajuste.
- Encontrar e interpretar la correlación entre variables.
- Utilizar el concepto de probabilidad en la toma de decisiones administrativas en condiciones de incertidumbre.
- Modelar situaciones administrativas mediante las distribuciones especiales de probabilidades.
- Efectuar estimaciones y probar hipótesis estadísticas para apoyar científicamente la toma de decisiones administrativas.

NOCIONES PRELIMINARES DE ESTADÍSTICA

Se empezará por definir algunos conceptos elementales que son básicos para una comprensión integral de la estadística.

Estadística

Es un conjunto de métodos científicos que estudian la recolección, resumen, análisis e interpretación de datos, para ayudar en la toma de decisiones informadas e inteligentes, o para explicar comportamientos de algún fenómeno o estudio, principalmente en aquellos casos en que hay incertidumbre.

La estadística permite obtener resultados en cualquier tipo de estudio cuyos movimientos y relaciones, por su variabilidad intrínseca, no puedan ser abordados desde la perspectiva de las leyes deterministas. Podríamos, desde un punto de vista más amplio, definir la estadística como la ciencia que estudia cómo sacar provecho de la información y cómo sugerir una guía de acción en situaciones que implican incertidumbre.

Anderson, Sweeney y Williams (2008) dicen que:

> Especialmente en los negocios y en la economía, la información obtenida al reunir datos, analizarlos, presentarlos e interpretarlos proporciona a directivos, administradores y personas que deben tomar decisiones una mejor comprensión del negocio o entorno económico, permitiéndoles así tomar mejores decisiones con base en mejor información (p.3).

La estadística se divide en:

- E. Descriptiva, cuyo objetivo es el logro de un orden lógico que logre revelar rápida y fácilmente el mensaje que contienen los datos, ya que grandes masas de datos desorganizados son de poco valor; la estadística descriptiva proporciona técnicas para organizar este tipo de datos. En este caso, los resultados del análisis no pretenden ir más allá del conjunto de datos.

- E. Inferencial, cuyo objetivo es la elaboración de estimaciones y pruebas de hipótesis acerca de las características de una población, a partir de los datos obtenidos de una muestra; apoyándose en el cálculo de probabilidades efectúa estimaciones, decisiones, predicciones u otras generalizaciones sobre un conjunto mayor de datos.

¡PIENSA!

¿Por qué y para qué la estadística en tu profesión?

Población

Conjunto de elementos de interés en un determinado estudio; deben tener una característica común. Se trabaja con una población cuando nuestro estudio se basa en datos reunidos para todos los elementos que cumplen la característica.

El tamaño de la población se representa como N.

Muestra

Subconjunto de la población; debe ser una porción representativa, es decir, debe ser el mejor reflejo posible del conjunto del cual proviene, tanto en número como en calidad (debe tener las mismas características de la población). Algunas razones por las cuales es necesario seleccionar una muestra son:

- El costo de estudiar a todos los integrantes de una población con frecuencia es excesivo.
- Frecuentemente el ponerse en contacto con toda la población supondría mucho tiempo.
- La idoneidad de los resultados de una muestra, ya que generalmente puede proporcionar información muy precisa sobre las principales propiedades de la población.
- La naturaleza destructiva de ciertas pruebas.
- La imposibilidad física de verificar todos los artículos de la población.

El tamaño de la muestra se representa como n.

Censo

Evaluación de todos los elementos de la población; es razonable si la población es pequeña o ante determinadas situaciones que lo ameriten, pero no siempre es viable por escasez de tiempo, o de recursos humanos o financieros.

Muestreo

Proceso de selección de una muestra dentro de una población; lo preferido es hacerlo aleatoriamente para no incurrir en sesgo, pero cuando la población es muy heterogénea es mejor hacerlo estratificado, que consiste en dividir la población en

varios subconjuntos diferenciables y de cada uno de ellos seleccionar una muestra aleatoria.

Parámetros

Valores característicos de una población. Un parámetro es un valor fijo (no aleatorio) que caracteriza a una población en particular; en general, un parámetro es una cantidad desconocida y rara vez se puede determinar exactamente su valor, debido a la dificultad práctica de observar todas las unidades de una población.

Estadígrafos

Valores característicos de una muestra. Un estadígrafo no es un valor fijo, ya que puede tener varios resultados posibles según la muestra seleccionada.

Variable

Característica que al ser medida en diferentes unidades elementales puede adoptar valores diferentes.

Las variables pueden ser:

- Cualitativas: etiquetas o nombres que se usan para identificar un atributo de un elemento. Pueden ser numéricas o no numéricas; para procesar datos cualitativos es frecuente codificar con números las diferentes categorías, pero esos números tienen únicamente un carácter representativo y, por lo tanto, las operaciones aritméticas carecen de sentido en este caso.

- Cuantitativas: siempre se expresan en escala numérica. Pueden ser: discretas (solamente pueden asumir ciertos valores –generalmente enteros– y hay "huecos" entre ellos; son el resultado de un conteo) o continuas (pueden tomar cualquier valor dentro de un rango específico y son resultado de una medición).

EJERCICIO GENERAL

Para el desarrollo de los diferentes objetos de aprendizaje, consideraremos el siguiente ejemplo:

Cierta compañía tiene 50 representantes de ventas que venden un producto suyo en todo el territorio nacional; dichos representantes tienen como ciudad sede a Medellín, Bogotá o Cali, ciudades donde se encuentran las fábricas.

A continuación se muestra el número de productos vendidos por cada representante durante el mes anterior, al igual que su ciudad sede y sus años de experiencia.

	Ciudad sede	Años exper.	Productos vendidos
1	B	9	215
2	B	6	186
3	M	6	240
4	C	4	105
5	B	4	156
6	B	8	225
7	M	3	133
8	C	6	120
9	B	4	85
10	M	8	185
11	B	4	76
12	B	6	190
13	B	8	256
14	C	3	102
15	M	12	350
16	C	8	198
17	B	7	235
18	B	13	303
19	B	11	250
20	C	6	178
21	M	6	196
22	B	6	204
23	M	11	303
24	B	15	356
25	M	5	256

	Ciudad sede	Años exper.	Productos vendidos
26	B	9	205
27	B	6	180
28	C	6	175
29	B	12	318
30	B	7	185
31	C	2	56
32	B	4	118
33	C	5	152
34	M	2	235
35	B	9	196
36	C	8	199
37	B	11	305
38	C	12	312
39	B	8	198
40	B	5	111
41	B	8	165
42	M	7	321
43	M	3	180
44	B	6	156
45	C	10	215
46	B	10	193
47	C	6	150
48	B	5	128
49	B	10	165
50	B	15	300

OBJETO DE APRENDIZAJE 1
TÉCNICAS DESCRIPTIVAS

El manejo de la información depende del tipo de variable que se evalúe, aunque siempre se reduce a la construcción de tablas, gráficos o medidas de resumen. Este tipo de resúmenes se ven a menudo en informes, artículos periodísticos y estudios investigativos; por eso es muy importante comprender cómo se elaboran y cómo interpretarlos.

1.1. VARIABLES CUALITATIVAS

1.1.1. TABLAS DE FRECUENCIA

El medio más simple para resumir un conjunto de observaciones es una tabla; el tipo de tabla que se utiliza para resumir datos se denomina distribución de frecuencias, que muestra la cantidad de elementos en cada una de varias clases que son por lo general mutuamente excluyentes (cada individuo pertenece únicamente a una categoría) y exhaustivas (cada individuo debe pertenecer a cualquiera de ellas); por ello, si es necesario debe agregarse alguna clase llamada "otros" o "no responde" o algo similar. El objetivo de tal distribución es proporcionar una perspectiva de los datos.

La distribución de frecuencias para una variable cualitativa está constituida por:

Frecuencia absoluta (n_i): indica cuántos elementos pertenecen a cada clase.

Frecuencia relativa (h_i): indica la proporción de elementos en cada clase, por lo tanto está dada por la razón entre la frecuencia absoluta de cada clase y el número total de observaciones. Obviamente, si se multiplica por 100 cada valor se obtiene la distribución de frecuencias porcentuales.

Si la variable cualitativa es ordinal, es decir, si sus valores tienen un orden lógico, pueden calcularse frecuencias acumuladas; cuando se mencionen las variables cuantitativas se hará más énfasis en ello.

EJEMPLO

Remítase al ejercicio general, que se encuentra después de las nociones preliminares. Construya la distribución de frecuencias de la variable "Ciudad sede".

Solución:

PROGRAMA	FRECUENCIA ABSOLUTA	FRECUENCIA RELATIVA	PORCENTAJE
Bogotá	28	28/50 (0,56)	56%
Cali	12	12/50 (0,24)	24%
Medellín	10	10/50 (0,2)	24%
	50	1	100%

Un caso en que las clases no son excluyentes y exhaustivas se da cuando una pregunta tiene múltiples respuestas posibles. En ese caso la suma de las frecuencias puede exceder el 100%.

1.1.2. GRÁFICOS

Una segunda manera de resumir y presentar información consiste en la utilización de gráficos; ellos deben diseñarse de tal forma que comuniquen de manera elemental los patrones generales de un conjunto de observaciones, para que puedan percibirse los hechos esenciales y sea fácil compararlos con otros.

Cada vez es más habitual el uso de gráficos o imágenes para representar la información obtenida. No obstante, hay que ser prudentes al elaborar o interpretar gráficos, puesto que una misma información puede representarse de formas muy diversas y no todas ellas son pertinentes, correctas o válidas.

El tipo de gráfico debe coincidir con el tipo de información o el objetivo que se persigue al representarla, de otra manera la representación gráfica se convierte en un instrumento ineficaz, que produce más confusión que otra cosa y podría ser innecesario o productor de interpretaciones equivocadas.

HAY QUE TENER EN CUENTA QUE LOS GRÁFICOS ESTADÍSTICOS SE UTILIZAN POR CONCISIÓN Y FACILIDAD DE INTERPRETACIÓN; SI ESTO NO ES ASÍ, ES MEJOR NO EMPLEARLOS.

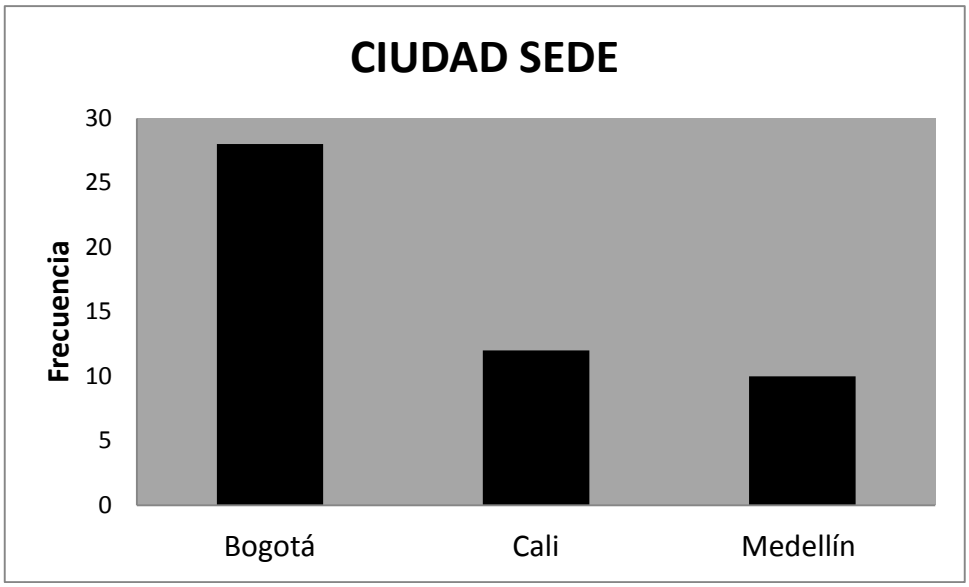
Si la variable que se está analizando es cualitativa se utilizan principalmente los siguientes gráficos:

- **Diagrama de columnas:** En el eje horizontal de un diagrama de columnas se especifican los indicadores o nombres de cada clase y en el eje vertical se representa una escala de frecuencias, bien sea absoluta, relativa o porcentual. Posteriormente, con una barra de un ancho fijo trazada sobre cada indicador de clase se llega a la altura correspondiente a la frecuencia respectiva; las barras se separan para señalar que cada clase es una categoría independiente.

  El eje que represente las frecuencias de las observaciones (generalmente el vertical) debe comenzar en cero (o), de otra manera podría dar impresiones erróneas al comparar la altura, longitud o posición de las barras.

  Las longitudes de los espacios entre las barras que representan cada clase en la gráfica deben ser iguales.

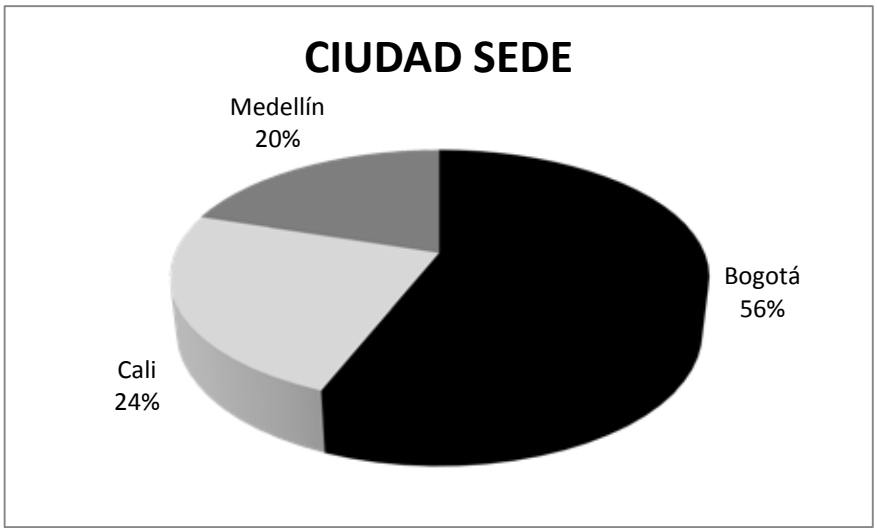
  Si quiere destacarse alguna clase, la barra respectiva debe diferenciarse por el color, mas no por su grosor.



  Las barras también pueden trazarse horizontalmente, pero la filosofía de la gráfica es la misma.

- **Diagrama circular o de sectores**: Es más apropiado para representar distribuciones de frecuencias relativas. Para trazarlo, se divide un círculo en tantas porciones como clases existan, de modo que a cada clase le corresponda un arco de círculo proporcional a su frecuencia relativa.

Para el ejemplo que se viene manejando, el gráfico de sectores quedaría así:



Existen igualmente otros tipos de gráficas, pero la mayoría se fundamentan en las mismas bases de los diagramas reseñados.

Nota: cuando hay muchas clases, deben reunirse las que tienen frecuencias menores en un solo grupo que se denomina "otros"; esto es conveniente, tanto para el gráfico de barras como para el de sectores.

En el Anexo 2 se encuentran las instrucciones para construir estos gráficos utilizando Excel.

1.1.3. ESTADÍSTICOS DE RESUMEN

Cuando la variable que se evalúa es cualitativa, el único estadístico de resumen útil es la moda, que es la clase que tiene mayor frecuencia; en el ejemplo bajo estudio, la moda es Bogotá porque es la clase que más se repite.

No puede utilizarse ninguna otra medida porque todas implican operaciones aritméticas.

1.2. VARIABLES CUANTITATIVAS

1.2.1. TABLAS DE FRECUENCIA

Si la variable que se evalúa es cuantitativa se hallan –además de las anteriores– las frecuencias acumuladas, que son las sumas de las frecuencias (absolutas o rela-

tivas) de cada clase con las de las clases anteriores; se representan como N_i (frecuencia absoluta acumulada del dato *i*) y H_i (relativa acumulada del dato *i*); se encuentran con el fin de mostrar la cantidad de elementos menores o iguales al límite superior de cada clase.

Si la variable es discreta y tiene pocos resultados posibles, debe trabajarse como las cualitativas.

Cuando la variable es continua o es discreta con muchos resultados posibles se debe hacer un agrupamiento de los datos. Para ello debe emplearse cierto criterio, de tal forma que pueda desarrollarse un diagrama razonable (Montgomery y Runger, 1996, p. 8). En general, se recomienda usar entre 3 y 12 clases de igual ancho, según el tamaño de la muestra (a mayor tamaño de muestra, más clases); el número de clases debe ser suficiente para mostrar la variabilidad en los datos, pero no deben ser muchas porque no se cumpliría el objetivo de resumir la información –ya que no se tendrían grandes ventajas en comparación a los datos sin procesar– ni muy pocas porque se perdería gran cantidad de información.

Como regla general se recomienda utilizar el mismo ancho para todas las clases (salvo casos excepcionales), con el fin de facilitar la comparación entre clases y reducir la probabilidad de una interpretación errónea. Para determinar un ancho aproximado de clase se empieza por identificar el rango del conjunto de datos –dado por la diferencia entre el valor máximo y el mínimo– y se divide por la cantidad de clases que se quiera construir.

Notas:

- También suele calcularse la marca de clase, que es el promedio entre los límites de cada intervalo y es, por lo tanto, un valor representativo de todo el conjunto.

- No hay distribución óptima de frecuencias para determinado conjunto, ya que distintas personas pueden formar distribuciones diferentes, aunque todas ellas sean igualmente correctas; el objetivo es mostrar el agrupamiento natural y la variabilidad en los datos. Existen algunas fórmulas para determinar el número apropiado de intervalos, pero el analista puede seleccionar ese número de manera subjetiva.

- Es necesario aclarar si los límites de los intervalos son abiertos o cerrados, de manera que quede claro a cuál intervalo pertenece un valor igual a uno de los límites. No es conveniente crear espacios entre cada intervalo y el siguiente si la variable es continua.

- Si hay datos muy extremos se acostumbra utilizar clases abiertas, es decir, con un solo límite ("*x* o más", "*y* o menos").

- Al hacer la distribución de frecuencias se gana cierto tipo de información pero se pierde otra. Por ejemplo, si hay cierta tendencia en el tiempo, esta información se pierde en el resumen.

EJEMPLO

Remítase al ejercicio general. Construya la distribución de frecuencias de la variable "Cantidad de productos vendidos" sin diferenciar la ciudad sede.

Solución:
Teniendo en cuenta que el valor mínimo es 56 y el máximo es 356, se establecerán cinco grupos con un ancho de 60 (eso es subjetivo, simplemente basados en la lógica)

$h_1 = 6/50 \quad h_2 = 11/50 \ldots$

$N_1 = 6 \quad N_2 = 6+11 \quad N_3 = 17+20 \ldots$

Marca de clase 1: (56+116)/2 = 86

En resumen, la distribución de frecuencias quedaría así:

Productos vendidos	n_i	h_i	N_i	H_i	*Marca clase*
[56 - 116]	6	0,12	6	0,12	86
(116 - 176]	11	0,22	17	0,34	146
(176 - 236]	20	0,4	37	0,74	206
(236 - 296]	4	0,08	41	0,82	266
(296 - 356]	9	0,18	50	1	326

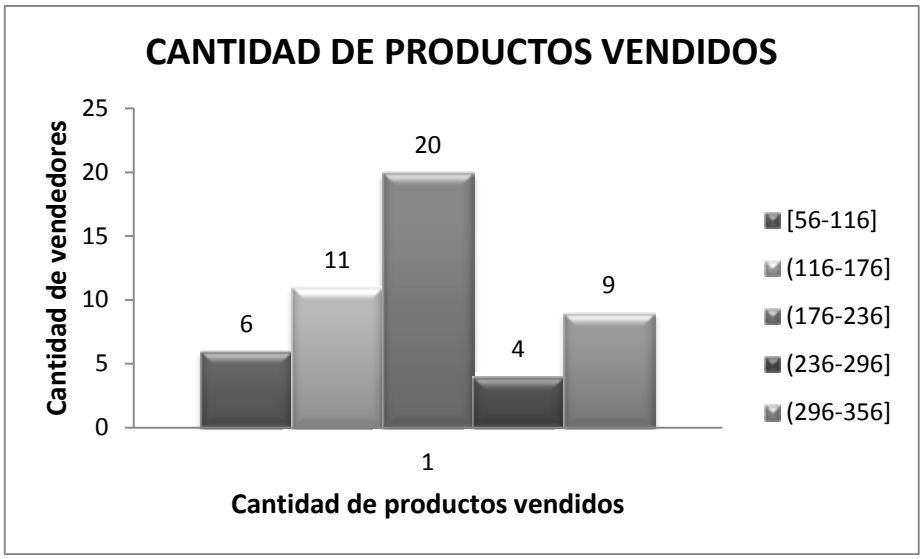
1.2.2. GRÁFICOS:

Cuando la variable es continua o es discreta con muchos resultados posibles, las gráficas utilizadas son:

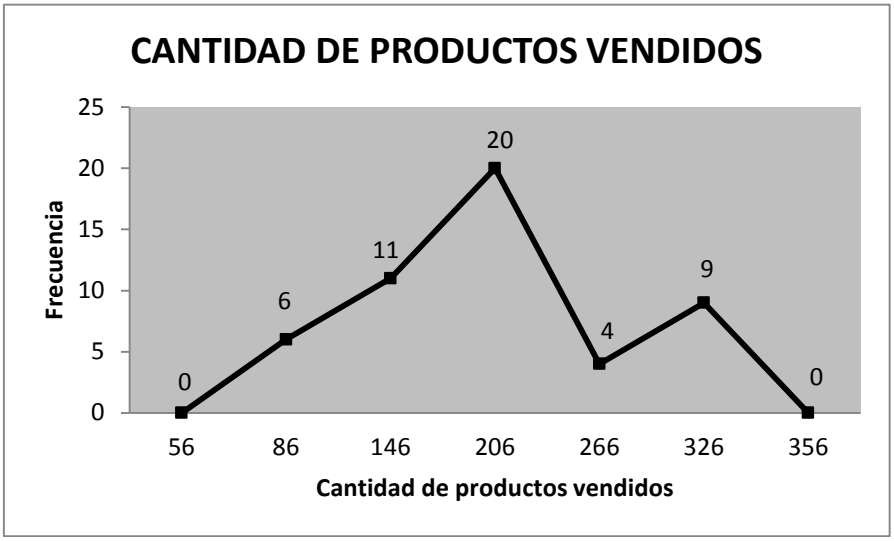
- **Histograma**, en el cual la variable de interés es colocada en el eje horizontal y la frecuencia de cada clase en el eje vertical. Posteriormente se traza una barra cuya base es el intervalo de clase sobre el eje horizontal y cuya altura

es la frecuencia correspondiente. En este caso, no hay discontinuidad natural entre las clases y, por lo tanto, esas barras van unidas. Al diseñar las escalas de los ejes, es importante que el eje vertical comience en cero; de no ser así pueden distorsionarse las comparaciones visuales entre los intervalos.

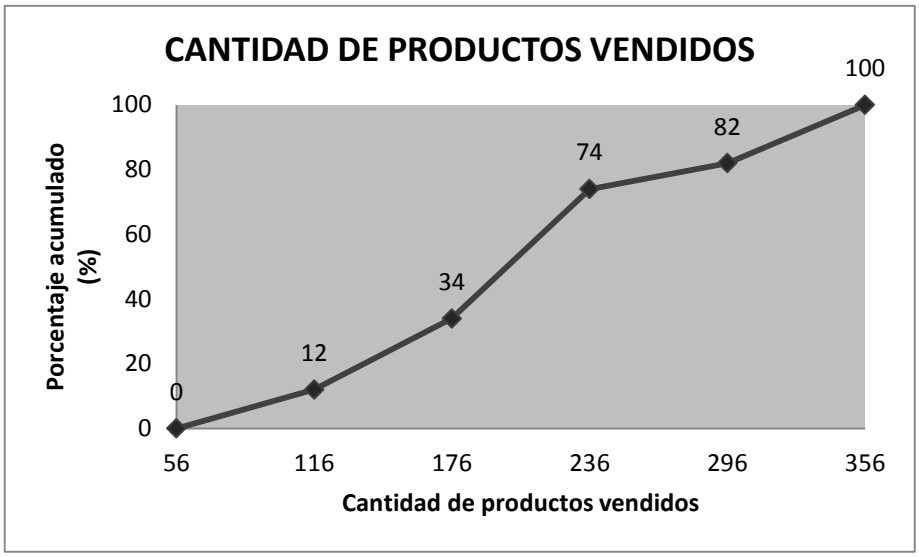
Para el ejemplo que se viene manejando, el histograma muestra la siguiente forma:



- **Polígono de frecuencias:** consta de segmentos de línea que conectan los puntos formados por la intersección de cada marca de clase y la frecuencia de la respectiva clase, es decir, puede obtenerse al unir los puntos medios del extremo superior de las barras del histograma. La escala en el eje x corresponde a los puntos medios de cada clase y la escala en el eje y corresponde a las frecuencias de clase. El polígono de frecuencias tiene la gran ventaja de que permite comparar directamente dos o más distribuciones de frecuencia.



- **Polígono de frecuencias acumuladas (ojiva):** los límites de cada intervalo van en el eje x y las frecuencias acumuladas se muestran en el eje y. La ojiva comienza en el límite inferior de la primera clase, se traza otro punto en el cruce del límite superior y la frecuencia de esa clase, y así sucesivamente para cada límite superior y la frecuencia acumulada correspondiente. Finalmente, los puntos graficados se unen con rectas y el resultado es la ojiva.



Esta gráfica se presenta con el porcentaje, pero puede hacerse con cualquier frecuencia acumulada.

1.2.3. ESTADÍSTICOS DE RESUMEN

La tabla de frecuencias y la representación gráfica nos permiten tener una idea global de la distribución de los datos; pero esta idea es netamente cualitativa y también es de interés un resumen cuantitativo, de tal manera que sirva para estudios o conclusiones posteriores y para comparación con otras distribuciones. Su objetivo es sintetizar las características dominantes del conjunto de datos mediante el cálculo de unas cifras que sean representativas de la muestra o de la población.

Los estadísticos de resumen pueden ser muy engañosos cuando se mezclan distintas poblaciones (o poblaciones con subgrupos muy marcados), por lo que debe tenerse mucho cuidado. Si hay más de una variable, los estadísticos de resumen deben calcularse por separado para cada variable.

A continuación se presentarán las principales medidas de localización y dispersión. Las medidas de localización se dividen en medidas de tendencia central (media, mediana y moda) y medidas de posición (percentiles y cuartiles); las principales medidas de dispersión son rango, varianza, desviación estándar y coeficiente de variación.

» **MEDIDAS DE LOCALIZACIÓN**: valores que contribuyen a la ubicación de un valor determinado en un conjunto de datos. Entre ellos están:

MEDIDAS DE TENDENCIA CENTRAL:

- **Media aritmética (Promedio):** es la medida de localización central más empleada. Se obtiene sumando todos los valores de los datos y dividiendo el resultado entre el total de mediciones. Si los datos proceden de una muestra, el promedio se representa con $\overline{X}$ y si proceden de una población se denomina μ.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

La forma de cálculo anterior se denomina promedio simple, pero cuando algunos datos tienen más importancia que otros se debe ponderar, así:

$$\mu = \frac{\sum_{i=1}^{n} x_i n_i}{n} = \sum_{i=1}^{n} x_i h_i$$ (donde $h_i = {}^{n_i}/_{n}$; también aplicable para $\overline{X}$)

EJEMPLO

Calcule la nota definitiva de un estudiante que tiene las siguientes notas:

Seguimiento (30%): 3,48
Examen parcial (20%): 3,8
Trabajo aplicativo (20%): 4,5
Examen final (30%): 2,5

Solución:

$$\mu = 3{,}48 \times 0{,}3 + 3{,}8 \times 0{,}2 + 4{,}5 \times 0{,}2 + 2{,}5 \times 0{,}3 = 3{,}45$$

La media ponderada también puede utilizarse cuando se tienen datos continuos agrupados y no se conocen los datos individuales; en dicho caso se usa cada marca de clase como x_i. Pero si se conocen datos individuales, es mejor calcular el promedio simple, aunque los datos sean muchos.

EJEMPLO

Calcule la media ponderada, según la distribución de frecuencias elaborada anteriormente.

Solución:

$$\mu = \frac{86 \times 6 + 146 \times 11 + 206 \times 20 + 266 \times 4 + 326 \times 9}{50} = 204{,}8$$

La media aritmética tiene varias propiedades importantes:

1. La media actúa como centro de gravedad, ya que equilibra desviaciones positivas y negativas de los datos. Una desviación es la distancia entre cada dato y la media, es decir,
2. Es muy sensible a valores extremos.
3. Para calcularse se tienen en cuenta todos los valores.
4. Es única, lo que significa que un conjunto de datos tiene solamente una media.
5. Si debe ponderarse, la media se ve muy afectada por los valores con mayor peso (tienden a arrastrar a los demás).

   Nota: Aunque una variable sea discreta, puede tener media (y, en general, cualquier medida de resumen) que no sea un valor entero.

- **Mediana** ($\tilde{x}$–muestral– o $\tilde{\mu}$ –poblacional–): es el valor que supera a la mitad de los datos y es superado por la otra mitad (valor central). Para calcularla, se deben ordenar los valores de medida; la mediana es el valor que se encuentra en el centro, pero si n es par, la mediana es el promedio de los dos valores centrales.

  Si unos pocos valores son extremadamente grandes o extremadamente pequeños, la media aritmética puede no ser un promedio apropiado para representar los datos. Por el contrario, la mediana no se ve afectada por valores extremos y por eso es preferida en casos donde existen valores de este tipo;

se dice que la mediana es robusta, lo que significa que es mucho menos sensible a los datos atípicos.

Cada una (media y mediana) tiene ventajas y desventajas según los datos y el objetivo perseguido; ambas medidas diferirán mucho cuando la distribución no es muy simétrica, lo que sugiere heterogeneidad en los datos.

- **Moda** (Mo): es el valor de los datos que tiene mayor frecuencia. Un conjunto de datos puede ser multimodal, unimodal o carecer de moda.

MEDIDAS DE POSICIÓN:

Determinan la ubicación de los valores que dividen un conjunto de datos en partes iguales.

- **Percentiles**: "el percentil p es un valor tal que por lo menos p% de las observaciones son menores o iguales que este valor y por lo menos (100 – p)% son mayores o iguales que este valor" (Anderson et al., 2008, p. 86). Por ejemplo, el valor de un percentil 80 es superado por el 20% de los datos y es igual o superior al 80% de ellos.

  Un percentil puede calcularse, de manera aproximada, al hallar un índice *i* representado por $i = p \times n/100$, donde *p* es el percentil de interés y *n* es la cantidad de elementos.

  Si *i* no es entero, el valor entero inmediato mayor que *i* indica la posición del percentil *p* (después de ordenar los datos de manera ascendente). Si *i* es entero, el percentil *p* es el promedio de los valores de los datos ubicados en los lugares *i* e *i+1*.

  Para hacer el cálculo exactamente, lo más apropiado es recurrir a un software. En el anexo 2 se encuentran indicaciones al respecto.

EJEMPLO

Calcule el percentil 65 de los siguientes datos:

2350	2450	2550	2380	2255	2210	2390	2630	2440	2825	2420	2380

Solución:

Los datos ordenados en forma ascendente son:

2210	2255	2350	2380	2380	2390	2420	2440	2450	2550	2630	2825

$$i = \frac{65 \times 12}{100} = 7{,}8$$

Eso implica que el percentil 65 es el dato que ocupa la octava posición (2440) (con Excel se halla que ese valor es exactamente 2441,5).

Nota: cuando se tienen pocos valores no es muy apropiado hablar de porcentajes; el ejemplo se hace con pocos datos para facilidad de comprensión.

- **Cuartiles**: son simplemente percentiles específicos, resultado de dividir la distribución en cuatro partes iguales porcentualmente; por lo tanto, los pasos para calcular percentiles se pueden aplicar en forma directa para calcular cuartiles.

  Por tal razón, el cuartil 1 (Q_1) equivale al percentil 25, el cuartil 2 (Q_2) corresponde al percentil 50 (mediana, y por lo tanto nunca se calcula) y el tercer cuartil (Q_3) al percentil 75.

» **MEDIDAS DE VARIABILIDAD O DISPERSIÓN:** no siempre las medidas de localización suministran toda la información necesaria para describir adecuadamente unos datos; con las medidas de dispersión se muestra qué tan esparcidos están los datos. Dos conjuntos pueden diferir tanto en tendencia central como en variabilidad, pero pueden tener medidas de tendencia central similares y ser diferentes en términos de dispersión; es por eso que una sola clase de medida es insuficiente.

  Un valor pequeño en una medida de dispersión indica que los datos están estrechamente agrupados alrededor de la media; esto implica, entonces, que una medida de dispersión puede considerarse como medida del riesgo en los casos en que tiene sentido hablar de ello (Dispersión alta implica alto riesgo).

Se considerarán varias medidas de dispersión:

- **Rango** (R): es la medida de dispersión más sencilla pues, como ya se había dicho, es simplemente la diferencia entre los datos mayor y menor. No es recomendable debido a que, al basarse únicamente en dos de los elementos, se ve muy influenciado por valores extremos.

$$R = X_{máx} - X_{mín}$$

La forma más natural de medir dispersión toma la media como punto de referencia. Aparentemente lo más lógico sería calcular la diferencia de cada valor con respecto a la media (desviaciones) y promediarlos, pero la sumatoria de las desviaciones siempre es cero; por ello se han buscado otras alternativas para calcular la dispersión. Entre ellas están las medidas que se mencionan a continuación.

- **Varianza:** un método consiste en eliminar signos, empleando valores absolutos, pero esto no es lo mejor. Una segunda idea consiste en eliminar signos, pero sin emplear valores absolutos, sino elevando al cuadrado; de esa idea se deriva la varianza, que es el dato promedio de las desviaciones cuadráticas respecto a la media. La varianza de una población se representa como σ^2 y la varianza muestral se representa con S^2 ó σ_{n-1}^2.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N} \qquad \sigma_{n-1}^{\ 2} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}$$

- **Desviación típica o estándar:** como las unidades de la varianza son cuadradas, es difícil formarse una idea intuitiva e interpretar su valor numérico. Por ello es más conveniente usar su raíz cuadrada positiva para medir variabilidad, lo cual se conoce como desviación estándar. La desviación estándar poblacional se representa como σ y la desviación estándar muestral se representa como σ_{n-1} ó S.

  Al comparar dos grupos de datos, el grupo con la menor desviación estándar tiene las observaciones más homogéneas. Sin embargo, la magnitud real de la desviación estándar depende de los valores del conjunto de datos –lo que puede ser grande para un grupo de datos puede ser pequeño para otro–; además, como la desviación estándar tiene unidades de medición, comparar las desviaciones típicas de dos cantidades no relacionadas carece de significado.

- **Coeficiente de variación (C_v):** es una medida relativa de variabilidad, que evalúa qué tan grande es la desviación estándar en relación con la media.

$$C_v = \text{Desviación estándar} / \text{Media} \times 100$$

  Como la desviación estándar y la media tienen las mismas unidades de medición, se cancelan y el coeficiente de variación es adimensional. Si se multiplica por 100 queda expresado en porcentaje.

  Si se evalúa una sola muestra: Si C_v es pequeño (más o menos inferior a 10%), puede afirmarse que dicha muestra es muy homogénea; si es grande (superior al 30%) la muestra es muy heterogénea.

Es particularmente útil al comparar la variabilidad de dos o más grupos de datos que se expresan en diferentes unidades de medida o si la media es muy distinta (en estos casos las desviaciones estándar no son comparables).

EJEMPLO

Calcular los estadísticos de resumen de la cantidad de productos vendidos por los representantes de cada ciudad y establecer comparaciones:

Solución:

Bogotá:

$$\mu = \frac{215+186+156+.....+300}{28} = 202{,}1 \ \ productos$$

$\tilde{X}$ = 194,5 (correspondiente al promedio de los datos 14º y 15º des pués de ordenarlos)

Modas: 156 y 165 (Dos personas vendieron cada una de esas cantidades)

Cuartil 1 (Q_1) = 162,8 (28 x 25% = 7 –por lo tanto es un dato intermedio entre los datos séptimo y octavo–. Se calcula exactamente en un computador)

Cuartil 3 (Q_3) = 238,8 (28 x 75% = 21 –por lo tanto es un dato intermedio entre los que ocupan las posiciones 21 y 22–. Se calcula de manera exacta en un computador)

Rango = 356 - 76 = 280 (diferencia entre máximo y mínimo)

$$\sigma^2 = \frac{(215-202{,}1)^2 + \cdots + (300-202{,}1)^2}{28} = 4759{,}3 \ \text{productos}$$

$$\sigma = \sqrt{4759{,}3} = 69 \ \text{productos}$$

$$CV = \frac{69}{202{,}1} \times 100 = 34{,}1\%$$

Por ser el ejemplo, se explica la manera de calcular manualmente estas medidas; lo más acertado es utilizar un software estadístico o por lo menos las funciones estadísticas de una calculadora.

Los estadísticos de resumen tienen como unidad de medida la de los datos originales (en este caso, productos), excepto la varianza (cuya unidad de medida es el cuadrado de la original) y el coeficiente de variación (dado en porcentaje).

Debe tenerse en cuenta que se está manejando una población, ya que dice que esos son todos los vendedores de la empresa.

Trabajando de igual forma se obtienen las siguientes medidas para Cali y Medellín (también se incluye Bogotá para comparar).

	BOGOTÁ	CALI	MEDELLÍN
Media	202,1	163,5	239,9
Mediana	194,5	163,5	237,5
Moda	156	No hay	No hay
Cuartil 1	162,8	116,3	187,8
Cuartil 3	238,8	198,3	291,3
Desviación estándar	69	66,4	65,5
Varianza	4759,3	4038,8	4292,1
Coef. de variación (%)	34,1	38,9	27,3
Rango	280	256	217
Mínimo	76	56	133
Máximo	356	312	350
Suma	5660	1962	2399
Cantidad	28	12	10

Comparaciones:

Lo anterior implica que los representantes que más productos venden, en promedio, son los de Medellín; por el contrario, los de Cali son los que menos venden, en promedio. También puede notarse que, en cantidad, los que más vendieron fueron los representantes de Bogotá (pero son muchos) y los que menos vendieron fueron los de Cali (aún menos que los de Medellín, que son más pocos vendedores).

Además, como el coeficiente de variación de los de Cali es el mayor significa que sus ventas tienen una dispersión grande con respecto al promedio, es decir, hay unos representantes que venden mucho pero otros venden muy poco. Por el contrario, entre los vendedores de Medellín hay mayor homogeneidad, aunque los coeficientes de variación son relativamente altos en todos los casos, lo que implica que hay diferencias significativas entre los vendedores de cada ciudad.

También puede deducirse que por lo menos la cuarta parte de los vendedores de Bogotá vendió 162,8 productos o menos y la cuarta parte de ellos vendió 238,8 o más productos. Lo mismo se aplica para los vendedores de las otras ciudades, aunque no tiene mucho sentido hablar de porcentajes cuando los datos son tan pocos.

Es importante destacar que no necesariamente los que tienen una desviación absoluta mayor (desviación estándar) tienen una desviación relativa mayor (coeficiente de variación).

El mayor rango se presentó en Bogotá, ya que hubo una persona que vendió muy pocos productos en el mes (76) y otra que vendió muchos (356).

En los anexos 1 y 2 se encuentra la manera de calcular los estadísticos de resumen con la calculadora y con Excel.

1.3. TABULACIONES CRUZADAS

Si el interés principal consiste en comprender la relación entre dos variables, para resumir los datos se utiliza la tabulación cruzada; en ella, los encabezados de los márgenes izquierdo y superior definen las clases de las dos variables. Si por lo menos una de las variables es cuantitativa, deben crearse intervalos.

Si ello se quiere representar gráficamente, debe utilizarse un diagrama de barras compuestas (si por lo menos una de las variables es cualitativa) o un diagrama de dispersión (si ambas variables son cuantitativas).

Cuando los tamaños de las dos poblaciones son diferentes, puede ser conveniente utilizar las frecuencias relativas, ya que en otro caso podrían resultar engañosas. Todo depende de lo que se quiera mostrar.

EJEMPLO

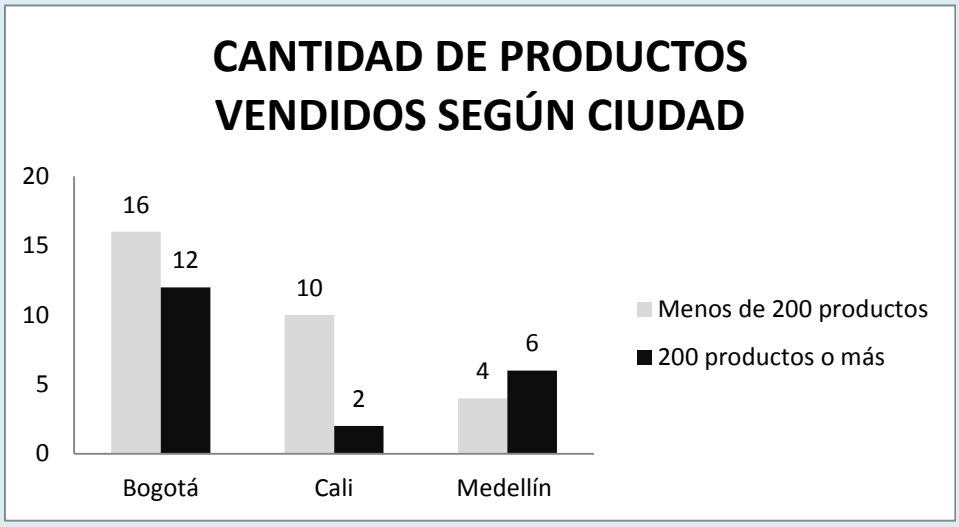
Realizar el cruce de las variables "ciudad sede" y "cantidad de productos vendidos", tomando como base de referencia las ventas inferiores a 200 productos y los iguales o mayores a él.

Solución:

CIUDAD SEDE	CANTIDAD DE PRODUCTOS VENDIDOS		Total
	Menos de 200 productos	200 productos o más	
Bogotá	16	12	28
Cali	10	2	12
Medellín	4	6	10
Total	30	20	50

Se observa, por ejemplo, que el 42,9% (12 de 28) de los representantes de Bogotá vendieron 200 productos o más, al igual que el 16,7% de los de Cali y el 60% de los de Medellín. También se concluye que el 60% de los vendedores del país (30 de 50) vendió menos de 200 productos y el 40% restante vendió 200 productos o más.

A continuación se presenta la gráfica correspondiente:



Finalmente, si las dos variables de interés son cuantitativas debe realizarse un diagrama de dispersión, que es una representación gráfica de la relación entre dos variables cuantitativas, en la cual cada par (x_i, y_i) es representado con un punto en un sistema de coordenadas bidimensional. Para representarlo, la variable independiente (si la hay) se indica en el eje x y la variable dependiente en el eje y; en el desarrollo del objeto de aprendizaje 2 se hará mayor énfasis en este gráfico.

EJERCICIOS PROPUESTOS

1. Un informe de Skyscraper Life (2011) detalla el Producto Interno Bruto PPA per cápita –en dólares– de los países latinoamericanos en 2010. A continuación se presentan los valores:

PAÍS	PIB PER CÁPITA (dólares)
Argentina	15604
Chile	14983
Uruguay	14342
México	14266
Panamá	12398
Venezuela	11889
Brasil	11289
Costa Rica	10732
Colombia	9445
Perú	9281
Rep. Dominicana	8648
Ecuador	7952
El Salvador	7442
Paraguay	4915
Guatemala	4871
Bolivia	4584
Honduras	4405
Nicaragua	2970
Haití	1122

Calcule e interprete medidas de resumen.

Se conoce que el cuartil 1 es $4893 y el cuartil 3 es $12 143,5. Interprete estos valores en contexto y compare el caso colombiano con el resto de la zona.

2. En una encuesta realizada a los habitantes de Sabaneta se preguntó lo siguiente: "¿Está usted satisfecho con la gestión de los miembros del Concejo Municipal?". De 365 personas encuestadas, 313 manifestaron estar satisfechos, 23 insatisfechos y los demás no saben o no responden.

Represente la información con una tabla de frecuencias lo más completa posible y el bosquejo de un diagrama circular.

3. Se seleccionó una muestra aleatoria de 25 municipios antioqueños y se verificó su tasa de analfabetismo (en %):

9,2	15,8	22,5	14,1	15,2	8,8	27,2	32,7	15,1	11,4	23,6	15,8	12,4
9,9	20,8	30,6	33,0	19,8	10,7	17,6	22,9	14,7	12,8	15,6	27,7	

   a. Halle e interprete media y coeficiente de variación.
   b. El cuartil 3 es 22,9%. ¿Qué indica ese valor en el contexto de los datos?
   c. Elabore una tabla de frecuencias apropiada.

4. A continuación se presentan los índices de Libertad Económica de los países suramericanos en 2011, otorgados por Skyscraper Life (2011). Dicha calificación se da a los países teniendo en cuenta la carga fiscal, la intervención del gobierno en la economía, las restricciones a las inversiones extranjeras, el peso de las regulaciones, la protección de los derechos de propiedad, el grado de corrupción, etc.

PAÍS	Chile	Uruguay	Perú	Colombia	Paraguay	Brasil
Índice lib. económica	77,4	70,0	68,6	68,0	62,3	56,3

PAÍS	Surinam	Argentina	Bolivia	Guyana	Ecuador	Venezuela
Índice lib. económica	53,1	51,7	50,0	49,4	47,1	37,6

Calcule medidas de resumen e interprételas.

5. La siguiente tabla presenta el monto de las exportaciones mensuales de Colombia (en millones de dólares) durante los últimos 24 meses (Banco de la República (2011)):

2427,6	2742,5	2738,1	2898,6	2673,5	2796,0	2948,7	2861,3
3189,6	2913,4	2876,4	3345,2	3491,0	3565,3	3058,0	3154,9
3298,9	3291,2	3553,6	3426,1	3845,5	3782,0	3947,6	4899,4

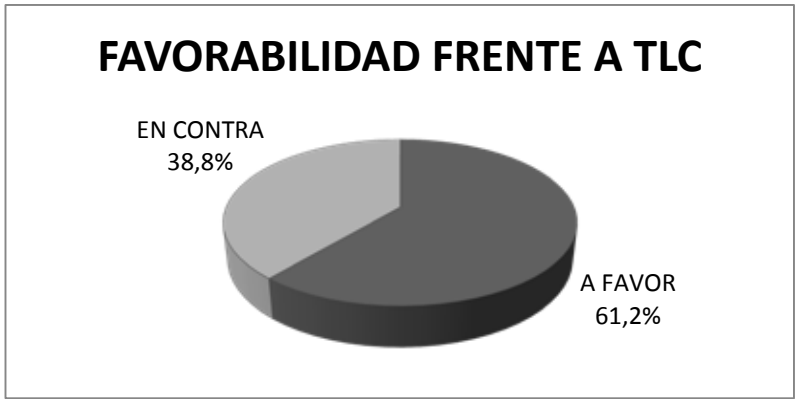
Igualmente, se presentan los datos de exportaciones mensuales (en la misma unidad) para un país vecino:

3055,3	1875,2	2635,3	4088,6	2655,6	3555,1	3237,0	4155,8
1958,7	3005,6	4321,2	3995,3	4215,2	3514,3	3305,2	3255,3
5656,2	1896,3	3236,4	4821,0	3656,3	3714,1	4902,5	4014,6

Complete la siguiente tabla y establezca comparaciones según cada medida:

	COLOMBIA	PAÍS VECINO
MEDIA (mill. de dólares)		
MEDIANA (mill. de dólares)		
COEF. DE VARIACIÓN (%)		

6. Se quiere evaluar el grado de favorabilidad de los grandes ejecutivos colombianos frente a los tratados de libre comercio. Se hizo una encuesta a 500 de ellos y se resumió la información en el siguiente gráfico:



Elabore una tabla de frecuencia a partir de esa información.

7. El siguiente ejercicio es tomado de las pruebas Ecaes 2006 (Scribd (2008)). La junta directiva de una empresa textil está considerando adquirir una compañía y se le presentan dos alternativas de compra. Para esto se examinan minuciosamente los resultados de estas dos compañías con el fin de realizar una mejor inversión. Durante los pasados 5 años la compañía A tuvo una recuperación promedio anual de lo invertido del 21% con una desviación estándar de 3,9%; la compañía B tuvo una recuperación promedio anual de lo invertido de 37,8% con una desviación estándar de 4,8%. Si se considera riesgoso invertir en una compañía que tenga una alta dispersión con respecto a la media anual de recuperación, entonces

   a. Las dos compañías han desempeñado estrategias igualmente riesgosas.
   b. La compañía B ha estado desempeñando una estrategia más riesgosa.
   c. La compañía A ha estado desempeñando una estrategia más riesgosa.
   d. Ninguna de las dos compañías tiene grandes riesgos.

8. Se tiene duda de invertir en acciones, CDT o en una cuenta de ahorros; para decidirlo se evaluaron las tasas de rentabilidad mensual promedio de cada uno durante los últimos meses y se obtuvieron los siguientes resultados:

Acciones	3,25	5,80	0,65	1,14	3,20	8,40	3,15	2,60	1,59	0,12	9,65	6,88
CDT	3,15	3,80	3,96	2,88	3,12	3,45	3,55	4,00	3,00	2,65	3,88	3,12
Cuentas	1,85	1,90	1,93	2,05	1,82	1,80	1,75	1,90	1,78	1,95	1,90	1,87

Halle media y coeficiente de variación de cada uno y con base en ello responda:

a. Si decide su inversión teniendo en cuenta únicamente la rentabilidad promedio, ¿cuál método preferiría y por qué?
b. Si quiere hacer su inversión sin correr riesgos, ¿cuál método preferiría y por qué?

9. A continuación se presentan los índices de desarrollo humano 2010 de los países de América Latina; datos aportados por el Programa de las Naciones Unidas para el Desarrollo –PNUD– (2011).

PAÍS	IDH 2010
Chile	0,783
Argentina	0,775
Uruguay	0,765
Panamá	0,755
México	0,75
Costa Rica	0,725
Perú	0,723
Brasil	0,699
Venezuela	0,696
Ecuador	0,695
Colombia	0,689
Rep. Dominicana	0,663
El Salvador	0,659
Bolivia	0,643
Paraguay	0,64
Honduras	0,604
Nicaragua	0,565
Guatemala	0,56
Haití	0,404

Realice un estudio estadístico basado en técnicas descriptivas para analizar los niveles del desarrollo humano de los países latinoamericanos. Argumente cada uno de los resultados.

10. En el siguiente cuadro puede verse el monto anual de las exportaciones colombianas de petróleo y sus derivados (en millones de dólares FOB) desde 2000 hasta 2010.

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
4775,5	3285,1	3275,2	3383,2	4227,4	5559,0	6328,3	7317,9	12212,6	10267,5	16485,1

Halle e interprete media y coeficiente de variación

11. Se va a hacer un análisis descriptivo de la tasa de inflación en 2008 en los países de América Latina. El informe se basa en datos tomados de Indexmundi.com (2010); tales datos son:

PAÍS	TASA DE INFLACIÓN 2008 (%)
Venezuela	30,4
Nicaragua	19,8
Haití	15,5
Bolivia	14
Costa Rica	13,4
Honduras	11,4
Guatemala	11,4
Rep. Dominicana	10,6
Paraguay	10,2
Panamá	8,8
Chile	8,7
Argentina	8,6
Ecuador	8,3
Uruguay	7,9
El Salvador	7,3
Colombia	7
Perú	5,8
Brasil	5,7
México	5,1
Cuba	3,4

a. Calcule e interpreta media, mediana y coeficiente de variación.

b. Construya una tabla de frecuencias para la variable en cuestión.

12. En un muestreo se seleccionaron 25 instituciones colombianas de educación superior que ofrecen el programa de Administración de empresas o uno de sus derivados y se registran algunos datos que corresponden a ellos, así:

IES	CARÁCTER IES	SECTOR IES	METODOLOGÍA	NÚMERO CREDITOS	PERIODICIDAD	PROMEDIO DE SUS EGRESADOS EN SABER PRO*
1	Universidad	Privada	Distancia (virtual)	146	Semestral	103,5
2	Institución Universitaria	Privada	Presencial	166	Anual	99,6
3	Institución Tecnológica	Privada	Presencial	132	Semestral	95,4
4	Institución Universitaria	Privada	Presencial	160	Semestral	105,3
5	Universidad	Oficial	Distancia (tradicional)	172	Semestral	101,2
6	Institución Tecnológica	Privada	Presencial	144	Semestral	99,7
7	Universidad	Privada	Presencial	161	Semestral	98,7
8	Universidad	Oficial	Presencial	157	Sin definir	100,2
9	Institución Universitaria	Oficial	Distancia (tradicional)	156	Semestral	107,6
10	Institución Universitaria	Privada	Distancia (virtual)	169	Bimensual	100,8
11	Institución Universitaria	Privada	Presencial	164	Anual	98,3
12	Institución Universitaria	Privada	Presencial	163	Anual	92,5
13	Universidad	Privada	Presencial	130	Semestral	103,4
14	Institución Universitaria	Privada	Distancia (tradicional)	160	Semestral	106,2
15	Universidad	Privada	Presencial	169	Semestral	98,0
16	Universidad	Privada	Presencial	156	Trimestral	102,4
17	Institución Universitaria	Privada	Presencial	165	Anual	99,8
18	Universidad	Oficial	Distancia (tradicional)	196	Semestral	103,9

19	Institución Universitaria	Privada	Presencial	166	Anual	99,6
20	Institución Tecnológica	Privada	Presencial	169	Semestral	96,8
21	Universidad	Privada	Distancia (virtual)	160	Semestral	103,5
22	Institución Universitaria	Privada	Distancia (tradicional)	173	Semestral	102,1
23	Institución Técnica	Privada	Distancia (tradicional)	141	Semestral	93,3
24	Institución Universitaria	Privada	Distancia (tradicional)	152	Semestral	100,0
25	Universidad	Privada	Presencial	170	Semestral	108,7

* Promedio de sus egresados en las pruebas Saber Pro en el último período evaluado.

a. Clasifique las variables evaluadas.

b. Construya tablas de frecuencias para una de las variables cualitativas y una de las cuantitativas.

c. Halle medidas de resumen para las variables cuantitativas.

d. Haga un cruce de variables de "Sector y "Promedio en Saber Pro". Construya un gráfico apropiado.

OBJETO DE APRENDIZAJE 2
REGRESIÓN Y CORRELACIÓN

A menudo, las decisiones gerenciales se basan en la relación entre dos o más variables. Puede emplearse un procedimiento estadístico llamado análisis de regresión para plantear una ecuación que muestre la manera como se interrelacionan dichas variables; por ejemplo, cuando se evalúa cómo se comportan algunas variables del negocio de acuerdo al comportamiento de los indicadores financieros.

La forma más simple para mostrar dicha relación es la construcción de un diagrama de dispersión, que es una gráfica en la que cada par (x_i, y_i) está representado con un punto en un sistema de coordenadas bidimensional. Este método puede ofrecer una idea base y por ello siempre es conveniente graficar los datos, pero es demasiado subjetivo y se limita exclusivamente a dos variables.

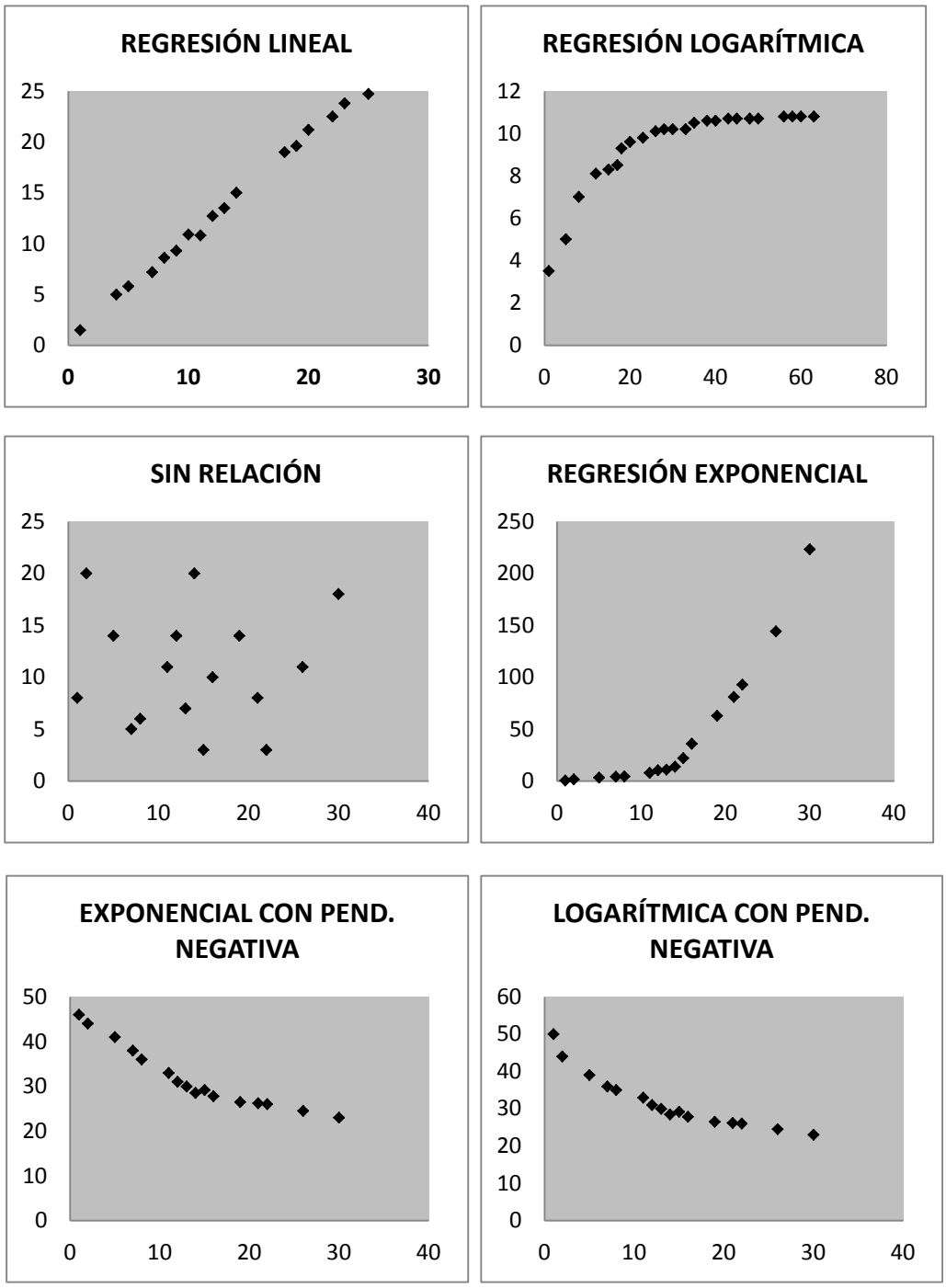
En la mayoría de aplicaciones se debe hacer una distinción entre las variables en lo que respecta a su papel en el experimento, así:

- Variable dependiente o respuesta (y): no se controla en el experimento y es el resultado producido por la acción de una variable independiente. Se representa en el eje vertical de un diagrama de dispersión.

- Variable independiente o regresora (x): se controla en el proceso y es la supuesta causa; los experimentos manipulan variables independientes para ver sus efectos sobre variables dependientes. Se representa en el eje horizontal del diagrama de dispersión.

La relación que se ajusta a un conjunto de datos experimentales se caracteriza con una ecuación, que se denomina ecuación de regresión y describe en forma razonable el comportamiento de la variable respuesta, dados los valores de las variables regresoras.

Según sea la dispersión de los datos (nube de puntos) en el plano cartesiano, pueden darse alguna de las siguientes relaciones: Lineal, Logarítmica, Exponencial, entre otras.



2.1. REGRESIÓN LINEAL SIMPLE

La regresión es muy utilizada para interpretar situaciones reales, pero comúnmente se hace de mala forma, por lo cual es necesario realizar una selección adecuada de las variables que van a construir las ecuaciones de la regresión, ya que tomar variables que no tengan relación en la práctica arrojará un modelo carente de sentido.

El caso más sencillo de una regresión simple ocurre cuando se da una relación lineal entre ambas variables. El valor estimado de *y* está dado por:

$$\hat{y} = A + Bx,$$

donde *A* es el punto de corte de la recta estimada con el eje *y* y B es la pendiente, que mide el cambio promedio en *y* cuando *x* aumenta en una unidad.

Los valores de A y B pueden calcularse con las siguientes ecuaciones:

$$B = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \qquad A = \frac{\sum y - B\sum x}{n}$$

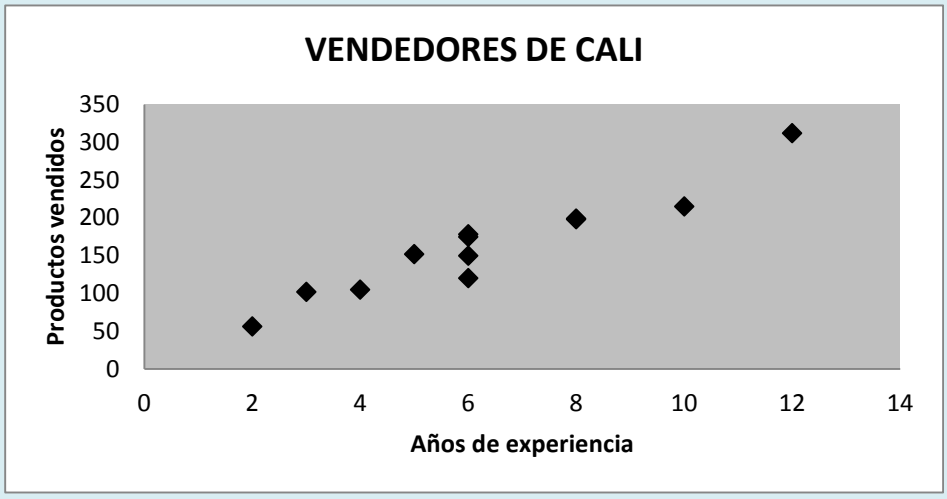
EJEMPLO

Los representantes de Cali vendieron las siguientes cantidades de productos en el mes y tienen los siguientes años de experiencia:

Años de exp.	4	6	3	8	6	6	2	5	8	12	10	6
Prod. vendidos	105	120	102	198	178	175	56	152	199	312	215	150

a. Evaluar la correlación entre ambas variables, tomando la cantidad de productos vendidos como variable dependiente.
b. ¿Cuánto se espera que venda un representante caleño con una experiencia de 7 años?

Solución:



a. Se observa que existe una relación directa entre ambas variables, ya que tienden a vender más productos aquellos representantes que tienen más años de experiencia.

$$\sum x = 76 \qquad \sum y = 1962 \qquad \sum xy = 14406 \qquad \sum x^2 = 507$$

$$B = \frac{12(14406) - (76)(1962)}{12(570) - (76)^2} = 22{,}33 \qquad A = \frac{1962 - (22{,}33)(76)}{12} = 22{,}08$$

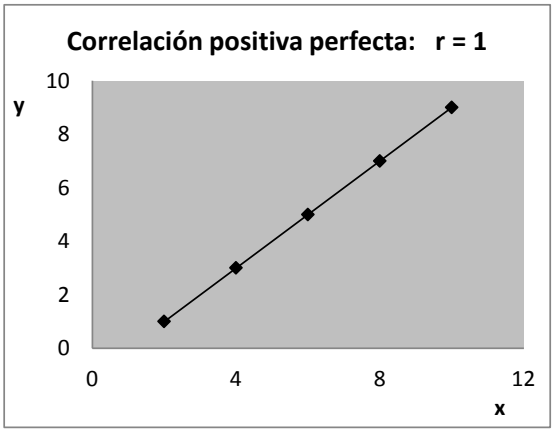
En consecuencia, $\hat{y} = 22{,}08 + 22{,}33x$, lo cual implica que por cada año de experiencia adicional, un representante de Cali vende, en promedio, 22,3 productos más (pendiente); además, se esperaría que alguien sin experiencia vendiese aproximadamente 22 productos en el mes (punto de corte).

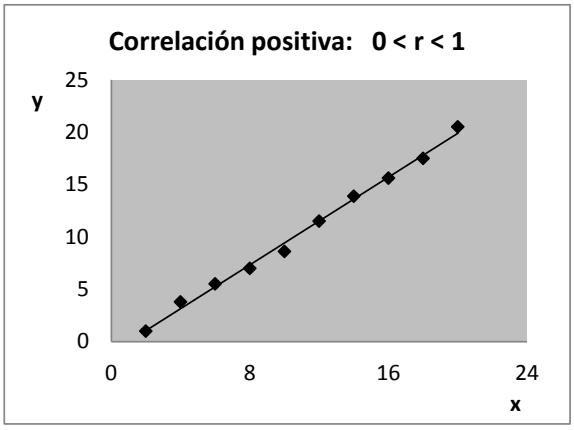
b. Para el caso específico en estudio: $\hat{y} = 22{,}08 + 22{,}33(7) = 178{,}4$
Se espera que quien tenga 7 años de experiencia venda en el mes aproximadamente 178 productos.

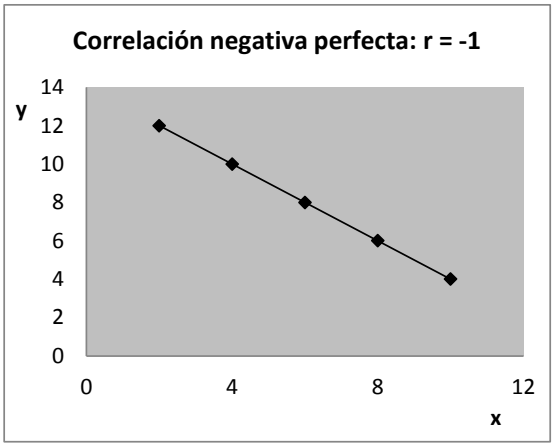
La adecuación del modelo de regresión se puede medir con el coeficiente de determinación (R^2), que expresa la proporción de la variación total en los valores de la variable *y* que se puede explicar mediante una relación lineal con los valores de *x*. El grado de solidez de la relación lineal también puede medirse con el coeficiente de correlación (r), que es la raíz cuadrada de R^2 con el signo de la pendiente; el coeficiente puede tener cualquier valor entre -1 y 1, inclusive; pero su valor absoluto debe ser cercano a 1 para poder considerar un buen ajuste (preferiblemente superior a 0,8).

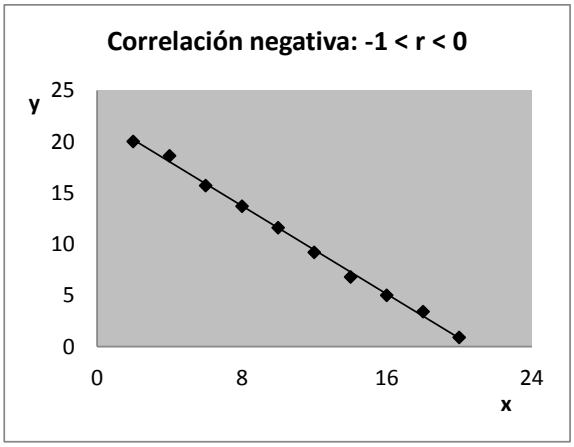
Si se tiene duda sobre varios modelos, se debe preferir aquel cuyo coeficiente de correlación es más cercano a 1 o a -1.

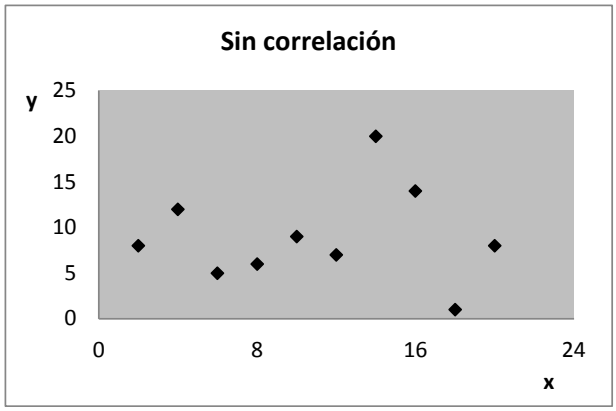
VALOR DE r	SIGNIFICADO
1	Correlación positiva perfecta
-1	Correlación negativa perfecta
Cercano a 0	Sin correlación
0 < r < 1	Relación directa entre variables (a medida que aumenta *x*, también aumenta *y*). La relación es más fuerte mientras mayor sea *r*
-1 < r < 0	Relación inversa entre variables (a medida que aumenta *x*, *y* disminuye). La relación es más fuerte mientras más cercano sea *r* a -1





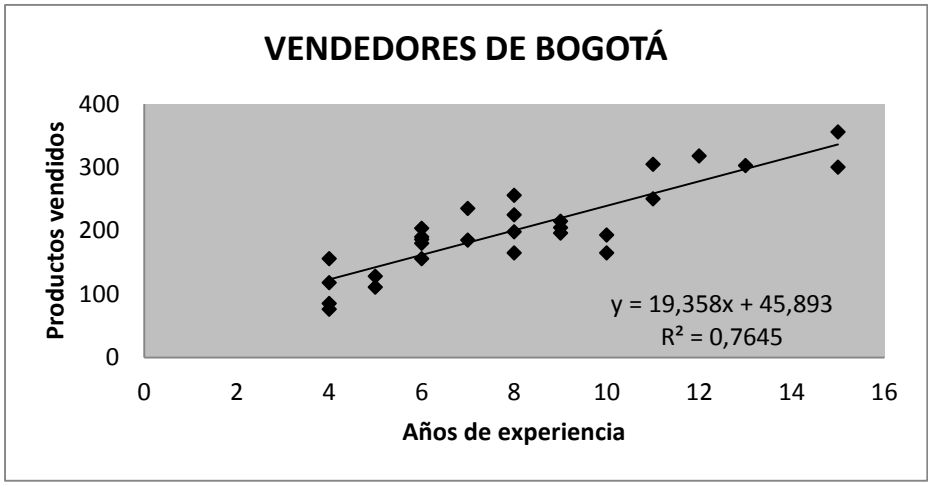


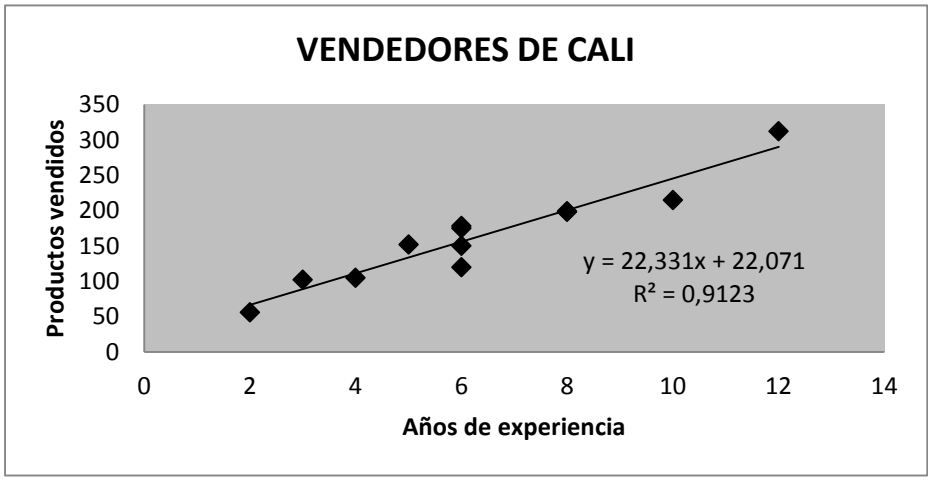


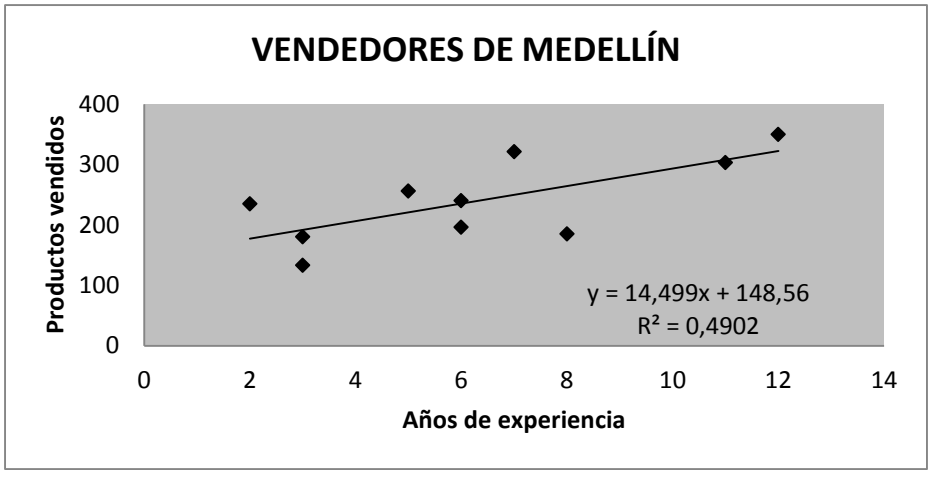


Los parámetros de regresión (A, B y r) se pueden calcular directamente con una calculadora o con un programa computacional. En los anexos 1 y 3 se explica cómo.

A continuación se presentan los diagramas de dispersión y sus correspondientes ecuaciones de regresión y coeficientes de determinación para los diferentes programas contemplados:







Puede observarse que la correlación es apropiada para Bogotá y Cali, pero no para Medellín porque el coeficiente de correlación es relativamente bajo (0,70, ya que R^2 es 0,4902).

2.2. OTROS MODELOS

Un diagrama de dispersión de los datos o un valor de R^2 no muy alto o consideraciones teóricas inherentes al estudio científico pueden sugerir la necesidad de utilizar un modelo diferente al lineal.

Se mostrarán únicamente las ecuaciones para algunos de los más utilizados:

Regresión	Forma funcional
Exponencial	$y = Ae^{Bx}$
Logarítmica	$y = A + B\ln x$
Polinomial de 2º grado	$Y = Ax^2 + BX + C$

EJEMPLO

En el siguiente ejemplo se muestra el costo por unidad de cierto producto, según la cantidad de unidades que se produzcan:

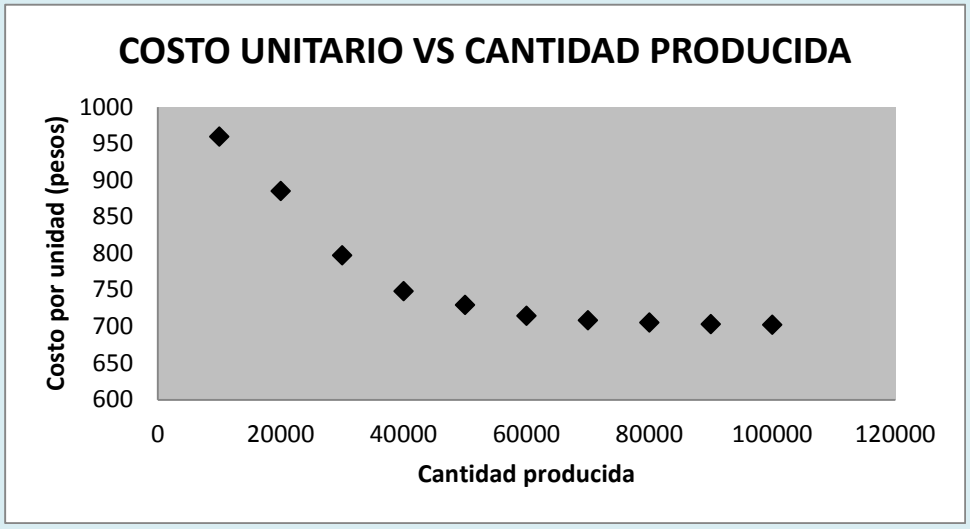
Cantidad de unidades producidas	10000	20000	30000	40000	50000
Costo unitario (pesos)	959	885	797	748	729

Cantidad de unidades producidas	60000	70000	80000	90000	100000
Costo unitario (pesos)	714	708	705	703	702

Realice una regresión apropiada para predecir el costo por unidad de acuerdo a una cantidad específica de unidades producidas.

Solución:

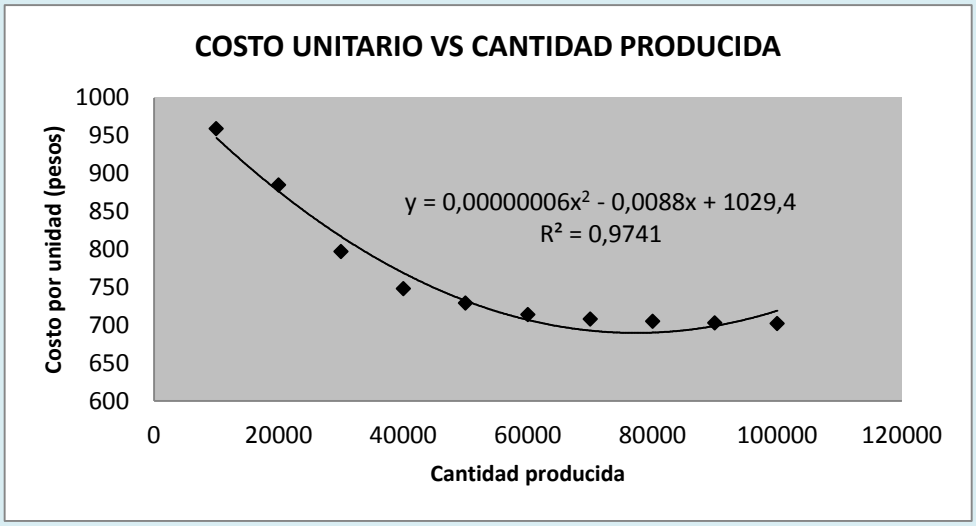
Este es el diagrama de dispersión:

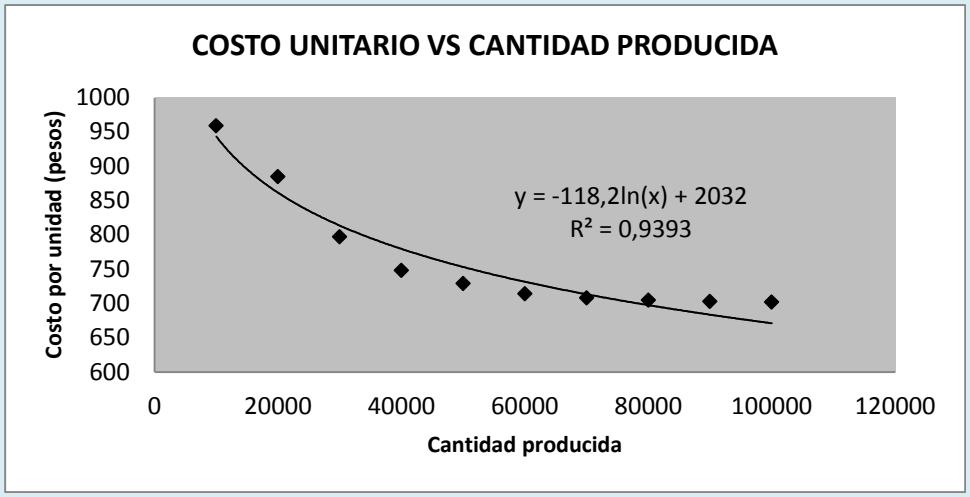


Por la forma de la nube de puntos, podrían ser apropiadas las regresiones Logarítmica, Exponencial o Polinomial de segundo grado.

Después de hacer ajustes de regresión, puede detectarse que la regresión Exponencial no es apropiada, pues el coeficiente de correlación es bajo.

Cualquiera de las dos siguientes regresiones puede resultar apropiada. El primer gráfico corresponde a una regresión Polinomial de segundo grado y el segundo corresponde a una regresión Logarítmica.





Nota: debe tenerse presente que al hacer un análisis de regresión no debe extrapolarse, es decir, no es conveniente hallar valores esperados para valores fuera del rango de los datos y, si decide hacerse, deben ser valores no muy retirados. Al hacer el análisis hay que ser muy lógicos; por ejemplo, en el caso bajo estudio, si se quiere hallar el valor esperado del costo unitario si se producen 120 000 unidades, es más apropiado utilizar la regresión Logarítmica, aunque el coeficiente de determinación sea menor, ya que el gráfico de una ecuación polinómica de segundo grado es una parábola y en este caso no se conoce el vértice.

2.3. SERIES CRONOLÓGICAS

Una serie cronológica o serie de tiempo es el resultado de observar el comportamiento de una variable a través del tiempo en intervalos regulares, que pueden ser días, meses, años, etc.

Para el análisis es importantísimo distinguir entre datos transversales y series temporales de datos; los datos transversales se recogen en el mismo (o aproximadamente el mismo) punto de tiempo, mientras que las series de tiempo son grupos de datos cuantitativos que se obtienen en períodos con intervalos regulares en el transcurso del tiempo.

Los histogramas y algunas otras gráficas estadísticas son muy útiles para mostrar la variabilidad presente en los datos; sin embargo, frecuentemente el tiempo es un factor importante que contribuye a la variabilidad dada y no debería desconocerse.

El primer paso para analizar una serie de tiempo consiste en graficarla. En la representación gráfica de las series temporales se utilizan los diagramas de líneas; en el eje de las abscisas se representa el tiempo *t* y en el eje de las ordenadas, los

valores de la magnitud observada *x(t).* Se obtiene así una serie de puntos (*t, X(t)*) que, unidos, proporcionan una visión dinámica de la evolución del fenómeno que se esté analizando.

Se trata, por tanto, de un caso particular de variable estadística bidimensional, donde la variable independiente es el tiempo (*t)* y la variable dependiente es la magnitud de interés (*y)*.

La suposición básica fundamental en el análisis de series de tiempo es que los factores que han influido en el pasado o en el presente en los patrones continuarán haciéndolo más o menos en la misma forma en el futuro. Eso significa que los datos históricos son usados para pronosticar posibles valores futuros.

Una presentación del Centro de Investigación Estadística y Mercadeo –CIEM– (2010) dice:

> El tratamiento numérico de las *series temporales* es variado y la metodología a utilizar en cada caso depende de los objetivos planteados por el analista. En general, se puede decir que de una secuencia cronológica puede interesar adquirir un conocimiento *descriptivo o diagnóstico*, en el sentido de poder detectar la dinámica generadora del fenómeno bajo estudio (si este es el objetivo, solamente se hace un análisis gráfico, que incluye gráfico y evaluación de los componentes); o un conocimiento *predictivo o pronóstico*, si se pretende deducir de los datos registrados hasta el momento, cómo será el comportamiento futuro.

Usualmente, una serie cronológica muestra una correlación entre observaciones adyacentes; a partir de eso se usan técnicas especiales para hacer modelos y análisis.

2.3.1. IMPORTANCIA DE LAS SERIES CRONOLÓGICAS

- Sirven para hacer pronósticos, ya que pueden hacerse estimados y predicciones del futuro con base en tendencias y modelos del pasado. La planeación a futuro es un aspecto primordial para cualquier administración, ya que es muy importante poder "anticipar el futuro" y desarrollar estrategias apropiadas; a partir de las proyecciones podría, por ejemplo, planificarse compras y demanda.

- Útiles para planeación y control.

- Facilitan la comparación de datos que ocurren en momentos diferentes en el tiempo.

- Pueden servir como registros históricos valiosos. Por ejemplo, para estudiar el comportamiento a lo largo del tiempo de una variable macroeconómica o el comportamiento de una acción.

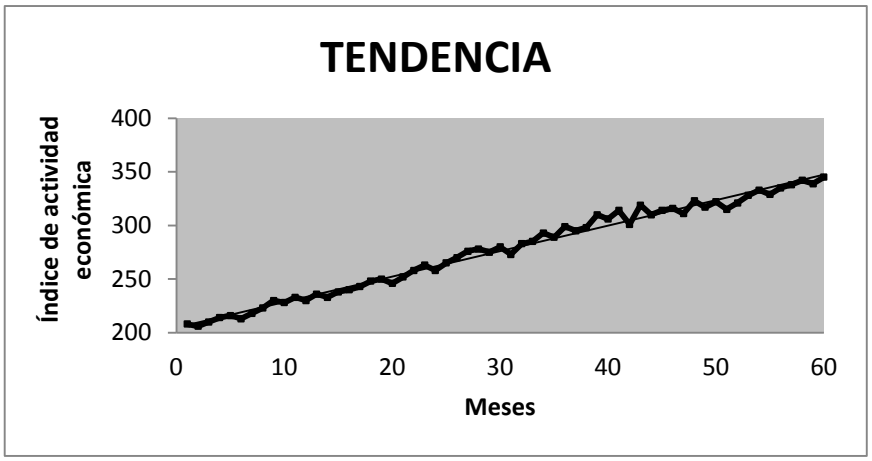
¡PIENSA!

¿Para qué más le podría servir una serie cronológica a un profesional de tu área?

2.3.2. COMPONENTES DE UNA SERIE DE TIEMPO:

- **Tendencia**: patrón de movimientos general o persistente, a largo plazo; puede ser alcista, a la baja o neutra (horizontal).

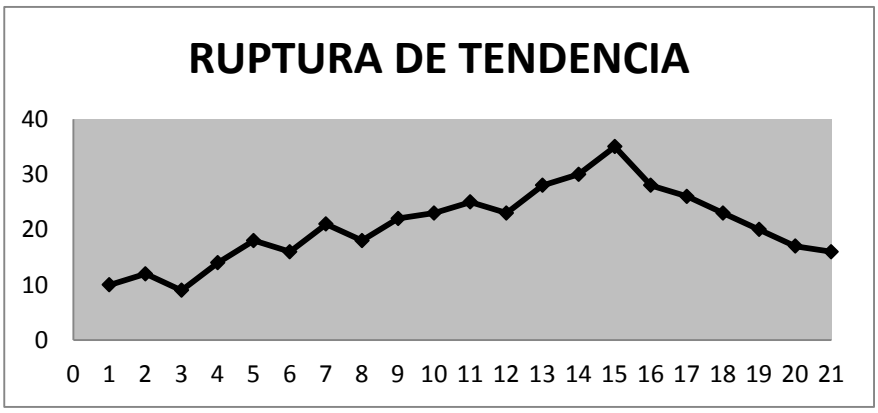
  La tendencia representa el comportamiento predominante de la serie. Esta puede ser definida vagamente como el cambio de la media a lo largo de un periodo.



  Ese desplazamiento gradual es, generalmente, el resultado de factores a largo plazo, como cambios en la población, tecnología o preferencias del consumidor.

  Para establecer una línea de tendencia que sea significativa para una serie cronológica, el período tiene que ser suficientemente largo.

  Algunas veces se da una ruptura de tendencia, que es cuando la serie trae una tendencia definida y cambia bruscamente. En ese caso, si se quiere hacer pronósticos hay que tener en cuenta únicamente el último tramo.



(Se observa una ruptura de tendencia después del dato 15)

- **Componente cíclico:** fluctuaciones a largo plazo más o menos periódicas que se repiten regularmente cada cierto período.

- **Estacionalidad:** patrón regular que se repite en un período específico. Se trata de un componente causal debido a la influencia de ciertos fenómenos que se repiten periódicamente (las estaciones, los fines de semana, las horas pico, etc); por ejemplo, ventas altas en el mes de diciembre de todos los años evaluados.

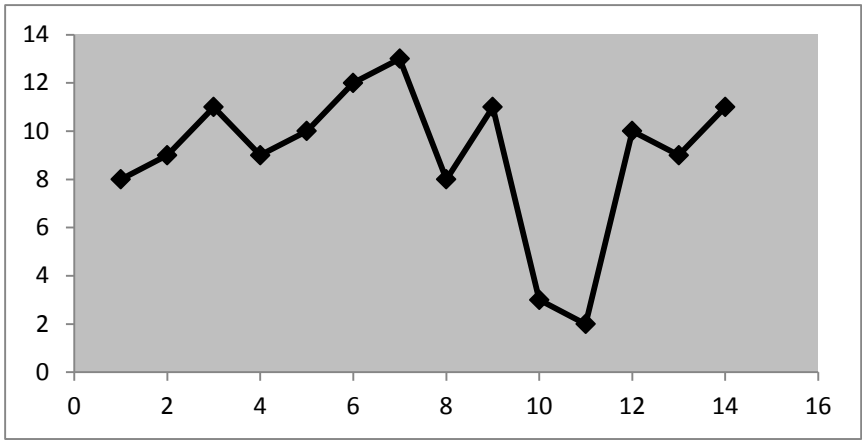
- **Componente irregular:** son movimientos de tipo causal que no muestran ninguna regularidad. Se debe a factores a corto plazo, imprevisibles y no recurrentes que afectan la serie de tiempo: huelgas, terremotos, inundaciones, problemas políticos circunstanciales, entre otros; recoge la influencia que ejercen sobre la serie circunstancias aleatorias o accidentales.

  Como no presenta un comportamiento sistemático a corto, mediano o largo plazo, no se puede predecir de ninguna forma. Si la serie que se esté estudiando no es muy larga, se recomienda la eliminación de las observaciones anómalas y se evita así la influencia de factores esporádicos en los resultados del análisis final de la serie.

El análisis gráfico de una serie temporal consiste en describir las pautas de regularidad que sigue cada uno de sus componentes, con el fin de conocer la serie para el período para el cual se tienen datos y poder predecir así su evolución futura.

Se debe detectar puntos de la serie que se escapan de lo normal (*outliers*). Un *outlier* es una observación de la serie que corresponde a un comportamiento anormal del fenómeno (sin incidencias futuras) o a un error de medición y por lo tanto se debe omitir antes de analizar la serie.

Por ejemplo, en un estudio de la producción diaria en una fábrica se presentó la siguiente situación:



Los puntos 10 y 11 parecen corresponder a un comportamiento anormal de la serie. Al investigar estos dos puntos se vio que correspondían a dos días de paro, lo que naturalmente afectó la producción en esos días; el problema, desde el punto de vista estadístico, fue solucionado eliminando las observaciones e interpolando.

2.3.3. EMPLEO DEL ANÁLISIS DE REGRESIÓN EN PRONÓSTICOS

Cuando se hacen pronósticos pueden utilizarse métodos cualitativos, es decir, basados en el juicio y la intuición. En muchos casos, estos son los únicos métodos disponibles, sobre todo en aquellos casos en los cuales no es posible conseguir datos certeros; su naturaleza no científica hace que sean difíciles de estandarizar y de validar su precisión.

Por ello, siempre que sea posible, es mucho más adecuado basarse en datos reales. En ese caso, como valores de la variable tiempo es más adecuado tener en cuenta el orden y no propiamente el período de tiempo; es recomendable tomar el primer dato temporal como o ó 1, debido a la imposibilidad, en algunos casos, de hacer regresión con los datos originales (por ejemplo si los datos son enero, febrero, marzo…) o a lo ilógico que resultaría interpretar punto de corte y pendiente (con datos como 2010, 2011…).

Pueden presentarse varios casos; de cada uno de los tres primeros se mostrarán ejemplos:

1. Si se presenta una tendencia durante todo el recorrido, se hace un análisis de regresión simple, según el caso más adecuado (lineal, exponencial, cuadrática…)

2. Si hay una ruptura de tendencia y se quiere hacer pronósticos hay que tener en cuenta únicamente el último tramo.

3. Si en la serie se detecta estacionalidad o ciclos, deben tenerse en cuenta esos aspectos al hacer un pronóstico. Por ejemplo, si se nota que en una serie cronológica existe estacionalidad en los diciembres y desea establecerse un pronóstico para un diciembre próximo, deben tenerse únicamente en cuenta los datos correspondientes a ese mes.

4. Si no hay una tendencia definida deben emplearse otros métodos para hacer proyecciones, como por ejemplo media móvil, suavizamiento exponencial, Índice de fuerza relativa, variación porcentual, métodos subjetivos, entre otros. Eso sucede, por ejemplo, con el precio del dólar y con el precio diario de las acciones; estos métodos no serán tenidos en cuenta en el desarrollo de estas notas.

   Uno de los métodos más usados es el de media móvil, que consiste en promediar los *n* datos más recientes; *n* es un número que el analista determina subjetivamente (generalmente tres o más, pero no un valor alto).

EJEMPLO

La siguiente tabla contiene datos anuales para la población colombiana (en miles de habitantes) desde 1985 hasta 2010

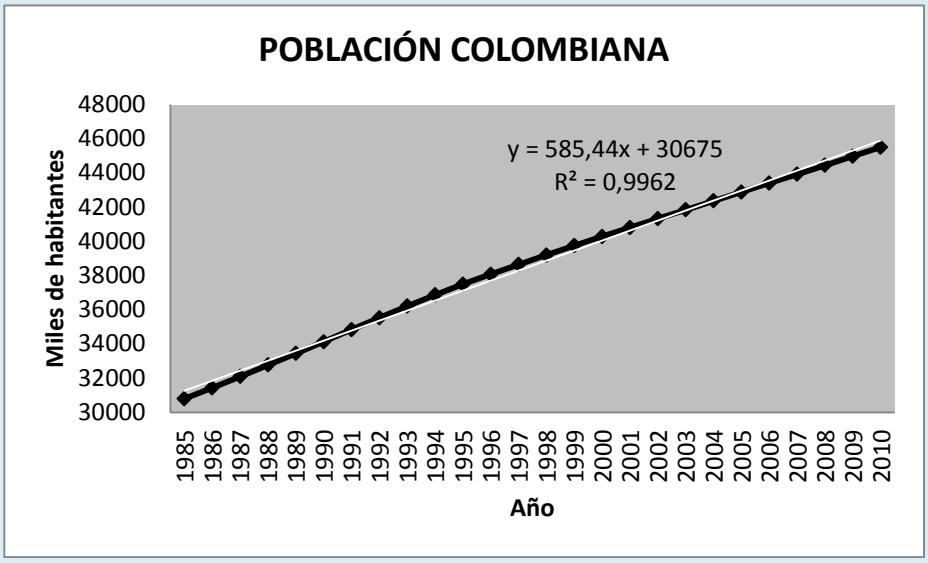
1985	1986	1987	1988	1989	1990	1991	1992	1993
30794	31433	32092	32764	33443	34125	34834	35530	36208

1994	1995	1996	1997	1998	1999	2000	2001	2002
36863	37490	38077	38646	39201	39746	40282	40806	41327

2003	2004	2005	2006	2007	2008	2009	2010
41847	42368	42889	43405	43926	44450	44978	45508

Haga un análisis de regresión y pronostique la población colombiana en 2011 y 2020.

Solución:



El análisis anterior muestra que la población colombiana ha aumentado durante los últimos años con una tendencia lineal y que el aumento corresponde a un promedio de 585.440 habitantes por año.

El pronóstico para 2011 es: $\hat{y}(27) = 585440(27) + 30675000 = 46\ 481\ 880$ hab.

El pronóstico para 2020 es: $\hat{y}(36) = 585440(36) + 30675000 = 51\ 750\ 840$ hab.

(1985 es el dato 1, 1986 es 2, 1987 es 3... 2011 es 27, 2020 es 36)

Nota: Por lo observado, podría pensarse en hacer una regresión logarítmica, dado que se nota que el tamaño de la población tiende a aumentar cada vez en menor proporción. Cualquiera de las dos situaciones es correcta.

EJEMPLO

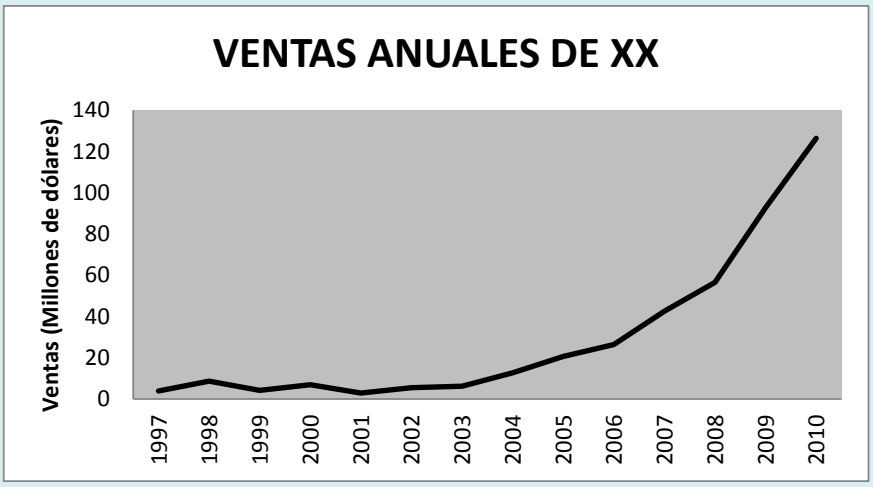
Cierta empresa desarrolla, fabrica y vende productos para acceso a redes. Los siguientes datos son las ventas anuales (en millones de dólares) de 1997 a 2010:

1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
3,8	8,6	4,1	6,9	2,8	5,4	6,2	12,7	20,6	26,4	42,6	56,5	92,8	126,4

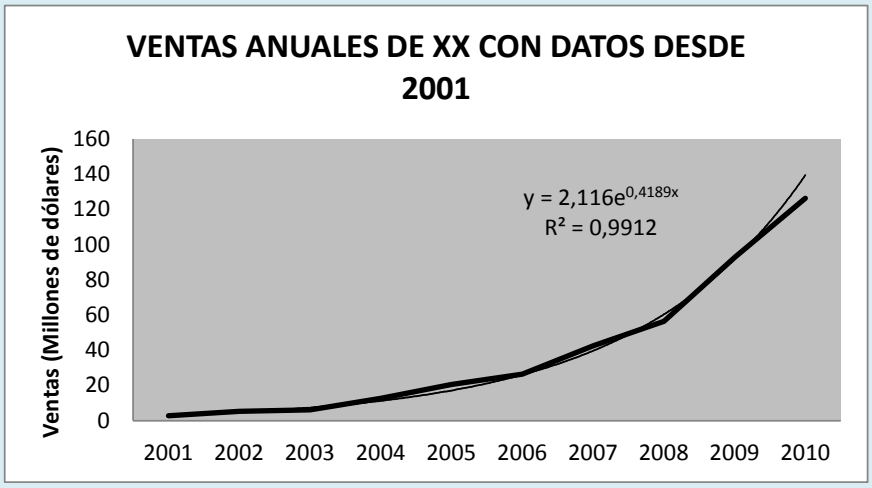
a. Deduzca una ecuación de tendencia para esta serie de tiempo.
b. Pronostique las ventas en 2011 y 2012.

Solución:

A continuación puede observarse el diagrama de líneas correspondiente:



Es fácil detectar que hasta el año 2001 no se presenta una tendencia definida, pero a partir de allí se revela una tendencia exponencial; por eso se hará nuevamente la gráfica con los datos tomados desde ese momento y se hará el análisis de regresión únicamente con esos datos:



De acuerdo a la ecuación de regresión obtenida, las proyecciones para 2011 y 2012 son:

Proyección para 2011:
$y(11) = 2{,}116\, e^{\,0{,}4189*11} = 212{,}2$ *millones de dólares*

Proyección para 2012:
$y(12) = 2{,}116\, e^{\,0{,}4189*12} = 322{,}6$ *millones de dólares*

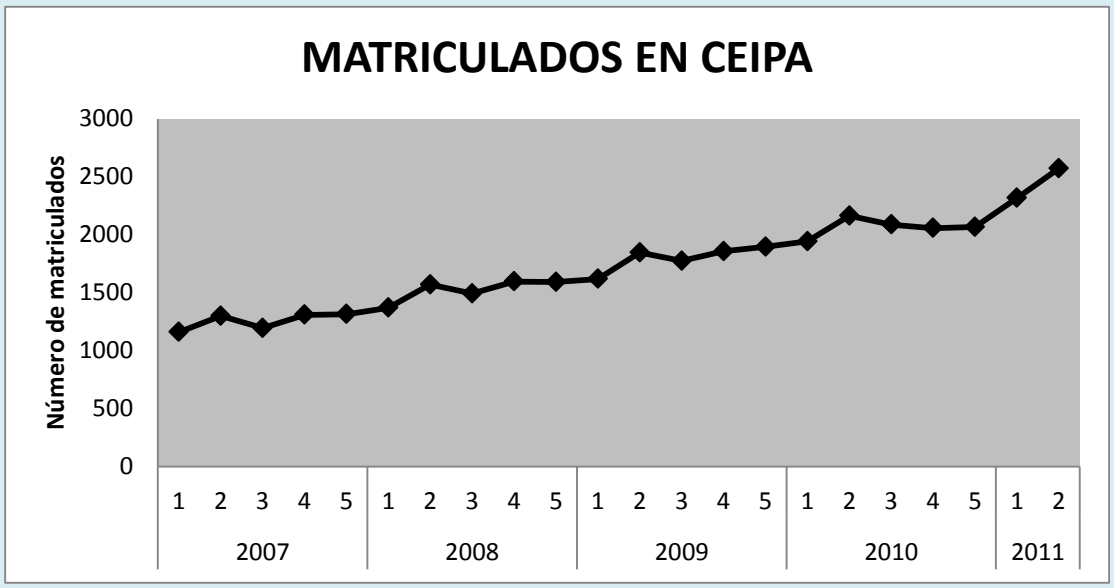
EJEMPLO

La siguiente tabla muestra el número de matriculados en Ceipa desde el período 1 de 2007 hasta el período 2 de 2011:

PERÍODO	2007	2008	2009	2010	2011
1	1160	1370	1619	1943	2317
2	1300	1570	1846	2164	2573
3	1194	1493	1774	2088	
4	1309	1597	1858	2058	
5	1314	1592	1896	2068	

a. Haga un bosquejo del gráfico. Evalúe tendencia, estacionalidad y ciclos.
b. Haga un pronóstico para el segundo período de 2012.

Solución:



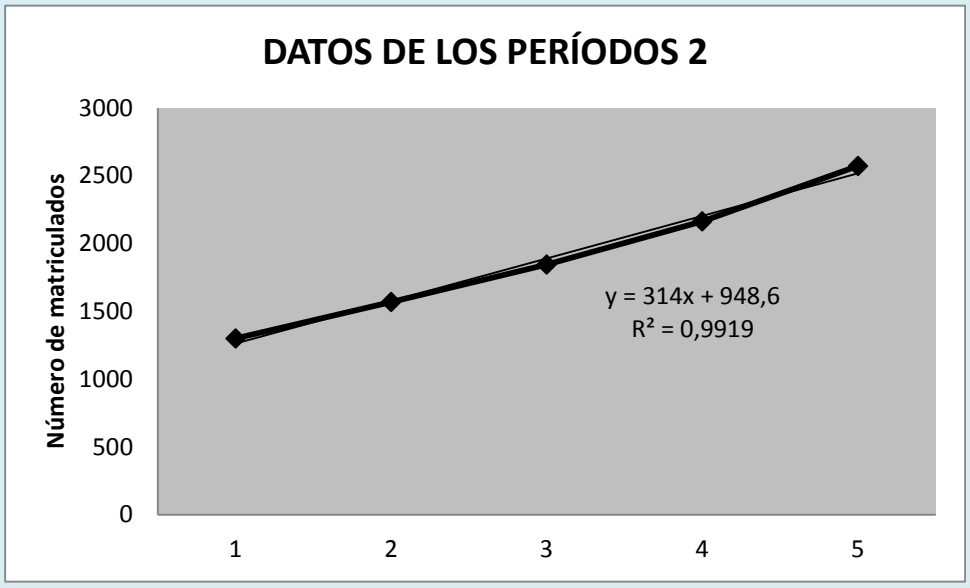
Es claro que el número de estudiantes matriculados por período académico y año ha venido creciendo, por lo que se detecta una tendencia alcista.

Igualmente, es claro que hay una estacionalidad, caracterizada por puntos máximos en los períodos 2 y mínimos en los períodos 3.

Se observan, además, ciclos anuales.

Debido a la estacionalidad y los ciclos, se tendrán en cuenta únicamente los datos de los segundos períodos de cada año. Dichos datos son:

2007	2008	2009	2010	2011
1 300	1 570	1 846	2 164	2 573



En Excel, mediante la herramienta **Pronóstico,** puede hacerse una proyección de 2833 matriculados para el segundo período de 2012; dicho cálculo también podría hacerse al reemplazar x = 6 en la ecuación de regresión.

Nota: Como se observan ciclos, podría hacerse cualquier proyección a partir de los datos del período respectivo; por ejemplo, si se quiere hacer una proyección para el período 5 de cualquier año deben utilizarse los datos de dicho período a través de los diferentes años.

La regresión múltiple (cuando se estudian una variable dependiente y varias independientes) será considerada únicamente en el anexo 3, debido a lo tedioso que resultaría calcularlo manualmente.

EJERCICIOS PROPUESTOS

1. Se pretende explicar los cambios en la variable "índice de desarrollo humano" a partir de los cambios en "Producto Interno Bruto per cápita". Para ello se tuvieron en cuenta los datos más recientes de los países de América Latina, tomados de Skyscraper Life (2011):

PAÍS	Argentina	Chile	Uruguay	México	Panamá	Venezuela	Brasil
PIB PER CÁPITA (dólares)	15604	14983	14342	14266	12398	11889	11289
IDH	0,775	0,783	0,765	0,75	0,755	0,696	0,699

PAÍS	Costa Rica	Colombia	Perú	República Dominicana	Ecuador	El Salvador	Paraguay
PIB PER CÁPITA (dólares)	10732	9445	9281	8648	7952	7442	4915
IDH	0,725	0,689	0,723	0,663	0,695	0,659	0,64

PAÍS	Guatemala	Bolivia	Honduras	Nicaragua	Haití
PIB PER CÁPITA (dólares)	4871	4584	4405	2970	1122
IDH	0,56	0,643	0,604	0,565	0,404

   a. Halle la ecuación de regresión más apropiada, si la hay.
   b. Según los datos, ¿cuál debería ser el PIB de un país para que su índice de desarrollo humano ascienda a 0,73?

2. Para proyectar las unidades demandadas de un producto en relación con el precio del producto en el mercado se construye una ecuación de regresión en la que se relaciona la cantidad de unidades demandadas (x) con el precio en el mercado (y). La ecuación obtenida es y = -x + 120. De acuerdo con esta ecuación es correcto concluir que

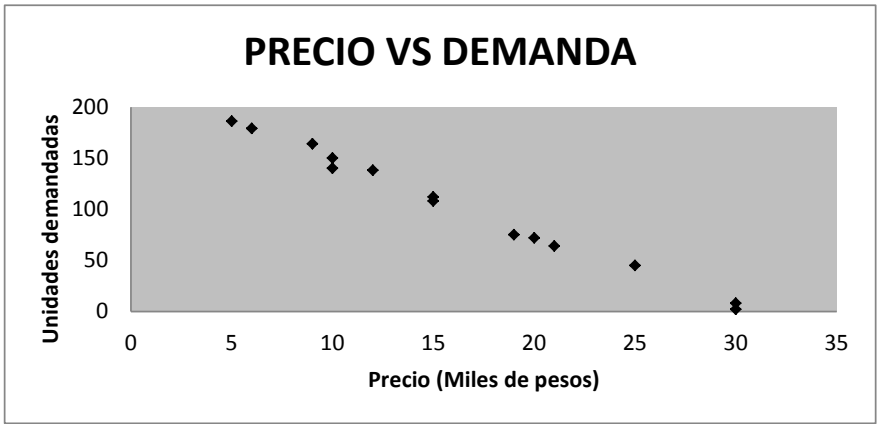
   a. entre el precio y las unidades demandadas hay una relación directa
   b. la diferencia entre el precio y las unidades demandadas es 120
   c. entre el precio y las unidades demandadas hay una relación inversa
   d. entre el precio y las unidades demandadas no existe relación alguna

3. Se desea utilizar el Índice de Producción (en porcentaje) como variable explicativa de la variable "Índice de pobreza en Colombia" (en porcentaje). Para ello se tomaron diferentes datos a través de la historia; esos datos son los siguientes:

Producción	10	20	30	40	50	60	70	80	90	100
Pobreza	75	66	46	41	32	27	26	26	25	24

a. Encuentre la ecuación de regresión más apropiada.
b. ¿Qué tipo de relación se da entre las variables?
c. Según esos datos –que no son reales-, ¿cuál debería ser el índice de producción para bajar el índice de pobreza a un 15%?

4. Para proyectar las unidades demandadas de un producto en relación con su precio en el mercado se hace un análisis de regresión en el que se relaciona la cantidad de unidades demandadas con el precio en el mercado. El siguiente gráfico representa la situación:



La ecuación que más se ajusta es:

$a)\ \hat{y} = 100 - 100x$ $b)\ \hat{y} = 100 + 100x$

$c)\ \hat{y} = 220 - 7x$ $d)\ \hat{y} = 220 + 7x$

JUSTIFIQUE SU RESPUESTA

5. Una aplicación importante del análisis de regresión es la estimación de costos. Al reunir datos sobre volumen y costo, un administrador puede estimar el costo asociado con determinada operación de manufactura, Se obtuvo la siguiente muestra de volúmenes de producción y costo total para una operación de manufactura:

Volumen de producción (unidades)	Costo total (millones de pesos)
400	8
450	8,6
550	9,6
600	9,9

700	10,5
750	10,7
830	11,1
900	11,4

a. Use una regresión lineal para deducir una ecuación de regresión con la que se pueda predecir el costo total según determinado volumen de producción.
b. ¿Cuál es el costo variable o costo adicional por unidad producida?
c. Calcule el coeficiente de determinación. ¿Qué porcentaje de la variación en el costo total puede explicar el volumen de producción?
d. ¿Es el modelo lineal el más adecuado para hacer un análisis de regresión en este caso?

6. Se quieren explicar las ventas de una empresa únicamente en función de los gastos en publicidad (ambas variables en millones de pesos). Para ello se tomó una muestra de diez tiendas y se encontró lo siguiente:

Gastos public.	300	280	390	460	320	150	80	360	140	90
Ventas	5000	4500	6200	7500	6500	2600	1800	4700	3200	2500

a. Haga un diagrama de dispersión.
b. Halle una ecuación de regresión apropiada.

7. El Gerente de Recursos Humanos de una empresa grande debe presentar un informe anual al Gerente General de la empresa en el que presente un estudio de los salarios de los empleados de oficina. Pretende utilizar el tiempo de servicio (en meses) como variable explicativa del salario (en millones de pesos). Se tomaron los siguientes datos para una muestra de diez elementos:

Tiempo	20	15	12	3	8	10	6	60	22	12
Salario	3,2	1,5	1	2,6	3,8	2,3	1,8	2	2,6	1,5

Haga un análisis de regresión. Si ningún modelo se ajusta, explique qué otras variables piensa que se deberían tener en cuenta.

8. Se desea utilizar el Producto Interno Bruto (dado en millones de dólares) como variable explicativa de la variable "Desempleo en Colombia" (en porcentaje). Se probaron diferentes tipos de regresión simple y el coeficiente de correlación más alto en valor absoluto es el lineal (-0,854). Además se sabe que A = 86,5 y B = -0,0009

a. Interprete A, B y r.
b. ¿Cuál es la tasa de desempleo que se esperaría si el Producto Interno Bruto es de 90 000 millones de dólares?

9. La mentalidad emprendedora implica perseverancia y esfuerzo, especialmente si se tienen en cuenta las estadísticas en creación y desarrollo de empresas. De acuerdo con el centro de emprendimiento "*Bogotá Emprende*" el 75% de las empresas sobreviven el primer año y únicamente el 29% sobreviven los primeros 10 años. Los datos completos pueden verse en la tabla siguiente:

Años	0	1	2	3	4	5	6	7	8	9	10
% de empresas que sobreviven	100	75	64	56	50	45	40	37	34	31	29

Halle una ecuación de regresión que se ajuste adecuadamente a los datos y proyecte el porcentaje de empresas que sobreviven al duodécimo año.

El ejercicio anterior fue extraído de Ángel, B. (2010)

10. Un supermercado registra sus ventas diariamente. He aquí los datos (en millones de pesos) para 4 semanas comprendidas entre el 21 de febrero y el 20 de marzo de 2011:

F21 (L)	F22 (M)	F23 (M)	F24 (J)	F25 (V)	F26 (S)	F27 (D)	F28 (L)
750	689	657	715	796	1595	1008	202

M1 (M)	M2 (M)	M3(J)	M4(V)	M5(S)	M6(D)	M7(L)	M8(M)	M9(M)	M10(J)
564	688	699	805	1758	1164	655	584	566	758

M11(V)	M12(S)	M13(D)	M14(L)	M15(M)	M16(M)	M17(J)	M18(V)	M19(S)	M20(D)
861	1995	1114	456	555	650	768	960	2510	1385

a. Haga un bosquejo de la gráfica y analice componentes.
b. Haga pronóstico de ventas para el domingo siguiente.

11. Se quiere analizar la relación entre la cantidad de hijos y el nivel de pobreza. Para desarrollar el análisis se hizo una regresión entre la tasa de fertilidad (medida en infantes nacidos/mujer) y la población bajo el nivel de pobreza (%); el estudio se hizo con base en todos los países del mundo que aportaron información de ambas variables a www.indexmundi.com.

Para el análisis se tomó como variable independiente la tasa de fertilidad y el mejor ajuste se logró para el modelo lineal. La ecuación de ajuste es $y = 10{,}464x + 4{,}7333$ y el coeficiente de determinación es 0,6849.

a. ¿Cuáles son los valores de A y B? ¿Qué significan en el contexto?
b. Discuta sobre la conveniencia del uso de la tasa de fertilidad como variable independiente.
c. ¿Cuál es el nivel de pobreza esperado para un país cuya tasa de fertilidad es de 6,88 hijos por mujer?

12. Se intenta analizar la influencia del nivel de pobreza de un país sobre la tasa de mortalidad infantil; una parte del estudio consiste en hacer un análisis de regresión para ambas variables en los países suramericanos. El estudio se basa en los datos aportados por Indexmundi.com:

	Población bajo el nivel de pobreza (%)	Tasa de mortalidad infantil
Surinam	70	19
Bolivia	60	45
Colombia	49	19
Perú	45	29
Ecuador	38	21
Venezuela	38	22
Paraguay	32	25
Brasil	31	23
Uruguay	27	11
Argentina	23	11
Chile	18	8

Tasa de mortalidad infantil: esta variable da el número de muertes de niños menores de un año de edad en un año determinado por cada 1000 niños nacidos vivos en el mismo año.

a. Si se tienen en cuenta todos los datos, ¿cuál es el coeficiente de correlación lineal? ¿Es la ecuación de regresión válida?
b. Si se eliminan los datos referentes a Surinam, ¿cuál es el coeficiente de correlación lineal? ¿Es la ecuación de regresión válida? Si lo es, ¿cuál es esa ecuación? ¿Cree que el análisis podría hacerse extensivo a otras regiones diferentes a Suramérica?

OBJETO DE APRENDIZAJE 3
PROBABILIDADES

Los fenómenos pueden ser determinísticos (si no hay incertidumbre acerca de su resultado cuando se repita o que puedan predecirse por ecuaciones matemáticas) o aleatorios (que no pueden predecirse exactamente, aunque se repitan en forma similar).

En estadística se manejan datos aleatorios; en ellos no es posible hacer predicciones exactas mediante el uso de modelos matemáticos, pero cuando son estudiados muchas veces bajo condiciones similares se encuentra que los resultados presentan cierta regularidad. Por lo tanto, nunca puede estarse seguro de lo que vaya a pasar, pero con base en la información del pasado puede predecirse con fundamentos.

El concepto de probabilidad ocupa un lugar importante en el proceso de toma de decisiones bajo incertidumbre, independientemente de que se trate del campo de los negocios, de la ingeniería, de las ciencias sociales, o simplemente en nuestras vidas diarias. En muy pocas situaciones de toma de decisiones la información perfecta está disponible –todos los factores o hechos necesarios–; la mayoría de las decisiones se toman encarando la incertidumbre.

El objetivo de la teoría de probabilidades es poder hacer predicciones y tener más elementos de juicio para tomar decisiones más fundamentadas (si se pronostica qué tan probable es que ocurra algo, se puede determinar si se toma el riesgo o no).

Con la teoría de probabilidades se pueden construir modelos que describen adecuadamente la regularidad de los resultados aleatorios, de tal forma que se puedan hacer predicciones.

3.1. CONCEPTOS BÁSICOS

3.1.1. ¿Qué es?:

La probabilidad es una herramienta para medir la posibilidad de ocurrencia de un evento. Dicho de otra forma, es la medición de la incertidumbre acerca de la ocurrencia de determinada situación. Puede tomar un valor entre 0 y 1 (0 si nunca se presenta y 1 si siempre lo hace).

Si un experimento puede tener como resultado cualquiera de N resultados diferentes, igualmente probables, y si exactamente *n* de esos resultados corresponden al evento A, entonces:

$$P(A) = n/N$$

La probabilidad de que no ocurra es: q = 1 – p.

Una probabilidad también puede expresarse como p/q (p:q o de p a q). Aparte de su valor en apuestas, esta manera de expresarla permite especificar una probabilidad pequeña (cerca de cero) o una probabilidad grande (cerca de uno) usando números enteros grandes (mil a uno ó un millón a uno) para magnificar probabilidades pequeñas (o probabilidades grandes) con el objetivo de hacer las diferencias relativas visibles.

Dicho de otra manera, la probabilidad clásica de que un evento ocurra se calcula dividiendo el número de resultados favorables entre el número de posibles resultados.

El enfoque anterior supone que todos los resultados experimentales tienen la misma probabilidad de ocurrir; eso es razonable si el caso es completamente aleatorio, pero tiene muchos problemas cuando intentamos aplicarlo a los problemas de decisión menos ordenados que encontramos en la realidad. Por eso, en general, lo mejor es calcular la probabilidad experimentalmente, determinando la frecuencia con que algo ha sucedido en el pasado y mediante esa cifra predecir la probabilidad de que vuelva a ocurrir en el futuro; debido a ello, la probabilidad de un resultado puede interpretarse como el valor límite de la proporción de veces que el resultado aparece en *n* repeticiones del experimento aleatorio, a medida que *n* crece sin cota alguna.

Si *n* tiende a infinito, se da una estabilización de la frecuencia relativa (tiende a un límite fijo). Por ejemplo, al lanzar un dado, es casi imposible que un valor determinado resulte en 1/6 de las observaciones; tampoco significa que si hacemos 600 observaciones, vamos a obtener 100 de cada especie; pero si se repite muchas veces, en promedio, los seis resultados posibles se presentarán con frecuencias prácticamente iguales. Si esto no sucede, debemos sospechar que otro factor está interviniendo en lo que observamos.

La probabilidad puede también tener un enfoque subjetivo, es decir, basada en el grado de creencia de que ocurra el resultado, por lo que personas distintas pueden asignar distintas probabilidades a un mismo resultado. Para aplicar este método se puede usar cualquier dato disponible o la experiencia e intuición de la persona que evalúa.

3.1.2. Experimento:

Es cualquier acción cuyo resultado se registra como un dato. En estadística, la noción de experimento es distinta de la noción en ciencias físicas; en ellas, por lo

general, un experimento se lleva a cabo en un laboratorio o en un ambiente controlado, para aprender acerca de un hecho científico, y cuando se repiten los experimentos bajo condiciones idénticas se espera obtener el mismo resultado. En los experimentos estadísticos, los resultados están determinados por el azar y aunque el experimento se repita exactamente en la misma forma, puede obtenerse un resultado diferente.

3.1.3. Espacio muestral (S):

Es el conjunto de todos los resultados posibles de un experimento estadístico.

3.1.4. Punto muestral:

Es cada resultado de un espacio muestral. Se pueden especificar por un diagrama de árbol, que es un dispositivo gráfico útil para visualizar un experimento de varias etapas y enumerar los resultados experimentales.

3.1.5. Evento o suceso (E):

Es cualquier subconjunto del espacio muestral (colección de puntos muestrales). Un evento es el elemento básico al cual se puede aplicar la probabilidad; un evento sucede o no sucede.

Con los eventos se pueden efectuar las operaciones comunes de conjuntos (intersección, unión y complemento) para describir cualquier caso en términos de eventos simples.

3.1.6. Evento aleatorio:

Se dice que un evento es aleatorio cuando no se tiene certeza de si ocurrirá o no en el momento de la observación.

Siempre deben satisfacerse dos requisitos básicos (axiomas de probabilidad):

(1) $0 \leq P(E) \leq 1$

(2) P(S) = 1, lo que implica que la suma de todas las probabilidades de resultados experimentales debe ser 1. Por lo tanto, si un espacio muestral tiene *k* resultados experimentales:

$$P(E_1) + P(E_2) + \ldots\ldots\ldots + P(E_k) = 1$$

3.2. MANERAS DE DESCRIBIR UN ESPACIO MUESTRAL

3.2.1. DIAGRAMA DE VENN:

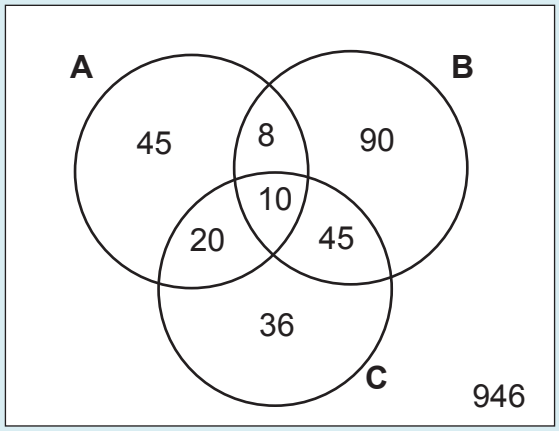
La relación entre eventos y el espacio muestral correspondiente se puede expresar gráficamente por este medio (espacio muestral con un rectángulo y los eventos con círculos dentro del rectángulo).

EJEMPLO

Ceipa piensa ofrecer tres nuevos programas de especialización: Gestión del mejoramiento y la productividad (A), Gestión empresarial (B) y Especialización en administración hospitalaria (C). En una investigación de mercados realizada, los programas fueron ofrecidos a 1200 estudiantes de pregrado de administración: 18 se mostraron interesados en A y B, 30 en A y C, 55 en B y C, 10 mostraron interés por los tres programas, 45 solamente en A, 90 únicamente en B y 36 solamente en C.

a. ¿Qué tan probable es que un estudiante de los encuestados no muestre interés por ninguno de los programas ofrecidos?
b. ¿Cuál es la probabilidad de que muestre interés en por lo menos dos de los programas?

Solución:



a. P(ninguno) = 946/1200 = 78,83%
b. P(X≥2) = 83/1200 = 6,92%

- **Eventos mutuamente excluyentes:** dos o más sucesos son considerados mutuamente excluyentes si ellos no pueden ocurrir simultáneamente, o sea que la ocurrencia de uno cualquiera de ellos excluye la ocurrencia de los otros. Dos eventos mutuamente excluyentes no tienen elementos en común, lo que implica que:

$$P(A \cap B) = 0$$

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

- **Eventos no excluyentes:** si es posible que ambos ocurran simultáneamente.

  En este caso:

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

$P(A \cap B)$ es diferente de 0,

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

EJEMPLO

Cierta empresa desarrolló una investigación de mercados para proyectar la probabilidad de éxito de dos productos que piensa lanzar al mercado. La probabilidad de que el producto A tenga éxito es 0,68 y la probabilidad de que el producto B tenga éxito es 0,84; además encontró que la probabilidad de que al menos uno de los productos se convierta en un éxito es 0,98.

a. ¿Cuál es la probabilidad de que ambos productos tengan éxito?
b. ¿Qué tan probable es que ambos productos fracasen?

Solución:

a. $P(A \cap B) = P(A) + P(B) - P(A \cup B)$
$= 0{,}68 + 0{,}84 - 0{,}98$
$= 0{,}54$

b. P (ambos fracasen) = $1 - 0{,}98 = 0{,}02$

3.2.2. TABLA DE CONTINGENCIA:

Si en el experimento están incluidas dos variables, el espacio muestral puede detallarse mediante una tabla de contingencia, bien sea con porcentajes o cantidades absolutas. Para poder emplearla, todas las clases deben ser mutuamente excluyentes.

EJEMPLO

En una encuesta entre los estudiantes de administración de una universidad se obtuvieron los datos siguientes acerca del principal motivo del estudiante para solicitar su ingreso, según la jornada en que estaba inscrito (diurna o mixta):

		Motivo de la solicitud			
		Calidad	Costo	Otros	Total
Jornada	Diurna	64	36	78	178
	Mixta	16	45	11	72
	Total	80	81	89	250

(información detallada por cantidades)

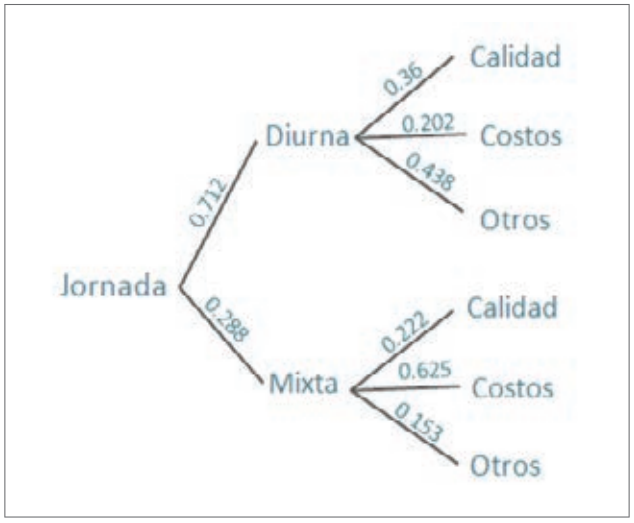
		Motivo de la solicitud			
		Calidad	Costo	Otros	Total
Jornada	Diurna	25,6	14,4	31,2	71,2
	Mixta	6,4	18	4,4	28,8
	Total	32	32,4	35,6	100

(información detallada por porcentajes)

Nota: En la última tabla se sacaron todos los porcentajes con base en los 250 estudiantes. Por ejemplo, los estudiantes de jornada diurna que escogieron la universidad por calidad fueron 64/250x100

3.2.3. DIAGRAMA DE ÁRBOL:

Es una representación gráfica que permite visualizar un experimento de pasos múltiples o un experimento que incluya dos o más variables.



3.2.4. COMBINACIONES Y PERMUTACIONES:

- **Combinaciones**: para contar el número de resultados experimentales cuando se seleccionan *r* elementos de un conjunto de tamaño *n* y no importa el orden. Eso se simboliza $_nC_r$ y puede calcularse como:

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

Dicho cálculo también puede desarrollarse directamente en una calculadora o en un programa estadístico. En los anexos 1 y 4 se explica la forma.

Cuando se hace un muestreo de una población finita de tamaño *N*, el número de posibles muestras diferentes de tamaño *n* que podrían seleccionarse se calcula mediante combinaciones.

EJEMPLO

¿Qué tan probable es ganar el premio mayor del Baloto si se compra un boleto? Tenga en cuenta que este juego se deben seleccionar seis números de un conjunto de 45 y, para ganar, ellos deben coincidir con los elegidos.

Solución:

$$_{45}C_6 = \frac{45!}{6!39!} = 8145060$$ (Hay 8 145 060 combinaciones posibles)

$$\Rightarrow P(ganar) = \frac{1}{8145060} = 0{,}000000123$$

- **Permutaciones:** para contar el número de resultados experimentales cuando se seleccionan *r* elementos de un conjunto de tamaño *n* y el orden de selección es importante. Esto se representa como ${}_nP_r$ y puede calcularse como:

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

Este cálculo también puede desarrollarse directamente en una calculadora o en un programa estadístico. En los anexos 1 y 4 se explica la forma.

EJEMPLO

¿Qué tan probable sería ganar el premio mayor del Baloto si se compra un boleto y los números se marcan de tal manera que tienen que coincidir con el orden en que salen en el sorteo?

Solución:

${}_{45}P_6 = \frac{45!}{39!} = 5\,864\,443\,200$ (Hay casi 6000 millones de permutaciones posibles)

$\Rightarrow P(ganar) = 1/\,5864443200 = 0{,}00000000017$

3.3. PROBABILIDAD MARGINAL, CONDICIONAL Y CONJUNTA:

- Probabilidad marginal es la probabilidad de un evento simple [P(A), P(B)]; es decir, es la probabilidad de que se presente un evento correspondiente a una variable, sin considerar la otra u otras variables.

- Probabilidad conjunta es la probabilidad de que varios eventos ocurran simultáneamente. Se expresa como P(A∩B) o P(AB).

- La probabilidad condicional se representa como P(B\A), lo cual indica la probabilidad de que un evento B ocurra si ya se sabe que ocurrió el evento A. De la misma forma P(A\B) indica la probabilidad de que ocurra A dado que ocurrió B.

Las anteriores definiciones han sido expresadas para el caso de dos eventos, pero las mismas ideas pueden aplicarse para cualquier número de eventos.

La noción de probabilidad condicional proporciona la capacidad de reevaluar la idea de probabilidad de un evento a la luz de la información adicional.

Las probabilidades pueden calcularse fácilmente mediante reglas de tres simples; para mayor facilidad puede asumirse un gran total de 100% y organizar la información en forma tabular (como tabla de contingencia).

EJEMPLO

La empresa que se viene evaluando tiene un número grande de representantes de ventas; el 56% de los representantes están en Bogotá; 24%, en Cali; y 20%, en Medellín.

El 70% de los representantes de Bogotá vendió 250 productos o menos en el mes anterior, al igual que el 75% de los de Cali y el 38% de los de Medellín.

a. ¿Qué porcentaje de los representantes de ventas de la empresa vendió 250 productos o menos en el mes anterior?
b. Si un representante de ventas de la empresa vendió menos de 250 productos en el mes anterior, ¿qué tan probable es que sea de Bogotá?
c. ¿Qué porcentaje de los vendedores de la empresa opera en Bogotá y vendió menos de 250 productos en el mes anterior?

Solución:

MODELO: (Menos de 250 productos en Bogotá) = 56 x 70% = 39,2 productos

CIUDAD SEDE	CANTIDAD DE PRODUCTOS VENDIDOS		**Total**
	Menos de 250 productos	250 productos o más	
Bogotá	39,2	16,8	56
Cali	18	6	24
Medellín	7,6	12,4	20
Total	64,8	35,2	100

a. 64,8% (correspondiente a la suma de todos los representantes que vendieron menos de 250 productos).

b. $P(A \setminus \overline{B}) = \frac{39,2}{64,8} = 0,605$ (correspondiente a los vendedores de Bogotá que vendieron menos de 250 productos del total de los que vendieron esa cantidad).

c. 39,2% (celda que corresponde al cruce de los valores).

EJEMPLO

El 80% de las empresas pequeñas del sector de alimentos no están preparadas para un tratado de libre comercio con USA, al igual que el 63% de las empresas medianas y 30% de las grandes.

El 80% de las empresas del sector son pequeñas, el 15% son medianas y el resto grandes.

a. ¿Cuál es la probabilidad de que una empresa no esté preparada para el TLC y sea pequeña?

b. ¿Qué porcentaje está preparada?

c. Si una empresa no está preparada para el TLC, ¿cuál es la probabilidad de que sea pequeña?

d. Si una empresa es pequeña, ¿cuál es la probabilidad de que esté preparada?

Solución:

		Tamaño			
		Pequeña (P)	Mediana (M)	Grande (G)	Total
Nivel de preparación	No	64	9,45	1,5	74,95
	Sí	16	5,55	3,5	25,05
	Total	80	15	5	100

a. $P(P \cap no) = 64/100 = 0,64$

b. P (sí) = 25,05%

a. P (P\no) = 64/74,95 = 0,854

b. P(sí\P) =16/80 = 0,2

EJERCICIOS PROPUESTOS

1. Cierta empresa planea exportar sus productos a otros países. Para aprovechar un tratado firmado recientemente se ha presupuestado exportar el 60% de su producción a países del Lejano Oriente y se ha presupuestado exportar el resto a Estados Unidos y a Europa por partes iguales. Es sabido que la posibilidad de éxito se ve influenciada directamente por la cultura de cada país; análisis que se han hecho demuestran que si se hacen exportaciones al Oriente, la probabilidad de que haya éxito es 0,45; si se exporta a Estados Unidos la probabilidad de éxito es 0,75 y si se exporta a Europa, dicha probabilidad es 0,55.

   a. Si se lleva a cabo lo presupuestado, ¿cuál es la probabilidad de que la exportación no sea un éxito?
   b. Si una exportación tiene éxito, ¿cuál es la probabilidad de que sea al Lejano Oriente?
   c. ¿Cuál es la probabilidad de que una exportación tenga éxito y sea al Lejano Oriente?

2. El 49% de las exportaciones antioqueñas tienen como destino Estados Unidos, 20% va a países cercanos de Suramérica, 16% a Europa y el resto a otros países. El 45% de lo que se exporta a Estados Unidos hace parte del sector agropecuario, al igual que 38% de lo que se exporta a países cercanos, 60% de lo que se exporta a Europa y 50% de lo que se exporta a otros países.

   a. ¿Qué proporción de las exportaciones antioqueñas es de productos agropecuarios?
   b. Si una exportación es de productos que no son agropecuarios, ¿cuál es la probabilidad de que sea a Europa?

3. Una empresa tiene tres plantas: A, B y C. La planta A origina el 50% de la producción total, B produce el 35% y C, el resto. El 3% de la producción de A es defectuosa, mientras que el 4% de B y el 5% de C también lo son.

   a. Si se elige al azar un artículo producido por la empresa, ¿cuál es la probabilidad de que ese artículo sea defectuoso?
   b. Si el artículo elegido resulta ser defectuoso, ¿cuál es la probabilidad de que provenga de C?

4. Una compañía de seguros clasifica a sus clientes como de alto, mediano y bajo riesgo; ellos reclaman el pago de un seguro con probabilidades 0,02, 0,01 y 0,0025, respectivamente. El 10% de los clientes es de alto riesgo, el 20% de

mediano y el 70% de bajo riesgo. Si uno de los clientes reclama el pago de un seguro, ¿cuál es la probabilidad de que sea de bajo riesgo?

5. En una encuesta entre los estudiantes de administración de una universidad se obtuvieron los datos siguientes acerca del principal motivo del estudiante para solicitar su ingreso:

		Motivo de la solicitud		
		Calidad	Costo	Otros
Horario	Diurno	64	36	105
	Mixto	16	45	11

   a. Si un alumno es de horario mixto, ¿cuál es la probabilidad de que la calidad sea el motivo principal para haber elegido la universidad?
   b. Si un estudiante de administración eligió la universidad por costo, ¿cuál es la probabilidad de que sea de horario diurno?

6. Por estadísticas anteriores se conoce que el 90% de los negocios que son franquicias pasan de los diez años en el mercado, mientras que el 29% de los negocios totalmente nuevos sobreviven por más de diez años.

   Para verificarlo, una empresa de estadísticas evaluó los casos de 395 empresas, de las cuales 108 eran franquicias; de ellas, 92 sobrevivían todavía a los diez años. De las que no eran franquicia, encontró que 204 habían cerrado antes de los diez años de funcionamiento.

   a. Según el último estudio, ¿cuál es la probabilidad de que una empresa -sea franquicia o no– sobreviva diez años o más en el mercado?
   b. Según los datos encontrados, ¿qué tan probable es que una franquicia sobreviva diez años o menos en el mercado? ¿Qué tan probable es que un negocio que no es franquicia sobreviva ese tiempo? ¿Parece estar eso en armonía con el estudio detallado previamente?

OBJETO DE APRENDIZAJE 4
DISTRIBUCIONES DE PROBABILIDADES

Una variable aleatoria es discreta si puede asumir únicamente valores enteros y es continua cuando puede tomar cualquier valor dentro de un intervalo de números reales. Las variables aleatorias cualitativas deben asumirse como discretas para su estudio como distribución, puesto que a cada posible valor podría asignársele un número entero.

Si la variable aleatoria es discreta, pero el rango es muy amplio, resulta más conveniente utilizar un modelo basado en variables aleatorias continuas.

MODELOS DE DISTRIBUCIONES

Con frecuencia, las observaciones que se generan en experimentos estadísticos tienen algunos tipos generales de comportamiento y por eso se pueden describir esencialmente con unas pocas distribuciones, las cuales se representan mediante una ecuación.

Jaramillo (2003) menciona: "Frente a la complejidad de los fenómenos bajo estudio, el experimentador aproxima y hace algunos postulados tentativos acerca del mecanismo aleatorio y deriva un modelo por el empleo de esos postulados en combinación con las leyes de probabilidad".

Un modelo de probabilidad para la variable aleatoria *X* es una forma específica de distribución de probabilidades que es asumida para reflejar el comportamiento de dicha variable. Las probabilidades son registradas en términos de parámetros desconocidos que relacionan las características de la población y el método de muestreo.

"EL MODELO DEBE SER COHERENTE CON LA REALIDAD"

En las páginas siguientes se examinarán detalladamente algunas distribuciones específicas de probabilidad que han demostrado, empíricamente, ser modelos útiles para diversos problemas prácticos. Pero dichas distribuciones son teóricas porque sus funciones de probabilidad se deducen matemáticamente con base en ciertas hipótesis que se suponen válidas para esos fenómenos aleatorios. Dichas distribuciones son idealizaciones del mundo real, por lo tanto sus resultados no siempre coinciden con la realidad.

La distribución de probabilidad se describe mediante una función de probabilidad, representada por f(x). Se denomina función de probabilidad en el caso de variables discretas; si la variable es continua, se denomina función de densidad.

4.1. DISTRIBUCIÓN BINOMIAL:

Se maneja cuando se satisfacen las siguientes características:

a. La variable que se estudia es discreta o cualitativa.

b. El experimento consiste en una sucesión de *n* ensayos o intentos idénticos.

c. El resultado de cada ensayo se clasifica dentro de dos categorías mutuamente excluyentes: éxito o fracaso. El uso de esos términos es por conveniencia, pero no tienen la misma connotación de la vida real (éxito no necesariamente es lo que convenga).

d. La probabilidad de éxito permanece constante en todos los ensayos.

e. Los ensayos son independientes, lo que significa que la ocurrencia de uno de ellos no afecta el resultado de cualquier otro.

La función de probabilidad binomial puede escribirse como:

$$f(x) = p^x q^{n-x}\, {}_nC_x$$

donde *x* es el número de éxitos, *n* el número de ensayos, *p* es la probabilidad de éxito y q es la probabilidad de fracaso.

${}_nC_x$ se denomina número de combinaciones de *x* en *n* y corresponde al número de grupos distintos de x elementos que se pueden formar a partir de un conjunto de *n* elementos. Las calculadoras tienen dicha función; o puede calcularse así:

$${}_nC_x = \frac{n!}{x!(n-x)!}$$

Por ejemplo, si se quieren conformar diferentes parejas de un grupo de 10 elementos podrían conformarse 45 parejas diferentes, ya que:

$${}_{10}C_2 = \frac{10!}{2!*8!} = 45$$

EJEMPLO

El presidente de una compañía planea contactar a otras 18 compañías en busca de nuevos socios para su negocio. Sus analistas han estimado que la probabilidad de que una firma contactada al azar acepte incorporarse como socio es de 0,25.

a. ¿Qué tan probable es que ninguna de las empresas contactadas acepte incorporarse como socio?
b. ¿Cuál es la probabilidad de que acabe reclutando máximo dos socios de entre las compañías contactadas?
c. ¿Cuál es el número esperado de socios que se incorporarán al proyecto?

Solución:

a. P(x=0) = $0{,}25^{0} * 0{,}75^{18} *_{18} C_{0}$ = 0,00564

b. P(x≤2) = P(x=0) + P(x=1) + P(x=2)

$$= 0{,}25^{0} * 0{,}75^{18} *_{18} C_{0} + 0{,}25^{1} * 0{,}75^{17} *_{18} C_{1} + 0{,}25^{2} * 0{,}75^{16} *_{18} C_{2}$$
$$= 0{,}1353$$

c. $\mu = 18 * 0{,}25 = 4{,}5$

Lo anterior implica que se esperaría que 4 ó 5 empresas de las contactadas terminen incorporándose al proyecto.

Mediante un programa estadístico pueden calcularse fácilmente esas probabilidades (en el anexo 4 está explicado cómo hacerlo con Excel).

Para ilustrar su uso, veamos el siguiente ejemplo:

EJEMPLO

Un examen de selección múltiple contiene 20 preguntas, cada una con cuatro posibles respuestas, de las cuales solamente una es correcta. Suponga que un estudiante trata de adivinar las respuestas.

a. ¿Cuál es la probabilidad de que el estudiante conteste correctamente más de la mitad de las preguntas?
b. ¿Cuál es la probabilidad de que el estudiante conteste correctamente menos

de cinco preguntas?

c. ¿Cuál es la probabilidad de que el estudiante gane el examen?
d. ¿Cuál es el número esperado de respuestas correctas?
e. Responder las preguntas a) y c) si cada pregunta tiene cinco opciones.

Solución:

p = ¼ n = 20

a. P(X>10) = 1 – P(X≤10) = 1 – 0,9961 = 0,0039
b. P(X<5) = P(X≤4) = 0,4148
c. P(X≥12) = 1 – P(X≤11) = 1 – 0,9991 = 0,0009
d. μ = n x p ⇒ μ = 20x¼ = 5 respuestas correctas
e. La probabilidad de éxito sería ya de 1/5, por tanto:
   P(X>10) = 1 – P(X≤10) = 1 – 0,9994 = 0,0006
   P(X≥12) = 1 – P(X≤11) = 1 – 0,9999 = 0,0001

4.2. DISTRIBUCIÓN HIPERGEOMÉTRICA:

Como el muestreo sin reemplazo viola las condiciones de un experimento binomial si la muestra no es grande, algunas veces es necesario plantear un tipo diferente de distribución.

Cuando se selecciona sin reemplazo una muestra aleatoria de tamaño *n* de una población de tamaño *N* y el interés recae en la probabilidad de seleccionar *x* éxitos de los *k* artículos considerados como éxitos en la población, se realiza un experimento hipergeométrico y su función de probabilidad viene determinada por:

$$f(x) = \frac{{}_{k}C_{x} * {}_{N-k}C_{n-x}}{{}_{N}C_{n}}$$

La distribución hipergeométrica requiere el conocimiento de *k* y *N*.

EJEMPLO

En Colombia se hace un sorteo, denominado Baloto, en el cual deben seleccionarse seis números de un conjunto de 45.

a. Existe un premio mayor; para hacerse acreedor a él, el apostador debe acertar todos los números. ¿Cuál es la probabilidad de ganar el premio mayor?
b. También está programado un premio más bajo para aquella persona que acierte cinco de los números. ¿Cuál es la probabilidad de ganar ese premio?
c. Un premio mucho más bajo es otorgado al que acierte cuatro de los números.

¿Cuál es la probabilidad de ganarlo?

d. ¿Cuál es la probabilidad de no acertar ninguno de los números o de acertar máximo uno?

Solución:

a. $P(x = 6) = \frac{{}_6C_6 * {}_{39}C_0}{{}_{45}C_6} = 0{,}000000123$

b. $P(x = 5) = \frac{{}_6C_5 * {}_{39}C_1}{{}_{45}C_6} = 0{,}000029$

c. $P(x = 4) = \frac{{}_6C_4 * {}_{39}C_2}{{}_{45}C_6} = 0{,}001365$

d. $P(x \leq 1) = P(x = 0) + P(x = 1) = \frac{{}_6C_0 * {}_{39}C_6 + {}_6C_1 * {}_{39}C_5}{{}_{45}C_6} = 0{,}825$

4.3. DISTRIBUCIÓN DE POISSON:

Este es el modelo de probabilidad más adecuado para eventos que ocurren aleatoriamente a través del tiempo o el espacio. La función de probabilidad de Poisson calcula la probabilidad de exactamente *x* ocurrencias independientes durante un período de tiempo dado, si los eventos ocurren independientemente y a una tasa constante. La función de la probabilidad de Poisson también representa el número de ocurrencias sobre áreas o volúmenes constantes.

De acuerdo con Walpole, Myers y Myers (1999), esta distribución supone:

a. Ritmo de llegada constante; eso significa que, si se dice que llegan en promedio 10 clientes por hora, se debe suponer que llega uno cada 6 minutos.
b. La posibilidad de dos ocurrencias simultáneas puede ser asumida como cero.
c. El número promedio de ocurrencias por unidad de tiempo o espacio se considera una constante.
d. La probabilidad de que suceda determinado número de eventos en un proceso de Poisson depende únicamente de la longitud del intervalo observado y no de su ubicación (p. 136).

La distribución de probabilidad de la variable aleatoria de Poisson *X*, que representa el número de resultados que ocurren en un intervalo de tiempo, área, espacio o volumen específico se calcula así:

$f(x) = \frac{e^{-\mu}(\mu)^x}{x!}$, donde μ es el número promedio de resultados en el intervalo evaluado.

EJEMPLO

A la sucursal *Aves María* de cierto banco llegan en promedio 36 clientes en una hora pico. Un cajero puede atender, en promedio 8 clientes en esa hora.

a. ¿Cuál es la probabilidad de que en un espacio determinado de 5 minutos lleguen 2 clientes como máximo?
b. ¿Qué tan probable es que un cajero alcance a atender entre 3 y 4 personas en una hora?

Solución:

a. 36 clientes que llegan por hora es equivalente a 3 llegadas en 5 minutos (Esto implica que $\mu = 3$)

$$P(x \leq 2) = P(x = 0) + P(x = 1) + P(x = 2) = \frac{e^{-3}(3)^0}{0!} + \frac{e^{-3}(3)^1}{1!} + \frac{e^{-3}(3)^2}{2!} = 0{,}4232$$

b. 8 clientes son atendidos por hora por un cajero

$$P(3 \leq x \leq 4) = P(x = 3) + P(x = 4) = \frac{e^{-8}(8)^3}{3!} + \frac{e^{-8}(8)^4}{4!} = 0{,}0859$$

Al igual que para la distribución binomial, las probabilidades se pueden establecer mediante el uso de Excel (ver anexo 4). Veamos un ejemplo:

EJEMPLO

Los mensajes que llegan a un computador utilizado como servidor lo hacen de acuerdo con una distribución Poisson con una tasa promedio de 10 mensajes por hora.

a. a) ¿Cuál es la probabilidad de que lleguen más de 3 mensajes en un espacio de 15 minutos?

b. ¿Cuál es la probabilidad de que lleguen entre 15 y 20 mensajes en un espacio de una hora?

Solución:

a. μ = 2,5

$$P(X > 3) = 1 - P(X \leq 3)$$
$$= 1 - 0{,}7576$$
$$= 0{,}2424$$

b. μ = 10

$$P(15 \leq X \leq 20) = P(X \leq 20) - P(X \leq 14)$$
$$= 0{,}9984 - 0{,}9165$$
$$= 0{,}0819$$

4.4. DISTRIBUCIÓN NORMAL:

Es la más empleada para modelar experimentos aleatorios continuos porque describe ajustadamente muchos fenómenos naturales, industriales e investigativos; los errores en mediciones científicas se aproximan bien mediante la distribución normal.

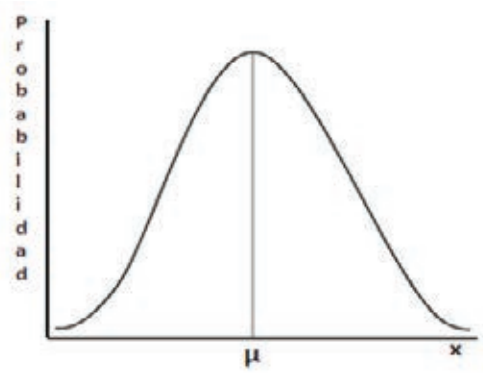
Esta distribución puede obtenerse al considerar el modelo de una variable binomial cuando el número de ensayos tiende a infinito.

La función de densidad de una variable aleatoria normal *X*, con media μ y desviación estándar σ es:

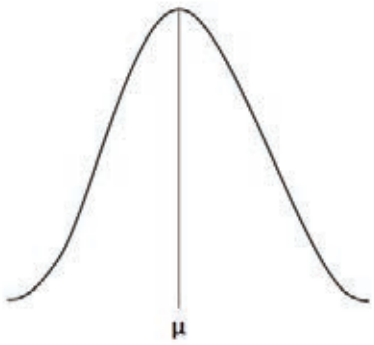
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]}, \quad -\infty < x < \infty$$

La distribución normal presenta las siguientes características:

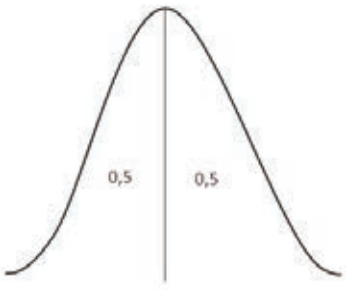
- Su gráfica tiene forma de campana; eso significa que valores muy pequeños o muy grandes tienen poca probabilidad de ocurrir, mientras que valores intermedios tienen una mayor probabilidad.



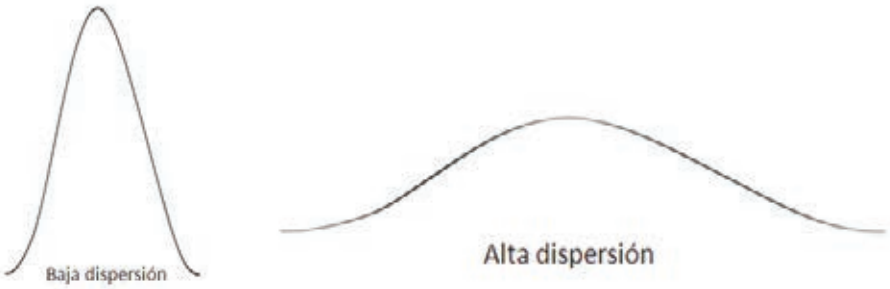
- El punto más alto de la curva normal es la media, que también es la mediana y la moda de la distribución.



- El área total bajo la curva es 1.
- La curva es simétrica alrededor de la media; por lo tanto, el área a la izquierda de la media es 0,5 e igual a su derecha.



- El eje *x* es una asíntota horizontal, es decir, los extremos de la curva se prolongan al infinito en ambas direcciones.
- La desviación estándar determina el ancho de la curva. A mayores valores de σ se obtienen curvas más anchas y bajas (mayor dispersión de los datos).



La función de densidad de la distribución normal no es de fácil solución, pero dicho problema puede obviarse mediante la estandarización de la variable, que consiste en transformar todas las observaciones de cualquier variable aleatoria normal *X* a un nuevo conjunto de observaciones de una variable aleatoria normal *Z* con media 0 y varianza 1; todas las distribuciones normales pueden convertirse a "distribuciones normales estándar" restando la media de cada observación y dividiendo por la desviación estándar, así:

$$Z = \frac{x - \mu}{\sigma}$$

Para la utilización de la distribución normal en problemas prácticos existen ciertas tablas donde se encuentran los valores *F*(*x*) –funciones de distribución acumulada– para una serie limitada de valores *Z* dados; igualmente, dichos valores pueden establecerse mediante el uso de Excel (explicado en anexo 4).

EJEMPLO

El volumen que una máquina de llenado automático deposita en latas de una bebida gaseosa tiene una distribución normal con media 150 mililitros y desviación estándar de 0,63 mililitros.

a. ¿Cuál es la probabilidad de que el volumen depositado sea menor de 148,5 mililitros?
b. ¿Cuál es la probabilidad de que el volumen depositado sea superior a 151,3 mililitros?
c. Si se desechan todas las latas que contienen menos de 148 o más de 152 mililitros de bebida, ¿cuál es la proporción de latas desechadas?
d. Calcule especificaciones que sean simétricas alrededor de la media, de tal forma que se incluya al 99% de todas las latas.

Solución:

a. $P(X<148{,}5)=P\left(Z<\dfrac{148{,}5-150}{0{,}63}\right)=P(Z<-2{,}38)=0{,}0087$

b. $P(X>151{,}3)=P\left(Z>\dfrac{151{,}3-150}{0{,}63}\right)=P(Z>2{,}06)=1-P(Z<2{,}06)=1-0{,}9803=0{,}0197$

c. La probabilidad de que la lata no sea desechada es:

$$P(148<x<152)=P\left(\frac{148-150}{0{,}63}<Z<\frac{152-150}{0{,}63}\right)=P(-3{,}17<Z<3{,}17)$$
$$=0{,}9992-0{,}0008=0{,}9984$$

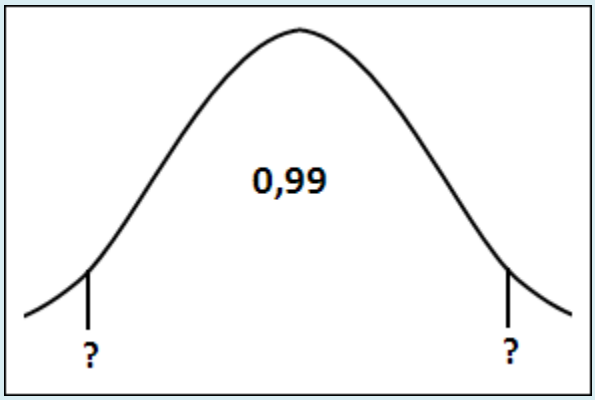
Por lo tanto, la probabilidad de que sea rechazada es 1 – 0,9984 = 0,0016

La probabilidad de que sean rechazadas también podría hallarse como:

$$2\times\ P\,(x<148)=2\ *0.0008=0{,}0016$$

Nota: La segunda forma solo puede utilizarse porque los dos valores están igualmente alejados de la media.

d.



$$P(Z>z)=0{,}005\Rightarrow P(Z<z)=0{,}995\Rightarrow z=2{,}575$$

$$z=\frac{x-\mu}{\sigma}\Rightarrow x_1=\mu+z\sigma=150-2{,}575(0{,}63)=148{,}38ml$$

$$x_2=\mu+z\sigma=150+2{,}575(0{,}63)=151{,}62\,ml$$

EJERCICIOS PROPUESTOS

1. Cierto producto tiene una demanda que exhibe una variabilidad que sigue aproximadamente una distribución normal con media 26,5 unidades demandadas por día y una desviación estándar de 7,8 unidades demandadas por día.

    a. ¿Qué tan probable es que en un día se demanden más de 50 unidades?
    b. ¿Qué tan probable es que en un día se demanden entre 20 y 30 unidades?
    c. ¿A partir de qué cantidad de pedidos se encuentra el 20% de los días en que el producto se demanda menos?

2. Se elige una muestra de 20 empresas colombianas exportadoras. Se sabe por estadísticas de años anteriores que aproximadamente el 40% de las empresas colombianas exportadoras registra operaciones en varias ciudades.

    a. ¿Qué tan probable es encontrar al menos 18 empresas que tengan operaciones en varias ciudades?
    b. ¿Qué tan probable es encontrar entre 10 y 15 empresas que operen en varias ciudades?

3. Un embarque contiene 1000 artículos, de los cuales hay 20 defectuosos (lo sabe la empresa productora). La empresa que realiza las inspecciones toma una muestra aleatoria de 10 artículos y si encuentra por lo menos uno defectuoso rechaza el embarque, ¿cuál es la probabilidad de que el embarque se rechace?

4. Al conmutador de una universidad llegan en promedio 120 llamadas/hora durante el período de actividad. El conmutador no puede hacer más de 5 conexiones por minuto; calcule la probabilidad de que:

    a. el conmutador se encuentre congestionado en un minuto dado.
    b. se pierdan 3 o más llamadas si la recepcionista salió 2 minutos de la oficina.

5. El 80% de los contadores graduados en cierta universidad trabajan en su área de estudios. Si se evalúan los 20 egresados de Contaduría de dicha universidad en este semestre,

    a. ¿cuál es la probabilidad de que todos se empleen en contaduría?
    b. ¿cuál es la probabilidad de que al menos el 80% de ellos se emplee en el área?
    c. ¿cuál es la probabilidad de que menos de 12 se empleen en el área?

6. La carretera que usa la empresa X para transportar sus productos hasta cierto pueblo del Departamento presenta, en promedio, 8 huecos grandes por kilómetro; si se presenta algún trayecto de un kilómetro que tenga 20 huecos o más, la carretera se considera intransitable. ¿Qué tan probable es que lo anterior se presente?

7. Se ha encontrado, por datos históricos, que el monto de negocios en las ferias de textiles que se han desarrollado en Colombia durante los últimos 5 años sigue aproximadamente una distribución normal con media 9,8 millones de dólares y desviación estándar de 8,7 millones de dólares.

    a. En la última feria de Colombiatex, realizada en Medellín, se efectuaron negocios por 33 millones de dólares; ¿hará parte del 5% de ferias textileras donde más negocios se han hecho?
    b. ¿Cuál es la probabilidad de que en una feria de este tipo se hagan negocios por más de 15 millones de dólares?

8. La empresa petrolera que lidera actualmente el sector de los hidrocarburos en Colombia extrajo, durante el primer semestre de 2011, una cantidad promedio de 707 300 barriles diarios de petróleo. La distribución del número de barriles extraídos diariamente se ajusta al modelo normal y tiene una desviación estándar de 105 565 barriles/día.

    a. ¿Qué tan probable es que mañana se extraigan más de un millón de barriles?
    b. ¿Qué tan probable es que mañana se extraigan entre 800 000 y 950 000 barriles?
    c. ¿Cuántos barriles de petróleo se extraen en el 50% de los días más "comunes"?

9. 3256 personas presentaron los exámenes de ingreso a una universidad, los cuales se calificaban con un puntaje máximo de 100. Las calificaciones se aproximan a una distribución normal, con una media de 67,8 y una desviación estándar de 9,1.

    a. El 12% de los postulantes con más alta calificación en el examen son aceptados en la universidad. ¿Un estudiante que obtiene un puntaje de 73 en el examen es aceptado o no?
    b. ¿Qué porcentaje de postulantes obtuvieron una calificación entre 60 y 80?

10. En un reporte presentado por el Observatorio Laboral el 2 de agosto de 2011 se revela que el salario promedio de un profesional en Colombia es $1 444 180, con desviación estándar de $369 690.

a. Según esto, ¿cuál es la probabilidad de que un profesional colombiano gane más de dos millones de pesos?
b. David, un joven profesional colombiano, gana $2 500 000. ¿Hace parte del 10% de los que más ganan? ¿Por qué?
c. ¿Por debajo de que valor está el salario del 5% de los profesionales colombianos más mal pagos?

11. Los costos de servicio por mes en una agencia de viajes tienen una distribución aproximadamente normal, con promedio de $150 millones y desviación estándar de $6 750 000.

a. ¿Cuál es la probabilidad de que en el próximo mes los costos de servicio estén entre 120 y 170 millones de pesos?
b. ¿Por encima de qué valor están los costos de servicio en el 20% de los meses con mayores costos?

12. En una universidad se encontró que 20% de los estudiantes no terminan el primer curso de estadística. Al curso se inscriben 15 estudiantes.

a. ¿Cuál es la probabilidad de que 2 o más no terminen?
b. ¿Cuál es la probabilidad de que no terminen entre 5 y 10?
c. ¿Cuál es la probabilidad de que nadie termine el curso?

OBJETO DE APRENDIZAJE 5
ESTADÍSTICA INFERENCIAL

El objetivo de la estadística inferencial es la elaboración de estimaciones y pruebas de hipótesis acerca de las características de una población, a partir de los datos obtenidos de una muestra representativa; se apoya en la teoría de probabilidades para hacer generalizaciones sobre la población.

Para garantizar la generalización es necesario que la muestra sea representativa y que se haya hecho un adecuado diseño de muestreo.

5.1. ESTIMACIÓN

Se ha venido manejando información de muestras aleatorias tomadas de una población conocida; pero lo más importante es inferir información sobre la población a partir de muestras suyas. Eso es lo que se conoce como estimación.

Con base en estimaciones se determinan presupuestos, planes de negocios, inversiones y pronósticos.

A un valor calculado con los datos de una muestra se le llama **estadígrafo**. Al estadístico que se usa para predecir el valor de un parámetro de la población se le llama **estimador**; si para estimar el parámetro se usa un valor único dicho estimador es llamado **estimador puntual**.

Algunos estimadores puntuales son:

- La proporción muestral ($\hat{p}$), usado como estimador de la proporción poblacional (p).
- La media muestral ($\overline{X}$), usado como estimador del valor esperado poblacional (μ).
- La varianza de la muestra ($\sigma_{n-1}^{\ 2}$), usado como estimador de la varianza de la población ($\sigma_{n}^{\ 2}$).

Aunque esta es la manera más común de expresar una estimación, da lugar a muchas preguntas; por ejemplo, no indica la cantidad de información sobre la cual se basa la estimación y no dice nada acerca del posible tamaño del error (y siempre que se toman muestras existe algún error).

En consecuencia, se opta por un segundo método de estimación denominado **estimación por intervalos,** que indica la precisión de una estimación; dicha idea de precisión es definida por el error estándar, que es la desviación estándar de la distribución muestral de las medias de las muestras.

La técnica de este tipo de estimación consiste, entonces, en asociar a cada muestra un intervalo que se sospecha que debe contener al parámetro; se le denomina **intervalo de confianza**. Este parámetro será habitualmente una proporción en el caso de variables cualitativas y la media o la varianza para variables cuantitativas.

Evidentemente esta técnica no tiene porqué dar siempre un resultado correcto. A la probabilidad de que hayamos acertado al expresar que el parámetro estaba contenido en dicho intervalo se le denomina **nivel de confianza**.

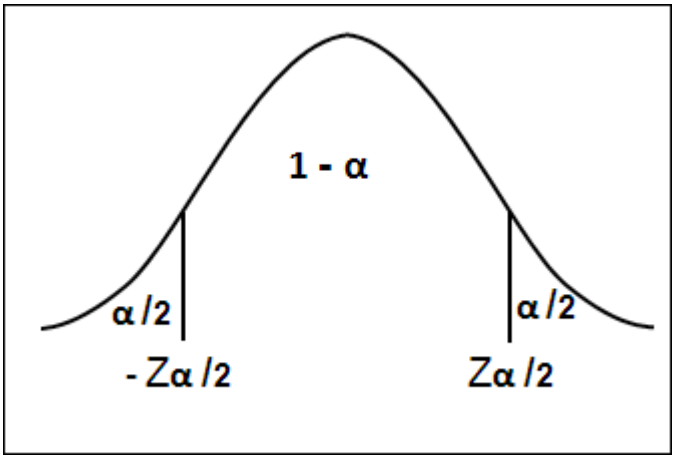
Un intervalo de confianza, si es bilateral, se expresa mediante dos valores: L (límite inferior) y U (límite superior), de tal manera que la siguiente proposición de probabilidad es verdadera:

$$P(L \leq \theta \leq U) = 1 - \alpha$$

donde α es un número entre o y 1 (arbitrario y preferiblemente pequeño –por ejemplo 0,05 o 0,1–).

El intervalo de confianza resultante se conoce como intervalo de confianza de 100(1-α)% para el parámetro desconocido θ. Las cantidades L y U reciben los nombres de **límites de confianza inferior y superior,** respectivamente, y 1-α es el **nivel de confianza**.

Se tiene una probabilidad de 1 – α de seleccionar una muestra que produzca un intervalo que contiene el valor verdadero de θ.



Interpretación de un intervalo de confianza: si se recopila un número infinito de muestras aleatorias y se calcula un intervalo de confianza del 100(1–α)% para θ

para cada una de las muestras, entonces el 100(1–α)% de esos intervalos contiene el verdadero valor de θ.

En la práctica se obtiene solamente una muestra aleatoria y se calcula un intervalo de confianza. Puesto que ese intervalo puede o no contener el valor verdadero de θ, no es razonable asociar un nivel de probabilidad a este evento específico. La proposición adecuada es que el intervalo observado [L,U] contiene el verdadero valor de θ con una confianza de 100(1-α). Esta proposición tiene una interpretación de frecuencia, es decir, no se sabe si es correcta para la muestra en particular, pero el método usado para obtener ese intervalo proporciona proposiciones correctas el 100(1 - α)% de las veces.

Obviamente, entre más amplio sea el intervalo de confianza, mayor es la seguridad de que realmente el intervalo contenga el verdadero valor de θ, pero menor información se tiene acerca de ese valor. Lo ideal es, entonces, un intervalo de confianza relativamente pequeño con una confianza grande, lo que se logra aumentando el tamaño de la muestra evaluada.

Se presentarán algunos métodos para encontrar intervalos de confianza para medias y proporciones.

5.1.1. ESTIMACIÓN DE LA MEDIA DE UNA POBLACIÓN

» **Varianza conocida:** este caso se plantea más a nivel teórico que práctico porque difícilmente vamos a poder conocer con exactitud la varianza mientras que la media sea desconocida. Sin embargo nos aproxima del modo más simple a la estimación de medias.

Para estimar μ, el estadístico que mejor nos va a ayudar es $\overline{X}$.

De este modo, fijado α (valor arbitrario), se toma un intervalo que contenga una probabilidad de 1 – α. Lo ideal es que este intervalo sea lo más pequeño posible; por ello lo mejor es tomarlo simétrico con respecto a la media ya que allí es donde se acumula más masa en una distribución normal. Así, las dos colas de la distribución (zonas más alejadas de la media) tendrán áreas iguales.

La construcción de un intervalo de confianza se basa en la distribución de las medias muestrales. Si la muestra tomada no es muy pequeña, tal distribución se fundamenta en el **Teorema del límite central**, que afirma que:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

La diferencia con el Z planteado en el Objeto de aprendizaje anterior es muy sencilla: es mucho más difícil hallar un valor de la media alejado de la media poblacional en una muestra que para un valor individual; mirémoslo con un ejemplo:

EJEMPLO

Una empresa fabrica focos que tienen una duración con media de 800 horas y desviación estándar de 35 horas.

a. Encuentre la probabilidad de que un foco que alguien compró dure menos de 780 horas.
b. Encuentre la probabilidad de que una muestra de 50 focos tenga una vida promedio de menos de 780 horas.

Solución:

a. $Z = \frac{780 - 800}{35} = -0{,}57 \quad \leftrightarrow \quad P(X < 780) = \ 0{,}2843$

b. $Z = \frac{780 - 800}{35/\sqrt{50}} = -4{,}04 \quad \leftrightarrow \quad P(\bar{X} < 780) = \ 2{,}67 * 10^{-5}$

Conclusión: que un foco de esa empresa tenga una duración tan corta es poco probable, pero que una muestra relativamente grande tenga un promedio de duración tan bajo es casi imposible.

El intervalo de confianza de $1-\alpha$ es:

$$\bar{X} - Z_{\alpha/2}\sigma/\sqrt{n} < \mu < \bar{X} + Z_{\alpha/2}\sigma/\sqrt{n}$$

EJEMPLO

Para vigilar la calidad de su servicio, una empresa que se dedica exclusivamente a satisfacer pedidos por internet selecciona una muestra aleatoria de clientes cada mes; se establece contacto con cada cliente muestreado, se le hace una serie de preguntas acerca de la calidad del servicio y se determina una calificación de satisfacción por cliente muestreado entre 0 y 100.

Las encuestas mensuales anteriores han demostrado que, aunque la media cambia mes a mes, la desviación estándar de las calificaciones ha tendido a estabilizarse en 20. La última encuesta se hizo a 100 clientes seleccionados aleatoriamente y arrojó una calificación promedio de 82. Estime la calificación para períodos posteriores.

Solución:

Como $\overline{X} = 82$, bajo la suposición de que las calificaciones recibidas son independientes, el intervalo de confianza de 90% es:

$$82 - 1,645\frac{20}{\sqrt{100}} < \mu < 82 + 1,645\frac{20}{\sqrt{100}}$$

$$78,71 < \mu < 85,29$$

De ello puede deducirse que existe un "90% de confianza" de que el valor promedio de las calificaciones que se reciben está entre 78,71 y 85,29. Dicho de otra forma, si se tomaran 100 muestras y a partir de ellas se establecieran intervalos de confianza, aproximadamente 90 de esos intervalos contendrían el verdadero valor de la media poblacional.

Nota: se prefirió hacer un intervalo de confianza de 90%, pero bien pudo ser de otro valor alto (generalmente 95 ó 99%).

Si la población es finita y de un tamaño conocido, debe emplearse un factor de corrección. En este caso, el intervalo de confianza de $1-\alpha$ es:

$$\overline{X} - Z_{\alpha/2}\sigma/\sqrt{n}\sqrt{\frac{N-n}{N-1}} < \mu < \overline{X} + Z_{\alpha/2}\sigma/\sqrt{n}\sqrt{\frac{N-n}{N-1}}$$

Algo muy importante es la elección del tamaño apropiado de la muestra que se ha de tomar, porque si es demasiado grande se desperdicia tiempo y dinero, y si es muy pequeña, las conclusiones resultantes no son muy confiables.

Si se quiere establecer qué tan grande debe ser una muestra para asegurar que el error al estimar μ sea menor a un error predeterminado, debe despejarse *n* del intervalo de confianza, teniendo en cuenta que el error es la parte que se suma o resta de la media muestral. Por lo tanto:

$$n = \left(\frac{Z_{\alpha/2}\sigma}{e}\right)^2$$

Lo anterior implica que el tamaño de la muestra a emplear depende de tres factores: el nivel de confianza que se desea, el margen de error que puede tolerar el investigador y la variabilidad en la población que se estudia.

De la relación, puede deducirse que:

- A medida que disminuye el margen de error, el tamaño requerido de la muestra (*n*) aumenta para un valor fijo de σ y una confianza especificada.

- A medida que σ aumenta, el tamaño requerido de la muestra aumenta para una longitud deseada y una longitud especificada. Dicho de otra forma, si la población tiene una dispersión grande se requerirá una muestra mayor que si la población es homogénea.

- Conforme aumenta el nivel de confianza, el tamaño requerido de la muestra aumenta para una longitud fija deseada y una desviación estándar determinada.

EJEMPLO

¿Qué tan grande se requiere una muestra en el ejemplo anterior si queremos tener 95% de confianza de que el error de estimación no excederá de 2?

Solución:

$$n = \left(\frac{1.96 * 20}{2} \right)^2 = 384,2$$

Esto implica que para satisfacer los requerimientos debe tomarse una muestra de 385 clientes.

» **Varianza desconocida:**

Como se ha mencionado, el caso anterior se presentará poco en la práctica, ya que lo usual es que el valor exacto de los parámetros μ y σ^2 no sean conocidos; de lo contrario, no interesaría buscar intervalos de confianza para ellos.

Si la muestra tomada es grande, un procedimiento aceptable consiste en reemplazar σ por el valor calculado de la desviación estándar muestral.

Cuando el tamaño de la muestra es pequeño (menos de 30 unidades) debe emplearse otro procedimiento; para producir un intervalo de confianza válido debe hacerse una hipótesis más fuerte con respecto a la población de interés y es que ella está distribuida normalmente. Esto conduce a intervalos de confianza basados en la distribución **t de Student,** que es una distribución continua que tiene una forma muy similar a la distribución normal estándar (tiene forma de campana y es

simétrica con una media de 0); una distribución *t* específica depende de un parámetro llamado *grados de libertad*, que para efectos prácticos equivale a *n – 1* (una unidad menos que el tamaño de la muestra). A medida que aumenta la cantidad de grados de libertad, la diferencia entre la distribución *t* y la distribución normal estándar se hace más y más pequeña.

$$T = \frac{\overline{X} - \mu}{\sigma_{n-1} / \sqrt{n}} \text{ con n – 1 grados de libertad}$$

De allí se obtiene un intervalo de confianza dado por:

$$\overline{X} - t_{\alpha/2,n-1}\sigma_{n-1} / \sqrt{n} \le \mu \le \overline{X} + t_{\alpha/2,n-1}\sigma_{n-1} / \sqrt{n}$$

EJEMPLO

La Asociación Americana de Agencias de Publicidad tiene un registro de datos sobre minutos de anuncios por cada media hora de programas principales de TV. En la tabla siguiente vemos una lista de datos de una muestra de programas preferentes en cadenas principales a las 8 P.M.:

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

Determine un intervalo de confianza de 95% para la cantidad promedio de minutos de anuncios en los principales programas de TV a las 8 de la noche.

Solución:

$\overline{X} = 6,525$ y $\sigma_{n-1} = 0,5437$; $t_{0.025,19} = 2,093$

$$6,525 - \frac{2,093 * 0,5437}{\sqrt{20}} \le \mu \le 6,525 + \frac{2,093 * 0,5437}{\sqrt{20}}$$

$$6,27 \le \mu \le 6,78$$

Debe tenerse muy presente que este intervalo de confianza supone que el muestreo se hace sobre una población normal; esta hipótesis es válida en muchas situaciones prácticas, pero si es muy alejada de la realidad deben emplearse métodos diferentes.

Lo anterior se cumple si la población es infinita o muy grande; pero si es finita, el intervalo de confianza se debe calcular como:

$$\overline{X} - t_{\alpha/2,n-1}\frac{\sigma_{n-1}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \overline{X} + t_{\alpha/2,n-1}\frac{\sigma_{n-1}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Si la muestra no es muy pequeña (cosa que usualmente sucede), el intervalo de confianza queda así:

$$\overline{X} - Z_{\alpha/2}\frac{\sigma_{n-1}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \overline{X} + Z_{\alpha/2}\frac{\sigma_{n-1}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Para determinar el tamaño apropiado de la muestra a utilizar es necesario encontrar una estimación preliminar de la desviación estándar de la población. Una forma posible consiste en disponer de la dispersión de otro estudio relacionado (obviamente, si se considera confiable); el método más común para hacerlo consiste en la realización de un estudio piloto, basado en la utilización de una muestra a la cual se le determina su desviación estándar y ese valor es usado para determinar el tamaño apropiado de la muestra.

El tamaño de la muestra para estimar una media poblacional es:

$$n = \left(\frac{Z_{\alpha/2}\sigma_{n-1}}{e}\right)^2$$

Si la población es finita, el tamaño de la muestra que debe seleccionarse es:

$$n = \frac{Z^2\sigma^2 N}{(N-1)e^2 + Z^2\sigma^2}$$

EJEMPLO

Se quiere estimar la media poblacional de los puntajes en Saber Pro de los estudiantes de cierta universidad, 356 estudiantes de esa universidad presentaron la prueba y se desea determinar un intervalo de confianza aproximado de 95%, con anchura máxima de 8 puntos. El año anterior se hizo el mismo estudio y se encontró que la desviación estándar de la muestra que se tomó es 15,5.

a. Calcule el tamaño de la muestra que se debe tomar.
b. Estime μ si se tomó una muestra aleatoria del tamaño apropiado y se halló una media para esa muestra de 105,3 y una desviación estándar de 14,4.

Solución:

$$n = \frac{1,96^2 * 15,5^2 * 356}{355 * 4^2 + 1,96^2 * 15,5^2} = 50$$

El intervalo de confianza es:

$$105,3 - 1,96 \times \frac{14,4}{\sqrt{50}} \times \sqrt{\frac{356-50}{355}} \leq \mu \leq 105,3 + 1,96 \times \frac{14,4}{\sqrt{50}} \times \sqrt{\frac{356-50}{355}}$$

$$101,6 \leq \mu \leq 109$$

5.1.2. ESTIMACIÓN DE LA PROPORCIÓN POBLACIONAL

Si queremos estimar el parámetro *p*, nos debemos basar en un experimento binomial.

Además, se debe tomar como estimador puntual de *p* la proporción de éxitos obtenidos en las *n* pruebas, es decir:

$$\hat{p} = \frac{X}{n}$$

La distribución del número de éxitos es binomial, y puede ser aproximada a la normal cuando el tamaño de la muestra *n* es grande y *p* no es una cantidad muy cercana a cero o uno.

Se sabe también que $\hat{p}$ tiene una distribución aproximadamente normal con

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Si se conocen *p* y *q* no tendría sentido estimarla, pero estas se pueden sustituir por los respectivos estadísticos muestrales, así:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}}$$

Por consiguiente, un intervalo de confianza para p está dado por:

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Un uso práctico al construir un intervalo de confianza basado en la información de una muestra se basa en la comparación de dicho valor con el valor propuesto para el parámetro poblacional; si el valor propuesto está dentro del intervalo, se llega a la conclusión de que el valor propuesto parece ser verdadero.

EJEMPLO

En una encuesta para determinar la popularidad presidencial se pidió a cada integrante de una muestra aleatoria de 1 000 personas que contestara si el presidente estaba haciendo un buen trabajo. Un total de 560 personas respondieron afirmativamente.

a. Construya un intervalo de confianza de 95% para la proporción de personas que consideran que el presidente está haciendo bien las cosas. Interprete.
b. Construya un intervalo de confianza de 90% para la proporción de personas que consideran que el presidente está haciendo bien las cosas. Interprete y compare con el intervalo anterior.

Solución:

a. $\hat{p} = 560/1000 = 0,56$

Por lo tanto, el intervalo de confianza es:

$$0,56 - 1,96\sqrt{\frac{0,56*0,44}{1000}} \leq p \leq 0,56 + 1,96\sqrt{\frac{0,56*0,44}{1000}}$$

$$0,529 \leq p \leq 0,591$$

Lo anterior indica que se puede tener un "95% de confianza" de que entre el 52,9 y el 59,1% de la población considera que el presidente está haciendo un buen trabajo.

b. El intervalo de confianza de 90% es:

$$0,56 - 1,645\sqrt{\frac{0,56*0,44}{1000}} \leq p \leq 0,56 + 1,645\sqrt{\frac{0,56*0,44}{1000}}$$

$$0,534 \leq p \leq 0,586$$

Es decir, se puede tener un "90% de confianza" de que entre el 53,4 y el 58,6% de la población considera que el presidente está haciendo un buen trabajo.

Conclusión: La precisión y el nivel de confianza están relacionados en forma inversa, lo que indica que al aumentar el nivel de confianza, disminuye la precisión.

Lo anterior equivale a señalar que cuando se incrementa el nivel de confianza, el intervalo de confianza se torna más amplio y, por lo tanto, dice menos sobre el parámetro que se quiere estimar.

En este caso tiene un interés particular la selección del tamaño apropiado de la muestra.

Como el error es la parte que se suma y se resta a $\hat{p}$ en el intervalo de confianza, el tamaño apropiado de la muestra es:

$$n = \left(\frac{Z_{\alpha/2}}{e} \right)^2 pq$$

Para utilizar la anterior ecuación se debe hacer una estimación de p. Para ello debemos basarnos en un valor $\hat{p}$ de una muestra anterior o en el establecimiento de $\hat{p}$ a partir de una muestra piloto (y se determina cuántas observaciones adicionales se necesitan para estimar p con una exactitud predeterminada) o haciendo una estimación subjetiva (en este caso, debe conocerse muy bien lo que se hace).

Otro enfoque para seleccionar el tamaño de muestra consiste en maximizar la ecuación, teniendo en cuenta que pq es máximo cuando p y q tienen un valor de 0,5.

Si la población es finita, el tamaño de la muestra que debe seleccionarse es:

$$\text{n} = \frac{Z^2 \hat{p}\hat{q}N}{(N-1)e^2 + Z^2 pq}$$

EJEMPLO

Se va a realizar una investigación sobre "Perfil y percepción del turista extranjero que visita la ciudad de Medellín". Una de las etapas del estudio comprende a los turistas que vinieron a Medellín durante la celebración de la feria de Colombiatex.

Según datos suministrados por los directivos de la Feria, la población esperada de turistas era de 1300 extranjeros; si en los resultados se quiere tener un nivel de confianza de 90% y un margen de error de 3,5%, ¿cuál debe ser el tamaño de la muestra a utilizar?

Solución:

$$n = \frac{1,645^2 * 0,5 * 0,5 * 1300}{1299 * 0,035^2 + 1,645^2 * 0,5 * 0,5} = 388$$

Lo que implica que debe seleccionarse una muestra de 388 turistas.

Nota 1: Generalmente una encuesta incluye variables cualitativas y cuantitativas; si esto sucede, lo más común es determinar el tamaño apropiado de muestra según proporciones.

Nota 2: Es posible obtener intervalos de confianza unilaterales para μ haciendo $l = -\infty$ o $u = \infty$ y reemplazando $Z_{\alpha/2}$ por Z_{α}.

El intervalo de confianza superior del 100(1-α)% para μ es:

$$\mu \leq \overline{X} + Z_{\alpha}\sigma / \sqrt{n}$$

Y el intervalo de confianza inferior del 100(1-α)% para μ es:

$$\overline{X} - Z_{\alpha}\sigma / \sqrt{n} \leq \mu$$

EJEMPLO

Una entidad bancaria seleccionó una muestra aleatoria de 30 días y registró el número de reclamos de sus usuarios en cada uno de ellos. Dichos registros fueron:

17	24	12	28	15	19	23	32	15	18
24	33	12	15	21	24	36	18	15	19
21	24	29	17	18	24	30	14	21	18

Establezca un intervalo de confianza unilateral, de tal manera que tenga una confianza aproximada de 90% de que el promedio de reclamos diarios en la entidad no supere un valor determinado.

Solución:

$\bar{X} = 21,2$ y $\sigma_{n-1} = 6,332$; $Z = 1,28$

$$\mu \leq 21,2 + \frac{1,28 * 6,332}{\sqrt{30}} \quad \Rightarrow \quad \mu \leq 22,7 \text{ reclamos}$$

Es decir, se tiene una confianza de 90% en que el promedio de reclamos no supere los 22,7 por día.

Si el tamaño de la población es finito, el intervalo de confianza para proporciones viene dado por

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}}$$

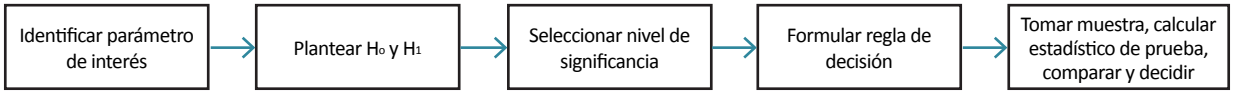
5.2. PRUEBAS DE HIPÓTESIS

Una hipótesis se define como una explicación tentativa o suposición adoptada provisionalmente para explicar ciertos hechos y guiar la investigación de otros.

La hipótesis es llamada estadística si consiste en un enunciado respecto a un parámetro de una o más poblaciones (μ, σ, p).

Con frecuencia es deseable probar la validez de tales hipótesis. A fin de lograrlo se lleva a cabo un experimento y la hipótesis es rechazada si los valores obtenidos para el equivalente muestral del parámetro son tan grandes o tan pequeños que la probabilidad de que ocurran si la hipótesis es cierta es bastante remota; si los datos son muy probables, la hipótesis es "no rechazada".

Para hacer una prueba de hipótesis deben llevarse a cabo ciertos procedimientos, que son detallados en el siguiente esquema:



- **Identificar parámetro de interés:** lo primero que debe hacerse es establecer claramente, según el contexto, cuál es el parámetro de interés (si es la proporción, la media o la desviación estándar).

- **Plantear la hipótesis nula y la alternativa:** en cualquier investigación deben plantearse dos hipótesis, que se denominan hipótesis nula (H_0) e hipótesis alternativa (H_1) y que de alguna manera reflejarán esa idea a priori que tenemos y que pretendemos contrastar con la "realidad".

  La hipótesis nula se denomina así por ser el punto de partida y siempre ha de incluir una igualdad; es la hipótesis que se trata de contrastar, de forma que al final del proceso, la rechazaremos o no. La hipótesis alternativa es el complemento de la nula; por lo tanto, el rechazo de la hipótesis nula supone el no rechazo de la hipótesis alternativa.

  Según lo anterior, la hipótesis nula puede ser con $=$, $\leq$ o $\geq$. En los respectivos casos, la hipótesis alternativa debe ser con $\neq$, $>$ o $<$.

  La única manera de asegurar la veracidad o falsedad de una hipótesis estadística con certeza absoluta consiste en evaluar toda la población. Como lo que se hace generalmente es tomar una muestra, solamente puede hablarse de sospechas y, por ende, siempre existe la posibilidad de una conclusión errónea. El rechazo de H_0 significa simplemente que hay una pequeña probabilidad de obtener la información muestral observada cuando, de hecho, la hipótesis es verdadera y no necesariamente que dicho planteamiento sea falso.

  La decisión de rechazar o no la hipótesis nula está basada en la elección de una muestra tomada al azar y por tanto es posible tomar decisiones erróneas. Los errores que se pueden cometer se clasifican como sigue:

  » Si se rechaza la hipótesis nula cuando realmente es verdadera. Esto se conoce como error tipo I; la probabilidad de cometerlo se simboliza con α y es lo que se conoce como nivel de significancia.

  » Si no se rechaza la hipótesis nula cuando realmente es falsa. Esto se conoce como error tipo II y la probabilidad de cometerlo se denota como β.

  En lenguaje estadístico:

  α = P(error tipo I) = P(rechazar H_0\ H_0 es V)

  β = P(error tipo II) = P(aceptar H_0\ H_0 es falsa)

  Los errores tipo I y II están relacionados: una disminución en la probabilidad de cometer uno de ellos siempre da como resultado un aumento en la probabilidad del otro, siempre que el tamaño muestral sea constante. En general, si el tamaño de muestra se aumenta, se reducen tanto α como β.

EJEMPLO

Una agencia de viajes planea desarrollar una propaganda para televisión si el promedio de clientes es por lo menos 15 al día en temporada fría. Para estimar dicho promedio, tomó una muestra aleatoria de 25 días de estos últimos tres meses (temporada fría); los datos que encuentre se usarán para probar las siguientes hipótesis: H_0: $\mu \geq 15$; H_1: $\mu < 15$

a. ¿Cuál es el error de tipo I en este caso? ¿Cuáles serían sus consecuencias?
b. ¿Cuál es el error de tipo II en este caso? ¿Cuáles serían sus consecuencias?

Solución:

a. Existe un error de tipo I cuando se concluye que dicho promedio no es mayor o igual a 15 clientes cuando realmente lo es. Podría perderse la oportunidad de proyectar el comercial y ganar nuevos clientes.

b. Existe un error de tipo II cuando se concluye que dicho promedio es mayor o igual a 15 clientes cuando realmente no lo es. En ese caso, se proyectará el comercial cuando el plan inicial indica que no se justificaría.

- **Seleccionar nivel de significancia:** refleja la probabilidad de que el estadígrafo caiga por fuera de región de aceptación. Se elige subjetivamente (no es producto de ningún cálculo), pero debe ser un valor bajo, aproximadamente 0,05.

- **Formular la regla de decisión:** con base en el nivel de significancia, en el valor hipotético del parámetro y asumiendo que los datos poblacionales siguen una distribución normal, deben establecerse una región crítica (o de rechazo) y una región de aceptación. Como se había dicho antes, si la muestra tomada es muy pequeña (de tamaño inferior a 30) no debe utilizarse la distribución normal, sino la distribución *t* de Student.

- **Tomar una muestra, calcular estadístico de prueba, comparar y decidir:** finalmente debe tomarse una muestra, preferiblemente aleatoria para garantizar una buena estimación de los parámetros poblacionales, y se calculan los datos necesarios para el contraste, es decir, la media de la muestra o su proporción (también podría ser su desviación estándar, pero no se está considerando el caso).

De acuerdo a la hipótesis que se esté probando, el estadístico de prueba puede ser:

$$Z_{muestra} = \frac{\bar{X} - \mu}{\sigma_{n-1}/\sqrt{n}} \quad \text{(para media)}$$

$$Z_{muestra} = \frac{\hat{p} - p}{\sqrt{\frac{p \times q}{n}}} \quad \text{(para proporción)}$$

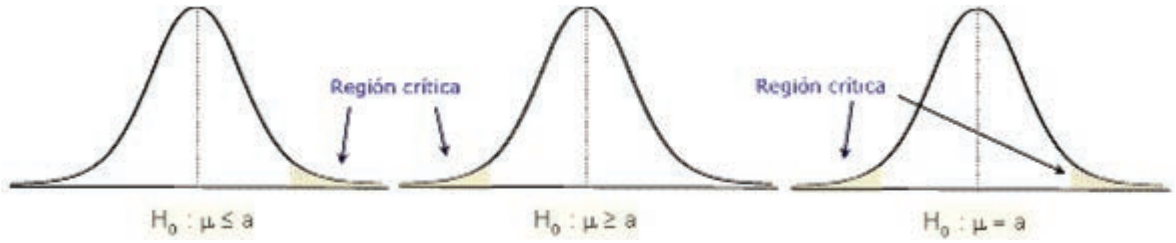
(en cualquiera de los casos puede emplearse su equivalente en *t*)

Debe comprobarse si el resultado del estadístico de prueba cae en la región crítica o en la de aceptación y, según eso, se rechaza o no la hipótesis nula, respectivamente.

HIPÓTESIS UNILATERALES Y BILATERALES

Una prueba de hipótesis se hace bilateral si es importante detectar diferencias a partir del valor hipotético del parámetro que se encuentren a su derecha o a su izquierda; en este caso la hipótesis alternativa se plantea con ≠, por ejemplo, $\mu \neq \mu_0$. En una prueba de este tipo, la región crítica se separa en dos partes, generalmente con la misma probabilidad en cada cola, por lo que el nivel de significancia se divide en ambos extremos.

Si la afirmación implica alguna dirección, es decir, si lo que se quiere demostrar es que el parámetro es mayor o menor que un determinado valor, lo apropiado es hacer una prueba unilateral. Si la hipótesis nula es $\mu > \mu_0$, la región de rechazo debe encontrarse en la cola inferior y, si es $\mu < \mu_0$, la región crítica se encontrará en la cola superior, así:



Fuente: educastur.princast.es

EJEMPLO

Una empresa comercializa una bebida refrescante en un envase en cuya etiqueta se puede leer: "Contenido: 250 centímetros cúbicos". El Departamento de Consumo toma aleatoriamente 25 envases y estudia el contenido medio, obteniendo una media de 234 cc y una desviación típica de 18 cc. ¿Puede afirmarse, con un nivel de significancia de 5%, que no se está vendiendo el contenido indicado?

(Ejercicio tomado de educastur.princast.es, se hicieron algunos ajustes)

Solución:

H_0: $\mu \geq 250$ cc $\qquad$ H_1: $\mu < 250$ cc

Si $\alpha = 0{,}05$ y $v = 24$, entonces la región de aceptación es: $[-1{,}711;\infty)$

$$t = \frac{\overline{X} - \mu}{\sigma_{n-1} / \sqrt{n}} \quad \Rightarrow \quad t = \frac{234 - 250}{18 / \sqrt{25}} = -4{,}44$$

Como el valor de t cae por fuera de la región de aceptación, parecería que no se está envasando lo que se dice.

EJEMPLO

Un contador cree que los problemas de flujo de efectivo de una empresa son resultado directo del lento proceso de cobro de las cuentas por cobrar. Dice que al menos el 70% de las actuales cuentas por cobrar tienen más de dos meses; una muestra de 120 cuentas por cobrar indica que hay 78 con más de dos meses. Pruebe la afirmación del contador con un nivel de significancia de 0,05.

Solución:

H_0: $p \geq 0{,}7$ $\qquad$ H_1: $p < 0{,}7$

$$\hat{p} = \frac{78}{120} = 0{,}65$$

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{0{,}65 - 0{,}7}{\sqrt{\frac{0{,}7 * 0{,}3}{120}}} = -1{,}20$$

Si $\alpha = 0{,}05$, la región de aceptación estará dada por $[-1{,}645, \infty)$. Como el valor de Z para la muestra cae dentro del rango, no hay suficiente evidencia para afirmar que lo que afirma el contador es falso.

5.3. OBSERVACIONES PAREADAS

Es un caso muy común, en el cual se pretende establecer diferencias entre dos poblaciones; en este caso, cada unidad experimental recibe ambas condiciones poblacionales. Por ejemplo, para probar una nueva dieta pueden compararse los pesos "antes" y "después" de cada uno de los individuos de una muestra.

Para la prueba de hipótesis, el interés recae sobre la distribución de las diferencias en el valor calculado. Para expresarlo más formalmente, se investiga si la media de la distribución de las diferencias de los valores calculados es o; por lo tanto, la muestra se construye de las diferencias para cada unidad experimental.

Para hacer la prueba: $$t = \frac{\bar{d}}{S_d / \sqrt{n}}$$

donde $\bar{d}$ es la media de las diferencias, S_d es su desviación estándar y n es el número de observaciones.

EJEMPLO

Ahorros y Préstamos X contrata dos firmas, A y B, para evaluar las propiedades sobre las que hace préstamos; es importante que estas dos firmas tengan avalúos semejantes. Para verificar la congruencia, X selecciona diez casas de manera aleatoria y pide a cada una de las firmas que haga su avalúo. Los resultados, en millones de pesos, son:

Propiedad	A	B
1	135	128
2	110	105
3	131	119
4	142	140
5	105	98
6	130	123
7	131	127
8	110	115
9	125	122
10	149	145

Con un nivel de significancia de 0,05 ¿puede concluirse que las dos firmas hacen avalúos diferentes?

Solución:

El primer paso consiste en establecer las hipótesis nula y alternativa; en este caso es apropiado utilizar una prueba de dos colas porque el interés estriba en determinar si existe diferencia entre los valores calculados.

$$Ho: \mu_d = 0 \qquad H_1: \mu_d \neq 0$$

Las diferencias son 7, 5, 12, 2, 7, 7, 4, -5, 3 y 4, cuya media es 4,6 y cuya desviación estándar es 4,402.

Por lo tanto el valor de t es: $t = \dfrac{4,6}{4,402/\sqrt{10}} = 3,305$

Como la t calculada cae en la región de rechazo, se rechaza la hipótesis nula y se concluye que hay diferencia entre los valores medios calculados de las propiedades.

EJERCICIOS PROPUESTOS

1. Se está desarrollando una investigación entre pequeñas y medianas empresas industriales del Valle de Aburrá. Fueron censadas 856 empresas de confecciones, 1057 empresas de alimentos, 165 empresas metalúrgicas y 1456 empresas de otros sectores. ¿Cuál debe ser el tamaño de la muestra seleccionada si en la investigación se quiere tener un nivel de confianza de 95% y un margen de error de 4%?

2. El siguiente ejercicio fue tomado de las pruebas Ecaes de 2006: Seleccione la respuesta correcta y justifique su elección:

   El gerente de una empresa que realiza estudios de mercado de diferentes productos presta gran atención a los detalles. En el informe final de uno de estos estudios se lee que la muestra fue de 200 personas y que se trabajó con nivel de confianza de 85% y con un margen de error del 10%. Una vez leído el informe, el gerente decide no aceptarlo porque:

   a. El nivel de confianza y el margen de error no suman 100%
   b. El menor nivel de confianza permitido es 95%
   c. El margen de error no supera el 10%
   d. El nivel de confianza es bajo y el margen de error es alto

3. Cierta compañía tiene 168 representantes de ventas que venden un producto suyo en todo el territorio nacional; dichos agentes tienen como ciudad sede a Medellín, Bogotá o Cali, ciudades donde se encuentran las fábricas.

   En la siguiente tabla se muestra el número de productos vendidos por una muestra aleatoria de algunos vendedores durante el mes anterior:

265	188	109	214	233	186	199	239
333	169	205	293	314	222	196	305
125	242	188	314	167	198	225	233
155	197	215	309	356	188	214	268

   Construya un intervalo de confianza de 95% para la cantidad promedio de productos vendidos por cada vendedor. ¿Qué pasará con el margen de error si el tamaño de la muestra se aumenta?

4. Se quiere evaluar el monto de las exportaciones de pequeñas empresas del sector de confecciones de Medellín mediante una muestra. ¿Qué tan grande debe ser la muestra que se tome para tener un nivel de confianza de 90% y un margen de error de 5? Por datos de Cámara de Comercio se sabe que en la

ciudad hay aproximadamente 2150 empresas de este tipo. ¿Tendría sentido tomar esa muestra o sería conveniente hacer un censo?

5. Desea estimarse el grado de favorabilidad de los grandes ejecutivos colombianos frente a los tratados de libre comercio.

   a. Se hizo una encuesta a 500 de ellos y el 41,2% mostró una actitud favorable. Un censo que se desarrolló a principios de este año mostró que en el país hay 27 250 grandes ejecutivos; estime con un 90% de confianza su grado de favorabilidad frente a los TLC.
   b. Con ese nivel de confianza y teniendo en cuenta el tamaño poblacional encuentre el tamaño de muestra apropiado para que el error muestral no sea mayor a 0,015.

6. Se quiere estimar la proporción de egresados de cierta universidad que consiguen empleo en su área de estudios.

   a. La universidad tiene 2250 egresados y se tomó una muestra del 10%; se encontró que 172 de esos egresados han conseguido empleo. Calcule un intervalo de confianza, con un nivel de confianza que considere pertinente, para la verdadera proporción de egresados que está trabajando.
   b. Con el mismo nivel de confianza, encuentra el tamaño de muestra requerido para que el error no sea superior a 0,02.

7. Va a realizarse una encuesta entre los estudiantes de pregrado del Ceipa para evaluar qué porcentaje de ellos piensa hacer una especialización en la misma institución.

   a. En el Ceipa hay 1371 estudiantes de pregrado; si en los resultados se quiere tener un nivel de confianza de 90% y un margen de error de 4%, ¿cuál debe ser el tamaño de la muestra a utilizar?
   b. Si ya se hizo esa encuesta al número de personas seleccionadas en el ítem a) y se encontró que el 31,5% de ellas quiere hacer una especialización en el Ceipa, establezca un intervalo de confianza de 90% para la proporción de estudiantes de pregrado del Ceipa que pretenden cursar una especialización allí.

8. Algunas pequeñas y medianas empresas exportan parte de su producción a países del Grupo Andino. Para estimar el monto de esas exportaciones se hizo un muestreo entre 100 empresas tomadas aleatoriamente y se investigó a cuánto ascendían sus exportaciones en el último mes.

   En dicha muestra se encontró un promedio de 7500 dólares y una desviación estándar de 3260 dólares.

a. Encuentre un intervalo de confianza del 90% para la verdadera media.
b. ¿Cuál sería ese intervalo si el nivel de confianza es 95% y el tamaño muestral no cambian? Concluya
c. ¿Cuál sería ese intervalo si el nivel de confianza sigue siendo 90%, pero el tamaño de muestra se duplica? Concluya
d. Si se pretende que el error no sea superior a 500 dólares, ¿cuántas empresas se deben muestrear si se quiere tener un nivel de confianza del 90% y en un sondeo anterior se encontró una desviación estándar de 2800 dólares?

9. En un estudio pretende evaluarse el grado de favorabilidad de los grandes ejecutivos colombianos frente a los tratados de libre comercio. Se hizo una encuesta a 500 de ellos y el 41,2% mostró una actitud favorable; un censo que se desarrolló a principios de este año mostró que en el país hay 27 250 grandes ejecutivos.

   En la revista Semana de la quincena anterior decía que por lo menos el 60% de los grandes ejecutivos colombianos estaba a favor del TLC. Según lo encontrado en la encuesta, ¿cree que es cierto lo que se dice allí? Para dar su respuesta, realice la prueba de hipótesis correspondiente.

10. Se quiere evaluar la proporción de egresados de cierta universidad que consiguen empleo en su área de estudios. La universidad tiene 2250 egresados y se tomó una muestra del 10%; se encontró que 172 de esos egresados han conseguido empleo.

    La publicidad de esa universidad asegura que por lo menos el 90% de los egresados consiguen empleo en el área. Según lo observado en la muestra, ¿parece que eso es cierto?

11. Aprovechando un tratado firmado recientemente, algunas pymes de alimentos han comenzado a exportar a países de Lejano Oriente. Las exportaciones de 25 de esas empresas seleccionadas aleatoriamente ascienden a los siguientes montos (en miles de dólares) en el período comprendido entre el 16 de octubre y el 15 de noviembre:

158	145	235	85	195	215	182	175	100	135
120	90	235	212	196	169	250	242	188	200
225	180	135	242	191					

    El Presidente de la República afirmó ayer que el promedio de exportaciones de empresas de las características señaladas era de 200 000 dólares. Según esa muestra, ¿parecería que eso es cierto?

12. El presidente de cierta compañía cree que el 30% de los pedidos a su empresa provienen de clientes nuevos. De 80 pedidos elegidos al azar, se encontró que 17 fueron hechos por clientes nuevos. A partir de un análisis bilateral, compruebe la veracidad de la hipótesis.

13. La Cámara de Comercio afirma que las microempresas de Medellín presentan ventas totales promedio de 12 millones de pesos por día. Para comprobarlo, se tomó una muestra aleatoria de 25 microempresas y se registraron sus ventas promedio (en millones), así:

15,5	7,6	12,5	23,1	8,2	11,4	15,3	9,1	6,9	10,5	12,4	9,3	12,5
21,1	12,8	15,5	13,9	6,5	20	15,5	11,1	10,5	9,6	15,2	12	

Con un nivel de significancia de 0,05, ¿parece que es cierto lo que afirma la Cámara de Comercio?

14. El vicepresidente de recursos humanos de una gran compañía industrial ha observado un aumento del ausentismo de sus empleados. Hace un año, tratando de mejorar la situación, comenzó con un programa de acondicionamiento físico. Para evaluar el programa seleccionó una muestra aleatoria de 12 participantes y registró el número de días que faltó cada uno de ellos antes y después del programa. A continuación se presentan los resultados; ¿puede concluirse que ha disminuido el número de ausencias?

EMPLEADO	ANTES DEL PROGRAMA	DESPUÉS DEL PROGRAMA
1	6	5
2	6	2
3	7	1
4	7	3
5	4	3
6	3	6
7	5	3
8	6	7
9	5	5
10	2	0
11	3	1
12	1	1

15. Se va a estimar el promedio del Producto Interno Bruto por persona de los países del mundo. En el mundo hay 225 países y se tomó para el estudio una muestra de 30 países.

    En los países seleccionados en la muestra se registraron los siguientes PIB per cápita en 2010 (en dólares):

12700	34000	15900	6600	30000	500	24200	36600	1900	5500
14700	23200	3400	11500	47200	35700	6200	17400	1500	9200
23200	8000	38500	7000	2700	3100	1800	300	1900	42600

    a. Estime, mediante un intervalo de confianza, el PIB promedio de los países del mundo.

    b. Una investigación de 2009 reveló un promedio de 16 500 dólares para el producto interno bruto per cápita. Según lo reportado por la muestra, ¿parecería que ese promedio se conserva?

BIBLIOGRAFÍA

Anderson, D., Sweeney D., Williams, T. (2008). *Estadística para Administración y economía.* México, D.F.: Cengage Learning Editores.

Ángel, B. (2010). Cartilla de Investigación y empresarimo. Sabaneta: Institución Universitaria Ceipa.

Banco de la República. Series estadísticas. Recuperado el 25 de julio de 2011 de www.banrep.gov.co

Centro de Investigación Estadística y Mercadeo. *Series, Enfoque clásico.* Recuperado el 5 de mayo de 2011 de http://www.slidefinder.net/s/series_clasico_ciem/20750047

Educastur. *Test de contraste de hipótesis*. Recuperado el 5 de agosto de 2008 de http://web.educastur.princast.es/ies/pravia/carpetas/recursos/mates/Descartes_antiguo/curso_inferencia_estadistica/tests_de_contraste_de_hipotesis.htm

Indexmundi. Mapa comparativo de países. Recuperado el 20 de agosto de 2010 de http://www.indexmundi.com/map/?t=0&v=71&r=sa&l=es

Jaramillo F. (2003). *Distribuciones de probabilidad*. Recuperado el 4 de junio de 2008 de estadisticaiue.awardspace.com/est1ce/t_dispro.doc

Montgomery D. y Runger G. (1996). *Probabilidad y estadística aplicadas a la ingeniería.* México, D.F.: McGraw-Hill Interamericana Editores.

Programa de las Naciones Unidas para el Desarrollo (2011). Informe sobre Desarrollo Humano 2010. Recuperado el 24 de junio de 2011 de http://www.pnud.org.co/sitio.shtml?x=63277

Scribd (2008). Recuperado en noviembre de 2009 de http://es.scribd.com/doc/30215787/ECAES-ADMON

Skyscraper Life (2011). Recuperado el 25 de julio de 2011 de http://www.skyscraperlife.com/latin-bar/51018-indice-de-libertad-economica-2011-a.html

Walpole R., Myers R. y Myers S. (1999). *Probabilidad y estadística para ingenieros*. México, D.F.: Pearson Prentice Hall.

ANEXO 1
USO DE LAS FUNCIONES ESTADÍSTICAS DE LA CALCULADORA

Las siguientes pautas se cumplen para la mayoría de modelos de calculadoras científicas sencillas, pero podría variar según el modelo. Si es el caso, debe acudirse al manual de la calculadora.

1. OBTENCIÓN DE MEDIA Y DESVIACIÓN ESTÁNDAR:

- ✓ Debe usarse el Modo SD (Mode SD). En algunas calculadoras aparece Mode Stat.
- ✓ Borrar datos anteriores: para ello debe usarse Scl. En algunas calculadoras se ve directamente en el teclado; de lo contrario, se encuentra con CLR (clear).
- ✓ Introducir nuevos datos: debe digitarse cada dato y confirmarlo con DT.

  Por ejemplo, si desea hallarse media y desviación estándar de los siguientes datos: 2, 5 y 8, debe escribirse 2 DT, 5 DT, 8 DT
- ✓ Para hallar media y desviación estándar: en algunas calculadoras se encuentran en el teclado $\bar{X}$ (media), $x\sigma_n$ (desviación estándar poblacional) y $x\sigma_{n-1}$ (desviación estándar muestral). Si no están en el teclado, deben buscarse por S-VAR.

2. ANÁLISIS DE REGRESIÓN

- ✓ Debe usarse el Modo REG (Mode REG). En ciertas calculadoras aparece Mode LR.
- ✓ Seleccionar el modelo adecuado, según el caso: Lin (lineal), Log (logarítmico), Exp (exponencial), Quad (cuadrático), Pow (potencial).

  Se escoge uno de ellos según la forma del gráfico de dispersión; si se tiene duda entre varios modelos, debe hallarse el coeficiente de correlación (r) para cada uno y elegir el modelo cuyo r sea más cercano a 1 en valor absoluto.

- ✓ Borrar datos anteriores: para ello debe usarse Scl. En algunas calculadoras se ve directamente en el teclado; de lo contrario, se encuentra con CLR (clear).
- ✓ Introducir nuevos datos: digita cada dato de la forma *x,y* (primero el respectivo valor de la variable independiente y luego el valor de la variable dependiente, separados por coma) y lo confirmas con DT.

  Por ejemplo, si deseas introducir los siguientes datos:

X	Y
5	2
8	6
10	8

  Debes escribir 5,2 DT, 8,6 DT, 10,8 DT

- ✓ Para hallar A, B, r, $\hat{x}$ o $\hat{y}$: en algunas calculadoras se encuentran en el teclado; de lo contrario, deben buscarse por S-VAR.

3. COMBINACIONES Y PERMUTACIONES:

Para las combinaciones se utiliza la tecla nCr; para buscar una combinación se debe ingresar primero el valor de n (tamaño del conjunto), luego la tecla mencionada y finalmente el valor de r (tamaño del subconjunto).

Para las permutaciones se utiliza la tecla nPr; para buscar una permutación se debe ingresar primero el valor de n (tamaño del conjunto), luego la tecla mencionada y finalmente el valor de r (tamaño del subconjunto).

Si la calculadora es de un modelo más reciente, debe ingresarse por el modo STAT.

1. OBTENCIÓN DE MEDIA Y DESVIACIÓN ESTÁNDAR:

- Luego de elegir el modo STAT, debe seleccionarse 1-VAR y luego teclear EXE.
- Después deben introducirse los valores de x en forma matricial, digitando EXE después de cada valor.
- Una vez introducidos todos los valores, se escoge Shift S-MENU y en ese menú se escoge VAR.

- Finalmente, se selecciona lo que se necesite: n (número de datos introducidos), $\bar{X}$ (media), $x\sigma_n$ (desviación estándar poblacional) y $x\sigma_{n-1}$ (desviación estándar muestral).

2. ANÁLISIS DE REGRESIÓN

- Luego de elegir el modo STAT, debe escogerse A+BX si el modelo es lineal, lnX si es logarítmico o e^x si es exponencial. Luego se teclea EXE.

- Después deben introducirse los valores de x en forma matricial, digitando EXE después de cada valor.

- Una vez introducidos todos los valores, debe seleccionarse Shift S-MENU y en ese menú escoger REG.

- Por último, seleccionar lo que se necesite: A, B, r, $\hat{x}$ o $\hat{y}$

3. COMBINACIONES Y PERMUTACIONES:

Si las funciones nCr y nPr no se encuentran directamente en el teclado, debe buscarse en el catálogo (CTLG); algunas veces se encuentran como C y P, respectivamente.

ANEXO 2
PRESENTACIÓN DE INFORMACIÓN CON EXCEL

1. VARIABLES CUALITATIVAS O CUANTITATIVAS DISCRETAS:

1.1. TABLA DE FRECUENCIAS:

- Introducir los datos en una columna de la hoja de cálculo (también podría ser en una fila o en varias columnas).

- Digitar las clases y estructurar la tabla, así:

	Frecuencia	Frecuencia relativa	Porcentaje
xxxxx			
yyyyy			

- Ubicado en la celda *Frecuencia de xxxxx,* debe seleccionarse el menú **Fórmulas, Insertar función,** posteriormente la categoría **Estadísticas** y la función **CONTAR.SI**

- En Contar.si aparece:

  Rango: son los datos que se quieren contar. Puede utilizarse cualquiera de dos formas para elegirlo:

  » Seleccionar rango de datos escribiendo las celdas de comienzo y fin, separadas por dos puntos.
  » O pueden asignarse los datos, seleccionándolos. Debe tenerse cuidado de estar ubicado en el rectángulo respectivo.

  Criterio: Son los resultados posibles.

En Excel deben fijarse las celdas que no cambian cuando se hace un arrastre de fórmula. Para ello, debe ubicarse el cursor en la barra de fórmulas al comienzo del espacio que muestra la ubicación de la celda que quiere cambiarse y pulsar F4; lo mismo se hace para la celda final. Queda, entonces, algo de la siguiente forma:

=C1/C5

(está fijada la celda C5)

Finalmente debe completarse la tabla de frecuencia, haciendo lo siguiente:

» Seleccionar las frecuencias absolutas existentes y hacer clic en el símbolo ∑ de la barra de herramientas.
» Luego, para completar la primera frecuencia relativa, debe ubicarse en la celda donde debe obtenerse el primer resultado, digitar el símbolo igual, señalar la casilla de la respectiva frecuencia absoluta y dividir por el total de las frecuencias absolutas (este número debe fijarse como se explicó antes).
» Para obtener las demás frecuencias relativas debe señalarse la casilla en la que se obtuvo la respuesta y ubicarse en el vértice inferior derecho. Cuando aparezca un signo positivo debe arrastrarse hacia abajo con el clic izquierdo hundido hasta la celda en la cual se quiera asignar algún valor.
» Por último, para completar los porcentajes, debe multiplicarse la primera frecuencia relativa por 100 y arrastrar.

EJEMPLO

Datos:

SI	**NO**	**SI**	**SI**	**SI**	**SI**	**NO**	**SI**	**SI**	**SI**
NO	**SI**	**SI**	**SI**	**NO**	**NO**	**SI**	**SI**	**SI**	**NO**

(Digitar en la columna 1 de la hoja de Excel)

A un lado crea la estructura de la tabla de frecuencia, así:

	Frecuencia absoluta	Frecuencia relativa	Porcentaje
SÍ			
NO			

Ubicado en la celda donde debe encontrarse la frecuencia absoluta de "Sí", ir a insertar función, Categoría **Estadísticas**, Función **Contar.si**. Aparece entonces un cuadro de diálogo con los siguientes ítems:

RANGO: Se asignan los datos.
CRITERIO: Asignar la celda que contiene SÍ.

Para completar la tabla:

- Fijar las celdas correspondientes al rango de datos.

- Arrastrar la frecuencia absoluta de "Sí" para completar las otras frecuencias; para ello debe señalarse la casilla en la que se obtuvo la respuesta y ubicarse en el vértice inferior derecho. Cuando aparezca un signo positivo debe arrastrarse hacia abajo con el clic izquierdo presionado hasta la celda en la cual se quiera asignar algún valor.
- Seleccionar las frecuencias absolutas existentes y hacer clic en el símbolo ∑ de la barra de herramientas.
- Luego, para completar la primera frecuencia relativa, debe ubicarse en la celda donde quiere obtenerse el primer resultado; digitar el símbolo igual, señalar la casilla de la respectiva frecuencia absoluta y dividir por el total de las frecuencias absolutas (este valor debe fijarse).
- Para obtener las demás frecuencias relativas debe arrastrarse, tal como se señaló antes. Por último, para completar los porcentajes, multiplicar cada frecuencia relativa por 100.

Debe obtenerse la siguiente tabla:

	Frecuencia absoluta	Frecuencia relativa	Porcentaje (%)
SI	14	0.7	70
NO	6	0.3	30

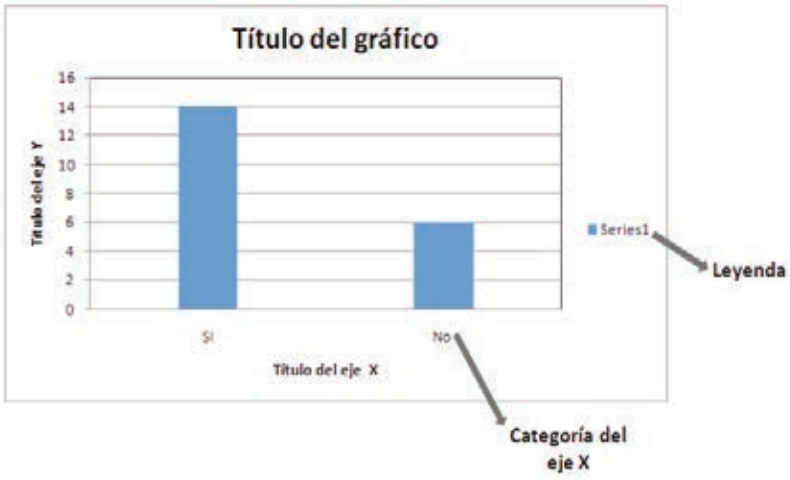
1.2. GRÁFICOS:

1.2.1. Gráfico de columnas:

Paso 1: Seleccionar los datos que se desea graficar (en este caso las clases y las frecuencias absolutas).

Paso 2: Dar clic en el menú **Insertar** y luego en la flecha que está en el inferior de Columnas y ahí escoger el subtipo que se prefiera.

Paso 3: Luego de que aparezca el gráfico, escoger en **Diseño del gráfico** aquel que más se ajuste (preferiblemente aquellos que tienen título y nombre de ejes).



Nota: Si las dos variables son numéricas, Excel graficará ambas variables. Para eliminar la que no deba graficarse, debe señalarse con clic derecho una de las barras correspondiente a la variable no deseada y se escoge Eliminar.

- Si se prefiere, pueden borrarse las líneas de división, señalando alguna y suprimiendo.

- Si aún se ve la leyenda, debe eliminarse si no aporta nada.

- Si se quiere cambiar el formato de las barras:
  - » Clic derecho sobre cualquiera de ellas (verificar que queden seleccionadas todas las barras).
  - » Dar formato a series de datos, relleno.
  - » Cambiar según lo deseado (color, efectos de relleno, etc).

- Si se quiere cambiar una sola barra:
  - » Hacer dos veces clic en la barra que quiere cambiarse (debe quedar seleccionada solamente la que se necesita).
  - » Clic derecho en la requerida.
  - » Formato de punto de datos y se hace el cambio deseado.

- Si desea hacerse un cambio en cualquiera de las secciones señaladas en el gráfico, debe darse clic derecho sobre la respectiva sección y seguir las indicaciones.

1.2.2. Gráfico de Sectores:

Paso 1: Seleccionar los datos que se van a graficar (las clases con las frecuencias absolutas o las relativas).

Paso 2: Insertar tipo de gráfico circular y escoger el que se prefiera.

Paso 3: Escoger en diseño de gráfico el que se prefiera (Es mejor con título, leyenda y rótulo de datos).

Para editar el gráfico:
» Doble clic sobre el área que quiera cambiarse.
» Clic derecho sobre esa área.
» Elegir **Formato de punto de datos**.
» Cambiar según lo deseado.

1.3. MEDIDAS DE RESUMEN:

- Digitar o pegar los datos en la columna A de la hoja de cálculo (solamente con numéricos).
- Seleccionar la barra de menús llamada **Datos**.
- Elegir la opción **Análisis de Datos**.
- Escoger **Estadística Descriptiva**.
- Cuando aparezca el cuadro de diálogo:
  - » Ubicados en el cuadro Rango de entrada ir y seleccionar los datos de la hoja de cálculo o teclear de donde a donde se ubican, por ejemplo: A1:A50
  - » Clic en la opción rango de salida, ubicados allí, seleccionar la celda donde quiere que se ubiquen los datos, ejemplo C1.
  - » Seleccionar **Resumen de Estadísticas**.
  - » Aceptar.
- Por defecto, Excel arroja resultados para las principales medidas; si se quiere adicionar otra medida, se debe insertar función y entre las funciones estadísticas se escoge la requerida (**Desvestp** para desviación estándar poblacional, **Cuartil** para cuartiles) o se hace la operación requerida (por ejemplo, para el coeficiente de variación).

2. VARIABLES CUANTITATIVAS CONTINUAS:

2.1. TABLA DE FRECUENCIAS:

- Digitar o pegar los datos en una columna de la hoja de cálculo.
- Si se prefiere, digitar en otra columna los límites superiores de los intervalos deseados. De lo contrario, el programa determina esos intervalos.
- Seleccionar la opción análisis de datos de la barra de menú datos
- Elegir la opción Histograma.
- Aceptar

- Asignar rango de entrada (los datos), rango de clases y rango de salida.
- Aceptar.

Nota: En la tabla de frecuencia aparecen únicamente los límites superiores del intervalo; es más recomendable digitar todo el intervalo.

2.2. GRÁFICO (Histograma):

Hacer diagrama de columnas, pero en el menú herramientas de gráficos **Diseño** debe seleccionarse **Cambiar entre filas y columnas**.

En caso de no tener habilitada la opción análisis de datos, debe hacer lo siguiente:

» Hacer clic en el **Botón Microsoft Office**  y, a continuación, hacer clic en **Opciones de Excel**.
» Hacer clic en **Complementos** y, en el cuadro **Administrar**, seleccionar **Complementos de Excel**.
» Hacer clic en **Ir**.
» En el cuadro **Complementos disponibles**, activar la casilla de verificación **Herramientas para análisis** y, a continuación, hacer clic en **Aceptar**.
» **Sugerencia:** Si **Herramientas para análisis** no aparece en la lista del cuadro **Complementos disponibles**, hacer clic en **Examinar** para buscarlo.
» Si se indica que Herramientas para análisis no está instalado actualmente en el equipo, hagcer clic en **Sí** para instalarlo.
» Una vez cargado Herramientas para análisis, el comando **Análisis de datos** estará disponible en el grupo **Análisis** de la ficha **Datos**. En algunos casos es necesario reiniciar el equipo.

ANEXO 3
REGRESIÓN Y CORRELACIÓN CON EXCEL

1. ELABORACIÓN DE UN DIAGRAMA DE DISPERSIÓN Y OBTENCIÓN DE UNA LÍNEA DE TENDENCIA:

a. Insertar o pegar los datos en la hoja de cálculo de Excel.
b. Seleccionar los datos a graficar.
c. **Insertar Gráfico.**
d. Seleccionar el gráfico de dispersión y escoger el primero (sin líneas).
e. En la ficha **Diseño de gráfico** se elige el diseño que se prefiera, se asigna título al gráfico y a los ejes, y se borra la leyenda.

Después de finalizar el gráfico:

- Si se quiere cambiar el tamaño del gráfico: Arrastrar desde una esquina con el ratón presionado.

- Si se quiere cambiar el formato de los ejes:
  - » Dar clic derecho sobre cualquier valor del eje que se quiera cambiar.
  - » Escoger **Dar formato a eje**: Entre otras cosas, permite cambiar escala (se puede alterar el mínimo y máximo requeridos, o el tamaño de la escala), fuente (cambia el tipo de letra y tamaño), número (cambia el número de decimales o la categoría del valor) o alineación.

- Si se quiere cambiar el formato de los títulos de eje o del título general:
  - » Dar clic derecho sobre lo que se quiere cambiar.
  - » Escoger **Formato de título de eje** o **Formato del título del gráfico**, según lo que se pretenda.
  - » Se puede cambiar tramas, fuentes o alineación y se pueden asignar efectos de relleno.

- Si se quiere alterar el fondo del gráfico:
  - » Dar clic derecho sobre alguna parte del fondo.
  - » Escoger **Formato de área del gráfico.**
  - » Se puede cambiar el color o dar efecto de relleno.

- Si se quiere cambiar marcadores ("puntitos"):
  - » Dar clic derecho sobre cualquiera de ellos.
  - » Escoger **Dar formato a serie de datos**. Se puede cambiar estilo, tamaño o color.

- Si se quiere agregar línea de tendencia:
  - » Dar clic derecho sobre cualquiera de los marcadores.
  - » Seleccionar **Agregar línea de tendencia**.
  - » Escoger lineal, logarítmica, exponencial, polinomial (con su grado u orden respectivo), potencial o media móvil.
  - » En opciones: Se puede agregar ecuación y R^2 en el gráfico.

- Si se quiere cambiar el marcador de un solo punto (para destacarlo):
  - » Dar clic en cualquiera para seleccionar la serie.
  - » Clic izquierdo en el deseado (comprobar que solo él quede seleccionado).
  - » Clic derecho en él.
  - » Cambiar (con **Formato de punto de datos**).

2. CÁLCULO DE UNA REGRESIÓN MÚLTIPLE

- Digitar o pegar datos en la hoja de cálculo en forma de columnas. La variable dependiente debe aparecer a la derecha.
- Seleccionar el menú **Datos**.
- Seleccionar la opción **Análisis de Datos** (si no está Análisis de datos, revisar la última parte del Anexo 2).
- En el cuadro de diálogo **Análisis de Datos**, escoger **Regresión**.
- En el cuadro de diálogo **Regresión**:
  - » Indicar Rango *y* de entrada (puede ser digitando la referencia correspondiente –primera y última celdas del rango respectivo, separadas por dos puntos–, o dando clic en la flechita roja del campo y seleccionando el rango respectivo).
  - » Indicar rango *x* de entrada (bloque de todas las variables independientes).
  - » Indicar rango de salida (teclear la dirección de la celda donde se quiere que aparezcan los resultados).
  - » Seleccionar Aceptar para obtener el análisis de regresión.

  En los resultados se obtienen, entre otros, los siguientes datos:
  Coeficiente de correlación múltiple (r)
  Coeficiente de determinación (R^2)
  Número de observaciones
  Punto de corte (A): Aparece en la última tabla en la columna Coeficientes, fila Intercepción
  Coeficiente B_1: Columna Coeficientes, fila Variable X_1
  Coeficiente B_2: Columna Coeficientes, fila Variable X_2
  Y así sucesivamente.
  De acuerdo con eso se forma la ecuación correspondiente.

3. GRÁFICO DE UNA SERIE DE TIEMPO:

- Insertar o pegar los datos en la hoja de cálculo de Excel (uno de las columnas corresponde al tiempo y el otro a la variable que es función del tiempo).

- Si los datos de tiempo no son numéricos en su totalidad, seleccionar todos los datos (tanto de tiempo como de la otra variable).

  Si los datos de tiempo son numéricos (por ejemplo, años) debe seleccionarse únicamente los datos de la variable que no es tiempo; una vez obtenido el gráfico, debe darse clic derecho en cualquiera de los valores de *x* y luego **Seleccionar datos**; en el espacio de **etiquetas del eje horizontal** elegir **Editar** y asignar el rango que se pretende que aparezca en el eje x.

Para editar el gráfico puede hacerse lo mismo que se explica en el ítem 2. Generalmente, por cuestión de estética y espacio, va a ser necesario cambiar la alineación de los elementos del eje x. Para eso debes dar clic derecho sobre alguno de esos elementos, seleccionar **Dar formato a eje** y la pestaña **Alineación**. Luego debe escogerse la **Dirección de texto preferida.**

Para hacer proyecciones debe reemplazarse el valor apropiado en la ecuación de regresión, pero si la tendencia es lineal puede utilizarse la función **Pronóstico,** que se encuentra en el grupo de las funciones estadísticas.

En dicha función se piden los parámetros **X** (es el valor de X para el cual se va a hacer el pronóstico), **Conocido_y** y **Conocido_x** (las dos últimas son las matrices de datos correspondientes). Por ejemplo, si se tienen los datos siguientes:

2007	2008	2009	2010	2011
1300	1570	1846	2164	2573

Y se quiere hacer una proyección para 2012, X es 2012, Conocido_x es la matriz que comprende los números entre 2007 y 2011, y Conocido_y es la matriz que comprende los números entre 1300 y 2573. La respuesta es 2832,6.

Nota importante: Si en lugar del diagrama de líneas, se construye el diagrama de dispersión se obtiene la misma gráfica, pero la ecuación de regresión daría con la misma pendiente, pero con un punto de corte errado.

EJEMPLOS

1. **Se pretende explicar los cambios en la variable "índice de desarrollo humano" a partir de los cambios en "Producto Interno Bruto per cápita". Para ello se tuvieron en cuenta los datos más recientes de los países de América Latina:**

PAÍS	Argentina	Chile	Uruguay	México	Panamá	Venezuela	Brasil
PIB PER CÁPITA (dólares)	15604	14983	14342	14266	12398	11889	11289
IDH	0,775	0,783	0,765	0,75	0,755	0,696	0,699

PAÍS	Costa Rica	Colombia	Perú	Rep. Dominicana	Ecuador	El Salvador	Paraguay
PIB PER CÁPITA (dólares)	10732	9445	9281	8648	7952	7442	4915
IDH	0,725	0,689	0,723	0,663	0,695	0,659	0,64

PAÍS	Guatemala	Bolivia	Honduras	Nicaragua	Haití
PIB PER CÁPITA (dólares)	4871	4584	4405	2970	1122
IDH	0,56	0,643	0,604	0,565	0,404

   a. **Hallar la ecuación de regresión más apropiada, si la hay.**
   b. **Hallar el coeficiente de correlación.**

<u>Solución</u>:

Primero deben pegarse los datos en Excel, de tal manera que queden en columnas; luego deben seleccionarse.

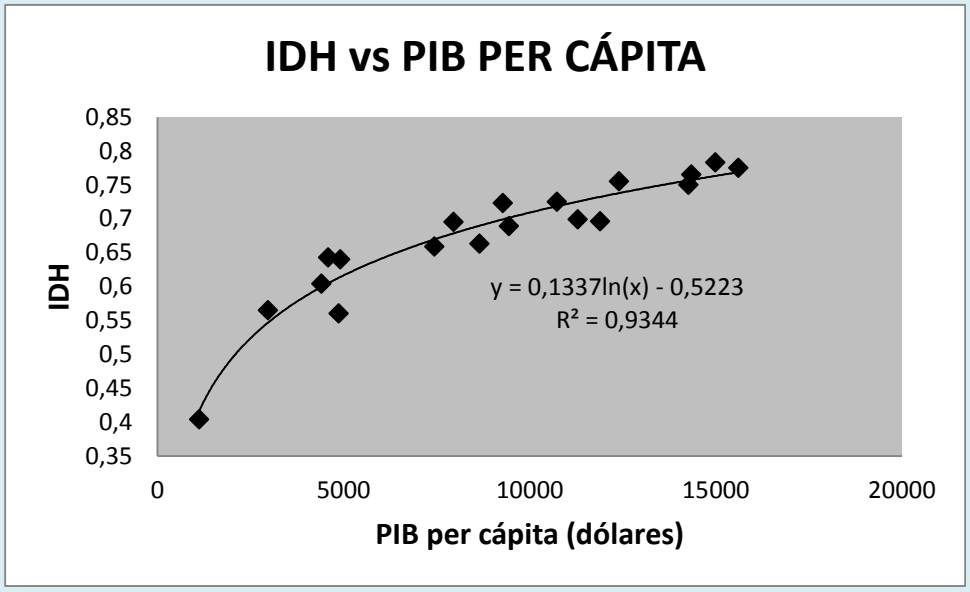
Posteriormente ir a **Insertar – Gráfico** y escoger el gráfico de dispersión.

Escoger el subtipo de gráfico que no une los puntos.

Asignar títulos, borrar leyenda y líneas de división.

Luego, después de dar clic derecho sobre uno de los marcadores (puntos), elegir **Agregar Línea de tendencia**; por la forma del gráfico se selecciona logarítmico y en la pestaña opciones se seleccionan **Presentar ecuación en el gráfico** y **Presentar el valor R cuadrado en el gráfico**.

El gráfico debe quedar de la siguiente forma, aunque puede editarse de tal manera que quede estéticamente más agradable:



Además, como R^2 es 0,9344, entonces el coeficiente de correlación es 0,9666 (su raíz cuadrada).

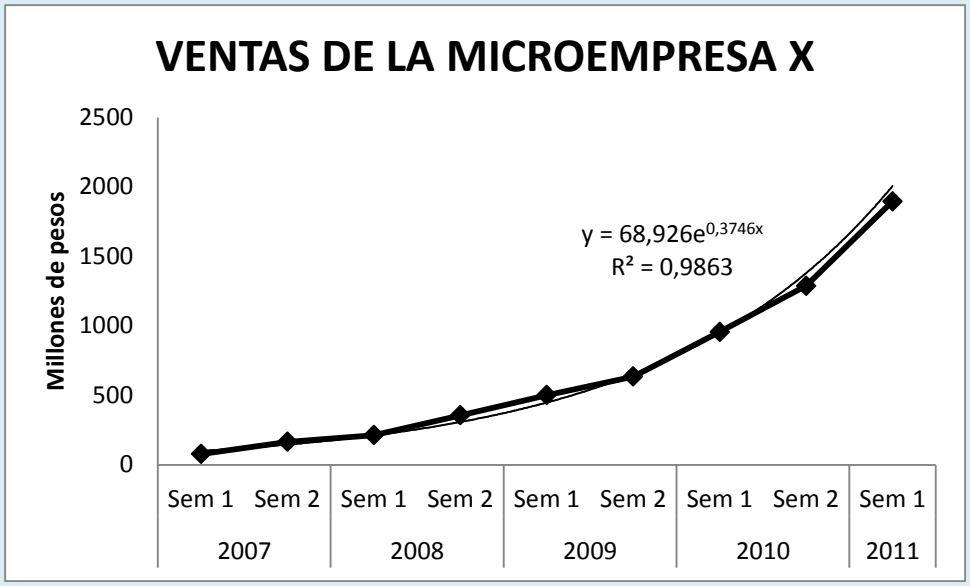
2. **Las siguientes son las cifras de ventas semestrales de una microempresa desde su creación (en millones de pesos):**

2007		2008		2009		2010		2011
Sem 1	Sem 2	Sem 1	Sem 2	Sem 1	Sem 2	Sem 1	Sem 2	Sem 1
78,5	165	214	356	502	635	956	1288	1896

Hacer una proyección de ventas para el segundo semestre de 2011.

Solución:

Luego de copiar los datos en Excel, debe elaborarse el diagrama de líneas y agregar una línea de tendencia que sea conveniente (por la forma del gráfico, exponencial); lo anterior queda de la siguiente manera:



Una proyección de ventas para el segundo semestre de 2011 es:

$$y(10) = 68{,}926e^{(0{,}3746)(10)} = 2919{,}1 \ millones\ de\ pesos$$

ANEXO 4
DISTRIBUCIONES DE PROBABILIDADES CON EXCEL

1. Seleccionar el icono f(x)
2. Elegir la categoría **Estadísticas**.
3. Allí se encuentran, entre otras, las siguientes funciones de utilidad para este Núcleo:
   DISTR.BINOM
   POISSON
   DISTR.NORM.ESTAND
   DISTR.NORM.ESTAND.INV
   DISTR.NORM.INV
   DISTR.NORM
   DISTR.HIPERGEOM

Funciones:

DISTR.BINOM: Evaluar la probabilidad de una variable aleatoria discreta que sigue una distribución binomial.

Si se selecciona DISTR.BINOM deben agregarse los siguientes parámetros:

Núm_éxito: es el número de éxitos en los ensayos.
Ensayos: es el número de ensayos independientes.
Prob_éxito: es la probabilidad de éxito en cada ensayo.
Acumulado: es un valor lógico: Para calcular la función de distribución acumulativa debe digitarse en el cuadro la palabra VERDADERO; para hallar la función de probabilidad no acumulativa debe escribirse FALSO.

POISSON: Evalúa la probabilidad de una variable aleatoria discreta que sigue una distribución de Poisson

Si se selecciona POISSON deben agregarse los siguientes parámetros:

x: es el número de sucesos.
Media: es el valor numérico esperado (media de la distribución).
Acumulado: igual a Binomial.

DISTR.NORM: Devuelve la distribución acumulativa normal para la media y la desviación estándar especificadas.

X es el valor cuya distribución se desea obtener.
Pide también media y desviación estándar.
Acum: igual a Binomial

DISTR.NORM.ESTAND: Evalúa la distribución normal estándar acumulativa (con media igual a cero y desviación estándar igual a 1). Eso implica que proporciona la misma información que las tablas estadísticas y, por lo tanto, el único parámetro que pide es el valor de Z.

DISTR.NORM.ESTAND.INV: Devuelve el inverso de la distribución normal estándar acumulativa. El único parámetro que pide es la probabilidad de que *X* sea menor o igual al valor predeterminado.

DISTR.NORM.INV: Devuelve el inverso de la distribución acumulativa normal para la media y la desviación estándar especificadas.

Pide probabilidad (de que *X* sea menor o igual al valor predeterminado), media y desviación estándar.

DISTR.HIPERGEOM: Devuelve la probabilidad para una variable aleatoria que sigue una distribución hipergeométrica.

Muestra_éxito: número de éxitos en la muestra.
Núm_de_muestra: tamaño de la muestra.
Población_éxito: número de éxitos en la población.
Núm_de_población: tamaño de la población.

Nota: Si desea hacerse alguna combinación en Excel, debe insertarse una función de la categoría **"Matemáticas y trigonométricas"**; posteriormente debe seleccionarse la función **COMBINAT**, allí **Número** es el número total de elementos **(*n*)** y **Tamaño** es el número de elementos en cada combinación **(*r*)**.

EJEMPLOS

1. **Se elige una muestra de 20 empresas colombianas exportadoras. Se sabe por estadísticas de años anteriores que aproximadamente el 40% de las empresas colombianas exportadoras registra operaciones en varias ciudades.**

    a. ¿Qué tan probable es encontrar al menos 18 empresas que tengan operaciones en varias ciudades?
    b. ¿Qué tan probable es encontrar entre 10 y 15 empresas que operen en varias ciudades?

Solución:

Se observa que se sigue una distribución binomial porque los eventos son independientes (el hecho de que una empresa opere en varias ciudades no tiene por qué influir en el hecho de que otra funcione en una o varias ciudades), la probabilidad de que opere en varias ciudades es constante (es 0,4 para todo el proceso), hay un número de ensayos determinado (20 empresas) y cada ensayo tiene 2 opciones (opera en una ciudad o en varias).

Eso implica que debe elegirse la opción DISTR.BINOM. Vamos a considerar que el éxito es funcionar en varias ciudades (aunque perfectamente podría hacerse si se elige que el éxito sea el caso contrario).

a. $P(x \geq 18) = 1 - P(x \leq 17)$

Para hallar $P(x \leq 17)$:
Núm_éxito: 17
Ensayos: 20
Prob_éxito: 0,4
Acumulado: VERDADERO
La respuesta obtenida es 0,99999, de donde se deduce que $P(x \geq 18)$ es aproximadamente cero.

b. $P(10 \leq x \leq 15) = P(x \leq 15) - P(x \leq 9)$

	$P(x \leq 15)$	$P(x \leq 9)$
Núm_éxito	15	9
Ensayos	20	20
Prob_éxito	0,4	0,4
Acumulado	VERDADERO	VERDADERO
Respuesta	0,9997	0,7553

Por lo tanto, $P(10 \leq x \leq 15) = 0{,}9997 - 0{,}7553 = 0{,}2444$

2. Al conmutador de una universidad llegan en promedio 120 llamadas/hora durante el período de actividad. El conmutador no puede hacer más de 5 conexiones por minuto;

a. ¿Cuál es la probabilidad de que el conmutador se encuentre congestionado en un minuto dado?

b. ¿Cuál es la probabilidad de que se pierdan 3 llamadas si la recepcionista salió 2 minutos de la oficina?

Solución:

Es claro que debe usarse la distribución de Poisson porque se busca la probabilidad de una variable que depende del tiempo, por lo tanto se elige la función POISSON.

a. Si en promedio llegan 120 llamadas por hora, se esperaría que en un minuto lleguen 2 llamadas (por regla de tres simple)

$P(X \geq 6) = 1 - P(X \leq 5)$

Para hallar $P(X \leq 5)$
X: 5
Media: 2
Acumulado: VERDADERO
La respuesta obtenida es 0,983436

Por lo tanto, $P(X \geq 6) = 1 - 0{,}983436 = 0{,}016564$

b. Si en promedio llegan 120 llamadas por hora, se esperaría que en 2 minutos lleguen 4 llamadas (por regla de tres simple)

X: 3
Media: 4
Acumulado: FALSO
La respuesta obtenida es 0,195367; esta es $P(X = 3)$

3. Durante los últimos años ha crecido el volumen de acciones negociadas en la Bolsa. Durante las dos primeras semanas de febrero, el volumen diario promedio fue de 586 000 acciones. La distribución de probabilidad del volumen es aproximadamente normal con desviación estándar de 115 000 acciones.

a. ¿Cuál es la probabilidad de que el volumen negociado en un día sea menor a 395 000 acciones?
b. ¿Cuál es la probabilidad de que en un día se negocien entre 500 000 y 600000 acciones?
c. Si la Bolsa quiere emitir un boletín de prensa sobre el 5% de los días más activos, ¿qué volumen activará la publicación?

Solución:

a. Puede resolverse con DISTR.NORM o con DISTR.NORM.ESTAND (más recomendable con el primero porque el segundo exige hacer antes la operación para Z).

Con DISTR.NORM:
X = 395 000
Media = 586 000
Desviación estándar = 115 000
Acumulado = VERDADERO
La respuesta es 0,0483698

Con DISTR.NORM.ESTAND:
Z = -1,66087 ((395000-586000)/115000)
La respuesta es 0,0483698

b. Puede resolverse con DISTR.NORM o con DISTR.NORM.ESTAND, aunque solamente va a ser hecha con DISTR.NORM:

X_1 = 500 000	X_2 = 600 000
Media = 586 000	Media = 586 000
Desviación estándar = 115 000	Desviación estándar = 115 000
Acum = VERDADERO	Acum = VERDADERO
La respuesta es 0,22728246	La respuesta es 0,54844723

$\Rightarrow$ P(500000 ≤ x ≤ 600000) = 0,54844723 – 0,22728246 = 0,32116477

c. Puede resolverse con DISTR.NORM.INV o con DISTR.NORM.ESTAND.INV

Si el 5% de los días se supera ese volumen, entonces el 95% es superado por él:

Con DISTR.NORM.INV:
Probabilidad = 0,95
Media = 586 000
Desviación estándar = 115 000

La respuesta es 775 158 acciones

Con DISTR.NORM.ESTAND.INV:
Probabilidad = 0,95

La respuesta es 1,644853
Posteriormente habría que despejar X de la fórmula de Z

4. **Un embarque contiene 1000 artículos, de los cuales hay 20 defectuosos (lo sabe la empresa productora). La empresa que realiza las inspecciones toma una muestra aleatoria de 10 artículos y si encuentra por lo menos uno defectuoso rechaza el embarque, ¿cuál es la probabilidad de que el embarque se rechace?**

Solución:

Debe usarse DISTR.HIPERGEOM. La única forma de no rechazar el embarque es que no se encuentren defectuosos en la muestra; para este último caso:

Muestra_éxito: 0
Núm_de_muestra: 10
Población_éxito: 20
Núm_de_población: 1000

La respuesta es 0,816318. Por lo tanto, la probabilidad de que el embarque se rechace es 1 – 0,816318, o sea 0,183682.

ANEXO 5
STATGRAPHICS PLUS

1. GENERALIDADES:

Statgraphics es un paquete estadístico que aborda ampliamente la mayoría de los temas estadísticos; es sencillo de aprender y manejar.

El menú principal contiene los siguientes comandos:

- **FILE** (Archivo): con órdenes que permiten abrir, cerrar o guardar ***Statfolio*** (grupos de trabajo) y ***Data file*** (ficheros de datos). Además permite imprimir.
- **EDIT** (Corregir): permite, entre otras cosas, copiar, cortar e insertar datos.
- **PLOT** (Diagramas): principalmente para elaborar gráficos, bien sea ***Scatterplots*** (gráficos de dispersión), ***exploratory plots*** (entre ellos el histograma de frecuencias) o ***business charts*** (***barchart*** o gráfico de barras y ***piechart*** o gráfico de sectores).
- **DESCRIBE** (Describir): por medio del cual se analizan y procesan datos de muestras individuales –una sola muestra–.
- **COMPARE** (Comparar): mediante él se puede hacer todo tipo de análisis que permita la comparación de muestras.
- **RELATE** (Relacionar): para hacer regresiones.
- **VIEW** (Mirar): para ver u ocultar las barras (***toolbar***, que contiene íconos directos y *Status bar,* que va reseñando lo que se hace) y el ***StatAdvisor***.
- **WINDOW**: para organizar las ventanas.
- **HELP** (Ayuda).

Otras herramientas importantes son:

- **STATADVISOR**: aparece cada que se hace una prueba; explica el significado de los resultados obtenidos.
- **STATFOLIO**: permite guardar y recuperar rápidamente grupos de trabajo –análisis previos–.

- **STATGALLERY**: permite almacenar gráficos.

2. ANÁLISIS DE UNA VARIABLE NUMÉRICA:

Al emplear ***One-variable analysis*** se pueden analizar datos de una variable cuantitativa o una cualitativa cuantificada –representada por números–.

Para usarlo hay que introducir previamente los datos a evaluar; para ello se crea una hoja electrónica maximizando ***Untitled***, en las celdas de dicha hoja se introducen los datos del nuevo fichero de datos (obviamente se usa una columna por muestra).

Nota: Los datos que se ingresen deben ser numéricos. Si se trabaja con una variable cualitativa debe representarse cada punto muestral por un número.

Si se quiere pueden salvarse los datos, usando ***File*** del menú principal y ***Save data file*** del submenú.

Una vez introducidos los datos, seleccionar **DESCRIBE - NUMERIC DATA** (datos numéricos) - **ONE VARIABLE ANALYSIS** (variable unidimensional). Si la variable es cualitativa, en vez de NUMERIC DATA se emplea CATEGORICAL DATA.

Aparece entonces una caja de diálogo donde debe asignarse la columna con la que se quiera trabajar; para ello debe escribirse la columna deseada (de la forma en que están escritas a la izquierda, es decir, Col_X) en el espacio correspondiente a data.

El ***One-variable analysis*** contiene 7 opciones tabulares (ícono amarillo de la barra de herramientas que acaba de aparecer) y 7 opciones gráficas (ícono azul).

Las opciones tabulares son:

- ***Analysis Summary*** (resumen): contiene simplemente el número de datos y los valores mayor y menor, principalmente para verificar que todos los datos hayan sido introducidos.
- ***Summary statics*** (resumen estadístico): proporciona los valores correspondientes a las medidas de tendencia central como media y mediana, medidas de variabilidad como desviación estándar y coeficiente de variación y medidas de forma. Estas últimas son *standardized skewness* (asimetría estandarizada) y *standardized kurtosis* (curtosis estandarizada) y pueden ser utilizadas para determinar si la muestra viene de una población distribuida normalmente; valores por fuera del rango [-1,1] indican desviaciones significativas de la normalidad.

Si se desea conocer algún valor estadístico que no aparece directamente en la lista desplegada se da clic derecho y se selecciona ***Pane option*** (panel de opciones), luego se seleccionan los estadísticos que se quiera conocer.

- ***Percentiles***: si desea conocerse un valor percentil que no aparezca en la lista desplegada, debe presionarse el botón alterno del ratón y seleccionar ***Pane option***; aparece entonces una caja de diálogo, donde debe escribirse el valor deseado. Los percentiles pueden verse gráficamente seleccionando ***Quantile plot*** de la lista de opciones gráficas.
- ***Frequency tabulation*** (tabla de frecuencia): donde se registran frecuencias absolutas y relativas de rangos de datos predeterminados.

  Por defecto, el programa toma como límite inferior el cero y un límite superior aleatorio; por lo tanto es preferible cambiar esa tabla, lo cual se logra dando clic derecho, seleccionando ***Pane option*** y escribiendo los valores deseados en la caja de diálogo. Pueden cambiarse el número de clases, el límite inferior (***lower limit***) y el límite superior (***upper limit***).

  Los resultados de la tabulación se observan gráficamente seleccionando ***Frequency Histogram*** (histograma de frecuencias) de la lista de opciones gráficas. Con ***pane option*** (luego de presionar el botón derecho del ratón) puede variarse la forma de la gráfica; puede obtenerse el histograma acumulado o el polígono de frecuencias, bien sea acumulado (ojiva) o no acumulado.
- ***Confidence Intervals*** (Intervalos de confianza): ver estimación.
- ***Hypothesis Tests*** (Pruebas de hipótesis): ver prueba de hipótesis.

3. DISTRIBUCIONES DE PROBABILIDAD:

Para establecer una distribución de probabilidades debe seleccionarse ***Plot*** del menú principal, posteriormente debe seleccionarse ***probability distributions*** (distribuciones de probabilidad) del submenú, que contiene 22 tipos de distribución, entre ellas Binomial, Poisson, Normal, entre otras.

Para dar los parámetros conocidos debe presionarse el botón alterno del ratón y seleccionar ***Analysis options*** (opciones de análisis). Aparece entonces una caja de diálogo que permite ingresar los parámetros deseados (media y desviación estándar de la población y número de sucesos o solamente uno o dos de ellos).

Probability distributions contiene tres opciones tabulares (ícono amarillo de la barra de herramientas específica) y cinco opciones gráficas (ícono azul).

Las opciones tabulares son:

- ***Cumulative distribution*** (distribución acumulada): aparece el área de las colas (probabilidad) mayor, menor o igual a un valor de la variable predeterminado.
- ***Inverse CDF*** (Función de densidad acumulada inversa): para hallar el valor de X correspondiente a una probabilidad predeterminada.
- ***Random numbers***: genera números aleatorios.

Las opciones gráficas más usadas son:

- ***Density/mass function*** (Función densidad/masa): expresada por una línea si la variable es continua. Si es discreta, se representa por puntos dispersos.
- ***CDF*** **(*cumulative density function*)**: distribución acumulada (ojiva), [p(x) ≤ cierto valor].
- ***Survivor function***: grafica la probabilidad de que x sea mayor o igual que un valor determinado.

4. ESTIMACIÓN:

4.1. Intervalos de confianza para una muestra:

Para estimar un parámetro de la población puede partirse de medias y varianzas muestrales o de los datos correspondientes a un muestreo aleatorio simple.

- Si se parte de los datos muestrales: Lo primero que debe hacerse es introducir dichos datos; para hacerlo maximizamos **U*ntitled***, para crear una hoja electrónica en cuyas celdas podamos escribir los datos. Esos datos pueden guardarse con ***Save data file*** (guardar archivo de datos).

  Después de introducir los datos se debe seleccionar ***Describe*** del menú principal, ***numeric data*** del submenú y posteriormente ***one-variable analysis***. Aparece una caja de diálogo, donde debe asignarse la variable deseada, lo cual se logra seleccionando la columna correspondiente y ubicándola con un clic en el espacio ***Data***.

  Aparece, entonces, una barra de herramientas que contiene las opciones tabulares (ícono amarillo) y las opciones gráficas (ícono azul).

  Entre las opciones tabulares está ***Confidence Intervals*** (intervalos de confianza), que crea intervalos de confianza para la media y la desviación estándar. El programa crea intervalos del 95% de confianza; si dicho nivel de confianza quiere cambiarse, debe presionarse el clic derecho y seleccionar ***Pane option***; finalmente se completa la caja de diálogo.

- Si se parte de los estadísticos muestrales (X, S o p): debe seleccionarse ***Describe*** del menú principal e ***Hypothesis Tests*** del submenú.

  Aparece una caja de diálogo donde primero debe escogerse uno de los siguientes parámetros: ***Normal mean*** (media normal), ***Normal sigma*** (desviación estándar normal), ***Binomial proportion*** (proporción binomial) o ***Poisson rate***.

  - » Si se selecciona ***normal mean***, el programa calcula intervalos de confianza para la media usando la distribución t. Asume que los datos vienen de una distribución normal, por eso debe verificarse previamente el grado de normalidad.
  - » Si se selecciona ***normal sigma***, calcula intervalos de confianza para la desviación estándar usando la distribución chi cuadrado. También asume que los datos están normalmente distribuidos.
  - » El análisis para una distribución binomial calcula intervalos de confianza para la proporción; debido a que usa una aproximación normal, no debe usarse para muestras muy pequeñas. Para proporciones cercanas a 0,5, el análisis proporciona aproximaciones útiles con muestras de 20 o más; para proporciones menores de 0,4 o mayores de 0,6, el análisis proporciona aproximaciones útiles con muestras de 50 o más.
  - » Para una distribución Poisson, el programa estima la media de la distribución. Como usa una aproximación normal, no debe usarse para muestras muy pequeñas. La aproximación es razonable cuando la media de la distribución Poisson es 10 o más (por ejemplo, cuando n = 10 y p = 0,2).

  En cualquiera de los casos puede calcularse el tamaño mínimo de la muestra que se necesita para obtener un intervalo de confianza de manera que se tenga una confianza determinada de que el error al hacer la estimación sea menor que un cierto error especificado. Para ello se selecciona ***Describe*** del menú principal y ***Sample size determination*** del submenú; aparece una caja de diálogo, donde se escoge el parámetro a seguir (*normal mean, normal sigma, binomial o Poisson*) y se suministran los valores de μ, σ y error absoluto.

  Si se desea cambiar el nivel de confianza debe presionarse el botón alterno del ratón y seleccionar la opción ***Analysis option***; luego se puede ver una caja de diálogo donde puede escogerse un valor de *Alpha* (α) y si se desea que el intervalo de confianza sea bilateral o unilateral.

4.2. Intervalos de confianza para la comparación de dos muestras:

Si se van a calcular intervalos de confianza para la diferencia de dos medias ($\mu_1 - \mu_2$), la diferencia de dos proporciones ($p_1 - p_2$) o el cociente de varian-

zas de dos distribuciones normales (σ_1^2/σ_2^2) debe seleccionarse ***Compare*** del menú principal, luego ***Two Samples*** (dos muestras) e ***Hypothesis Tests***; aparece entonces una caja de diálogo donde debe escogerse uno de los parámetros anteriormente reseñados:

- ***Normal means***, si se quiere determinar un intervalo de confianza para $\mu_1 - \mu_2$.
- ***Normal sigmas***, si se quiere determinar un intervalo de confianza para σ_1^2/σ_2^2; se basa en la distribución F.
- ***Binomial***, si se quiere determinar un intervalo de confianza para $p_1 - p_2$.

Dicha caja de diálogo también pide los estadísticos de la muestra ($\bar{X}$, S, p y n) y una hipótesis nula ($\bar{X}_1 - \bar{X}_2$, S_1^2/S_2^2 o $p_1 - p_2$), según el tipo de intervalo de confianza que se desee obtener.

En todos los casos puede calcularse el tamaño mínimo de muestra necesario para obtener un intervalo de confianza con un error absoluto menor que algún valor predeterminado, usando ***Sample sizes determination***.

Un caso especial lo constituye la estimación de $\mu_1 - \mu_2$ cuando las muestras son pareadas. En dicho caso debe seleccionarse ***Compare - Two samples - Paired samples comparison*** (comparación de muestras pareadas).

Previamente deben haberse introducido los datos en la hoja de cálculo creada al maximizar ***Untitled***. Las variables deben asignarse (de la manera antes descrita); de la barra de herramientas que aparece se eligen ***Tabular option*** y la opción ***Confidence intervals***.

Si se quiere cambiar el nivel de confianza que el programa determina por defecto (95%), se usa ***Pane option***.

5. PRUEBAS DE HIPÓTESIS

5.1. Prueba de hipótesis para una muestra:

Para hacer un test de hipótesis puede partirse de medias y varianzas muestrales o de los datos correspondientes a un muestreo.

- Si se parte de los datos muestrales: lo primero que debe hacerse es copiar los datos en la tabla y si se quiere, guardar el archivo (ver estimación).

  Luego seleccionar ***Describe - numeric data - one variable analysis*** y posteriormente asignar la variable.

Entre las opciones tabulares está ***Hypothesis Tests***, que produce valores de t (***computed t statics***) y p (***p-value***), según la hipótesis nula (*Null hypothesis*) y alternativa (*alternative*).

Si desea hacerse algún cambio, debe presionarse el botón alterno del ratón y seleccionar ***Pane options***. Aparece una caja de diálogo, donde pueden cambiarse la media (*mean*) y el nivel de significancia (*alpha*); también puede cambiarse la hipótesis alternativa, que puede ser *not equal* (≠), *less than* (<) o *greater than* (>),

La hipótesis nula debe rechazarse si $\alpha > p$.

- Si se parte de los estadísticos muestrales ($\bar{X}$, S o p): Seleccionar ***Describe - Hypothesis Tests***. En la caja de diálogo que se encuentra se puede escoger un parámetro, según el tipo de prueba de hipótesis que vaya a hacerse: *Normal mean* (para media), *normal sigma* (para desviación estándar) o *Binomial proportion* (para proporción).

  Deben suministrarse los datos que el programa pida, lo cual depende del tipo de prueba. Esos datos pueden ser *Null Hypothesis* (H_o), *sample mean* ($\bar{X}$), *sample sigma* (S), *sample size* (n) o *sample proportion* (p).

  El programa proporciona valores para t y p y si la hipótesis nula debe rechazarse o no con un determinado nivel de significancia. Si se quieren cambiar la hipótesis alternativa y/o el nivel de significancia, debe emplearse *Analysis options*.

5.2. Pruebas de hipótesis para comparación de dos muestras:

El procedimiento es exactamente el mismo que el realizado para hallar intervalos de confianza. Además del intervalo de confianza correspondiente, el programa también arroja valores de t y p y si la hipótesis debe rechazarse o no.

6. REGRESIÓN:

6.1 Análisis de regresión simple:

Estima principalmente una relación lineal entre dos variables; el modelo relaciona una variable independiente y una dependiente, minimizando la suma de los cuadrados de los errores respecto a la línea ajustada.

Inicialmente deben introducirse los datos. Luego se selecciona ***Relate - simple regression*** (regresión simple). Aparece una caja de diálogo que debe completarse; la variable independiente (X) debe corresponder a la columna 1 y la dependiente (Y) a la columna 2.

El programa devuelve valores para la pendiente (*slope*), el intercepto con el eje Y (*intercept*), la ecuación de la línea ajustada y el coeficiente de correlación.

Además brinda valores de la desviación estándar de la pendiente y el intercepto y sus valores t y p.

Además de una línea recta, también puede emplearse el análisis para estimar algunos modelos estandarizados no lineales: el análisis puede ajustar automáticamente cualquiera de doce modelos (entre ellos exponencial, logarítmico y multiplicativo).

El modelo preferido es el ajuste lineal a no ser que otro tipo de modelo incremente el valor de R cuadrado significativamente. Para seleccionar el mejor modelo se puede usar la opción tabular de comparación de modelos alternos. Si ningún modelo se ajusta (para ninguno el grado de correlación es cercano a uno) debe emplearse el análisis de regresión polinomial.

Principales opciones tabulares:

- ***Forecasts*** (pronósticos): muestra los valores esperados para la variable dependiente usando la ecuación de la recta ajustada.

  Si se desea conocer el valor de Y cuando X toma un valor determinado, se debe seleccionar ***Pane option*** y escribir dicho valor de X. También da intervalos de confianza para el intercepto y la pendiente.
  Con ***Analysis options*** se puede cambiar el tipo de modelo.

- ***Comparison of alternative models*** (comparación de modelos alternativos): da el grado de correlación para cada modelo; por lo tanto permite observar cual es el más adecuado para ajustar los datos.
  Principales opciones gráficas:

- **Plot of fitted model** (gráfico del modelo ajustado): devuelve la gráfica de la línea ajustada.
  Con ***Pane options*** se pueden incluir o no los límites de confianza y variar su nivel de confianza.
  Con ***Analysis options*** se puede mirar la gráfica para los diferentes modelos.

- ***Observed versus predicted*** (observados contra predichos): comparación de los valores de la variable independiente reales y esperados según el modelo que se haya seleccionado. Con ***Analysis options*** se puede cambiar el modelo.

- ***Residual versus X*****:** valores reales de X contra la diferencia entre valores de X observados y predichos.

- ***Residual versus predicted***: igual al anterior, pero con la variable dependiente.

6.2. Análisis de regresión múltiple:

Permite calcular un modelo de regresión entre una variable dependiente y dos o más variables independientes. Al igual que el análisis de regresión múltiple, la regresión múltiple utiliza los mínimos cuadrados para estimar la ecuación de la curva ajustada.

Para su empleo, debe seleccionarse ***Relate - multiple regression*** y llenar la caja de diálogo. Obviamente, primero hay que introducir los datos.

ANEXO 6
RESPUESTAS A EJERCICIOS PROPUESTOS

OBJETO DE APRENDIZAJE 1

1.

Media	9007,3 dólares
Mediana	9281 dólares
Moda	No hay
Rango	14 482 dólares
Desv.est	4233,4 dólares
CV	47%

En 2010, el PIB per cápita promedio de los países latinoamericanos es de 9007,3 dólares. La mitad de los países de la región tienen un PIB por persona superior a 9281 dólares y la otra mitad, inferior a ese valor. La diferencia entre el mayor PIB per cápita de la región –Argentina– y el menor –Haití– es de 14 482 dólares. El Coeficiente de variación es indicativo de las marcadas diferencias existentes en la zona en relación al PIB per cápita de los diferentes países.

2.

	ni	hi
Satisfechos	313	85,8%
Insatisfechos	23	6,3%
No sabe, no responde	29	7,9%
	365	1

3. a. Media: 18,4%, Coeficiente de variación: 40%. El promedio de analfabetismo en los municipios seleccionados es 18,4%; el valor del coeficiente de variación indica que en el grupo seleccionado hay unos municipios con una tasa de analfabetismo alta y otros con una tasa baja (hay un grado de dispersión más o menos alto).

b. Las tres cuartas partes de los municipios antioqueños seleccionados en la muestra tienen tasas de analfabetismo igual o inferior a 22,9%.

4.

Media	57,6
Mediana	54,7
Moda	No hay
Rango	39,8
Desv, estándar	11,1
Coeficiente de variación	19,3%

En 2011 el índice de libertad económica de los países suramericanos es, en promedio, 57,6; aunque la mitad de dichos países presenta un índice superior a 54,7; no hay moda en relación a esta variable; no hay diferencias tan marcadas entre los ILE de los diferentes países, aunque el valor se aumenta un poco por los índices de Venezuela y Chile. La diferencia en dicho índice entre esos dos países es 39,8.

5.

	COLOMBIA	PAÍS VECINO
MEDIA (mill. de dólares)	3238,5	3530,3
MEDIANA (mill. de dólares)	3172,3	3534,7
COEF. DE VARIACIÓN (%)	16,1	26,3

Colombia ha exportado menos durante los últimos 24 meses y presenta menor variabilidad. Durante ese período, en los 12 meses en que Colombia ha tenido más bajas exportaciones, ellas han sido por montos superiores a 3172,3 millones de dólares, mientras que en el país vecino, en más de la mitad de los meses evaluados se han hecho exportaciones por 3534,7 millones de dólares o más.

6.

POSICIÓN	Frecuencia	F. relativa	Porcentaje
A favor	306	0,612	61,2%
En contra	194	0,388	38,8%

7. La respuesta correcta es c) porque A presenta mayor coeficiente de variación.

8. a. Preferiría las acciones porque son las que tienen, en promedio, la mayor rentabilidad.
   b. Preferiría las cuentas bancarias porque presentan más baja dispersión.

9.

Media	0,673
Mediana	0,695
Moda	No hay
Rango	0,379
Desv. est	0,090
CV	13,4%
Q1	0,642
Q3	0,738

10. Media: 7.010,6 millones de dólares FOB Coeficiente de variación: 58,3%

Durante los últimos once años Colombia ha presentado exportaciones de petróleo y sus derivados por 7010,6 millones de dólares FOB, en promedio cada año; la variabilidad ha sido muy alta, pues ha presentado algunos años en que las exportaciones eran relativamente bajas, pero en los últimos años se han incrementado bastante.

11. Media: 10,7% Mediana: 8,8% Coeficiente de variación: 55,4%

Los países latinoamericanos presentaron una tasa de inflación promedio durante 2008 de 10,7%. La mitad de los países presentó una tasa inferior a 8,8%; la gran diferencia entre media y mediana es síntoma de la asimetría de la distribución.

Hay mucha variación entre las tasas de inflación de los países latinos, pero el valor del CV se infla mucho por la enorme tasa inflacionaria de Venezuela.

12. a. Carácter IES: Cualitativa.
Sector IES: Cualitativa.
Metodología: Cualitativa.
Número de créditos: Cuantitativa discreta.
Periodicidad: Cualitativa.
Promedio en Saber Pro: Cuantitativa continua.

b.

METODOLOGÍA	n_i	h_i
Presencial	15	60%
Distancia (tradicional)	7	28%
Distancia (virtual)	3	12%
TOTAL	25	

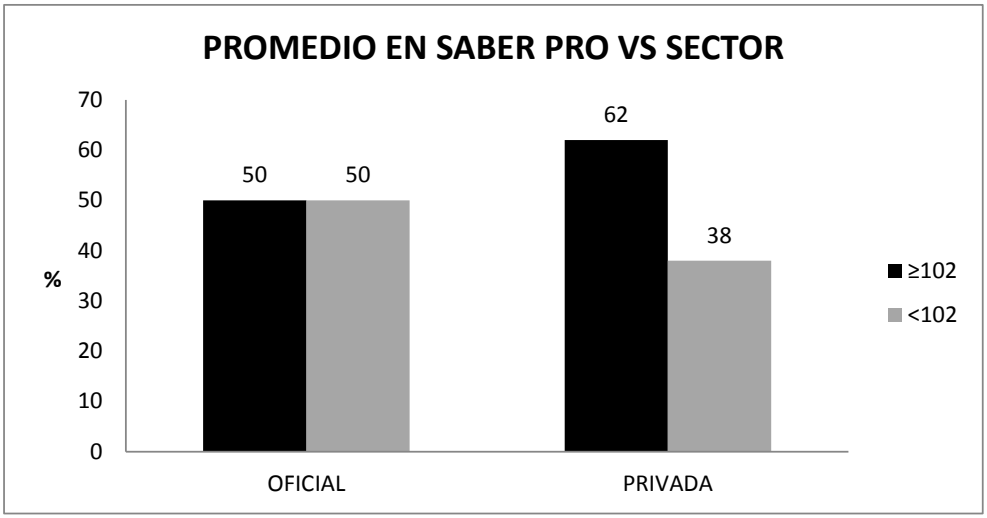
PROMEDIO EN SABER PRO	n_i	h_i
[¨92 – 96,5]	3	12%
(96,5 - 101]	11	44%
(101 – 105,5]	8	32%
(105,5 - 110]	3	12%
TOTAL	25	

c.

	CRÉDITOS	PROMEDIO SABER PRO
Media	160	100,82
Mediana	161	100,2
Moda	160	103,5
Desviación est.	14	4,02
CV	8,7%	4%
Q1	156	98,7
Q3	169	103,5

d.

PROMEDIO	SECTOR IES		**Total**
	Oficial	Privada	
≥ 102	2	13	15
< 102	2	8	10
Total	4	21	25



OBJETO DE APRENDIZAJE 2

1. a. $y = 0{,}1337lnx - 0{,}5223$ $(R^2 = 0{,}9344)$
   b. 11.690 dólares

2. La respuesta correcta es C, ya que el signo de la pendiente es negativo.

3. a. La más apropiada es $y = 0{,}0101x^2 - 1{,}6555x + 90{,}917$, que presenta un R^2 de 0,9813. También son apropiadas $y = -24{,}442\ lnx + 132$ y también $y = 302{,}15x - 0{,}5608$

   b. Relación inversa (a mayor índice de producción, menor nivel de pobreza)

4. La respuesta correcta es C, ya que el punto de corte es aproximadamente 220 y la pendiente es negativa.

5. a. $y = 0{,}0066x + 5{,}7015$
   b. 0,0066 millones de pesos, o sea $6.600
   c. R^2 = 0,9652, lo que indica que un 96,5% de los cambios en los costos son debidos al nivel de producción.
   d. No, sería más adecuado el logarítmico.

6. $y = 13{,}548x + 968{,}2, \quad R^2 = 0{,}898$

7. Ningún modelo se ajusta.

8. a. Si el PIB fuera cero se esperaría que el 86,5% de la población colombiana estuviera desempleada. Por cada millón de dólares que aumenta el PIB, el desempleo en Colombia baja en promedio en 0,0009%. La relación entre las variables es inversa y fuerte.

b. 5,5%

9. $y = -29{,}03lnx + 97{,}19 \quad (R^2 = 0{,}9961)$

Se espera que al duodécimo año sobreviva el 25%.

10. 1438 millones de pesos

11. a. $y = 10{,}464x + 4{,}733$

Si en un país las mujeres no tuvieran hijos se esperaría que la población bajo el nivel de pobreza fuera 4,7%.
Por cada hijo/mujer que aumenta la tasa de fertilidad, la población bajo nivel de pobreza aumenta 10,5% en promedio.

b. Parece que es conveniente, aunque también podríamos afirmar que estamos ante un círculo vicioso y que el nivel de pobreza podría tomarse como variable independiente.

c. 76,7%

12. a. $y = 0{,}3921x + 5{,}8205 \quad (R^2 = 0{,}3694)$ No es válida

b. $y = 0{,}7258x - 4{,}8011 \ (R^2 = 0{,}7393)$

La ecuación de regresión es válida, pero si solo se basa en datos para Suramérica solamente es válida para ellos y no debe hacerse extensiva a otras regiones.

OBJETO DE APRENDIZAJE 3

1. a. 0,47 b. 0,509 c. 0,27

2. a. 46,75% b. 0,12

3. a. 0,0365 b. 0,2055

4. 0,304

5. a. 0,222 b. 0,444

6. a. 0,443

b. Si es franquicia: 0,148 Si no es franquicia: 0,71
Sí parece estar en armonía con el estudio previo.

OBJETO DE APRENDIZAJE 4

1. a. 0,0013 b. 0,471 c. 20 unidades
2. a. 0,000005 b. 0,2443
3. 0,184
4. a. 0,0166 b. 0,762
5. a. 0,0115 b. 0,6296 c. 0,01
6. La probabilidad es 0,017
7. a. Sí b. 0,275
8. a. La probabilidad es 0,0028
   b. La probabilidad es 0,1792
   c. Entre 636 097,5 y 778 502,5 barriles.
9. a. No porque el puntaje mínimo para ser aceptado es 78,5
   b. 0,7143
10. a. 0,0664 b. Sí c. $836.094
11. a. 0,9985 b. $155 680 000
12. a. 0,8329 b. 0,1642 c. Casi imposible $(3{,}3 * 10^{-11})$

OBJETO DE APRENDIZAJE 5

1. 514
2. D
3. 208,3 – 245,7. El margen de error disminuiría
4. 241

5. a. 0,3764 < p <0,4476 b. 2708

6. a. 0,718 < p <0,811 (con nivel de confianza de 90%)
   b. 966

7. a. 324 b. 0,278 < p < 0,352

8. a. 6963,7 < μ < 8036,3
   b. 6861 < μ < 8139
   c. 7120,8 < μ < 7879,2
   d. 85

9. No

10. No

11. Parecería que no

12. Parece que es cierto

13. Parece que es cierto

14. Parece que sí ha disminuido

15. a. 10 762,6 < μ <20370,7
    b. Parece que se conserva