Tecnológico de Monterrey

Concentración en Analítica de Datos y Herramientas de Inteligencia Artificial II

Daniel Otero Fadul

Departamento de Ciencias
Escuela de Ingeniería y Ciencias

Estadística

La **estadística** es una rama de las matemáticas y una disciplina científica que se enfoca en la recopilación, organización, análisis, interpretación y presentación de datos. Su objetivo principal es proporcionar métodos y herramientas para comprender y describir fenómenos mediante la observación y el estudio de datos.

En algunos contextos también se puede entender como estadística los datos numéricos, tales como porcentajes, promedios, entre otros, que ayudan a comprender mejor una gran variedad de escenarios.

Estadística

La estadística desempeña un papel fundamental en la investigación científica, la planificación empresarial, la economía, la salud pública, la ingeniería, la sociología, la psicología y muchas otras disciplinas. Ayuda a resumir la información, a detectar patrones y a tomar decisiones informadas basadas en evidencia empírica.

Estadística

La estadística se divide an dos grandes áreas: la **estadística descriptiva** y la **estadística inferencial**.

- **Estadística descriptiva:** Se centra en la recopilación, organización, resumen y presentación de datos de una manera que facilite su comprensión e interpretación. Su objetivo principal es describir y resumir las características esenciales de un conjunto de datos, proporcionando una visión general de su distribución y tendencias, sin realizar inferencias o generalizaciones más allá de los datos observados.
- **Estadística inferencial:** Se enfoca en el proceso de hacer inferencias, tomar decisiones y/o realizar predicciones de una población basándose en la información contenida en una muestra de ésta.

Datos

Los **datos** son hechos/informaciones y cifras que se recogen, analizan y resumen para su presentación e interpretación. A todos los datos reunidos para un determinado estudio se les llama **conjunto de datos**.

Empresa	Bolsa de valores	Denominación abreviada Ticker	Posición en *BusinessWeek*	Precio por acción ($)	Ganancia por acción ($)
Abbott Laboratories	N	ABT	90	46	2.02
Altria Group	N	MO	148	66	4.57
Apollo Group	NQ	APOL	174	74	0.90
Bank of New York	N	BK	305	30	1.85
Bristol-Myers Squibb	N	BMY	346	26	1.21
Cincinnati Financial	NQ	CINF	161	45	2.73
Comcast	NQ	CMCSA	296	32	0.43
Deere	N	DE	36	71	5.77
eBay	NQ	EBAY	19	43	0.57
Federated Dept. Stores	N	FD	353	56	3.86
Hasbro	N	HAS	373	21	0.96
IBM	N	IBM	216	93	4.94
International Paper	N	IP	370	37	0.98
Knight-Ridder	N	KRI	397	66	4.13
Manor Care	N	HCR	285	34	1.90
Medtronic	N	MDT	53	52	1.79
National Semiconductor	N	NSM	155	20	1.03
Novellus Systems	NQ	NVLS	386	30	1.06
Pitney Bowes	N	PBI	339	46	2.05
Pulte Homes	N	PHM	12	78	7.67
SBC Communications	N	SBC	371	24	1.52
St. Paul Travelers	N	STA	264	38	1.53
Teradyne	N	TER	412	15	0.84
UnitedHealth Group	N	UNH	5	91	3.94
Wells Fargo	N	WFC	159	59	4.09

Fuente: Business Week (4 de abril de 2005).

Figura: Listado de 25 empresas que hacen parte del índice Standard & Poor's 500. Tabla tomada de [1].

Elementos, Variables y Observaciones

Los **elementos** son las entidades de las que se obtienen los datos. En el conjunto de datos de la tabla anterior, cada acción de una empresa es un elemento; los nombres de los elementos aparecen en la primera columna. Como se tienen 25 acciones, el conjunto de datos contiene 25 elementos.

Una **variable** es una característica de los elementos que es de interés. Para la tabla de la diapositiva anterior se puede ver que hay cinco variables.

Los valores encontrados para cada variable en cada uno de los elementos constituyen los datos. Al conjunto de mediciones obtenidas para un determinado elemento se le llama **observación**.

Escalas de Medición

La recolección de datos requiere alguna de las escalas de medición siguientes: nominal, ordinal, de intervalo o de razón.

- **Escala nominal:** Cuando el dato de una variable es una etiqueta o un nombre que identifica un atributo de un elemento, se considera que la escala de medición es una escala nominal.
- **Escala ordinal:** Una escala de medición para una variable es ordinal si los datos muestran las propiedades de los datos nominales y además existe un orden o una jerarquía en los datos.
- **Escala de intervalo:** Una escala de medición para una variable es una escala de intervalo si los datos tienen las características de los datos ordinales y el intervalo entre valores se expresa en términos de una unidad de medición fija. En estas escalas el cero no implica la ausencia de la variable medida.
- **Escala de razón:** Una variable tiene una escala de razón si los datos tienen todas las propiedades de los datos de intervalo y la proporción entre dos valores tiene significado. Además, el cero es absoluto, es decir, si una variable es igual a cero implica que está ausente.

Datos Cuantitativos y Cualitativos

Los datos se pueden clasificar en dos grupos: los **cualitativos** y los **cuantitativos**. Los datos cualitativos comprenden etiquetas o nombres que se usan para identificar un atributo de cada elemento. Los datos cualitativos emplean la escala nominal o la ordinal y pueden ser numéricos o no. Los datos cuantitativos requieren valores numéricos que hagan referencia a cuánto o cuántos. Los datos cuantitativos se obtienen usando las escalas de medición de intervalo o de razón.

Variables Cuantitativas y Cualitativas

Como es de esperarse, las **variables cuantitativas** solamente toman valores numéricos propios de los datos cuantitativos, mientras que las **variables cualitativas** solo están asociadas a datos cualitativos. Por cierto, a las variables cualitativas también se les conoce como **variables categóricas**.

Variables Discretas y Continuas

Las variables cuantitativas se dividen en dos categorías: las **variables discretas** y las **variables continuas**.

- **Variable discreta:** Una variable discreta es aquella que toma valores individuales aislados o separados, generalmente enteros, y no puede tomar valores intermedios.
- **Variable continua:** Una variable continua es aquella que puede tomar un conjunto infinito de valores dentro de un rango específico y puede tener valores fraccionarios o decimales.

Recapitulación

Dicho lo anterior, ¿qué podemos decir ahora acerca de las variables de la tabla del listado de 25 empresas que hacen parte del índice Standard & Poor's 500?

Visualización de Variables Cualitativas

Para empezar a hablar de la visualización de variables cualitativas es conveniente definir el concepto de **distribución de frecuencia**: una distribución de frecuencia es un resumen tabular de datos que muestra el número (frecuencia) de elementos en cada una de las diferentes clases observadas.

Visualización de Variables Cualitativas

Consideremos los siguientes datos de una venta de bebidas:

Coke Classic	Sprite	Pepsi
Diet Coke	Coke Classic	Coke Classic
Pepsi	Diet Coke	Coke Classic
Diet Coke	Coke Classic	Coke Classic
Coke Classic	Diet Coke	Pepsi
Coke Classic	Coke Classic	Dr. Pepper
Dr. Pepper	Sprite	Coke Classic
Diet Coke	Pepsi	Diet Coke
Pepsi	Coke Classic	Pepsi
Pepsi	Coke Classic	Pepsi
Coke Classic	Coke Classic	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coke Classic	Coke Classic
Coke Classic	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coke Classic	Pepsi	Sprite
Coke Classic	Diet Coke	

Figura: Datos tomados de [1].

Visualización de Variables Cualitativas

Calculando las frecuencias se obtienen los siguientes datos:

Bebida	**Frecuencia**
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

Visualización de Variables Cualitativas

Los datos anteriores se suelen presentar con una gráfica de barras.

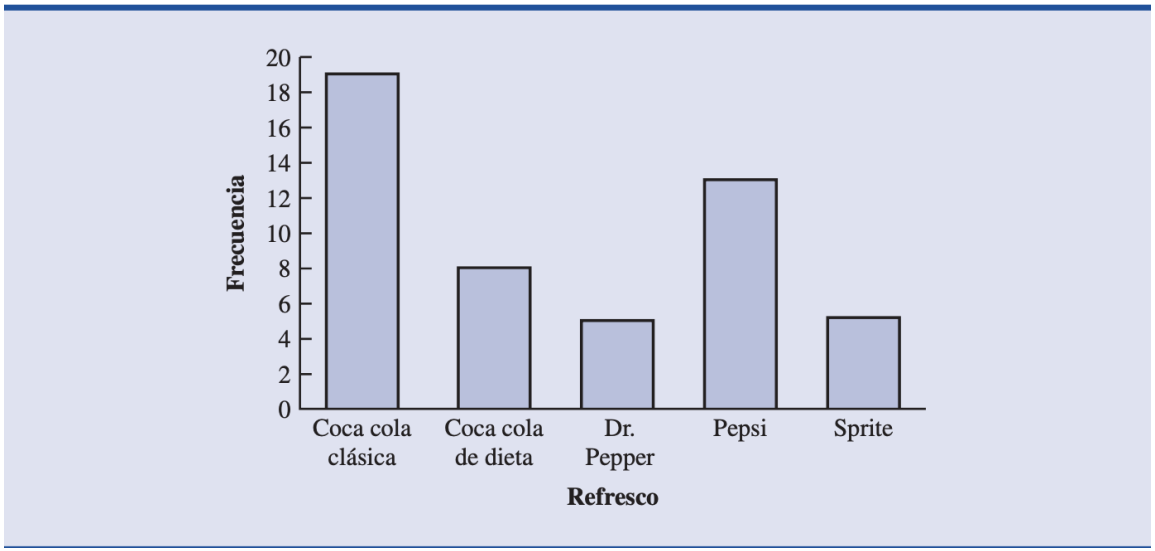

Figura: Gráfica de barras de la venta de bebidas. Imagen tomada de [1].

Visualización de Variables Cualitativas

Otro concepto útil para la visualización de variables cualitativas es el de **frecuencia relativa**: dado un conjunto de datos con n observaciones, la frecuencia relativa de una clase es igual a la proporción de los elementos que pertenecen de dicha clase:

$$\text{Frecuencia relativa de una clase} = \frac{\text{frecuencia de la clase}}{n}.$$

Visualización de Variables Cualitativas

También se tiene la **frecuencia porcentual**, la cual es simplemente la frecuencia relativa expresada como un porcentaje:

$$\text{Frecuencia relativa de una clase} = \left(\frac{\text{frecuencia de la clase}}{n} \right) \times 100\%.$$

Visualización de Variables Cualitativas

Refresco	Frecuencia relativa	Frecuencia porcentual
Coke Classic	0.38	38
Diet Coke	0.16	16
Dr. Pepper	0.10	10
Pepsi	0.26	26
Sprite	0.10	10
Total	1.00	100

Figura: Tabla de las distribuciones de frecuencia relativa y porcentual de la venta de bebidas. Imagen tomada de [1].

Visualización de Variables Cualitativas

Una opción común para visualizar este tipo de frecuencias es la **gráfica de pastel**

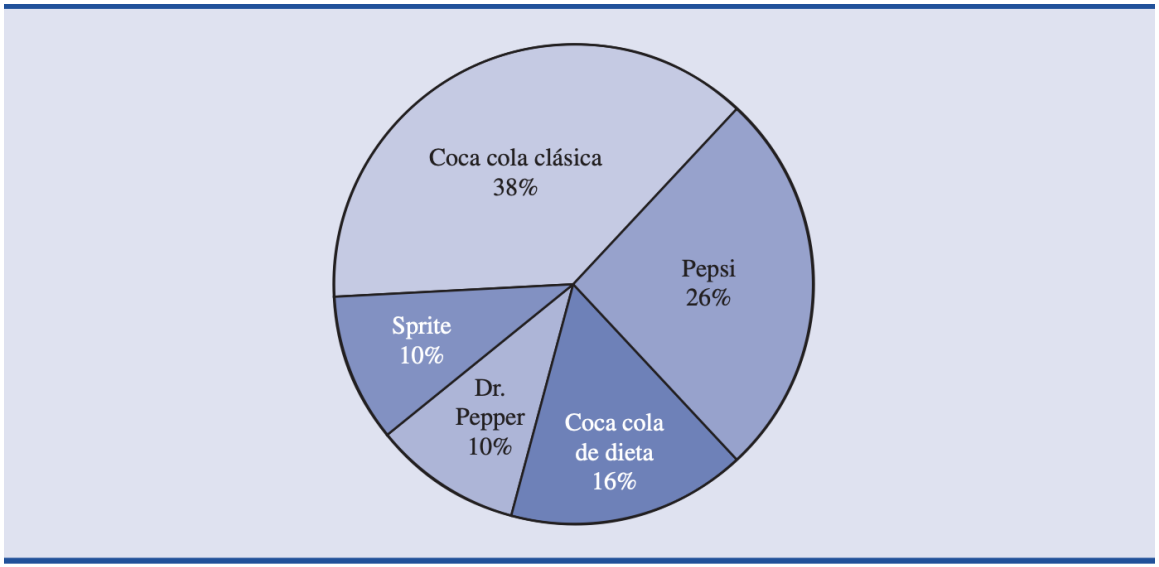

Figura: Gráfica de pastel de las frecuencias porcentuales de la venta de bebidas. Imagen tomada de [1].

Visualización de Variables Cuantitativas

Una distribución de frecuencia es un resumen de datos que presenta el número de elementos (frecuencia) en cada una de las clases disyuntas. Esta definición es válida tanto para datos cualitativos como cuantitativos. Sin embargo, cuando se trata de datos cuantitativos se debe tener más cuidado al definir las clases disyuntas que se van a usar en la distribución de frecuencia.

Visualización de Variables Cuantitativas

Consideremos los datos cuantitativos de la siguiente tabla.

AUDITORÍA ANUAL (DÍAS DE DURACIÓN)			
12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

Figura: En esta tabla se presenta la duración en días de una muestra de auditorías de fin de año de veinte clientes de una empresa pequeña de contadores públicos. Tabla tomada de [1].

Visualización de Variables Cuantitativas

Los tres pasos necesarios para definir las clases de una distribución de frecuencia con datos cuantitativos son los siguientes:

- Determinar el número de clases disyuntas.
- Determinar el ancho de cada clase.
- Determinar los límites de clase.

Visualización de Variables Cuantitativas

DISTRIBUCIÓN DE FRECUENCIA DE LAS AUDITORÍAS

Duración de las auditorías (días)	Frecuencia
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

Figura: En este caso, el número de clases es cinco, el ancho de cada clase también es cinco, y los límites de cada clase se pueden apreciar en la tabla. Tabla tomada de [1].

Visualización de Variables Cuantitativas

Uno de los más sencillos resúmenes gráficos de datos son las **gráficas de puntos**. En el eje horizontal se presenta el intervalo de los datos. Cada dato se representa por un punto colocado sobre este eje.

Visualización de Variables Cuantitativas

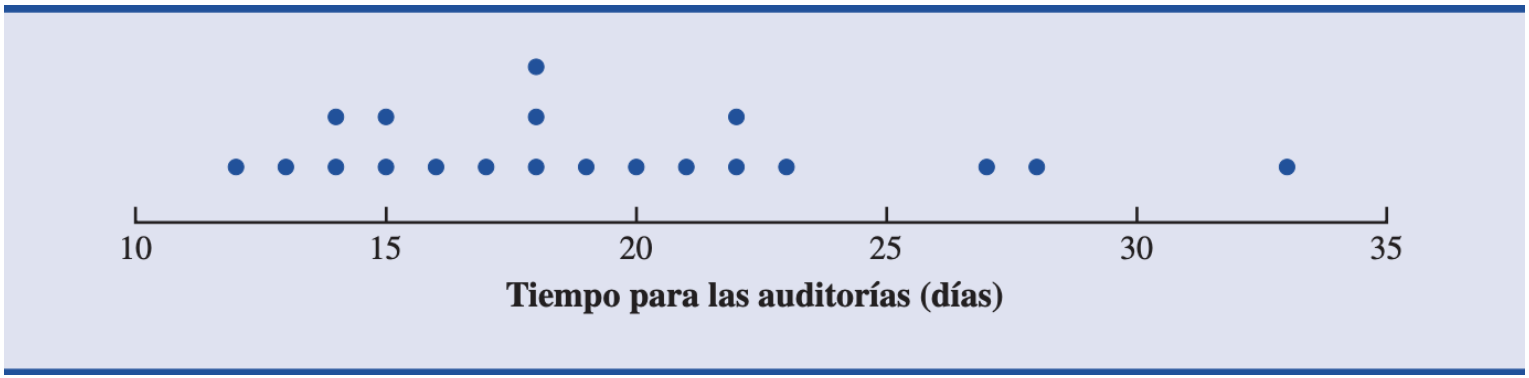

Figura: Gráfica de puntos de las duraciones de las auditorías. Nótese que los tres puntos que se encuentran sobre el 18 del eje horizontal indican que hubo tres auditorías de 18 días. Imagen tomada de [1].

Visualización de Variables Cuantitativas

Una presentación gráfica usual para datos cuantitativos es el **histograma**. Esta gráfica se hace con datos previamente resumidos mediante una distribución ya sea de frecuencias, de frecuencias relativas o de frecuencias porcentuales. Un histograma se construye colocando la variable de interés en el eje horizontal y la frecuencia, la frecuencia relativa o la frecuencia porcentual en el eje vertical.

Visualización de Variables Cuantitativas

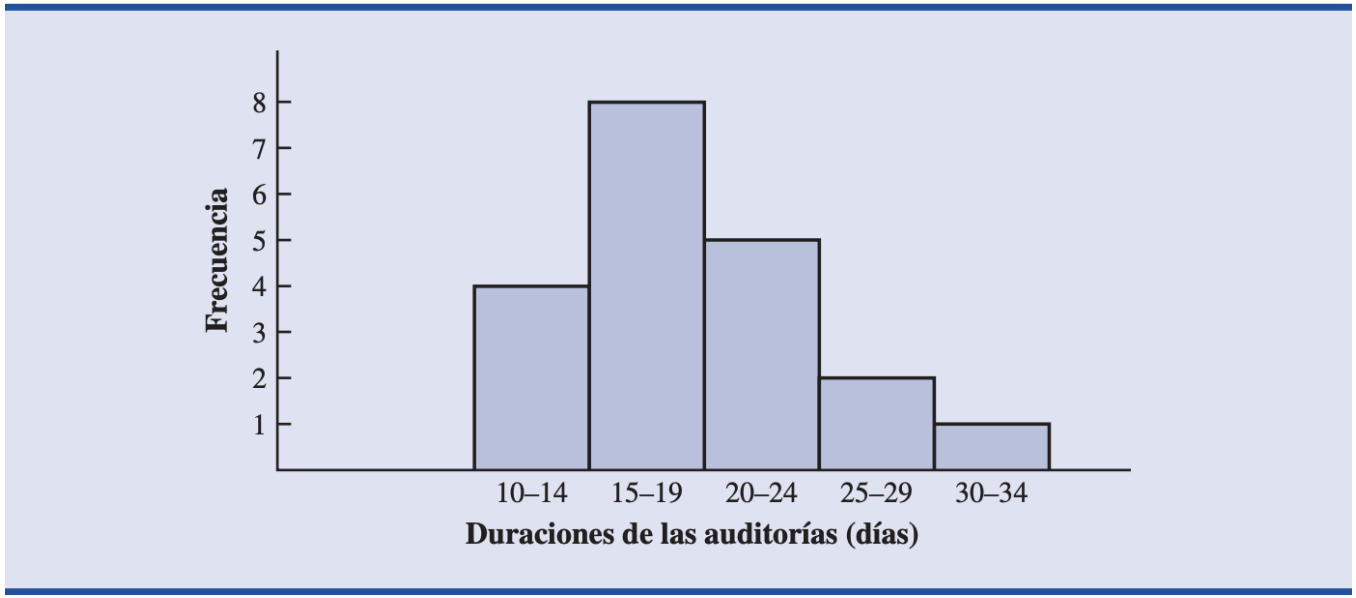

Figura: Histograma de las duraciones de las auditorías. Un histograma de las distribuciones de frecuencia relativa o porcentual de estos datos se vería exactamente igual, excepto que en el eje vertical se colocan los valores de frecuencia relativa o porcentual. Imagen tomada de [1].

Visualización de Variables Cuantitativas

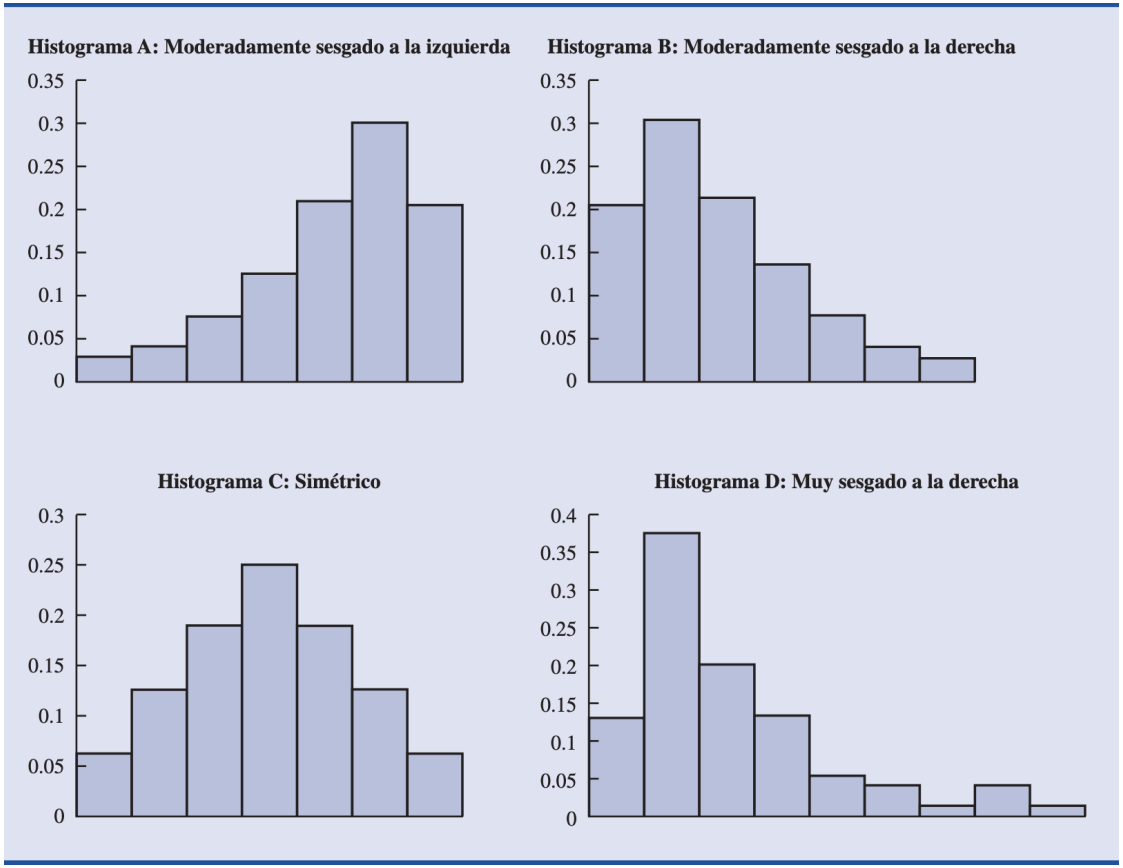

Figura: Uno de los usos más importantes de un histograma es proveer información acerca de la forma de la distribución. En esta figura se muestran cuatro histogramas construidos a partir de distribuciones de frecuencia relativa. En el histograma A se muestra un conjunto de datos moderadamente sesgado a la izquierda. Se dice que un histograma es sesgado a la izquierda si su cola se extiende más en esta dirección. Imagen tomada de [1].

Visualización de Variables Cuantitativas

Una variación de las distribuciones de frecuencia que proporcionan otro resumen tabular de datos cuantitativos es la **distribución de frecuencia acumulada**. La distribución de frecuencia acumulada usa la cantidad, las amplitudes y los límites de las clases de la distribución de frecuencia. Sin embargo, en lugar de mostrar la frecuencia de cada clase, la distribución de frecuencia acumulada muestra la cantidad de datos que tienen un valor menor o igual al límite superior de cada clase.

Visualización de Variables Cuantitativas

Duración de la auditoría en días	Frecuencia acumulada	Frecuencia relativa acumulada	Frecuencia porcentual acumulada
Menor o igual que 14	4	0.20	20
Menor o igual que 19	12	0.60	60
Menor o igual que 24	17	0.85	85
Menor o igual que 29	19	0.95	95
Menor o igual que 34	20	1.00	100

Figura: Tabla de distribuciones de frecuencias acumulas de las duraciones de las auditorías. Imagen tomada de [1].

Visualización de Variables Cuantitativas

Una gráfica que es útil para visualizar una distribución de frecuencia acumulada es la **ojiva**. Esta gráfica muestra los valores de los datos en el eje horizontal y las frecuencias acumuladas, las frecuencias relativas acumula- das o las frecuencias porcentuales acumuladas en el eje vertical.

Visualización de Variables Cuantitativas

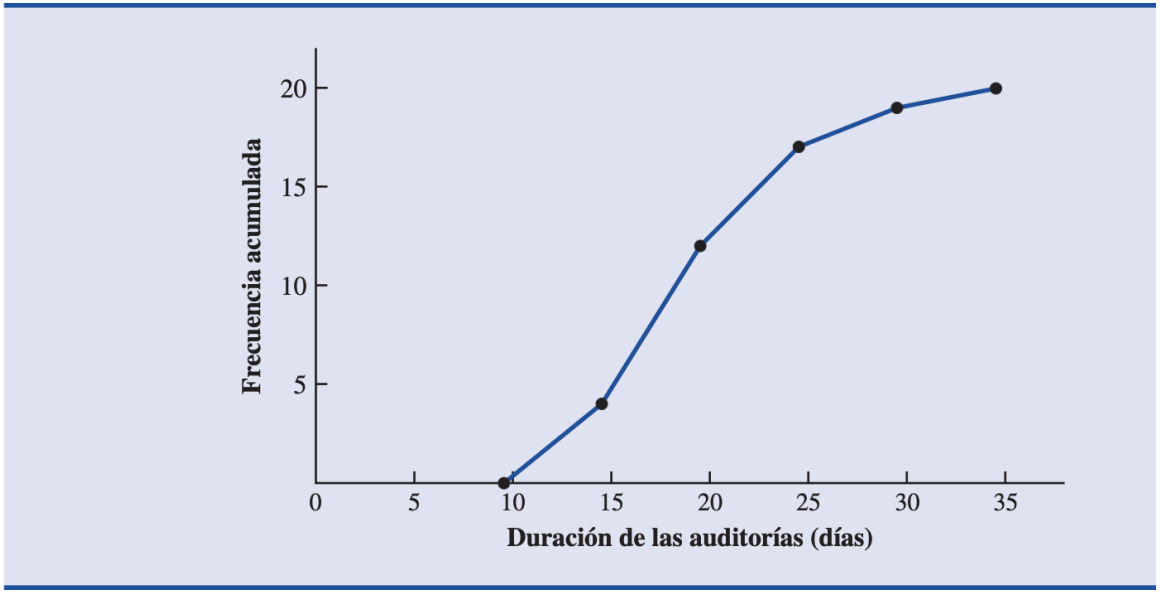

Figura: En esta figura se muestra la ojiva correspondiente a las frecuencias acumuladas de las duraciones de las auditorías. Imagen tomada de [1].

Visualización de Variables Cuantitativas

Hasta el momento nos hemos concentrado en los métodos tabulares y gráficos empleados para resumir datos de una sola variable. Sin embargo, hay ocasiones en las cuales es muy importante recurrir a métodos tabulares o gráficos que nos permitan entender la relación entre dos variables. La **tabulación cruzada** y los **diagramas de dispersión** son dos métodos que cumplen con este objetivo.

Visualización de Variables Cuantitativas

	Precio				
Calidad	**$10–19**	**$20–29**	**$30–39**	**$40–49**	**Total**
Buena	42	40	2	0	84
Muy buena	34	64	46	6	150
Excelente	2	14	28	22	66
Total	78	118	76	28	300

Figura: Una tabulación cruzada es un resumen tabular de los datos de dos variables. En la tabulación cruzada que se muestra en esta figura se muestran los datos que corresponden a la calidad y precios de 300 restaurantes en el área de Los Ángeles. Imagen tomada de [1].

Visualización de Variables Cuantitativas

Es posible combinar o agregar los datos de dos o más tabulaciones cruzadas para obtener una tabulación cruzada resumida que muestre la relación entre dos variables. En tales casos hay que tener mucho cuidado al sacar conclusiones acerca de la relación entre las dos variables de la tabulación cruzada agregada ya que es posible que las conclusiones obtenidas se invierten por completo al observar los datos no agregados. Esta situación se le conoce como **paradoja de Simpson**.

Visualización de Variables Cuantitativas

Consideremos el siguiente escenario: Los jueces Ron Luckett y Dennis Kendall presidieron los tres últimos años dos tipos de tribunales: de primera instancia y municipal. En la mayor parte de los casos los tribunales de apelación ratificaron las sentencias, pero en algunos casos fueron revocadas. Para cada juez se elabora una tabulación cruzada con las variables sentencia (ratificada o revocada) y tipo de tribunal (de primera instancia y municipal). Supongamos que se combinan las dos tabulaciones cruzadas agregando los datos de los dos tipos de tribunales.

Visualización de Variables Cuantitativas

	Juez		
Sentencia	**Luckett**	**Kendall**	**Total**
Ratificada	129 (86%)	110 (88%)	239
Revocada	21 (14%)	15 (12%)	36
Total (%)	150 (100%)	125 (100%)	275

Figura: La tabulación cruzada agregada que se obtiene tiene dos variables: sentencia (ratificada o revocada) y juez (Luckett o Kendall). En esta tabulación cruzada para cada uno de los jueces se da la cantidad de sentencias que fueron ratificadas y la cantidad de sentencias que fueron revocadas. En la tabla se presentan estos resultados junto a los porcentajes de columna entre paréntesis al lado de cada valor. Tabla tomada de [1].

Visualización de Variables Cuantitativas

Juez Luckett

Sentencia	Tribunal de primera instancia	Tribunal municipal	Total
Ratificada	29 (91%)	100 (85%)	129
Revocada	3 (9%)	18 (15%)	21
Total (%)	32 (100%)	118 (100%)	150

Juez Kendall

Sentencia	Tribunal de primera instancia	Tribunal municipal	Total
Ratificada	90 (90%)	20 (80%)	110
Revocada	10 (10%)	5 (20%)	15
Total (%)	100 (100%)	25 (100%)	125

Figura: Al comparar los porcentajes de columna de los dos jueces es claro que el juez Luckett tuvo un mejor desempeño en ambos tribunales que el Juez Kendall. Esto contradice las conclusiones obtenidas al agregar los datos de los dos tribunales en la primera tabulación cruzada. Tabla tomada de [1].

Visualización de Variables Cuantitativas

Un diagrama de dispersión es una representación gráfica de la relación entre dos variables cuantitativas y una línea de tendencia es una recta que da una aproximación de la relación de las variables. Como ejemplo consideremos la relación publicidad/ventas de una tienda de equipos de sonido.

Semana	Número de comerciales x	Ventas (en cientos de dólares) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

Figura: Datos recopilados durante diez semanas que muestran el comportamiento de dos variables cuantitativas: números de comerciales y ventas. Tabla tomada de [1].

Visualización de Variables Cuantitativas

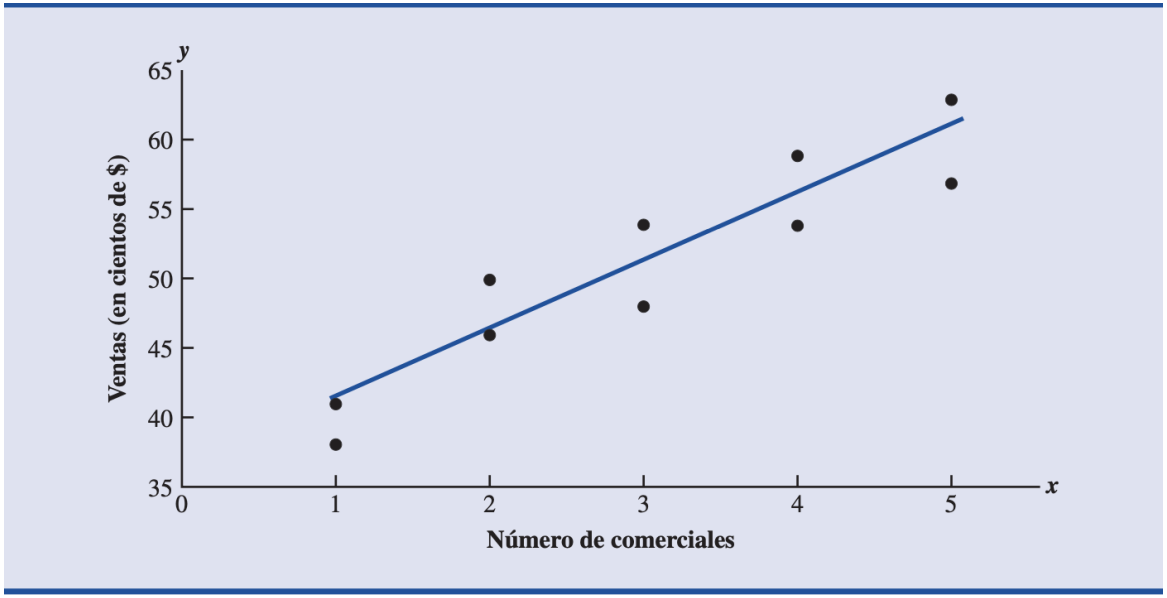

Figura: Diagrama de dispersión y línea de tendencia de la tienda de equipos de sonido. Imagen tomada de [1].

Visualización de Variables Cuantitativas

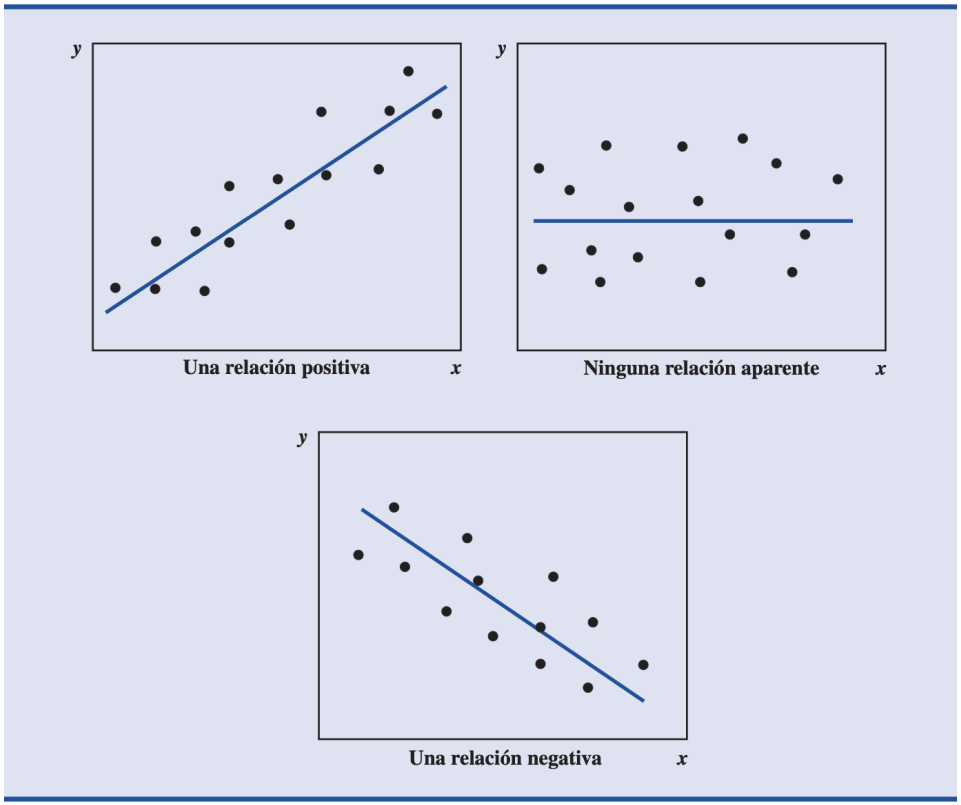

Figura: En esta figura se presentan algunos de los patrones de los diagramas de dispersión y el tipo de relación que sugieren. La gráfica de la primera fila a la izquierda representa una relación positiva parecida a la del ejemplo de la cantidad de comerciales y las ventas. En la gráfica que se ubica a la derecha en la fila superior no aparece ninguna relación entre las dos variables. La gráfica inferior representa una relación negativa en la que la variable y tiende a disminuir a medida que la variable x aumenta. Es relevante resaltar que los diagramas de dispersión exhiben una gran diversidad de patrones, los cuales no siempre muestran una relación apróximadamente lineal entre dos variables. Imagen tomada de [1].

Medidas de Tendencia Central

Antes de hablar de **medidas de tendencia central**, también conocidas como **medidas de localización**, es importante definir los conceptos de **población** y **muestra**.

- **Población:** La población se refiere al conjunto completo de elementos o individuos que comparten una característica común o son el objeto de estudio en una investigaci'on. Esta población puede ser finita o infinita dependiendo del contexto.
- **Muestra:** Una muestra es un subconjunto cuidadosamente seleccionado de elementos o individuos tomados de una población más grande. La muestra se elige de manera que sea representativa, es decir, que capture bien la diversidad interna de los datos de la de la población, lo cual es fundamental para realizar inferencias y generalizaciones de la población completa.

Medidas de Tendencia Central

Las medidas de localización, o tendencia central, son útiles para, dada una variable y su distribución, estimar alrededor de que punto los valores de dicha variable se localizan.

Si estas medidas se calculan con los datos de una muestra se llaman **estadísticos muestrales**; si son calculadas con los datos de una población se llaman **parámetros poblacionales**. En estadística inferencial, al estadístico muestral se le conoce como el **estimador puntual** del correspondiente parámetro poblacional.

Medidas de Tendencia Central

La **media** es la medida de localización más común. Si calculamos la media con todos los datos de la población decimos que está es igual a μ; si calculamos la media con una muestra de la población la denotamos como $\bar{x}$, la cual se conoce como la **media muestral**.

Sea $m = \{x_1, x_2, \ldots, x_n\}$ una muestra de n elementos. La media muestral se define como

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

La media muestral $\bar{x}$ es frecuentemente utilizada para estimar el valor de la media poblacional μ.

Medidas de Tendencia Central

La **mediana** es otra medida de tendencia central. Es el valor de intermedio de los datos ordenados de menor a mayor. Cuando el número de elementos de la muestra es impar, la mediana es el valor intermedio; si la cantida de elementos de la muestra es par, la mediana es definida como el promedio de las dos observaciones intermedias.

Por cierto, una de las características relevantes de la mediana es que su valor tiende a variar menos que el valor de la media ante la presencia de datos atípicos.

Medidas de Tendencia Central

La tercera medida de localización es la **moda**. La moda se define como el valor que se presenta con mayor frecuencia.

Medidas de Posición

El **percentil** es una medida de posición que indica aporta información acerca de la dispersión de los datos en el intervalo que va del menor al mayor valor de los datos. El percentil p es un valor tal que por lo menos p por ciento de las observaciones son menores o iguales a este valor y a lo sumo $(100 - p)$ por ciento de las observaciones son mayores o iguales que este valor.

Medidas de Posición

Sean nuestros datos $\{x_1, x_2, \ldots, x_n\}$ ordenados de manera ascendente. Para obtener el percentil p, calculamos primero la siguiente expresión:

$$i = \frac{p}{100} n.$$

Si $i \notin \mathbb{N}$, entonces $x_{\lceil i \rceil}$ es el percentil p, en caso contrario, el percentil p es igual a

$$\frac{x_i + x_{i+1}}{2}.$$

Medidas de Posición

El concepto de percentil nos permite definir los **cuartiles**: el primer cuartil Q_1 es el percentil 25, el segundo cuartil Q_2 es el percentil 50, el tercer cuartil Q_3 es el percentil 75. Nótese que Q_2 es igual a la mediana.

Medidas de Posición

Los valores en una muestra son 27, 25, 20, 15, 30, 34, 28 y 25. ¿A qué son iguales Q_1, Q_2 y Q_3?

Medidas de Variabilidad

La medida de variabilidad más simple es el **rango**:

$$\text{rango} = \text{valor mayor} - \text{valor menor}.$$

¿A qué es igual el rango de los datos de la diapositiva anterior?

Medidas de Variabilidad

Una medida robusta ante la presencia de valores extremos es el **rango intercuartílico** . Este sse define como la diferencia entre el tercer cuartil y el primer cuartil:

$$\text{rango intercuartílico} = Q_3 - Q_1.$$

¿Cuál es el rango intercuartílico de los datos que acabamos de ver?

Medidas de Variabilidad

La **varianza** es una medida de dispersión que utiliza todos los datos. Si tenemos acceso a todos los elementos de una población, la varianza es igual a

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2,$$

donde N es el número de elementos de la población, y μ es la media poblacional.

Medidas de Variabilidad

Si trabajamos con una muestra de la población, la varianza se define como

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

donde n es el número de elementos de la muestra, y $\bar{x}$ es la media muestral. Por cierto, s^2 es un estimador sin sesgo de la varianza poblacional.

Medidas de Variabilidad

La **desviación estándar** se define como la raíz cuadrada positiva de la varianza: si estamos trabajando con todos los datos de una población, la desviación estándar es igual a σ, si consideramos solamente los elementos de una muestra la desviación estándar es igual a s.

Una de las ventajas de la desviación estándar es que es más fácil de interpretar que la varianza debido a que la ésta se mide tiene las mismas unidades de los datos.

Medidas de la forma de la Distribución

Una medida numérica importante de la forma de una distribución es el **sesgo**. Una fórmula para calcular el sesgo de una muestra de n elementos está dada por

$$\text{sesgo} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 .$$

Nótese que $\bar{x}$ y s son la media muestral y la desviación estándar muestral, respectivamente.

Medidas de la forma de la Distribución

Por cierto, en una distribución simétrica, la media y la mediana son iguales. Si los datos están sesgados a la derecha, la media será mayor que la mediana; si los datos están sesgados a la izquierda, la media será menor que la mediana.

Visualización de Variables Cuantitativas

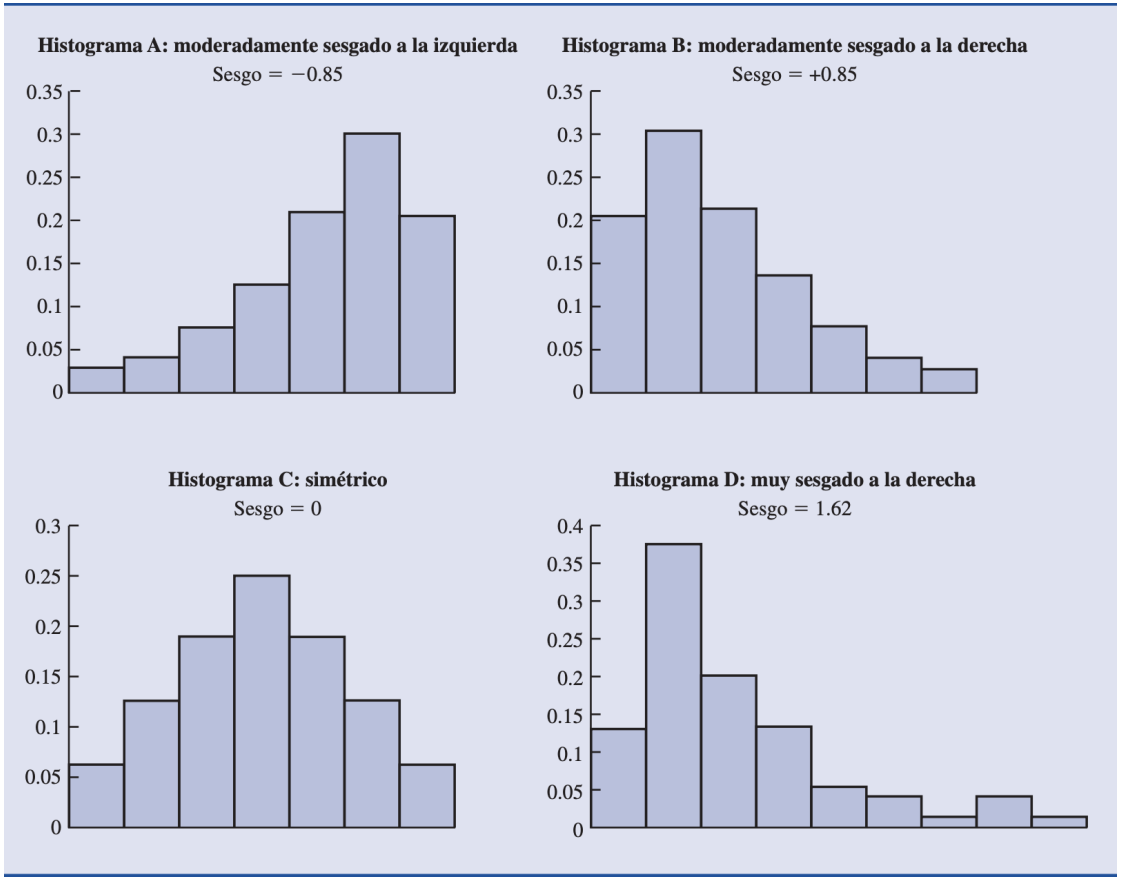

Figura: Sesgos de cuatro distribuciones. El histograma A es sesgado a la izquierda, su sesgo es -0.85; el histograma B es sesgado a la derecha, su sesgo es 0.85; el histograma C es simétrico, su sesgo es cero; el histograma D es muy sesgado a la derecha, su sesgo es 1.62. Imagen tomada de [1].

Medidas de la Asociación entre dos Variables

Hasta ahora se han examinado métodos numéricos que resumen datos en una sola variable. Con frecuencia es necesario conocer la relación entre dos variables. Una medida que nos permite medir esta asociación entre variables es la **covarianza**, la cual nos da un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias.

Medidas de la Asociación entre dos Variables

Sean X y Y dos variables aleatorias y sea $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ un conjunto de n observaciones de estas variables. La **covarianza muestral** se define como

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Medidas de la Asociación entre dos Variables

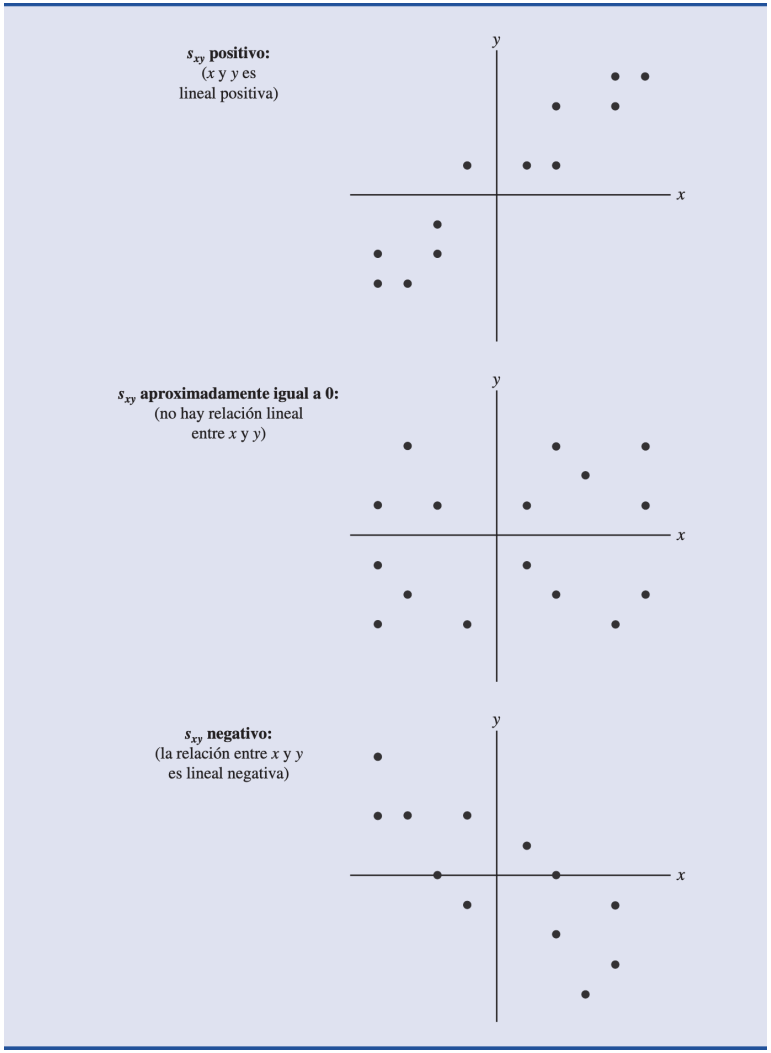

Figura: Tres situaciones que se pueden presentar en el cálculo de la covarianza. Si esta es positiva, entonces a medida que la variable X aumenta, la variable Y también lo hará; si la covarianza es negativa, esto implica que si X aumenta, Y disminuye; si la covarianza es cercana a cero, entonces no hay una relación fuerte entre las dos variables. Imagen tomada de [1].

Medidas de la Asociación entre dos Variables

La versión normalizada de la covarianza se le conoce como el **coeficiente de correlación**, el cual se define como

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y},$$

donde s_x y s_y son las desviaciones estándar muestrales de las variables X y Y, respectivamente.

Medidas de la Asociación entre dos Variables

Por cierto, ρ_{xy} mide qué tan lineal es la relación entre las variables X y Y: si

$$Y = aX + b,$$

donde $a, b \in \mathbb{R}$, entonces $\rho_{xy} = -1$ si y solo si $a < 0$, y $\rho_{xy} = 1$ si y solo si $a > 0$. Esto implica que

$$-1 \leq \rho_{xy} \leq 1.$$

BIBLIOGRAFÍA

1 Anderson, D. R., Sweeney, D. J., Williams, T. A., *"Estadística para Administración y Economía"*, Décima Edición, CENGAGE Learning, 2008.