ESTADÍSTICA DESCRIPTIVA

UNIDAD I

I.1 DEFINICIÓN Y CLASIFICACIÓN DE VARIABLES

La *estadística descriptiva* es la rama de las Matemáticas que recolecta, presenta y caracteriza un conjunto de datos (por ejemplo, edad de una población, altura de los estudiantes de una escuela, temperatura en los meses de verano, etc.) con el fin de describir apropiadamente las diversas características de ese conjunto.

Al conjunto de los distintos valores numéricos que adopta un carácter cuantitativo se llama *variable estadística*.

Las variables pueden ser de dos tipos:

- Variables *cualitativas o categóricas*: no se pueden medir numéricamente (por ejemplo: nacionalidad, color de la piel, sexo).

- Variables *cuantitativas*: tienen valor numérico (edad, precio de un producto, ingresos anuales).

Las variables también se pueden clasificar en:

- Variables *unidimensionales*: sólo recogen información sobre una característica (por ejemplo: edad de los alumnos de una clase).

- Variables *bidimensionales*: recogen información sobre dos características de la población (por ejemplo: edad y altura de los alumnos de una clase).

- Variables *pluridimensionales*: recogen información sobre tres o más características (por ejemplo: edad, altura y peso de los alumnos de una clase).

Por su parte, las variables cuantitativas se pueden clasificar en discretas y continuas:

- *Discretas*: sólo pueden tomar valores enteros (1, 2, 8, -4, etc.). Por ejemplo: número de hermanos (puede ser 1, 2, 3...., etc., pero, por ejemplo, nunca podrá ser 3.45).

- *Continuas*: pueden tomar cualquier valor real dentro de un intervalo. Por ejemplo, la velocidad de un vehículo puede ser 90.4 km/h, 94.57 km/h...etc.

Cuando se estudia el comportamiento de una variable hay que distinguir los siguientes conceptos:

- *Individuo*: cualquier elemento que porte información sobre el fenómeno que se estudia. Así, si estudiamos la altura de los niños de una clase, cada alumno es un individuo; si se estudia el precio de la vivienda, cada vivienda es un individuo.

- *Población*: conjunto de todos los individuos (personas, objetos, animales, etc.) que porten información sobre el fenómeno que se estudia. Por ejemplo, si se estudia el precio de la vivienda en una ciudad, la población será el total de las viviendas de dicha ciudad.

- *Muestra*: subconjunto que seleccionado de una población. Por ejemplo, si se estudia el precio de la vivienda de una ciudad, lo normal será no recoger información sobre todas las viviendas de la ciudad

(sería una labor muy compleja), sino que se suele seleccionar un subgrupo (muestra) que se entienda que es suficientemente representativo.

Las *variables aleatorias* son variables que son seleccionadas al azar o por procesos aleatorios.

I.2 DATOS. CLASIFICACIÓN, ORGANIZACIÓN Y CONSTRUCCIÓN DE BLOQUES ESTADÍSTICOS

Los *datos* son medidas y/o números recopilados a partir de la observación. Los datos pueden concebirse como información numérica necesaria para ayudar a tomar una decisión con más bases en una situación particular.

Existen muchos métodos mediante los cuales se pueden obtener datos necesarios. Primero, se puede buscar datos ya publicados por otras fuentes. Segundo, se puede diseñar un experimento. En tercer lugar, se puede conducir un estudio. Cuarto, se pueden hacer observaciones del comportamiento, actitudes u opiniones de los individuos en los que se está interesado.

Los datos se pueden clasificar en:

- Datos discretos. Son respuestas numéricas que surgen de un proceso de conteo.
- Datos continuos. Son respuestas numéricas que surgen de un proceso de medición.

I.2.1 ESCALAS DE MEDICIÓN

Medir en el campo de las ciencias exactas es comparar una magnitud con otra, tomada de manera arbitraria como referencia, denominada patrón y expresar cuántas veces la contiene. En el campo de las ciencias sociales medir es “el proceso de vincular conceptos abstractos con indicadores empíricos”. Al resultado de medir lo se le llama *medida*.

La medición de las variables puede realizarse por medio de cuatro escalas de medición: la nominal, ordinal, de intervalo y de razón. Se utilizan para ayudar en la clasificación de las variables, el diseño de las preguntas para medir variables, e incluso indican el tipo de análisis estadístico apropiado para el tratamiento de los datos.

Una característica esencial de la medición es la dependencia que tiene de la posibilidad de variación. La validez y la confiabilidad de la medición de una variable depende de las decisiones que se tomen para operarla y lograr una adecuada comprensión del concepto evitando imprecisiones y ambigüedades, en caso contrario, la variable corre el riesgo inherente de ser invalidada debido a que no produce información confiable.

a) Medición Nominal.

En este nivel de medición se establecen categorías distintivas que no implican un orden específico. Por ejemplo, si la unidad de análisis es un grupo de personas, para clasificarlas se puede establecer la categoría sexo con dos niveles, masculino (M) y femenino (F), los encuestados sólo tienen que señalar su género, no se requiere de un orden real.

Así, se pueden asignar números a estas categorías para su identificación: 1=M, 2=F o bien, se pueden invertir los números sin que afecte la medición: 1=F y 2=M. En resumen en la escala nominal se asignan números a eventos con el propósito de identificarlos.

b) Medición Ordinal.

Se establecen categorías con dos o más niveles que implican un orden inherente entre si. La escala de medición ordinal es cuantitativa porque permite ordenar a los eventos en función de la mayor o menor posesión de un atributo o característica. Por ejemplo, en las instituciones escolares de nivel básico suelen formar por estatura a los estudiantes, se desarrolla un orden cuantitativo pero no suministra medidas de los sujetos. Estas escalas admiten la asignación de números en función de un orden prescrito. Las formas más comunes de variables ordinales son ítems (reactivos) actitudinales estableciendo una serie de niveles que expresan una actitud de acuerdo o desacuerdo con respecto a algún referente. Por ejemplo, ante el reactivo: Pemex debe privatizarse, el respondiente puede marcar su respuesta de acuerdo a las siguientes alternativas:

__ Totalmente de acuerdo
__ De acuerdo
__ Indiferente
__ En desacuerdo
__ Totalmente en desacuerdo

Las anteriores alternativas de respuesta pueden codificarse con números que van del uno al cinco que sugieren un orden preestablecido pero no implican una distancia entre un número y otro.

c) Medición de Intervalo.

La medición de intervalo posee las características de la medición nominal y ordinal. Establece la distancia entre una medida y otra. La escala de intervalo se aplica a variables continuas pero carece de un punto cero absoluto. El ejemplo más representativo de este tipo de medición es un termómetro, cuando registra cero grados centígrados de temperatura indica el nivel de congelación del agua y cuando registra 100 grados centígrados indica el nivel de ebullición, el punto cero es arbitrario no real, lo que significa que en este punto no hay ausencia de temperatura.

d) Medición de Razón.

Una escala de medición de razón incluye las características de los tres anteriores niveles de medición (nominal, ordinal e intervalo). Determina la distancia exacta entre los intervalos de una categoría. Adicionalmente tiene un punto cero absoluto, es decir, en el punto cero no existe la característica o atributo que se mide. Las variables de ingreso, edad, número de hijos, etc. son ejemplos de este tipo de escala. El nivel de medición de razón se aplica tanto a variables continuas como discretas.

I.2.2 ORGANIZACIÓN DE DATOS

Muchas veces uno se pregunta, ¿para qué sirven las encuestas que a veces se hacen en la calle?, ¿Cómo saber si una estación de radio se escucha más que otra? , ¿Cuál candidato puede ganar? La respuesta se comienza con la recaudación de datos.

Los datos son información que se recoge, esto puede ser opinión de las personas sobre un tema, edad o sexo de encuestados, dónde viven, cuántas personas viven en una casa, qué tipo de sangre tiene un grupo de personas, etc.

Hay datos que pueden ser de mucha utilidad a diferentes profesionales en la toma de decisiones, para resolver problemas o para mostrar resultados de investigaciones. Una vez que se haya recogido toda la información, se procede a crear una base de datos, donde se registran todos los datos obtenidos. Algunas veces, si los datos son muy complicados, se codifican, esto quiere decir que se le coloca una palabra clave que identifica un título muy largo. Cuando ya está elaborada la base de datos se parece a una tabla.

Es importante recordar que nunca se colocan las tablas y las gráficas juntos, porque en realidad dicen lo mismo, corrientemente se utiliza o una tabla y su análisis, o una gráfica y su análisis[1].

Por ejemplo, supóngase que se ha preguntado a un conjunto de n personas: ¿qué opinión tienen acerca de la instalación de playas en la Ciudad de México en que el Gobierno del Distrito Federal ha hecho a partir de 2007? Las n respuestas se encuentran en una escala que va de 1 a 9, donde 1 representa un total desacuerdo con la medida mientras que 9 quiere significar un acuerdo total.

El resultado de la medición es el siguiente:

Tabla 1: Conjunto original de datos

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

Si se plantean las siguientes preguntas:

- ¿Cuántas personas fueron encuestadas?
- ¿Cuál fue la respuesta más frecuente?
- ¿Cuántas personas tienen, como máximo, una actitud de cuatro puntos en la escala? (es decir, ¿cuántas personas se encuentran en desacuerdo con la medida?)

Es difícil responder a las tres cuestiones. ¿Cuál es el problema?

Las personas tienen dificultades para procesar o tener en cuenta mucha información de forma simultanea. La tabla 1 muestra demasiados datos y es preciso contar con mucha paciencia y una buena vista para responder a las preguntas anteriores con seguridad.

Así pues, ¿qué se puede hacer? Una solución alternativa al repaso repetitivo de la tabla 1 es organizar los datos de tal forma que tengan una disposición que facilite la lectura. En este sentido, la primera acción a realizar es ordenar los datos desde el que posee el valor más pequeño hasta el que cuenta con el valor mayor.

Obsérvese el resultado:

Tabla 2: Conjunto ordenado de datos

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

Se logra una ganancia al pasar de la tabla 1 a la tabla 2. Parece que ésta es más fácil de interpretar. No ha desaparecido ninguna información, el único cambio está en su ordenación. No obstante, la solución es parcial, puesto que aún debe ser mejorada (sigue siendo difícil responder a las preguntas).

[1] Generalmente, el titulo de la tabla va encima de ésta, mientras que el título de una figura va por debajo. El título, de ambas, sólo lleva la primera palabra en mayúscula y no va subrayado.

Si se observa la tabla 2, contiene una sucesión de datos con valores repetidos. Por ejemplo, el valor 1 se encuentra presente en seis ocasiones. Una buena estrategia es mostrar una sola vez cada valor y hacerlo seguir por su frecuencia, es decir, por la cantidad de ocasiones en que aparece, tal como lo muestra la tabla 3:

Tabla 3: Conjunto ordenado de valores y frecuencias

1 (6), 2 (11), 3 (12), 4 (30), 5 (40), 6 (25), 7 (14), 8 (9), 9 (3)

Aún se puede disponer la información de tal forma que resulte fácil responder a preguntas que se han planteado. En la tabla 3 se ha mantenido la misma disposición que en la tabla 2. Esto es innecesario. Para disponer la información de manera óptima, se genera una tabla que tenga dos columnas. En la columna primera se presentarán los valores, que se representa con la letra x mientras que en la segunda columna se dispondrán las frecuencias, que se representa con la letra f. Obsérvese el resultado en la tabla 4:

Tabla 4: Tabla de frecuencias

x	f
1	6
2	11
3	12
4	30
5	40
6	25
7	14
8	9
9	3
Total:	150

Ahora sí, la tabla de frecuencias permite responder a las preguntas planteadas con facilidad:
¿Cuántas personas fueron encuestadas? Solución: 150
¿Cuál fue la respuesta más frecuente? Solución: 5 (40 datos)
¿Cuántas personas tienen, como máximo, una actitud de cuatro puntos en la escala? Solución: 59 (6+11+12+30)

I.2.3 ACUMULACIÓN DE FRECUENCIAS

No todas las preguntas que se han realizado sobre el mismo conjunto de datos han exigido el mismo esfuerzo. Así, mientras que las preguntas sobre el número de datos y el valor más frecuente se han respondido con una lectura de la tabla, la tercera pregunta ha necesitado de algunas operaciones:

¿Cuántas personas tienen, como máximo, una actitud de cuatro puntos en la escala? Solución: 59 (6+11+12+30). Para responder a esa pregunta se ha tenido que realizar una suma: la de todas las frecuencias comprendidas entre el primer valor de la tabla y el valor que interesa, ambos inclusive. Esta cantidad final recibe el nombre de frecuencia acumulada.

Muchas interrogantes requieren respuestas que se basan en las frecuencias acumuladas. Es recomendable escribir esta nueva información en la tabla, de tal forma que permita respuestas directas en el futuro. Véase el resultado:

Tabla 5: Tabla de frecuencias de tres columnas

x	f	F
1	6	6
2	11	17
3	12	29
4	30	59
5	40	99
6	25	124
7	14	138
8	9	147
9	3	150
Total:	150	

Imagínese ahora que se ha preguntado a 25 alumnos por el nombre de la colonia en la que habitan, obteniendo los siguientes resultados:

Álamos, Portales, Nápoles, Mixcoac, Mixcoac, Plateros, Nápoles, Florida, Álamos, Tepeaca, Copilco, Nápoles, Florida, Álamos, Portales, Copilco, Tepeaca, Portales, Tepeaca, Florida, Tepeaca, Álamos, Plateros, Copilco, Plateros

Si se construye una tabla de frecuencias con la información sobre las colonias en que tienen su domicilio, utilizando la siguiente equivalencia:

Colonia	Código
Álamos	1
Copilco	2
Florida	3
Mixcoac	4
Nápoles	5
Portales	6
Plateros	7
Tepeaca	8

Una posibilidad es esta:

Tabla 6: Distribución por colonias

Colonia	Código	f	F
Álamos	1	4	4
Copilco	2	3	7
Florida	3	3	10
Mixcoac	4	2	12
Nápoles	5	3	15
Portales	6	3	18
Plateros	7	3	21
Tepeaca	8	4	25

Sin embargo, al analizar esta tabla se concluye: ¿qué sentido tiene acumular frecuencias en el problema que se ha planteado sobre las colonias? Por ejemplo, no tiene ningún significado la cantidad 12 que acompaña al valor 4 (Mixcoac).

La diferencia esencial entre el problema de las colonias y el de las respuestas a la escala de acuerdo, se encuentra en el tipo de variable. En el caso de las colonias, éstas no pueden ordenarse en función de ser más o ser menos "colonia de domicilio" (se pueden ordenar según número de habitantes, extensión, altitud media, etc. Pero no en función de ser más o ser menos colonia).

Por lo tanto, la acumulación de frecuencias sólo procede si los valores de la variable que se está estudiando se pueden ordenar. Así, la respuesta correcta al problema debe ser:

Colonia	Código	f
Álamos	1	4
Copilco	2	3
Florida	3	3
Mixcoac	4	2
Nápoles	5	3
Portales	6	3
Plateros	7	3
Tepeaca	8	4

I.2.4 FRECUENCIAS RELATIVAS

Si se retoma el problema de las actitudes frente a la instalación de playas, la tabla de frecuencias no termina donde se dejó. Se puede añadir más información útil en la que se basan respuestas para otras preguntas. Por ejemplo ¿Cuántas personas han respondido con una actitud media (valor 5)? Solución: 40.

Ahora, considérese la siguiente tabla con datos nuevos:

Tabla 7: Nueva tabla de frecuencias

x	f
1	200
2	170
3	120
4	60
5	40
6	60
7	120
8	170
9	200
Total	1,140

Si se trata de responder a la misma pregunta, ocurre lo siguiente:

En la tabla 7 ha cambiado el conjunto de datos. Ahora son 1,140, frente a los 150 de la encuesta anterior. Una misma frecuencia, en este caso $f = 40$, no tiene la misma interpretación en ambas tablas. ¿Qué ha cambiado? La importancia relativa de la frecuencia, puesto que $f = 40$ frente a $N = 150$ es diferente a $f = 40$ frente a $N = 1{,}140$. De hecho, el valor 5 pasa de ser el más frecuente al menos frecuente.

La solución se encuentra en expresar las frecuencias en términos relativos en vez de absolutos. Esto es precisamente lo que consiguen las proporciones: expresar una cantidad con respecto al total. Así, se añade una nueva columna, conteniendo las frecuencias relativas (f_r) que surgen de hacer la operación: $f_r = \frac{f}{N}$. Obsérvese el resultado comparando el obtenido con cada una de las dos tablas afectadas en este problema (tablas 4 y 7):

Tabla 8: Comparación entre dos tablas de frecuencias

	Nuevos datos		Datos anteriores	
x	f	f_r	f	f_r
1	200	0.1754	6	0.0400
2	170	0.1491	11	0.0733
3	120	0.1053	12	0.0800
4	60	0.0526	30	0.2000
5	40	0.0351	40	0.2667
6	60	0.0526	25	0.1667
7	120	0.1053	14	0.0933
8	170	0.1491	9	0.0600
9	200	0.1754	3	0.0200
Total	1,140	1.0000	150	1.0000

Nótese que el valor 5 pasa de contar con una frecuencia relativa $f_r = 0.2667$ (más de la cuarta parte) a $f_r = 0.0351$ al ser comparado, respectivamente, con un total de $N = 150$ a $N = 1{,}140$.

Un aspecto de interés se encuentra en la fila de los totales. El resultado es 1.0000 en los dos casos. Esto debe ocurrir siempre. Lo que se hace al traducir las frecuencias absolutas a las relativas es *unificar* el referente. En el conjunto de datos de la tabla 4, el referente absoluto es 150. En el conjunto de datos de la tabla 7, el referente absoluto es 1,140. No se pueden comparar frecuencias de conjuntos de datos diferentes porque los referentes son diferentes. Para que la comparación sea factible es necesario unificar. Dado que las proporciones se expresan en tantos por uno, es posible comparar frecuencias entre tablas. En otros términos: para interpretar una frecuencia absoluta se necesita conocer el número total de datos puesto que, según se ha visto, el número de datos condiciona la importancia de una frecuencia. Pero para interpretar una frecuencia relativa expresada como una proporción no es necesario conocer el número total de datos, puesto que aquí el referente es constante de una tabla a otra: 1.0000.

Sin embargo, no se terminó el proceso de enriquecimiento de la tabla.

Las proporciones se expresan siempre en cantidades que se sitúan entre 0 y 1. Es decir, las proporciones son números decimales. Y las personas también se sienten incómodas con las cantidades decimales.

I.2.5 TABLAS DE FRECUENCIAS

Por lo general, cuando se exponen los resultados de una encuesta en un medio de comunicación, lo habitual es utilizar otro tipo de frecuencias relativas: los porcentajes.

El principio que rige la utilización de los porcentajes es el mismo que para las proporciones: utilizar un referente fijo de tal forma que no sea necesario contar con el número total de datos para interpretar una frecuencia. La diferencia entre los porcentajes y las proporciones es que los primeros utilizan el referente 100, mientras que las proporciones utilizan el 1.

Entonces, conseguir los porcentajes es muy fácil si se cuenta con las proporciones: bastará con multiplicar a éstas por 100:

Tabla 9: Tablas de frecuencias con porcentajes

x	f	f_r	%
1	6	0.0400	4.00
2	11	0.0733	7.33
3	12	0.0800	8.00
4	30	0.2000	20.00
5	40	0.2667	26.67
6	25	0.1667	16.67
7	14	0.0933	9.33
8	9	0.0600	6.00
9	3	0.0200	2.00
Total	150	1.0000	100.00

Por lo que se puede completar también la tabla que se refiere a las colonias de los domicilios:

Tabla 10: Distribución por colonias

Provincia	Código	f	f_r	%
Álamos	1	4	0.16	16
Copilco	2	3	0.12	12
Florida	3	3	0.12	12
Mixcoac	4	2	0.08	8
Nápoles	5	3	0.12	12
Portales	6	3	0.12	12
Plateros	7	3	0.12	12
Tepeaca	8	4	0.16	16

Regresando a la encuesta de las playas, la variable actitud frente al subsidio admite orden entre sus valores. Por lo tanto, para completar la tabla, bastará con acumular sus frecuencias:

Tabla 11: Tabla de Frecuencias completa

x	f	f_r	%	F	F_r	% acumulado
1	6	0.0400	4.00	6	0.0400	4.00
2	11	0.0733	7.33	17	0.1133	11.33
3	12	0.0800	8.00	29	0.1933	19.33
4	30	0.2000	20.00	59	0.3933	39.33
5	40	0.2667	26.67	99	0.6600	66.00
6	25	0.1667	16.67	124	0.8267	82.67
7	14	0.0933	9.33	138	0.9200	92.00
8	9	0.0600	6.00	147	0.9800	98.00
9	3	0.0200	2.00	150	1.0000	100.00
Total:	150	1.0000	100.00			

I.3 TIPOS DE GRÁFICAS

Una gráfica es la representación de datos, generalmente numéricos, mediante líneas, superficies o símbolos, para ver la relación que esos datos guardan entre sí. Sirven para analizar el comportamiento de un proceso, o un conjunto de elementos o signos que permiten la interpretación de un fenómeno. Las gráficas se pueden agrupar en cinco tipos:

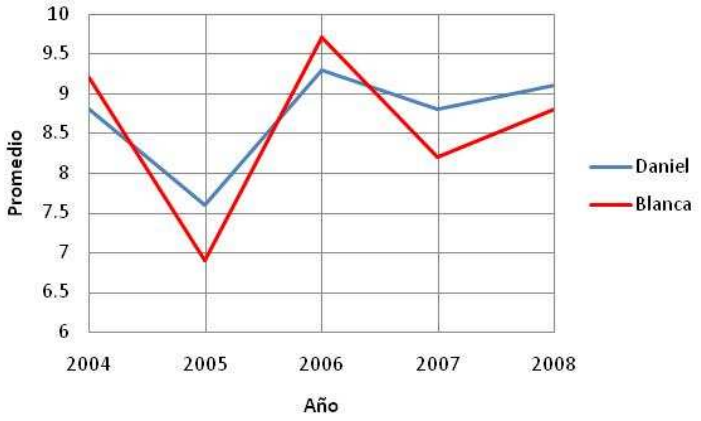
I.3.1 GRÁFICAS DE LÍNEAS

Gráfica simple de líneas

Muestran la relación entre dos variables cuantitativas. En el eje horizontal (x) se gráfica la variable independiente en el eje vertical (y). Las marcas de los cuadrantes en los ejes marcan las unidades de medida; las escalas en los ejes pueden ser lineales, logarítmicas o ambas.

Cuando los datos se relacionan entre sí, es decir, cuando podemos decir que existe cierta continuidad entre las observaciones (como por ejemplo el crecimiento poblacional, la evolución del peso o estatura de una persona a través del tiempo, el desempeño académico de un estudiante a lo largo de su instrucción escolar, las variaciones presentadas en la medición realizada en algún experimento cada segundo o minuto) se pueden utilizar las gráficas de líneas, que consisten en una serie de puntos trazados en las intersecciones de las marcas de clase y las frecuencias de cada una, uniéndose consecutivamente con líneas.

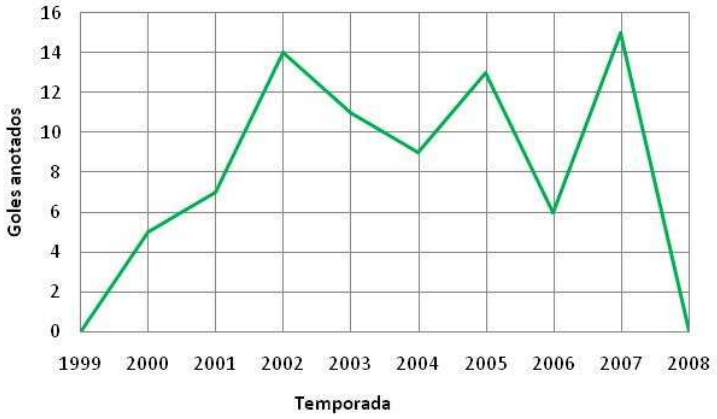
Ejemplo.
Aquí se muestra el comportamiento de los promedios escolares finales de dos alumnos (Daniel y Blanca) a lo largo de cinco observaciones anuales:



Polígono de frecuencias

Otra forma de representación de uso menos común, y muy parecida a las gráficas de líneas, es el polígono de frecuencias. La diferencia fundamental entre ambas es que en el polígono de frecuencias se añaden dos clases con frecuencias cero: una antes de la primera clase con datos y otra después de la última. El resultado es que se "sujeta" la línea por ambos extremos al eje horizontal y lo que podría ser una línea separada del eje se convierte, junto con éste, en un polígono.

Ejemplo.
El siguiente polígono de frecuencias muestra los goles anotados por un delantero en un equipo de fútbol en las temporadas de 2000 a 2007:



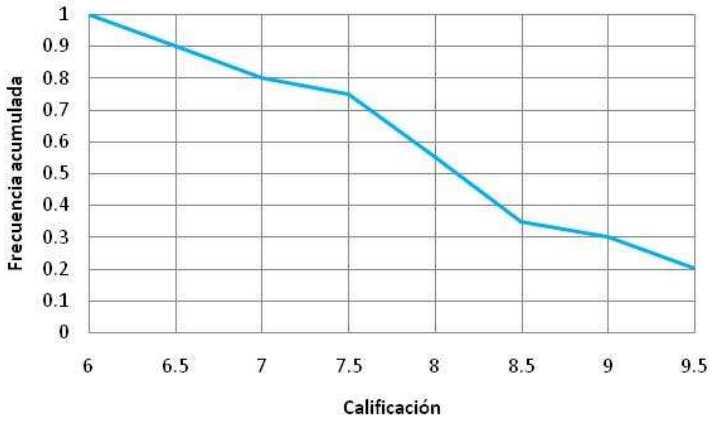
Una gráfica similar al polígono de frecuencias es la ojiva, pero ésta se obtiene de aplicar parcialmente la misma técnica a una distribución acumulativa y de igual manera que éstas, existen las *ojivas mayor que* y las *ojivas menor que*.

La diferencia fundamental entre las ojivas y los polígonos de frecuencias es que en el eje horizontal (x) en lugar de colocar las marcas de clase se colocan las fronteras de clase. Para el caso de la *ojiva mayor que* es la frontera menor y para la *ojiva menor que*, la mayor. Los dos casos posibles son:

- Caso 1. Para la *ojiva mayor que*, el extremo izquierdo de la ojiva no se "amarra" al eje x.

Ejemplo.
De un grupo de 30 estudiantes, 20 acreditaron la materia de Estadística y Probabilidad y se agruparon sus calificaciones desde 6 hasta 10 en intervalos de 0.5. Se obtuvo la frecuencia acumulada hasta el intervalo de clase mayor de 9.5. De la gráfica de ojiva mayor puede verse que el 30% de los estudiantes sacaron 9 o más de calificación.

Calificación mayor que	Número de estudiantes	Frecuencia acumulada
6	20	1
6.5	18	0.9
7.0	16	0.8
7.5	15	0.75
8.0	11	0.55
8.5	7	0.35
9	6	0.3
9.5	4	0.2



- Caso 2. Para la *ojiva menor que*, el extremo derecho no se “amarra” al eje x.

Ejemplo.
Se tomaron las estaturas de 50 estudiantes en un grupo del plantel 8 de la ENP y se agruparon por intervalos de 5 centímetros, iniciando en 1.45m y terminando en 1.90m. Se obtuvo la frecuencia acumulada hasta el intervalo de clase menor de 1.90m. De la gráfica de ojiva menor puede verse que el 90% de los estudiantes miden menos de 1.80 metros.

Altura (m) menor que	Número de estudiantes	Frecuencia acumulada
1.45	0	0
1.50	4	0.08
1.55	13	0.26
1.60	23	0.46
1.65	30	0.6
1.70	34	0.68
1.75	39	0.78
1.80	45	0.9
1.85	48	0.96
1.90	50	1

I.3.2 GRÁFICAS DE BARRAS O HISTOGRAMAS

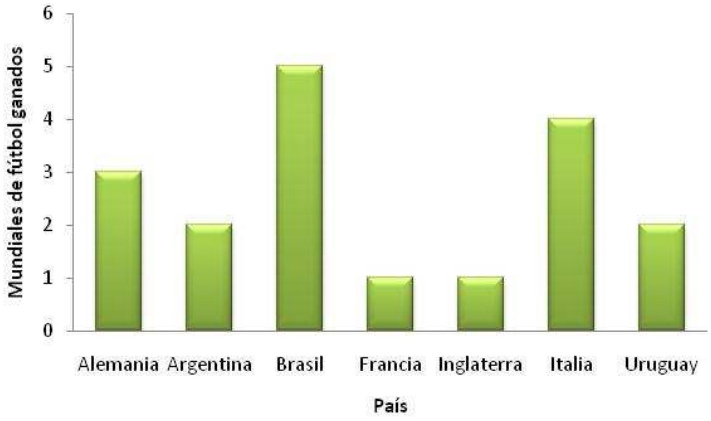
Se emplea cuando la variable independiente es categórica. Cada barra sólida, ya sea vertical u horizontal representa un tipo de dato. Cuando es necesario representar divisiones de datos se utiliza un gráfica de barras subdivididas. Los histogramas no muestran frecuencias acumuladas, son preferibles para el tratamiento de datos cuantitativos y la barra con mayor altura representa la mayor frecuencia. La sumatoria de las alturas de las columnas equivale al 100% de los datos.

Barras verticales

En el eje horizontal (o de las abscisas) se representan los intervalos de los datos, marcándose de manera continua las fronteras entre cada uno de éstos. De esta manera, el histograma está compuesto por rectángulos, cuyo número coincide con la cantidad de intervalos considerados, el ancho de la base de cada uno de esos rectángulos es la misma siempre y coincide con las fronteras de los intervalos, y la altura corresponde a la frecuencia de cada intervalo. En este tipo de gráficas es recomendable:

- El empleo de sombreado o colores facilita la diferenciación de las barras.
- El punto cero se indica en el eje de ordenadas y se deben establecer las unidades en los ejes.
- La longitud de los ejes debe ser suficiente para acomodar la extensión de la barra.

Ejemplo.
La gráfica siguiente representa el número de campeonatos de fútbol que han ganado los países en las 18 ediciones desde 1930 hasta 2006:



Es importante notar que resulta difícil utilizar este tipo de representación cuando existen intervalos abiertos o cuando los intervalos no son iguales entre sí.

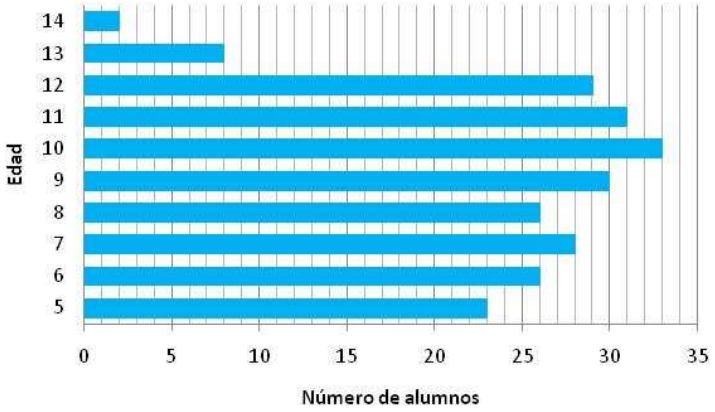
Barras horizontales

Se parecen mucho a las gráficas de columnas, con la salvedad importante de que la función de los ejes se intercambia y el eje horizontal queda destinado a las frecuencias y el eje vertical a las clases.

Es muy común que este tipo de gráficos se utilicen para ilustrar el tamaño de una población dividida en estratos como, por ejemplo, son sus edades.

Ejemplo.

La siguiente gráfica presenta la distribución de las edades de los 236 niños que estudian en una escuela primaria:



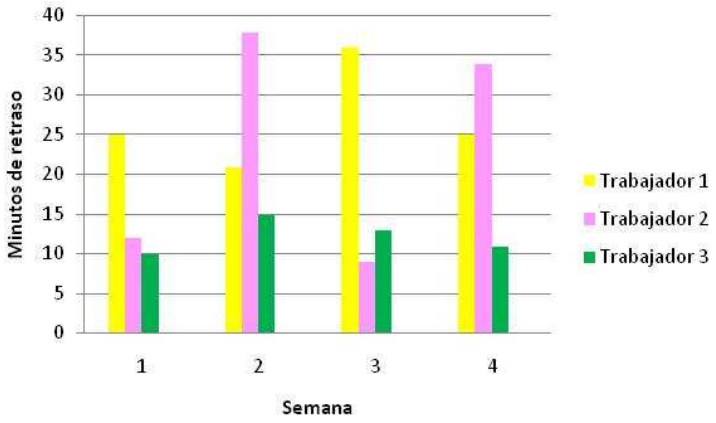
A este tipo de gráficos en particular se le llama *pirámide de edades* por su forma.

Gráficas de columnas bidimensionales

Un tipo de gráfico muy parecido al histograma es la gráfica de columnas. Para este tipo de gráfica, elaboradas con rectángulos también, se pide que sus bases sean del mismo ancho y sus alturas equivalentes con las frecuencias. Para este tipo, a diferencia del histograma, no es necesario tener una escala horizontal continua, por lo que los rectángulos (o barras) no tienen que aparecer juntas entre sí.

Otra observación pertinente es que se pueden representar en la misma gráfica, utilizando las mismas escalas horizontales y verticales, varios datos correspondientes a las mismas variables producto de varias observaciones. Esto produce una gráfica con varias series, correspondiendo cada una de ellas a cada observación de la muestra (o población), y teniéndose una gráfica compuesta. Es conveniente que cada serie de datos (u observaciones) sean coloreados de igual manera entre sí, pero distinta de las demás.

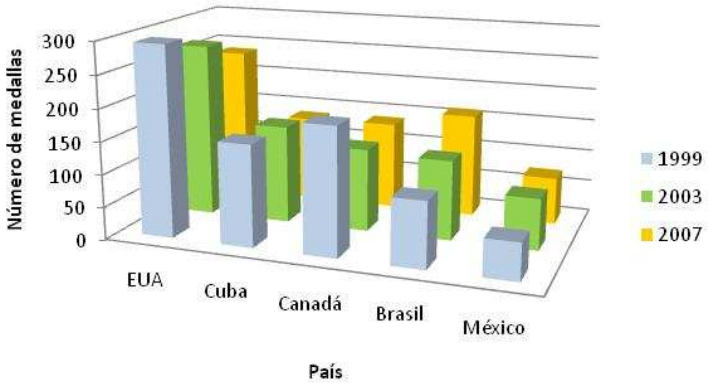
Ejemplo.
La gráfica siguiente muestra el comportamiento de los minutos de retraso que acumularon tres trabajadores de una tienda durante cuatro semanas. Las series están coloreadas con diferente color para mostrar el comportamiento tanto individual, como de cada uno de los trabajadores con respecto a los demás.



Gráficas de columnas tridimensionales

Existe la posibilidad, y si los recursos lo permiten, de representar gráficos compuestos de una manera "tridimensional", es decir, con gráficos que posean no sólo dos ejes, sino tres; y en los que los rectángulos son sustituidos por prismas de base rectangular (ocasionalmente el software en el mercado permite utilizar prismas cuya base son polígonos regulares de más de cuatro lados, pirámides o cilindros).

Ejemplo.
En la gráfica se puede apreciar el número de medallas que han ganado cinco países en las ediciones de los Juegos Panamericanos de 1999 a 2007:



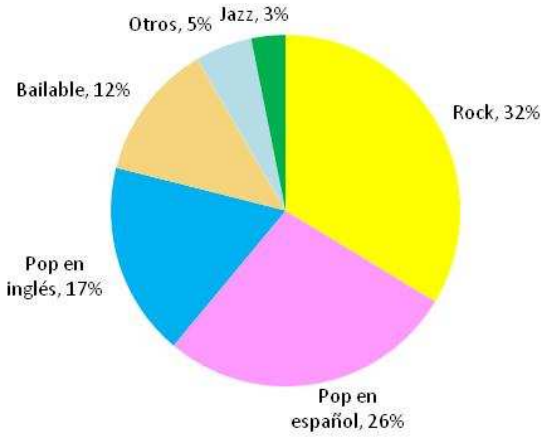
En este tipo de gráficas pueden complicarse mucho si hay demasiados datos ya que la información es menos legible.

I.3.3 GRÁFICAS CIRCULARES

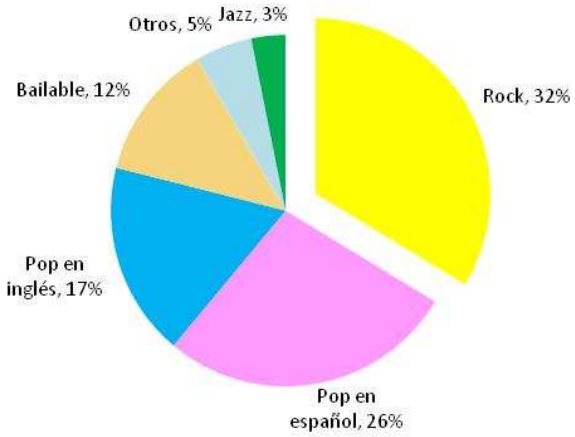
Denominadas también gráfica de pastel, se utilizan para mostrar porcentajes y proporciones. El número de elementos comparados dentro de un gráfico circular, no deben ser más de 7, ordenando los segmentos de mayor a menor, iniciando con el más amplio a partir de las 12 como en un reloj. Una manera sencilla de diferenciar los segmentos es sombreándolos con colores contrastantes.

Este tipo de gráficas es muy útil cuando lo que se desea es resaltar las proporciones que representan algunos subconjuntos con respecto al total, es decir, cuando se está usando una escala categórica.

Ejemplo.
La siguiente gráfica ilustra los gustos musicales de un grupo de jóvenes del sexto año de preparatoria:



De hecho, si se desea resaltar una de las categorías que se presentan, es válido tomar esa "rebanada" de la gráfica y separarla de las demás:

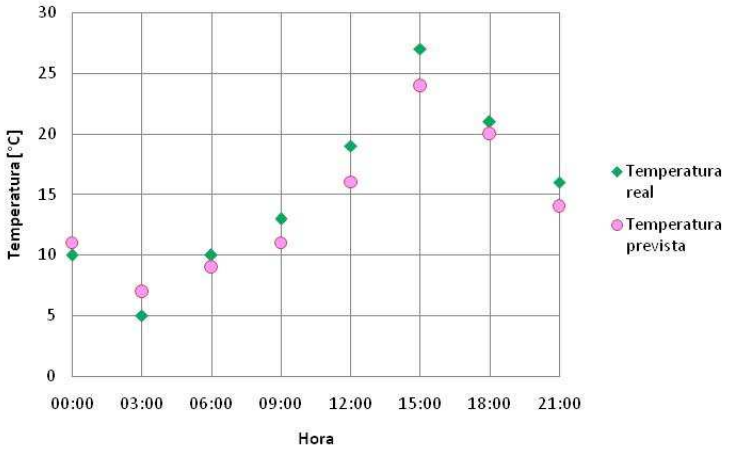


I.3.4 GRÁFICAS DE DISPERSIÓN

Una gráfica de dispersión tiene dos ejes de valores, mostrando un conjunto de datos numéricos en el eje x y otro en el eje y. Combina estos valores en puntos de datos únicos y los muestra en intervalos uniformes o agrupaciones. Los gráficos de dispersión se utilizan normalmente para mostrar y comparar valores numéricos, como datos científicos, estadísticos y de ingeniería. Este tipo de gráficas se usan cuando:

- Desea cambiar la escala del eje horizontal.
- Desea convertir dicho eje en una escala logarítmica.
- Los espacios entre los valores del eje horizontal no son uniformes.
- Hay muchos puntos de datos en el eje horizontal.

Ejemplo.
La siguiente gráfica de dispersión compara temperaturas en un día en la Ciudad de México. En el eje de horizontal mide la hora de medición y el eje vertical mide las temperaturas previstas y las temperaturas reales.

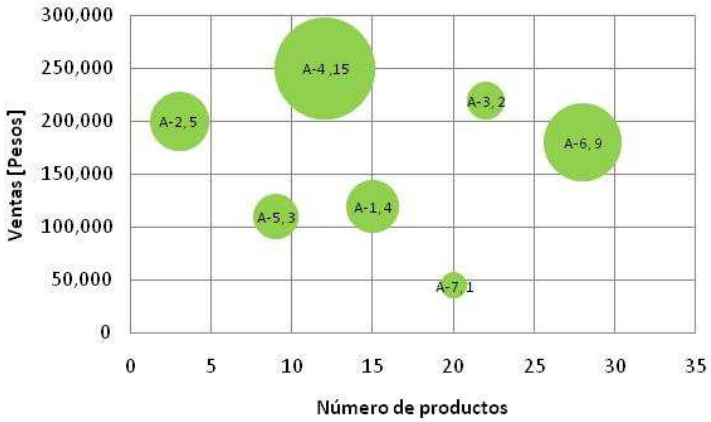


Es interesante observar que los puntos parecen seguir una cierta tendencia en una curva imaginaria. Uno de los usos de este tipo de gráficas es precisamente encontrar si las observaciones siguen algún patrón (lineal, exponencial, polinomial, logarítmica, etc.) o si existen valores atípicos.

I.3.5 GRÁFICAS DE BURBUJAS

Un tipo de gráfico similar a las gráficas de dispersión son las gráficas de burbujas, en las cuales se presenta la dispersión de las observaciones de la misma forma pero se le añade la posibilidad de visualizar otra variable representada en el tamaño del punto, pues éstos se convierten en círculos (burbujas) con radios proporcionales a las magnitudes que representan.

Ejemplo.
La gráfica siguiente se puede apreciar el volumen de ventas y el número de productos de siete artículos (A-1 a A-7) en una fábrica. Además, se puede ver fácilmente la participación o cuota de mercado de cada artículo a través del tamaño de cada burbuja, que corresponde a la cifra que está después de cada coma:

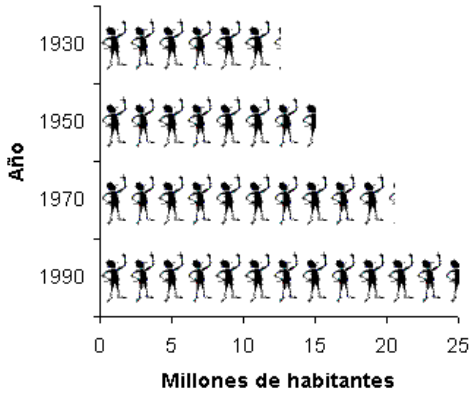


I.3.6 PICTOGRAMAS

Son gráficos con dibujos alusivos al carácter que se está estudiando y cuyo tamaño es proporcional a las frecuencias que representan. Se emplean para representar diferencias cuantitativas simples entre grupos. Los símbolos utilizados para representar valores idénticos deben ser de igual dimensión.

Actualmente muchos medios masivos de comunicación utilizan gráficos para ilustrar resultados de alguna investigación. Regularmente se utilizan dibujos llamativos para captar el interés del público.

Ejemplo.
El pictograma siguiente representa la población de los Estados Unidos de 1930 a 1990 (cada figura representa a dos millones de habitantes).



I.4 MEDIDAS DE TENDENCIA CENTRAL

I.4.1 INTRODUCCIÓN A LA SUMATORIA

Por sumatoria se entiende la suma de un conjunto finito de números, que se denota por la letra sigma mayúscula $\sum$:

$$S = \sum_{i=k}^{k+n} x_i = x_k + x_{k+1} + x_{k+2} + \cdots + x_{k+n-1} + x_{k+n}$$

donde:
S es magnitud resultante de la suma.
n es la cantidad de valores a sumar.
k es punto inicial de la sumatoria.
$k+n$ es punto final de la sumatoria.
i es el índice de la suma, que varía entre k y $k+n$.
x_k es el valor de la magnitud objeto de suma en el punto i.

En el caso particular en que la sumatoria empiece en con el primer valor de la serie y termine en el último, la expresión se simplifica a:

$$S = \sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_{n-1} + x_n$$

Ejemplo.
Dados los valores de X: 4, 3, 8, 5, 9, 2, 11, 1, 6, la sumatoria para:
Los primeros cuatro términos es:

$$S_1 = \sum_{i=1}^{4} x_i = 4+3+8+5 = 20$$

Entre el segundo término y el sexto:

$$S_2 = \sum_{i=2}^{6} x_i = 3+8+5+9+2 = 27$$

Todos los términos:

$$S_3 = \sum_{i=1}^{9} x_i = 4+3+8+5+9+2+11+1+6 = 49$$

Ejemplo.
Dada la siguiente tabla:

Valor de variable x_i	Valor de la variable y_i
$x_1 = 1$	$y_1 = -3$
$x_2 = 2$	$y_2 = -5$
$x_3 = 3$	$y_3 = 2$
$x_4 = 4$	$y_4 = 0$
$x_5 = 5$	$y_5 = 1$

Obtener:

a) $\sum_{i=3}^{5} x_i$

Solución.

$$\sum_{i=3}^{5} x_i = 3+4+5 = 12$$

b) $-\sum_{i=1}^{n} x_i$

Solución.

$$-\sum_{i=1}^{n} x_i = -(1+2+3+4+5) = -15$$

c) $\sum_{i=2}^{4} y_i$

Solución.

$$\sum_{i=2}^{4} y_i = -5+2+0 = -3$$

d) $\sum_{i=1}^{5} (y_i)^2$

Solución.

$$\sum_{i=1}^{5} y_i = (-3)^2 + (-5)^2 + (2)^2 + (0)^2 + (1)^2 = 9 + 25 + 4 + 0 + 1 = 39$$

e) $7 \cdot \sum_{i=1}^{5} x_i \cdot y_i$

Solución.

$$7 \cdot \sum_{i=1}^{5} x_i \cdot y_i = 7[1(-3) + 2(-5) + 3(2) + 4(0) + 5(1)] = 7(-3 - 10 + 6 + 0 + 5) = 7(-2) = -14$$

I.4.2 MEDIA, MEDIANA Y MODA

Cuando se tiene un grupo de observaciones, se desea describirlo a través de un sólo número. Para tal fin, no se usa el valor más elevado ni el valor más pequeño como único representante, ya que sólo representan los extremos. Una de las propiedades más sobresalientes de la distribución de datos es su tendencia a acumularse hacia el centro de la misma. Esta característica se denomina *tendencia central*.

Las medidas de tendencia central más usuales son: la media aritmética, la mediana y la moda.

MEDIA ARITMÉTICA

La *media aritmética* de n valores, es igual a la suma de todos ellos dividida entre n. Se denota por $\bar{x}$. Esto es:

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

Cuando los datos tienen más de una frecuencia, para obtener la media aritmética se agrega otra columna a la tabla estadística con el producto de las observaciones y sus frecuencias. Es decir, si se cuenta con una distribución de datos entonces se aplica la fórmula:

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

Ejemplo.
Con los datos: 10, 8, 6, 15, 10, 5, hallar la media aritmética.

Solución.

$$\bar{x} = \frac{10 + 8 + 6 + 15 + 10 + 5}{6} = \frac{54}{6} = 9$$

Ejemplo.
Mediante la siguiente distribución de frecuencias que muestra las estaturas en metros de los alumnos de un grupo de la prepa 8, hallar la media aritmética.

Estaturas [m]	f
1.52	1
1.54	5
1.55	4
1.58	5
1.60	2
1.62	4
1.64	7
1.66	3
1.70	5
1.71	8
1.73	6
1.74	5
1.77	3
1.80	1
1.83	1

Solución.
Construyendo una tabla:

Estaturas [m]	f	$f \cdot x$
1.52	1	1.52
1.54	5	7.70
1.55	4	6.20
1.58	5	7.90
1.60	2	3.20
1.62	4	6.48
1.64	7	11.48
1.66	3	4.98
1.70	5	8.45
1.71	8	13.68
1.73	6	10.38
1.74	5	8.70
1.77	3	5.31
1.80	1	1.80
1.83	1	1.83
Total:	60	99.61

$$\bar{x} = \frac{99.61}{60} = 1.6601$$

Las características de la media aritmética son:

1. Es una medida totalmente numérica o sea sólo puede calcularse en datos de características cuantitativas.
2. En su cálculo se toman en cuenta todos los valores de la variable.
3. Es lógica desde el punto de vista algebraico.
4. La media aritmética es altamente afectada por valores extremos.
5. No puede ser calculada en distribuciones de frecuencia que tengan clases abiertas.
6. La media aritmética es única, o sea, un conjunto de datos numéricos tiene una y sólo una media aritmética.

MEDIANA

La *mediana* es el punto central de una serie de datos ordenados de forma ascendente o descendente.

De acuerdo al número de casos o datos, hay dos formas para calcular la mediana: para número impar y para número par:

- Número impar de datos ordenados de menor a mayor o de mayor a menor: la mediana es el valor que queda justo al centro.

Ejemplo.
Obtener la mediana de los siguientes datos: 4, 7, 1, 9, 2, 5, 6.

Solución.
Ordenando de forma ascendente: 1, 2, 4, 5, 6, 7, 9.
El valor que queda al centro es el 5, porque hay tres datos antes y tres datos después de él, entonces la mediana es 5.

- Número de datos par: en este caso se busca la media aritmética entre los dos valores centrales.

Ejemplo.
Obtener la mediana de los siguientes datos: -3, 5, 18, 4, 11, -6, 9, 10, -1, 2.

Solución.
Ordenando de forma ascendente: -6, -3, -1, 2, 4, 5, 9, 10, 11, 18.
Los valores centrales son 4 y 5. Su media aritmética es:

$$\bar{x} = \frac{4+5}{2} = 4.5$$

En este caso, la mediana de este conjunto no pertenece al conjunto de datos.

Las características de la mediana son:

1. En su cálculo no se incluyen todos los valores de la variable.
2. La Mediana no es afectada por valores extremos.
3. Puede ser calculada en distribuciones de frecuencia con clases abiertas.
4. No es lógica desde el punto de vista algebraico.

MODA

La *moda* de un conjunto de datos numéricos es el valor que más se repite, es decir, el que tiene el mayor número de frecuencias absolutas. La moda puede ser no única e inclusive no existir.

La moda es una medida de tendencia central muy importante, porque permite planificar, organizar y producir para satisfacer las necesidades de la mayoría.

Ejemplo.
Obtener la moda de los siguientes datos: -3, 3, -2, 0, 3, -1, -2, 4, 5, -2, 0, 1.

Solución.
Ordenando de forma ascendente: -3, -2, -2, -2, -1, 0, 0, 1, 3, 3, 4, 5.
El valor que más se repite es el -2, por lo tanto ese valor es su moda.

Ejemplo.
Obtener la moda de los siguientes datos: 6, 2, -1, -5, 3, -3, -2, 5, 0, -4, 4, 1.

Solución.
Ordenando de forma ascendente: -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6.
Ningún valor se repite es, decir su moda no existe.

Ejemplo.
En una tienda, 18 empleados presentan la siguiente información:

Horas laboradas por día	Frecuencia
6	3
7	2
8	5
9	5
10	2
11	1

¿Cuál es la moda de las horas laboradas por los empleados?

Solución.
Hay dos valores con frecuencia 5. Entonces, se concluye que hay más de una moda. La mayor frecuencia son 8 y 9 horas diarias de trabajo.

Las características de la moda son:

1. En su cálculo no se incluyen todos los valores de la variable.
2. El valor de la moda puede ser afectado grandemente por el método de designación de los intervalos de clases.
3. No está definida algebraicamente.
4. Puede ser calculada en distribuciones de frecuencia que tengan clases abiertas.
5. No es afectada por valores extremos.

MEDIA PONDERADA

La *media ponderada* de un conjunto de valores de una variable x a los que se han asignado, respectivamente, una ponderación se calcula mediante la fórmula:

$$\overline{x_p} = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i} = \frac{x_1 p_1 + x_2 p_2 + x_3 p_3 + \cdots x_n p_n}{p_1{}' + p_2 + p_3 + \cdots p_n}$$

Los valores $p_1, p_2, p_3, \cdots p_n$ indican la importancia que se quiere dar a cada uno de los valores que toma la variable x.

Ejemplo.
Un profesor de la prepa 8 decide que la calificación final de un alumno constará del 60% del promedio de los exámenes, el 30% de promedio de tareas y el 10% de participación en clase a lo largo del año escolar. Si un alumno tiene 5.3 de promedio de exámenes, 7.1 de tareas y 7.8 promedio de participaciones. ¿Cuál será su calificación final?

Solución.

$$x_p = \frac{5.3(0.6) + 7.1(0.3) + 7.8(0.1)}{0.6 + 0.3 + 0.1} = 6.09$$

Si el profesor sólo tomara en cuenta los exámenes, el alumno no aprobaría. Sin embargo al darle importancia a las tareas y a su participación en clase, esto hace que al final consiga aprobar con la media ponderada.

Su característica principal es que su resultado depende de la importancia o “peso” de cada uno de los valores asignado por quien efectúa el cálculo.

MEDIA GEOMÉTRICA

La *media geométrica* de un conjunto de n observaciones es la raíz enésima de su producto. El cálculo de la media geométrica exige que todas las observaciones sean positivas:

$$\overline{x_g} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n}$$

Ejemplo.
Obtener la media geométrica de los datos: 3, 8, 9.

Solución.
$\overline{x_g} = \sqrt[3]{3(8)(9)} = \sqrt[3]{216} = 6$

Ejemplo.
Las siguientes temperaturas han sido tomadas de un experimento químico: 13.4℃, 12.8℃, 11.9℃, 13.6℃. Determinar la temperatura geométrica media de este proceso.

Solución.
$\overline{x_g} = \sqrt[4]{13.4(12.8)(11.9)(13.6)} \approx \sqrt[4]{2{,}7758.79} \approx 12.90\,℃$

Las características de la media geométrica son:

1. Se toman en cuenta todos los valores de la variable.
2. Es afectada por valores extremos aunque en menor medida que la media aritmética.
3. Si un dato es cero, su resultado será cero.
4. No puede ser calculada en distribuciones con clase abiertas.
5. Es mayormente usada para promediar tasas de intereses anuales, inflación razones y valores que muestren una progresión geométrica (efecto multiplicativo sobre el de los años anteriores).

MEDIA ARMÓNICA

La *media armónica* se define como el recíproco de la media aritmética. Esto es:

$$\overline{x_a} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots + \frac{1}{x_n}}$$

Ejemplo.
Obtener la media armónica de los datos: 6, 9, 7, 2.

Solución.
$$\overline{x_a} = \frac{4}{\frac{1}{6}+\frac{1}{9}+\frac{1}{7}+\frac{1}{2}} = \frac{4}{\frac{116}{126}} = \frac{504}{116} \approx 4.34$$

Las características de la media armónica son:

1. No se influye por la existencia de determinados valores mucho más grandes que el resto.
2. Presenta cambio sensible a valores mucho más pequeños que el conjunto.
3. No está definida en el caso de la existencia de valores nulos.

CENTRO DE AMPLITUD

Es el valor que queda en medio de los valores mínimo y máximo. Esto es:

$$C_a = \frac{x_{max} + x_{min}}{2}$$

Ejemplo.
Obtener el centro de amplitud de los datos siguientes: 2, -1, 8, 7, -3, 4, 9, 2, 0, 5.

Solución.
Ordenando los datos para obtener los valores extremos: -3, -1, 0, 2, 2, 4, 5, 7, 8, 9. Entonces:

$$C_a = \frac{9+(-3)}{2} = \frac{6}{2} = 3$$

I.5 MEDIDAS DE DISPERSIÓN

La *dispersión* mide que tan alejados están un conjunto de valores respecto a su media aritmética. Así, cuanto menos disperso sea el conjunto, más cerca del valor medio se encontrarán sus valores. Este aspecto es de vital importancia para el estudio de investigaciones.

Se llaman *medidas de dispersión* aquellas que permiten retratar la distancia de los valores de la variable a un cierto valor central, o que permiten identificar la concentración de los datos en un cierto sector del recorrido de la variable. Se trata de coeficientes para variables cuantitativas.

RANGO

El *rango* de una distribución es la diferencia entre el valor máximo (M) y el valor mínimo (m) de la variable estadística. Para su cálculo, basta con ordenar los valores de menor a mayor m de M.

Ejemplo.
Si se conoce que el valor promedio de días de espera para obtener una licencia de manejo, es de 5 días en la oficina A, y de 7 días en la oficina B, con esta única información no es posible hacer una elección adecuada. Sin embargo, si se sabe que en la oficina A, el número mínimo de días de espera es de 3 y el máximo de 15, mientras que en la oficina B, los valores son 3 y 8 días respectivamente, se podrá tomar una decisión más adecuada para acudir a obtener la licencia, gracias a esta información adicional.

Características del rango:

1. A medida que el rango es menor, el grado de representatividad de los valores centrales se incrementa.
2. A medida que el rango es mayor, la distribución está menos concentrada o más dispersa.
3. Su cálculo es extremadamente sencillo.
4. Tiene gran aplicación en procesos de control de calidad.
5. Tiene el inconveniente de que sólo depende de los valores extremos. De esta forma basta que uno de ellos se separe mucho para que el recorrido se vea sensiblemente afectado.

I.5.1 MEDIDAS DE POSICIÓN PARA DATOS AGRUPADOS Y NO AGRUPADOS: PERCENTILES, DECILES Y CUARTILES

Los *cuantiles* son los valores de la distribución que la dividen en partes iguales, es decir, en intervalos que comprenden el mismo número de valores. Cuando la distribución contiene un número alto de intervalos o de marcas y se requiere obtener un promedio de una parte de ella. Generalmente, se divide la distribución en cuatro, en diez o en cien partes.

Los cuantiles más usados son los percentiles, cuando dividen la distribución en cien partes, los deciles, cuando dividen la distribución en diez partes y los cuartiles, cuando dividen la distribución en cuatro partes.

PERCENTILES

Los *percentiles* son números que dividen en 100 partes iguales un conjunto de datos ordenados. Es decir, El percentil k es un valor que deja aproximadamente el k por ciento de los datos por abajo de él. Se denota por medio de P(k%).

Ejemplo.
En un estudio de ingresos mensuales de la población económicamente activa, revela que el percentil 90 (P90) es $20,000. Esto significa que aproximadamente el 90% de las personas tienen ingresos que son menores o iguales a $20,000, y por supuesto, el 10% tiene ingresos mayores o iguales a dicho valor.

En el ejemplo anterior se tomo el percentil 90 pero se podría haber considerado cualquier valor, por ejemplo, 70, 80 entre otros. Fundamentalmente cuando la distribución de frecuencia es asimétrica, puede ser más útil e informativo, resumir la distribución de la variable en estudio, mediante los percentiles.

DECILES

Los *deciles* son números que dividen la sucesión de datos ordenados en diez partes porcentualmente iguales. Son los nueve valores que dividen al conjunto de datos ordenados en diez partes iguales, son también un caso particular de los percentiles. Los deciles se denotan D(1), D(2),..., D(9), que se leen primer decil, segundo decil, etc.

Ejemplo.
Dada la siguiente distribución de frecuencias en el número de recámaras en 75 casas en una colonia de la delegación Coyoacán, calcular sus deciles.

x_i	n	$f=\frac{n}{75}$	f_a
1	9	12	12
2	17	22.66	34.66
3	21	28	62.66
4	11	14.66	77.33
5	7	9.33	86.66
6	5	6.66	93.33
7	3	4	97.33
8	2	2.67	100
		100	

Solución.

El primer decil es el primer valor de x que cumple con $\frac{f_a}{10}>10$, por lo tanto, $D(1)=1$

El segundo decil es el primer valor de x que cumple con $\frac{f_a}{10}>20$, por lo tanto, $D(2)=1$

El tercer decil es el primer valor de x que cumple con $\frac{f_a}{10} > 30$, por lo tanto, $D(3) = 2$

De manera sucesiva se pueden obtener los otros deciles. La tabla siguiente concentra sus valores:

D_i	$\frac{f_a}{10}$	x_i
D(1)	>10	1
D(2)	>20	1
D(3)	>30	2
D(4)	>40	2
D(5)	>50	2
D(6)	>60	3
D(7)	>70	4
D(8)	>80	5
D(9)	>90	6

Esto significa que el 50% de las casas en esa colonia tienen una o dos habitaciones.

CUARTILES

Los *cuartiles* se definen como los tres valores que dividen la distribución en cuatro partes iguales.

En términos de percentiles el primer cuartil Q(1) coincide con el P(25) (percentil 25); el segundo cuartil Q(2) con el P(50) o mediana, y el tercer cuartil Q(3) con el P(75).

Entre el primer y el tercer cuartil se encuentra el 50% central de las observaciones.

Ejemplo.
Dada la siguiente distribución de frecuencias en el número de hijos de cien familias, calcular sus cuartiles.

x_i	f	f_a
0	14	14
1	10	24
2	15	39
3	26	65
4	20	85
5	15	100
	100	

Solución.

El primer cuartil es el primer valor de x que cumple con $\frac{f_a}{4} > 25$, por lo tanto, $Q(1) = 2$

El segundo cuartil es el primer valor de x que cumple con $\frac{f_a}{4} > 50$, por lo tanto, $Q(2) = 3$

El tercer cuartil es el primer valor de x que cumple con $\frac{f_a}{4} > 75$, por lo tanto, $Q(3) = 4$

Esto significa que el 75% de las familias tienen cuatro o menos hijos.

RANGO INTERCUARTIL

Para un cálculo de rangos más eficiente, se eliminan los valores extremadamente alejados aplicando el *rango intercuartil* que es una medida de variabilidad adecuada cuando la medida de posición central

empleada ha sido la mediana y él se define como la diferencia entre el Tercer Cuartil superior y el Primer Cuartil, es decir: Rango Intercuartil = Q(3) – Q(1).

Ejemplo.
Dados los siguientes valores ordenados:
26, 33, 36, 39, 40, 40, (41), 42, 44, 45, 47, 47, 47, (48), 50, 51, 51, 53, 54, 54, (55), 57, 59, 61, 63, 66, 71.
Obtener su rango y su rango intercuartil.

Solución.
El rango es: 71 – 26 = 45
Los valores cuartiles se muestran entre paréntesis, es decir, 41, 48 y 55, donde, el segundo cuartil es simplemente la mediana. La dispersión calculada a través del rango intercuartil, es en este caso será:
Q(3) – Q(1)= 55 - 41 = 14. Nótese como el la dispersión es mucho menor aplicando el rango intercuartil.

DESVIACIÓN MEDIA

La *desviación media* es la división de la sumatoria del valor absoluto de las distancias existentes entre cada dato y su media aritmética y el número total de datos:

$$D_m = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{n}$$

Este indicador muestra que tan disperso se encuentran un conjunto de datos a un punto de concentración.

Ejemplo.
Sean los siguientes datos: 4, 5, 3, 5, 3, 2, 2, 2, 2, 3, 5, 1, 4, 1, 4. Obtener su desviación media:

Solución.

Se calcula la media aritmética: $D_m = \frac{\sum_{i=1}^{15} x_i}{15} = 3.0666$

El primer dato (4), se aleja de la media en 0,9334 hacia la derecha. Para el segundo dato (5), se aleja de la media 1,9333 también hacia la derecha. Para el tercer dato (3), se aleja de la media en 0,0667 pero hacia la izquierda. La suma de las distancias absolutas es 17.2, así que los datos se separan de la media en:

$$D_m = \frac{\sum_{i=1}^{n} \left| x_i - 3.066 \right|}{n} = \frac{17.2}{15} = 1.1466$$

Ejemplo.
Hallar la desviación media en la siguiente distribución de frecuencias.

Clases	f
8-10	3
11-13	6
14-16	9
17-19	11
20-22	5
	n=34

Solución.

Calculando los puntos medios de cada clase y obteniendo $f \cdot x$:

Clases	f	x	$f \cdot x$
8-10	3	9	27
11-13	6	12	72
14-16	9	15	135
17-19	11	18	198
20-22	5	21	105
	n=34		537

La media es: $\bar{x} = \dfrac{\sum_{i=1}^{5}(f \cdot x)}{n} = \dfrac{537}{34} = 15.794$. Por lo tanto:

$$\sum_{i=1}^{5} f \cdot |x_i - \bar{x}| = 3(9-15.794)+6(12-15.794)+9(15-15.974)+11(18-15.794)+5(21-15.794)$$

$$\sum_{i=1}^{5} f \cdot |x_i - \bar{x}| = 100.6$$

$$\therefore \quad D_m = \frac{\sum_{i=1}^{n} |x_i - 3.066|}{34} = \frac{100.6}{34} = 2.958$$

DESVIACION ESTÁNDAR

La *desviación estándar* o *desviación típica* se define como la raíz cuadrada de los cuadrados de las desviaciones de los valores de la variable respecto a su media. Esto es:

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

La desviación estándar es una medida estadística de la dispersión de un grupo o población. Una gran desviación estándar indica que la población esta muy dispersa respecto de la media. Una desviación estándar pequeña indica que la población está muy compacta alrededor de la media.

Para el caso de datos agrupados, la desviación estándar se calcula por medio de:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f \cdot (x_i - \bar{x})^2}{n}}$$

VARIANZA

La *varianza* mide la mayor o menor dispersión de los valores de la variable respecto a la media aritmética. Cuanto mayor sea la varianza mayor dispersión existirá y por tanto, menor representatividad tendrá la media aritmética. La varianza se expresa en las mismas unidades que la variable analizada, pero elevadas al cuadrado.

La varianza de un conjunto de datos se define como el cuadrado de la desviación estándar y está dada por:

$$v = \sigma^2$$

Ejemplo.
Hallar la desviación estándar y la varianza de la siguiente serie de datos: 10, 18, 15, 12, 3, 6, 5, 7
Solución.

$$\bar{x} = \frac{\sum_{i=1}^{8} x_i}{8} = 9.5$$

$$\sum_{i=1}^{8}(x-\bar{x})^2 = (10-9.5)^2 + (18-9.5)^2 + (15-9.5)^2 + (12-9.5)^2 + (3-9.5)^2 + (6-9.5)^2 + (5-9.5)^2 + (7-9.5)^2$$

$$\sum_{i=1}^{8}(x-\bar{x})^2 = 0.25 + 72.25 + 30.25 + 6.25 + 42.25 + 12.25 + 20.25 + 6.25 = 190$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}} = \sqrt{\frac{190}{8}} = \sqrt{23.75} = 4.873$$

La varianza es: $v = \sigma^2 = 23.75$

Ejemplo.
Hallar la desviación estándar y la varianza para la siguiente distribución de frecuencias.

Clases	f
10-15	2
16-21	8
22-27	13
28-33	10
34-39	6
	n=39

Solución.
Calculando los puntos medios de cada clase y obteniendo $f \cdot x$:

Clases	f	x	$f \cdot x$
10-15	2	12.5	25
16-21	8	18.5	148
22-27	13	24.5	318.5
28-33	10	30.5	305
34-39	6	36.5	219
	n=39		1,015.5

La media es: $\bar{x} = \frac{\sum_{i=1}^{5}(f \cdot x)}{n} = \frac{1{,}015.5}{39} = 26.038$. Por lo tanto:

$$\sum_{i=1}^{5} f(x_i-\bar{x})^2 = 2(12.5-26.038)^2 + 8(18.5-26.038)^2 + 13(24.5-26.038)^2 + 10(30.5-26.038)^2$$

$$+ 6(36.5-26.038)^2 = 366.55 + 454.57 + 30.75 + 199.09 + 656.46 = 1{,}707.42$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{1{,}707.42}{39}} = \sqrt{43.78} = 6.616$$

La varianza es: $v = \sigma^2 = 43.78$

COEFICIENTE DE VARIACIÓN

Cuando se quiere comparar el grado de dispersión de dos distribuciones que no vienen dadas en las mismas unidades o que las medias no son iguales se utiliza el *coeficiente de variación de Pearson* que se define como el cociente entre la desviación estándar y el valor absoluto de la media aritmética:

$$\%\ CV = \frac{\sigma}{\bar{x}} \cdot 100$$

Este coeficiente, representa el porcentaje que la desviación estándar contiene a la media aritmética y por lo tanto cuanto mayor es CV mayor es la dispersión y menor la representatividad de la media.

Ejemplo.
Hallar el coeficiente de variación del ejemplo anterior.

Solución.

$$CV = \frac{6.616}{26.038} \cdot 100 = 25.40\ \%$$

I.6 ANÁLISIS DESCRIPTIVO DE DATOS BIVARIADOS. CORRELACIÓN

Hasta ahora se han estudiado los índices y representaciones de una sola variable por individuo. Son del tipo distribución unidimensional.

Cuando sobre una población se estudian simultáneamente los valores de dos variables estadísticas, el conjunto de los pares de valores correspondientes a cada individuo se denomina distribución bidimensional.

DIAGRAMAS DE DISPRESIÓN

La distribución conjunta de dos variables puede expresarse gráficamente mediante un diagrama de dispersión: en un plano se representa cada elemento observado haciendo que sus coordenadas sobre los ejes cartesianos sean los valores que toman las dos variables para esa observación.

Ejemplo.
La siguiente tabla muestra los datos correspondientes a un conjunto de diez pares de observaciones de estaturas de padres e hijos:

Padre (m)	1.70	1.77	1.68	1.75	1.80	1.75	1.69	1.72	1.71	1.73
Hijo (m)	1.74	1.78	1.72	1.77	1.78	1.77	1.71	1.76	1.73	1.74

El diagrama de dispersión de ese grupo de datos es:

Se representa la variable dependiente en el eje de las ordenadas y la independiente en el eje de las abscisas. Cuando se estudia la relación entre dos variables, una puede considerarse causa y la otra resultado o efecto de la primera, siendo ésta una decisión teórica. Se conoce como variable exógena, o variable independiente a la que causa el efecto y variable endógena, o variable dependiente a la que lo recibe.

Por supuesto que diferentes conjuntos de datos ofrecerán diagramas diferentes. Sin embargo, se pueden considerar cuatros tipos de diagramas de dispersión, que son los más típicos:

1. Relación tal que al aumentar los valores de la variable independiente aumenta (en promedio) el valor de la variable dependiente. Cuando esto ocurre hay una relación lineal positiva.
2. Relación tal que al aumentar los valores de la variable independiente se reduce (en promedio) el valor de la variable dependiente. Cuando esto ocurre hay una relación lineal negativa.
3. No hay relación entre ambas variables. Esto significa que las variables son independientes.
4. Relación entre ambas, pero no lineal.

COVARIANZA

La *covarianza* es una medida de la asociación lineal entre dos variables que resume la información existente en un gráfico de dispersión. Es un indicador de si los valores están relacionados entre sí, se simboliza por σ_{xy} y se calcula por medio de:

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

Esta medida, refleja la relación lineal que existe entre dos variables. El resultado numérico fluctúa entre los rangos de $[-\infty, \infty]$. Al no tener unos límites establecidos no puede determinarse el grado de relación lineal que existe entre las dos variables, sólo es posible ver la tendencia.

- Una covarianza positiva significa que existe una relación lineal positiva entre las dos variables. Es decir, los valores bajos de la variable x se asocian con los valores bajos de la variable y, mientras los valores altos de x se asocian con los valores altos de la variable y.
- Una covarianza de negativa significa que existe una relación lineal inversa (negativa) entre las dos variables. Lo que significa que los valores bajos en x se asocian con los valores altos en y, mientras los valores altos en x se asocian con los valores bajos en y.
- Una covarianza de cero se interpreta como la no existencia de una relación lineal entre las dos variables estudiadas.

Ejemplo.
Dada la tabla de estaturas de 10 padres y 10 hijos, calcular su covarianza e interpretarla.

Padre (m)	1.70	1.77	1.68	1.75	1.80	1.75	1.69	1.72	1.71	1.73
Hijo (m)	1.74	1.78	1.72	1.77	1.78	1.77	1.71	1.76	1.73	1.74

Solución.

La estatura media para los padres es: $\bar{x} = \dfrac{\sum_{i=1}^{10} x_i}{n} = \dfrac{17.3}{10} = 1.73\,\text{m}.$

La estatura media para los hijos es: $\bar{y} = \dfrac{\sum_{i=1}^{10} y_i}{n} = \dfrac{17.5}{10} = 1.75\,\text{m}.$

Por lo tanto:

$$\sum_{i=1}^{10}(x_i-\bar{x})(y_i-\bar{y}) = (1.70-1.73)(1.74-1.75)+(1.77-1.73)(1.78-1.75)+(1.68-1.73)(1.72-1.75)$$
$$+(1.75-1.73)(1.77-1.75)+(1.80-1.73)(1.78-1.75)+(1.75-1.73)(1.77-1.75)+(1.69-1.73)(1.71-1.75)$$
$$+(1.72-1.73)(1.76-1.75)+(1.71-1.73)(1.73-1.75)+(1.73-1.73)(1.74-1.75)=0.0078$$

$$\sigma_{xy} = \frac{\sum_{i=1}^{10}(x_i-\bar{x})(y_i-\bar{y})}{n} = \frac{0.0078}{10} = 0.00078$$

Como la covarianza es positiva significa que existe una relación lineal positiva entre las dos variables. Es decir, a valores grandes de x (estaturas de los padres) se asocian valores altos de y (estaturas de los hijos).

CORRELACIÓN

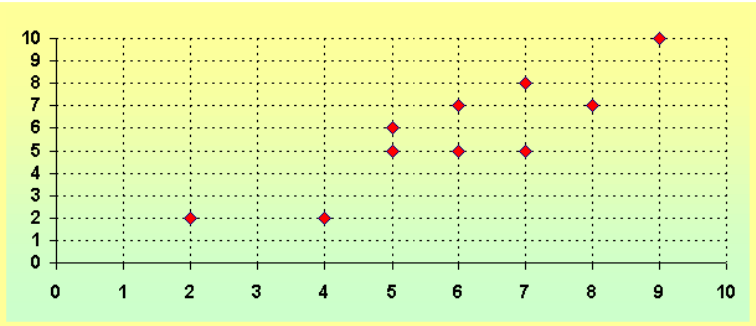
Es frecuente que se estudie sobre una misma población los valores de dos variables estadísticas distintas, con el fin de ver si existe alguna relación entre ellas, es decir, si los cambios en una de ellas influyen en los valores de la otra. Si ocurre esto se dice que las variables están correlacionadas o bien que hay *correlación* entre ellas.

Ejemplo.
Las calificaciones de 10 alumnos en Matemáticas y Física vienen dadas en la siguiente tabla:

Matemáticas	2	4	5	5	6	6	7	7	8	9
Física	2	2	5	6	5	7	5	8	7	10

Los pares de valores { (2,2), (4,2), (5,5), ⋯, (8,7), (9,10) }, forman la distribución bidimensional en la que hay cierta tendencia a que cuanto mejor es la calificación en Matemáticas, mejor es la de Física.

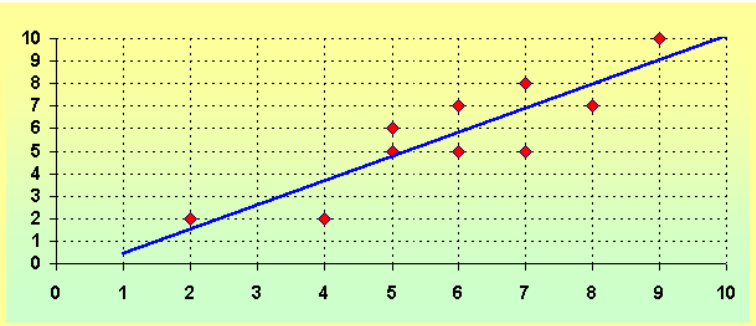
Representando los pares de valores en el plano cartesiano se obtiene su diagrama de dispersión:



Cuando se puede apreciar si los puntos se distribuyen alrededor de una recta entonces se dice que hay *correlación lineal*.

Una correlación lineal fuerte es cuando la nube (conjunto de puntos) se parece mucho a una recta y será cada vez más débil (o menos fuerte) cuando la nube vaya diseminándose con respecto a la recta.

En el ejemplo se aprecia que la correlación es bastante fuerte, ya que si se traza una recta, ésta se ubica muy próxima a los puntos de la nube.



La correlación indica la fuerza y la dirección de una relación lineal entre dos variables aleatorias. Se considera que dos variables cuantitativas están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los valores homónimos de la otra: si se tienen dos variables (x y y) existe correlación si al aumentar los valores de x lo hacen también los de y y viceversa. La correlación entre dos variables no implica, por sí misma, ninguna relación de causalidad

La relación entre dos variables cuantitativas queda representada mediante la línea de mejor ajuste, trazada a partir de la nube de puntos. Los tres principales componentes elementales de una línea de ajuste y, por lo tanto, de una correlación, son la fuerza, el sentido y la forma:

1. La *fuerza* mide el grado en que la línea representa a la nube de puntos: si la nube es estrecha y alargada, se representa por una línea recta, lo que indica que la relación es fuerte; si la nube de puntos tiene una tendencia elíptica o circular, la relación es débil.

2. El *sentido* mide la variación de los valores de y con respecto a x: si al crecer los valores de x lo hacen los de y, la relación es positiva; si al crecer los valores de x disminuyen los de y, la relación es negativa.

3. La *forma* establece el tipo de línea que define el mejor ajuste: la línea recta, cuadrática, polinomial, etc.

La apreciación visual de la existencia de correlación no es suficiente. Así que se define como *coeficiente de correlación de Pearson* al índice estadístico que mide la relación lineal entre dos variables cuantitativas. Se denota por r:

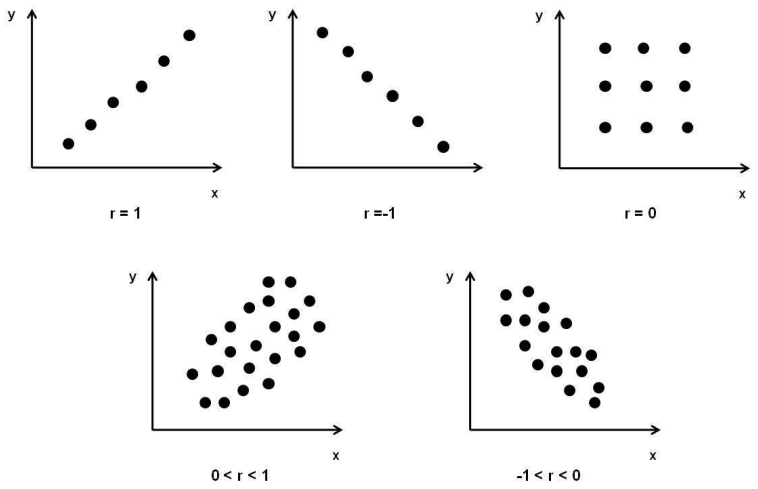
$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Este coeficiente de correlación lineal divide la covarianza por el producto de las desviaciones estándar de ambas variables. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables.

El valor del índice de correlación varía en el intervalo [-1, 1] y se interpreta de la siguiente forma:

- Si r = 0, no existe ninguna correlación. El índice indica, por lo tanto, una independencia total entre las dos variables, es decir, que la variación de una de ellas no influye en absoluto en el valor que pueda tomar la otra.
- Si r = 1, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada relación directa: cuando una de ellas aumenta, la otra también lo hace en idéntica proporción.
- Si 0 < r < 1, existe una correlación positiva.
- Si r = -1, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada relación inversa: cuando una de ellas aumenta, la otra disminuye en idéntica proporción.
- Si -1 < r < 0, existe una correlación negativa.

Gráficamente es:



Ejemplo.
Obtener la correlación que existe entre la estatura y el peso de 10 jugadores del equipo Leopardos de fútbol americano de la prepa 8.

Estatura (m)	1.72	1.79	178	1.75	1.80	1.79	1.81	1.70	1.68	1.73
Peso (kg)	74	81	76	77	87	86	92	67	76	74

Solución.
Considerando que la estatura es la variable x y que el peso es la variable y se tiene:

La estatura media es: $\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{17.55}{10} = 1.755\,\text{m}.$

El peso medio es: $\bar{y} = \frac{\sum_{i=1}^{10} y_i}{n} = \frac{790}{10} = 79\,\text{kg}.$

Por lo tanto la covarianza es:

$$\sum_{i=1}^{10}(x_i-\bar{x})(y_i-\bar{y}) = (1.72-1.755)(74-79)+(1.79-1.755)(81-79)+(1.78-1.755)(76-79)$$
$$+(1.75-1.755)(77-79)+(1.80-1.755)(87-79)+(1.79-1.755)(86-79)+(1.81-1.755)(92-79)$$
$$+(1.70-1.755)(67-79)+(1.68-1.755)(76-79)+(1.73-1.755)(74-79)=2.51$$

$$\sigma_{xy} = \frac{\sum_{i=1}^{10}(x_i-\bar{x})(y_i-\bar{y})}{n} = \frac{2.51}{10} = 0.251$$

Calculando la desviación estándar de las estaturas:

$$\sum_{i=1}^{10}(x_i-\bar{x})^2 = (1.72-1.755)^2+(1.79-1.755)^2+(1.78-1.755)^2+(1.75-1.755)^2+(1.8-1.755)^2$$
$$+(1.79-1.755)^2+(1.81-1.755)^2+(1.70-1.755)^2+(1.68-1.755)^2+(1.73-1.755)^2=0.01865$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{10}(x_i-\bar{x})^2}{n}} = \sqrt{\frac{0.01865}{10}} = \sqrt{0.001865} = 0.043185$$

Calculando la desviación estándar de los pesos:

$$\sum_{i=1}^{10}(y_i-\bar{y})^2 = (74-79)^2+(81-79)^2+(76-79)^2+(77-79)^2+(87-79)^2+(86-79)^2$$
$$+(92-79)^2+(67-79)^2+(76-79)^2+(74-79)^2=502$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{10}(x_i-\bar{x})^2}{n}} = \sqrt{\frac{502}{10}} = \sqrt{50.2} = 7.085195$$

Por lo tanto el coeficiente de correlación entre las dos variables es:

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{0.251}{(0.043185)(7.085195)} = 0.8203$$

Como el valor está cercano a uno, entonces existe una correlación positiva. Este índice indica que a mayor estatura de los jugadores, mayor es su el peso.

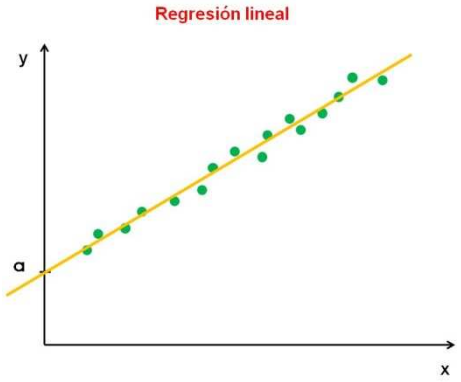
REGRESIÓN LINEAL POR COVARIANZA

En múltiples ocasiones se requiere analizar la relación entre dos variables cuantitativas. Los dos objetivos fundamentales de este análisis son:

1. Determinar si dichas variables están asociadas y en qué sentido se da dicha asociación (es decir, si los valores de una de las variables tienden a aumentar –o disminuir- al aumentar los valores de la otra).
2. Estudiar si los valores de una variable pueden ser utilizados para predecir el valor de la otra.

La forma correcta de abordar el primer problema es recurriendo a coeficientes de correlación. Sin embargo, el estudio de la correlación es insuficiente para obtener una respuesta a la segunda pregunta, ya que se limita a indicar la fuerza de la asociación mediante un único número, tratando las variables de modo simétrico, mientras que lo que se busca es modelar dicha relación y usar una de las variables para explicar la otra. Para tal propósito se recurrirá a la técnica de regresión[2].

La *regresión lineal* permite definir la recta que mejor se ajusta a esta nube de puntos. Gráficamente:



La recta está definida por la siguiente expresión:

$$y = a + bx$$

donde y es la variable dependiente y x es la variable independiente. Sus coeficientes representan:

- b determina la *pendiente* de la recta, es decir, su grado de inclinación. Se calcula como la covarianza de las dos variables, dividida por la varianza de la variable x:

$$b = \frac{\sigma_{xy}}{\sigma_x^{\,2}}$$

- a es el valor que toma y cuando la variable independiente x vale cero. Es el punto donde la recta cruza el eje vertical, llamado ordenada al origen de la recta. Se calcula como la media de la variable y, menos la media de la variable x multiplicada por el parámetro b que se ha calculado:

$$a = \bar{y} - b \cdot \bar{x}$$

Ejemplo.
Obtener y graficar la recta de regresión con los datos de estatura y peso de los 10 jugadores del equipo Leopardos de fútbol americano de la prepa 8 del ejemplo anterior.

Solución.
Los valores obtenidos de la covarianza y de la desviación estándar de las estaturas son:
$\sigma_{xy} = 0.251$
$\sigma_x = 0.043185$

[2] Aquí se analizará el caso más sencillo en el que se considera únicamente la relación entre dos variables. Asimismo, se limitará al caso en el que la relación que se pretende modelar es de tipo lineal.

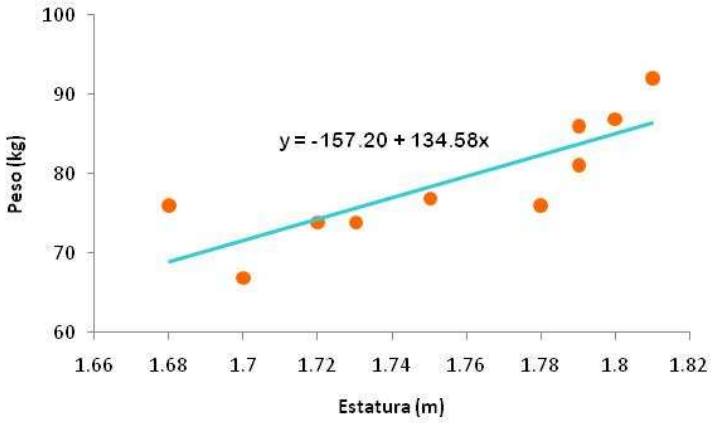
$$\Rightarrow \quad b = \frac{\sigma_{xy}}{\sigma_x^{\,2}} = \frac{0.251}{(0.043185)^2} = 134.58$$

Considerando que $\bar{x} = 1.755$ m. y que $\bar{y} = 79$ kg:

$a = \bar{y} - b \cdot \bar{x} = 79 - (134.58)(1.755) = -157.20$

Por lo que la recta de regresión lineal es: $y = -157.20 + 134.58x$

Su gráfica es:



Nótese como la gráfica es congruente con el coeficiente de correlación $(r = 0.8203)$. Muestra una pendiente positiva y se ajusta a una recta lo que ratifica que a mayor estatura de los jugadores, mayor es su peso.

REGRESIÓN LINEAL POR MÍNIMOS CUADRADOS

Una estrategia adicional para ajustar adecuadamente el comportamiento o la tendencia general de los datos a través de una recta que minimice la suma de los cuadrados de las distancias verticales de los puntos a la recta. Este método se conoce como *regresión por mínimos cuadrados*.

Para obtener una recta de la forma:

$$y = a + bx$$

este método se basa en la aplicación de las siguientes expresiones:

$$a = \frac{\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \cdot \sum_{1=1}^{n} (x_i)^2}{\left(\sum_{i=1}^{n} x_i\right)^2 - n\sum_{i=1}^{n} (x_i)^2}$$

$$b = \frac{n \cdot \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} (x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Ejemplo.
Aplicando el método de mínimos cuadrados, obtener la recta de regresión para los siguientes datos:

Estatura (m)	1.72	1.79	178	1.75	1.80	1.79	1.81	1.70	1.68	1.73
Peso (kg)	74	81	76	77	87	86	92	67	76	74

Solución.

Acomodando la tabla convenientemente y obteniendo $x \cdot y$ y x^2 se tiene:

Estatura x	Peso y	$x \cdot y$	x^2
1.72	74	127.28	2.9584
1.79	81	144.99	3.2041
1.78	76	135.28	3.1684
1.75	77	134.75	3.0625
1.8	87	156.6	3.24
1.79	86	153.94	3.2041
1.81	92	166.52	3.2761
1.7	67	113.9	2.89
1.68	76	127.68	2.8224
1.73	74	128.02	2.9929
17.55	790	1388.96	30.8189

$$a = \frac{\sum_{i=1}^{10} x_i \cdot \sum_{i=1}^{10} x_i y_i - \sum_{i=1}^{10} y_i \cdot \sum_{1=1}^{10} (x_i)^2}{\left(\sum_{i=1}^{10} x_i\right)^2 - 10\sum_{i=1}^{10} (x_i)^2} = \frac{17.55(1388.96) - 790(30.8189)}{(17.55)^2 - 10(30.8189)} = -157.19$$

$$b = \frac{10 \cdot \sum_{i=1}^{10} x_i y_i - \sum_{i=1}^{10} x_i \sum_{i=1}^{10} y_i}{10\sum_{i=1}^{10} (x_i)^2 - \left(\sum_{i=1}^{10} x_i\right)^2} = \frac{10(1388.96) - 17.55(790)}{10(30.8189) - (17.55)^2} = 134.58$$

Por lo que la recta de regresión lineal usando mínimos cuadrados es: $y = -157.19 + 134.58x$

Puede advertirse que la esta ecuación es prácticamente igual que la obtenida por medio del método de covarianza. Eso significa que ambos métodos son aplicables siempre que el coeficiente de correlación sea cercano a la unidad.