Estadística

La estadística se ocupa de los métodos científicos que se utilizan para recolectar, organizar, resumir, presentar y analizar datos así como para obtener conclusiones válidas y tomar decisiones razonables con base en este análisis. El término estadística también se usa para denotar los datos o los números que se obtienen de esos datos; por ejemplo, los promedios. Así, se habla de estadísticas de empleo, estadísticas de accidentes, etcétera

La ciencia de datos utiliza la estadística para extraer conocimiento de los datos. En esta unidad veremos algunos conceptos y técnicas de estadística comunes.

.1 Población vs Muestra

Cuando se recolectan datos sobre las características de un grupo de individuos o de objetos, puede ser imposible observar todo el grupo, en especial si se trata de un grupo grande. En vez de examinar todo el grupo, al que se le conoce como población o universo, se examina sólo una pequeña parte del grupo, al que se le llama muestra. Las poblaciones pueden ser finitas o infinitas.

.2 Estadística descriptiva o deductiva vs Estadística inductiva o inferencial

A la parte de la estadística que únicamente trata de describir y analizar un grupo dado, sin sacar ninguna conclusión ni hacer inferencia alguna acerca de un grupo más grande, se le conoce como estadística descriptiva o deductiva.

Si, en cambio, a partir de una muestra, logramos inferir conclusiones que son válidas para toda la población, estamos aplicando estadística inductiva o inferencial.

.3 Variables cuantitativas y cualitativas: discretas y continuas. Dominio

Una variable **cualitativa o categórica** describe cualidades, circunstancias o características de un objeto o persona. No pueden ser medidas en números y se pueden distinguir dos tipos:

- variable cualitativa ***nominal***: no admiten un criterio de orden (ej: estado civil)
- variable cualitativa ***ordinal***: tiene una modalidad no numérica, pero existe un orden (ej: calificación conceptual – desaprobado, aprobado, notable, sobresaliente; medallas en una prueba deportiva – oro, plata, bronce; puesto conseguido enuna prueba deportiva – primero, segundo, tercero)

Una variable **cuantitativa o numérica** puede tomar una serie de valores y admiten operaciones aritméticas.

Una variable que puede tomar cualquier valor entre dos números cualquiera es una variable ***continua*** (por ejemplo: longitud o temperatura); de lo contrario es una variable ***discreta*** (como cantidad de alumnos o unidades vendidas). Una variable discreta sólo puede tomar un número finito de valores entre dos valores cualesquiera, una variable continua, puede tomar infinitos valores entre dos de la escala.

A los valores válidos que puede tomar una variable, se le denomina **dominio** de esa variable. Por ejemplo: La cantidad de agua que entra en un recipiente es una variable cuantitativa continua, cuyo dominio es desde 0, hasta la capacidad total del recipiente.

Países de Europa es una variable cualitativa nominal, pero aún así tiene un dominio: España, Portugal, Francia, Italia, Alemania, etc.

Medidas de tendencia central

.1 Media

Es el valor promedio de un conjunto de datos numéricos.

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum_{j=1}^{N} X_j}{N} = \frac{\sum X}{N}$$

$\bar{X} = media$
$N = cantidad\ de\ números$
$X_1, \ldots X_N = serie\ de\ N\ números$

.2 Mediana

La mediana de un conjunto de números **ordenados** es el valor central o la media de los dos valores centrales.

$$\begin{cases} x_{\frac{N+1}{2}} \text{ si } N \text{ impar} \\ \frac{1}{2} \cdot \left(x_{\frac{N}{2}} + x_{\frac{N}{2}+1} \right) \text{ si } N \text{ par} \end{cases}$$

Ejemplos:

Buscamos la mediana en dos series de conjuntos A y B:

A = [56, 5, 88, 2, -10, 73, 11, 8, 33]

B = [22, 9, 5, 32, 14, -3, 1, 12, 2, 9]

El primer paso es ordenar los conjuntos:

A_{ord} = [-10, 2, 5, 8, 11, 33, 56, 73, 88]

B_{ord} = [-3, 1, 2, 5, 9, 7, 12, 14, 22, 32]

Busco el elemento central de A_{ord} porque tiene una cantidad de elementos impar.

A_{ord} = [-10, 2, 5, 8, 11, 33, 56, 73, 88]

Entonces el *11 es la mediana del conjunto A $_{ord}$*, y determina una partición en el conjunto con la misma cantidad de elementos a izquierda y derecha.

En el conjunto B_{ord} tenemos una cantidad par de elementos, por lo que debemos hallar los dos centrales, y calcular su media:

B_{ord} = [−3, 1, 2, 5, 7, 9, 12, 14, 22, 32]

$media_{(7,9)}$ = (7+9)/2 = 8

Decimos entonces que la *mediana del conjunto* B_{ord} *es 8.*

.3 Moda

Es el valor que aparece con mayor frecuencia. En un conjunto de datos puede no haber moda, y si la hay, puede que no sea única.

Ejemplos:

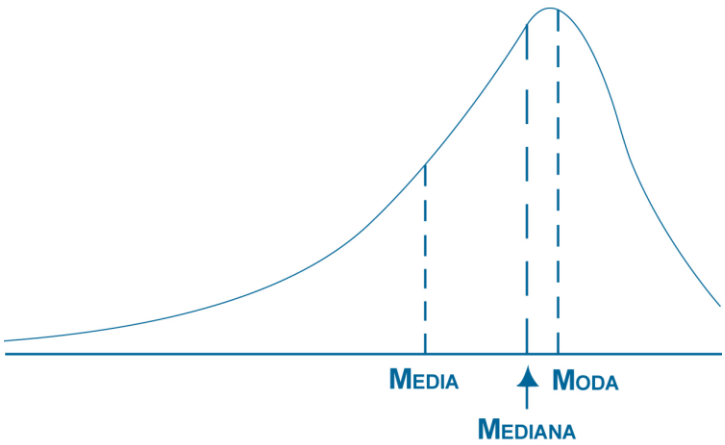
La moda del conjunto [2, 2, 5, 7, **9, 9, 9**, 10, 10, 11, 12, 18] es 9. Tiene sólo una, por lo que se llama *unimodal.*

El conjunto [3, 5, 8, 10, 12, 15, 16] no tiene moda.

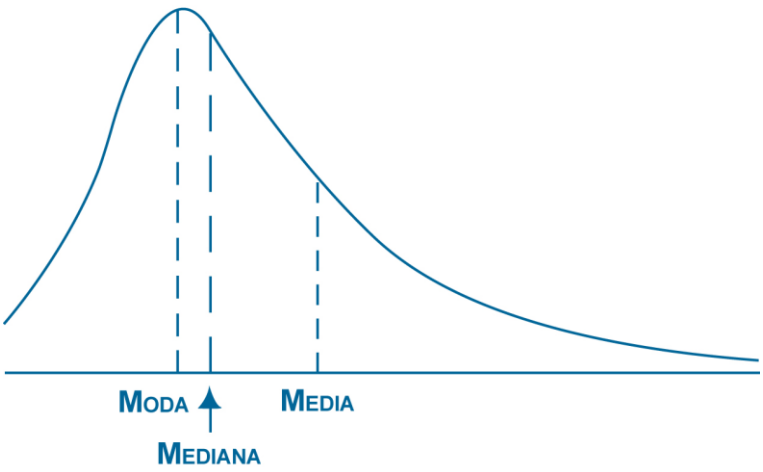
El conjunto [2, 3, **4, 4, 4**, 5, 5, **7, 7, 7**, 9] tiene dos modas, 4 y 7, por lo que se le llama *bimodal*.

.4 Significado

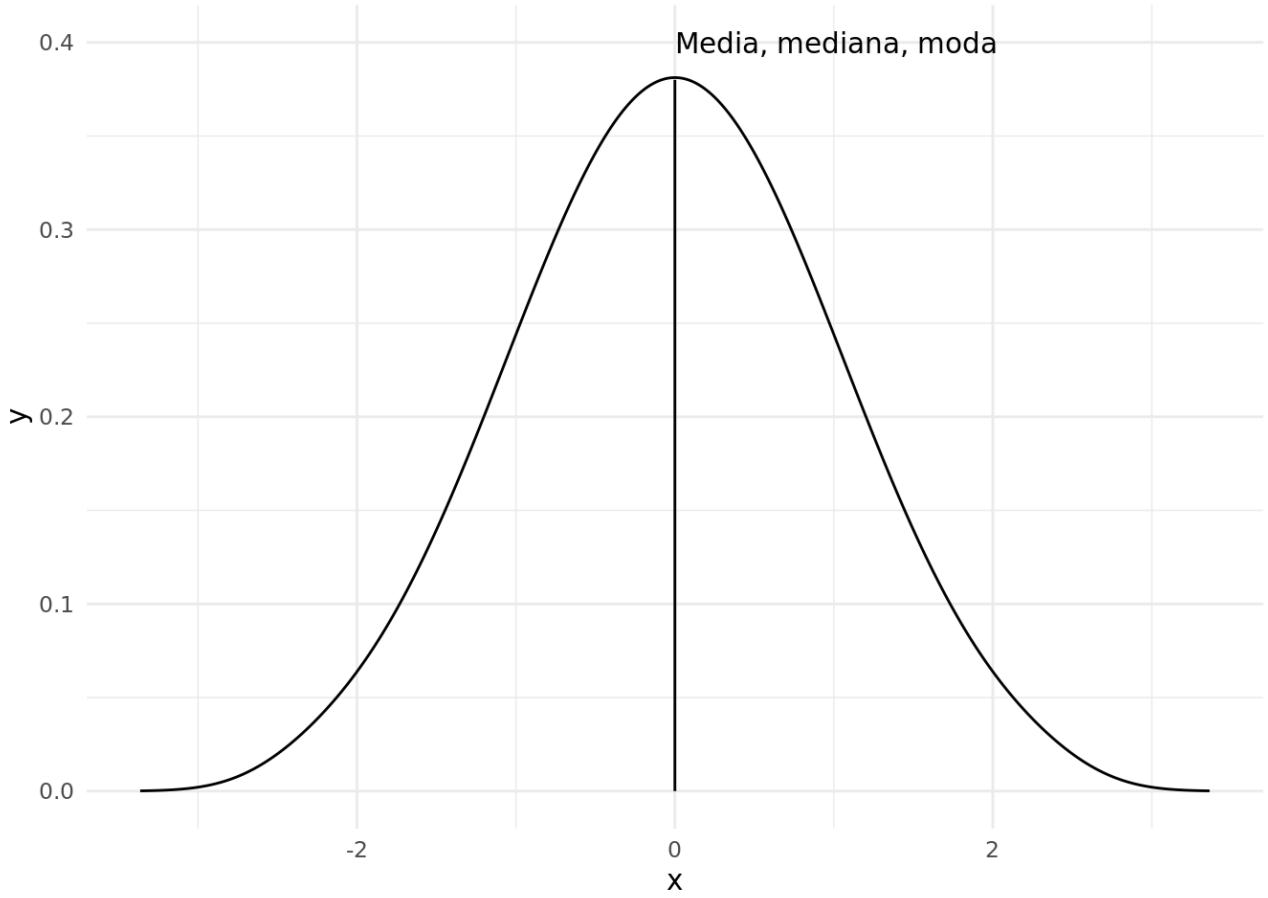
Estas tres medidas nos indican la forma en que están distribuidos nuestros datos



Distribución sesgada a la izquierda



Distribución sesgada a la derecha



Distribución normal, gaussiana, de Laplace-Gauss o distribución de Gauss. Coinciden media, mediana y moda. La curva es acampanada y simétrica respecto de estos parámetros..

Distribuciones de frecuencia

.a) Datos en bruto

Los **datos en bruto** son los datos recolectados que aún no se han organizado. Por ejemplo, las edades de 100 miembros practicantes de tai-chi para adultos mayores.

.b) Ordenaciones

Ordenación se le llama a los datos numéricos en bruto, dispuestos en orden creciente o decreciente.

La diferencia entre el número mayor y el número menor, se le llama **rango de los datos**. Por ejemplo, si la mayor edad de entre los 100 miembros es de 74 años, y la menor de 60 años, el rango es 74 – 60 = 14 años.

.c) Distribuciones de frecuencia

Al organizar gran cantidad de datos, se suele distribuir en clases o categorías. Llamamos **frecuencia de clase** a la cantidad de datos que pertenecen a cada clase.

Si presentamos esta información en forma de tabla, obtenemos una ***distribución de frecuencias*** o ***tabla de frecuencias***.

Edad (años)	Cantidad de practicantes
60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 – 74	8
	Total: 100

Tabla 1 – Tabla de frecuencias

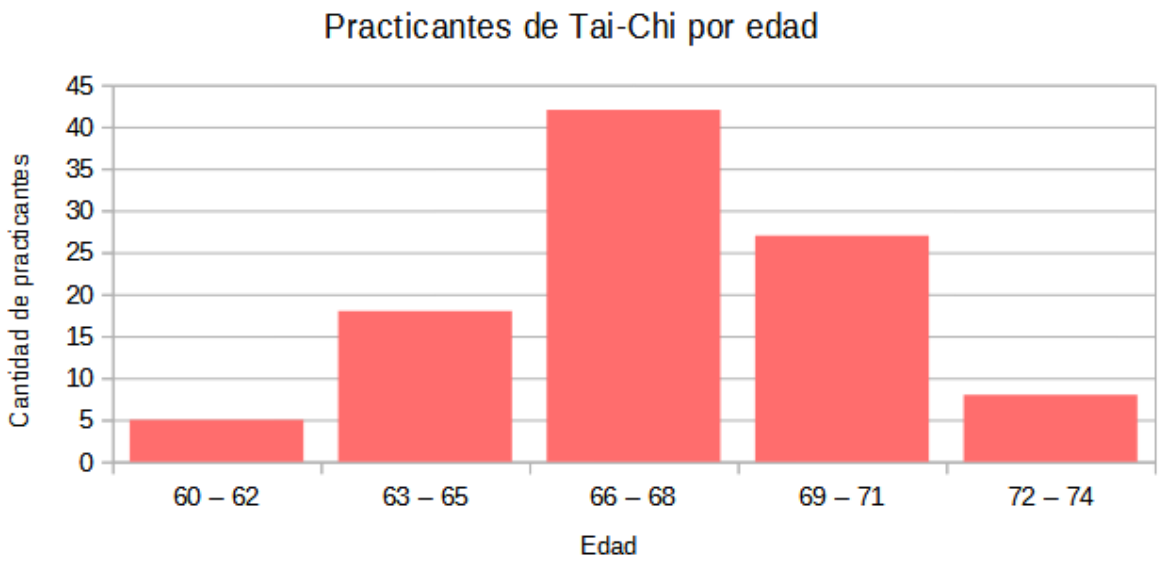


Figura 1- Diagrama de frecuencias

.d) Intervalos de clase y límites de clase

Al rango que representa cada clase, por ejemplo 60 – 62, en la tabla 1, se le conoce como *intervalo de clase.* A los números de los extremos, se los conoce como *límites de clase:* el número menor (60) es el *límite inferior de la clase,* y el número mayor (62) es el *límite superior de la clase.*

Un intervalo de clase que no tenga indicado el límite de clase superior o el límite de clase inferior, se conoce como *intervalo de clase abierto.* Por ejemplo, si hubiésemos definido el primer grupo como “personas menores a 60 años”, o el grupo del otro extremo como “personas mayores a 62 años”.

Tamaño o amplitud de un intervalo o clase: El tamaño o amplitud de un intervalo de clase es la diferencia entre sus fronteras superior e inferior. Se lo conoce como *amplitud de clase, tamaño de clase* o *longitud de clase.*

Marca de clase: Es el punto medio de la clase y se obtiene suamndo los límites de clase inferior y superior y dividiendo entre 2. A la marca de clase se la conoce también como *punto medio de clase.*

Reglas generales para formar una distribución de frecuencias

1. Se determinan el mayor y menor valor para hallar el rango.
2. Se divide el rango en una cantidad adecuada de intervalos de clase de la misma amplitud, si es posible. La cantidad de intervalos suele ser de 5 a 20, dependiendo de los datos.
3. Se encuentran las frecuencias de clase.

.e) Histogramas y polígonos de frecuencias

Los histogramas y los polígonos de frecuencias son dos maneras de representar gráficamente las distribuciones de frecuencias.

1. Un histograma o histograma de frecuencias consiste en un conjunto de rectángulos que tienen:

a) sus bases sobre un eje horizontal (el eje X), con sus centros coincidiendo con las marcas de clase de longitudes iguales a la amplitud del intervalo de clase, y
b) áreas proporcionales a las frecuencias de clase

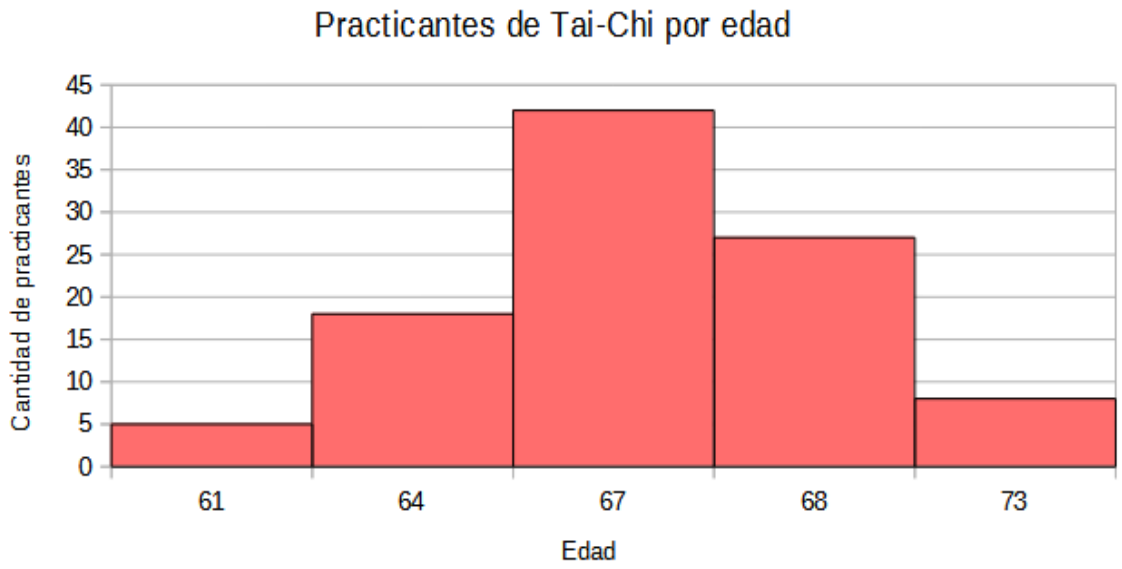


Figura 2- Histograma

2. Un polígono de frecuencias es una gráfica de línea que presenta las frecuencias de clase graficadas contra las marcas de clase. Se puede obtener conectando los puntos medios de las partes superiores de los rectángulos de un histograma.

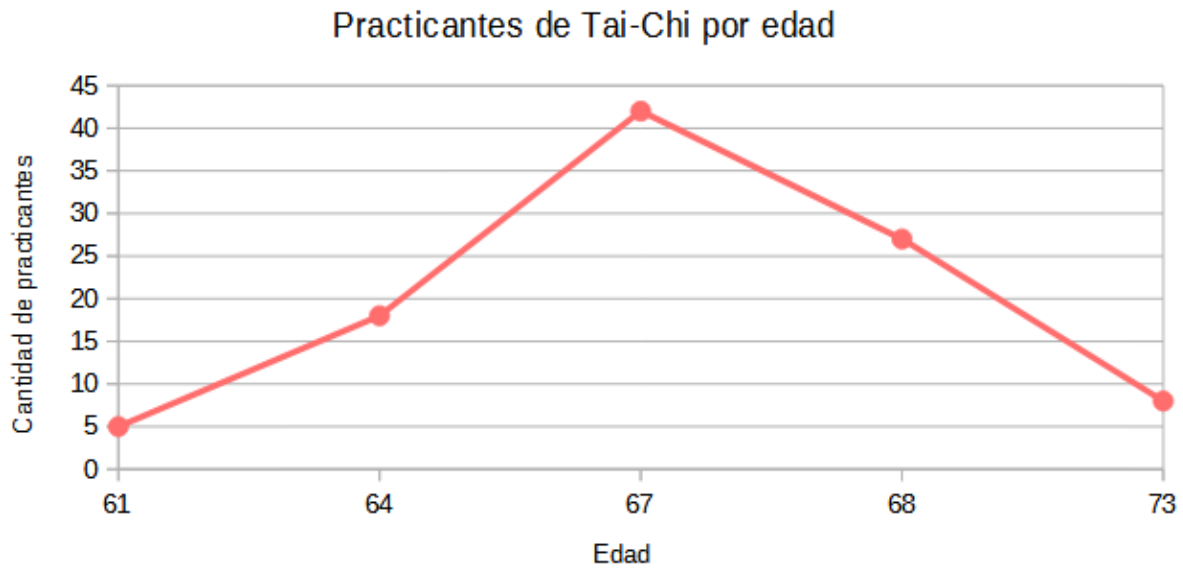


Figura 3 - Polígono de frecuencias

.f) Relación entre media, mediana y moda

Según la distribución de los datos, podemos definir la relación entre las tres medidas de tendencia central.

Cuando la distribución de los datos es simétrica, la línea de la media, la mediana y la moda, coinciden.

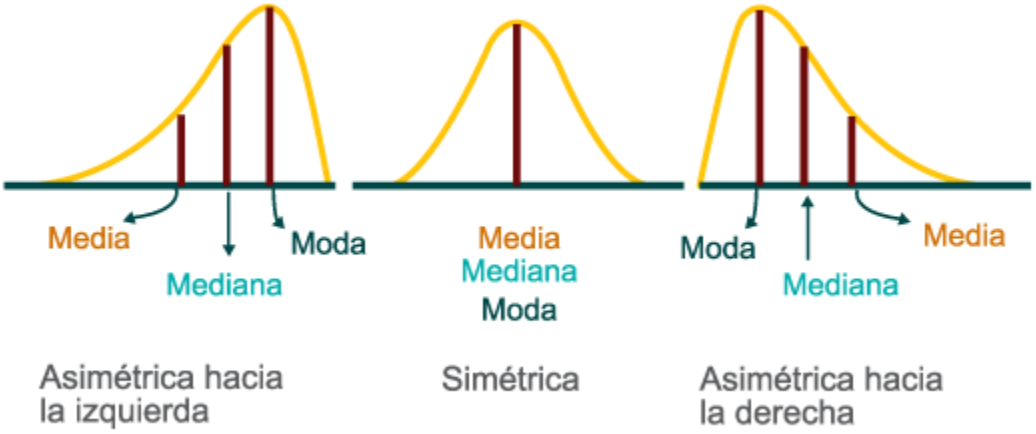


Figura 4 – Relación entre las medidas de tendencia central