Unidad 9: Estadística Descriptiva

Introducción

Estadística

Algunas definiciones:

- Disciplina científica que se ocupa de la obtención, orden y análisis de un conjunto de datos con el fin de obtener explicaciones y predicciones sobre fenómenos observados.
- Disciplina que estudia la variabilidad, recolección, organización, análisis, interpretación y presentación de los datos, así como el proceso aleatorio que los genera siguiendo las leyes de la probabilidad.
- Ciencia que se ocupa de la recolección, resumen, análisis, e interpretación de hechos o datos numéricos.

Algunos objetivos

- Extraer conocimiento a partir de un conjunto de datos, con el fin de tomar decisiones en base a la mejor información posible (siempre existe incertidumbre).
- Obtener información de una población, sin necesidad de estudiar todos los elementos que la componen.
- Sacar conclusiones sobre alguna característica de una población en base a una muestra representativa de la misma.

Introducción

Estadística Descriptiva

Se refiere a los métodos de recolección, organización, resumen y presentación de un conjunto de datos. Se trata principalmente de describir las características fundamentales de los datos y para ellos se suelen utilizar indicadores, gráficos y tablas.

Estadística Inferencial

Se refiere a los métodos utilizados para poder hacer predicciones, generalizaciones y obtener conclusiones a partir de los datos analizados teniendo en cuenta el grado de incertidumbre existente.

En esta unidad veremos conceptos básicos de Estadística Descriptiva.

Objetivo Básico: Describir lo más simplemente posible los resultados obtenidos en un experimento/encuesta/censo, etc.

Organización y Presentación de resultados:

- Representaciones tabulares (tablas): 1er paso en la organización de datos, que se ordenan en filas y columnas para documentar y comunicar la información.
- Representaciones gráficas (histogramas, gráficos de barras, circulares, etc): brindan un resumen visual de los datos.
- Dependiendo del tipo de datos e información a comunicar se elegirá qué tipo de representación utilizar

Conceptos Estadísticos Básicos

- **Unidad Elemental:** Es cualquier objeto real o ideal sobre el cual pueden hacerse mediciones.
  Ejemplos: un alumno, un paciente de determinado hospital, etc.
- **Población:** Conjunto de unidades elementales que satisfacen una definición común. Debe estar bien definida en tiempo y espacio. Denotamos N : tamaño de la población.
  Ejemplos: alumnos ingresantes a FaCENA en 2023, pacientes ingresados en Hospital Escuela durante 2020-2022, etc.

Población:

Denotamos con N al tamaño de la población. La población puede ser:

- **Finita:** cuando todas las unidades elementales que la componen pueden ser físicamente listadas o individualizadas, es decir, que está constituida por un número finito de elementos.
  Ejemplos: alumnos de una universidad, habitantes de una ciudad, etc.
- **Infinita:** cuando en la práctica no se puede individualizar o listar los elementos que componen la población. Es la que está compuesta por un número indefinidamente grande de unidades elementales.
  Ejemplos: estrellas en el universo, granos de arena en una playa, población actual de mosquitos en el planeta, etc.

Muestra:

Es un subconjunto de la población. Debe ser:

- **Representativa:** debe representar los aspectos más importantes de la población de la cual se extrae. Para ello, deben tenerse en cuenta las variables que importan para el estudio que se encare.
- **Aleatoria:** todos los elementos de la población deben tener las mismas chances (probabilidad) de ser extraídos en la muestra.

Denotamos $n \leq N$ tamaño de la muestra (número de elementos observados).
Ejemplos: 10 alumnos por carrera de FaCENA, elegidos al azar, 30 pacientes por mes , elegidos al azar, ingresados al Hospital Escuela durante el período 2020-2022.

Diseño del Estudio

- **Censo:** conjunto de actividades destinadas a medir y/u observar ciertas características de todas las unidades elementales que componen la población objeto de estudio.
- **Muestreo:** conjunto de actividades destinadas a observar una parte o subconjunto de las unidades elementales que componen una muestra.

Factores

- **Exactitud y Precisión:** Censo $\rightarrow$ parámetro exacto. Muestreo $\rightarrow$ parámetro estimado.
- **Costo-Tiempo:** El Muestreo es más barato y demanda un tiempo menor de recolección de datos que un Censo.
- **Imposibilidad:** de acceder a todos los elementos que componen la población objeto de estudio.
- **Destrucción:** Existen estudios que para ser desarrollados, terminan destruyendo a las unidades elementales, por lo que debe trabajarse con muestras.

Variables:Clasificación

Variable: Cualquier característica susceptible de tomar distintos estados entre unidades elementales, o que varían dentro de una misma unidad elemental a través del tiempo.

Cualitativas

Expresan una cualidad o propiedad que el objeto en estudio tiene o no, o bien lo tiene en distinto grado.Pueden ser **DICOTÓMICAS:** dos categorías o clases (ej: fumador o no) ó **POLICOTÓMICAS:** más de dos categorías (ej:nivel de estudios alcanzado)

Cuantitativas

Asumen valores numéricos. Expresan una cantidad.

- **Discretas:** Surgen de contar. Sólo toman valores discretos dentro de su campo de variación (ej: cant de hijos).
- **Continuas:** Surgen de medir. Toman cualquier valor dentro de su rango de variación (ej: altura de un alumno).

Escala de Medición - Variables Cualitativas

Nominal o Clasificatoria:

- Los elementos se clasifican en clases o categorías, de modo que esta clasificación sea mutuamente excluyente y exhaustiva, verificando así una relación de equivalencia.
- La operación más elemental que se puede realizar es contar cuántos elementos pertenecen a cada categoría.
- Esta escala constituye el nivel de medición más bajo.

Ordinal o Jerárquica:

- Es posible ordenar los datos cualitativos de acuerdo a una jerarquía preestablecida de sus valores según el grado o rango que poseen.
- Las relaciones lógicas propias de esta escala son, Relación de Equivalencia (dentro de cada categoría) y Relación de Orden Estricto (entre categorías).

Escala de Medición - Variables Cuantitativas

Escala de Intervalos

- Al punto de origen de esta escala se le asigna arbitrariamente el valor cero, llamado cero arbitrario, que no necesariamente indica ausencia de la característica medida.
- Al no haber un cero absoluto en la zona de medición, permite valores negativos.
- En este nivel, se pueden realizar operaciones matemáticas entre los valores de las categorías, constituyendo un nivel de medición superior al de la Escala Ordinal.

Propiedades:

- La razón entre dos intervalos de dos escalas en que se mida la misma variable es la misma.
- Se puede pasar de una escala a otra mediante una transformación lineal de la forma $aX + b; a > 0; b > 0$.

Escala de Medición - Variables Cuantitativas

Escala de Razón o Proporción

- El punto de origen de esta escala es realmente cero, llamado cero real o cero absoluto, que indica ausencia de la característica medida.
- Se puede establecer una distancia, o una proporcionalidad, entre dos entes cualesquiera.
- Esta escala constituye el nivel de medición más alto.

Propiedades:

- La razón entre dos puntos de dos escalas en que se mida una misma variable es constante.
- El 0 es el mismo para todas las escalas en que se mida esa variable.
- Se puede pasar de una escala a otra mediante una transformación lineal de la forma $aX; a > 0$.

Tablas de frecuencias

Supondremos que X es una variable cualitativa o cuantitativa discreta. Sean $x_1, \ldots, x_k$; $k \leq n$ los distintos valores que adopta la variable.

Frecuencias Simples:

- f_i: frecuencia absoluta simple de x_i. Es el número de veces que se repite ese valor en las n observaciones.
- $r_i = \frac{f_i}{n}$: frecuencia relativa simple de x_i.
- $p_i = 100 \times r_i$: frecuencia porcentual simple de x_i.

Frecuencias Acumuladas

Supongamos que X es cuantitativa discreta, y $x_1 < x_2 < \cdots < x_k$.

- $F_i = \sum_{j=1}^{i} f_j$: frecuencia absoluta acumulada, suma de las frecuencias absolutas anteriores hasta el dato actual.
- $R_i = \frac{F_i}{n}$: frecuencia relativa acumulada.
- $P_i = 100 \times R_i$: frecuencia porcentual acumulada.

Tablas de frecuencias para datos en Agrupación Simple

- Representación tabular de los datos correspondientes a una variable.
- **Variable Cualitativa**: En 1^{er} columna modalidades que adopta la variable, en las siguientes columnas: frecuencias absolutas, relativas y porcentuales **simples**.
- **Variable Cuantitativa Discreta**: En 1^{er} columna los k distintos valores de la variable ordenados en forma creciente, siguientes columnas: frecuencias absolutas, relativas y porcentuales **simples y acumuladas**.

Datos Agrupados en Intervalos de Clase

La variable X puede ser: Cuantitativa continua, o Cuantitativa discreta, pero:

- Hay demasiados datos;
- Hay pocos datos pero muy dispersos;
- Se busca una clasificación particular.

Intervalos de clase:

- Deben ser **disjuntos**: cada observación debe estar contenida en un, y sólo un intervalo de clase.
- Se pierde información individual, pero la variable se vuelve más manejable.

Intervalos de Clases: Cantidad de intervalos

- Algunos consideran $\sqrt{n}$ como primera aproximación.
- Recomendación: no inferior a 5, no superior a 20.
- Puede definirse según el criterio del investigador.

Intervalos de Clase: Amplitud

Amplitud:

- Datos ordenados: $x_1 \leq x_2 \leq \cdots \leq x_n$.
- Rango: $R = x_n - x_1$.
- Cantidad de intervalos: $I \Rightarrow A = \frac{R}{I}$.
- Redondear al entero superior si es necesario.
- Primer intervalo inicia en x_1 y el último contiene a x_n.

Consideraciones:

- No debe haber intervalos con frecuencia 0.
- La agrupación no debe distorsionar la distribución original.

Intervalos de Clase: Marca de Clase

Sean L_{inf} y L_{sup} los límites del intervalo i-ésimo, con $i = 1, \ldots, k$:

- Intervalos deben ser **disjuntos**:
  - $L_{inf} \leq x < L_{sup}$ (incluye inferior).
  - $L_{inf} < x \leq L_{sup}$ (incluye superior).
- **Marca de clase:**

$$Mc_i = \frac{L_{inf} + L_{sup}}{2}$$

- Es el punto medio del intervalo i y es el *representante* de dicho intervalo.

Tablas de Frecuencia para Datos Agrupados

- 1^{er} columna: Intervalos de clase.
- 2^{da} columna: Marca de clase.

El resto es como en agrupación simple:

$$f_i = \text{frecuencia absoluta simple del intervalo } i$$

$$r_i = \frac{f_i}{n} \quad \text{frecuencia relativa simple}$$

$$p_i = 100 \times r_i \quad \text{frecuencia porcentual simple}$$

$$F_i = \sum_{j=1}^{i} f_j \quad \text{frecuencia absoluta acumulada}$$

$$R_i = \frac{F_i}{n} \quad \text{frecuencia relativa acumulada}$$

$$P_i = 100 \times R_i \quad \text{frecuencia porcentual acumulada}$$

EJEMPLO

A continuación se muestran las mediciones sobre la tasa de flujo (libras/hora) de una torre de destilación (A Self-Scalling Distillation Tower, Chem. Eng. Prog., 1968, pp. 79-84) 1170,

1350, 1640, 1800, 1800, 1260, 1440, 1730, 1710, 1350, 1440, 1710, 1530, 1800,1530, 1170, 1440, 1350, 1260, 1530, 1350, 1440, 1170, 1350, 1530, 1620, 1440, 1170,1440, 1800, 1260, 1170, 1260, 1710, 1710,1350, 1530, 1440, 1440, 1530, 1170, 1350, 1620, 1495, 1440, 1260, 1540, 1520, 1170, 1170, 1440.

Variable en estudio: tasa de flujo de una torre de destilación.
Tipo de variable: cuantitativa discreta.

En esta lista los datos (mediciones) aparecen según se fueron registrando, no están ordenados ni clasificados, tampoco **agrupados**.

Ordenando las mediciones

1170, 1170, 1170, 1170, 1170, 1170, 1170, 1170, 1170, 1260, 1260, 1260, 1260, 1350, 1350, 1350, 1350, 1350, 1350, 1350, 1440, 1440, 1440, 1440, 1440, 1440, 1440, 1440, 1495, 1530, 1530, 1530, 1530, 1530, 1540, 1620, 1620, 1640, 1710, 1710, 1710, 1710, 1730, 1800, 1800, 1800, 1800, 1800.

Si bien pareciera que de esta manera la lectura de los datos es más simple de interpretar, la presentación puede mejorarse...

Frecuencia absoluta y relativa - Agrupación Simple

Tasa	f_i	r_i	p_i	F_i	R_i	P_i
1170	9	0.19	19	9	0.19	19
1260	5	0.10	10	14	0.29	29
1350	7	0.14	14	21	0.43	43
1440	8	0.17	17	29	0.60	60
1495	1	0.02	2	30	0.62	62
1530	5	0.10	10	35	0.72	72
1540	1	0.02	2	36	0.74	74
1620	2	0.04	4	38	0.78	78
1640	1	0.02	2	39	0.80	80
1710	4	0.08	8	43	0.88	88
1730	1	0.02	2	44	0.90	90
1800	5	0.10	10	49	1.00	100

Medidas Descriptivas Numéricas

Objetivo: Caracterizar una distribución de frecuencias por medio de un número reducido de medidas numéricas, las cuales complementan la información aportada por tablas de frecuencias y gráficos.

Estas medidas están rigurosamente definidas y brindan en forma resumida información del conjunto total de datos y una idea del comportamiento global de la población o muestra en estudio.

Medidas Descriptivas Numéricas

- **Medidas de Tendencia Central:** Valores numéricos que se obtienen de variables cuantitativas y cuyos resultados se localizan por el centro de la distribución. Ej: Media (promedio aritmético), Mediana, Moda.
- **Medidas de Posición:** Valores numéricos que permiten dividir la distribución de datos en partes iguales. Ej: Cuartiles, Deciles, Percentiles.
- **Medidas de Dispersión:** Valores numéricos que proporcionan una idea sobre cuán esparcidos o concentrados están los datos correspondientes a una variable. Ej: Rango, Rango intercuartílico, varianza, desviación estándar, coeficiente de variación.
- **Medidas de Forma:** Dan una idea de la forma de la distribución de la variable. Ej: Coeficiente de asimetría, de kurtosis (apuntamiento).

Notación

Recordemos: n: tamaño de la muestra, X: variable.

- **Datos en agrupación simple:** $x_1, \ldots, x_k$, k valores distintos que asume la variable, y supongamos $x_1 < \cdots < x_k$. Sean $f_1, \ldots, f_k$ sus respectivas frecuencias absolutas.
- **Datos agrupados en intervalos:** Supongamos datos agrupados en intervalos $([L_{inf1}, L_{sup1}), \ldots, ([L_{infk}, L_{supk})$, cuyas marcas de clase son $M_{C1}, \ldots, M_{Ck}$ y f_i frecuencias absolutas del intervalo i-ésimo, $i = 1, \ldots, k$.

Medidas de Tendencia Central - Media

Datos sin agrupar

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

Datos en agrupación simple

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

Datos Agrupados en Intervalos

$$\bar{x} = \frac{\sum_{i=1}^{k} M_{Ci} f_i}{n}$$

Medidas de Tendencia Central - Media Aritmética

Ventajas:

- En su cálculo se emplea toda la información disponible.
- Se expresa en las mismas unidades que la variable en estudio.
- Es el centro de gravedad de toda la distribución, representando a todos los valores observados.
- Es un valor único.
- Se trata de un valor familiar para la mayoría de las personas.
- Es útil para llevar a cabo procedimientos estadísticos como la comparación de medias de varios conjuntos de datos.

Desventajas:

- Es muy sensible a los valores extremos de la variable.
- No es recomendable usar la media como medida central en las distribuciones muy asimétricas.

Medidas de Tendencia Central - Mediana

Es aquel valor de la variable que divide al conjunto de valores observados en dos partes de modo que el 50 % de los valores observados son menores o iguales que la mediana y el 50 % restante son mayores o iguales a ella. Ocupa el lugar central del conjunto de datos, ordenados en forma creciente (y repetidos tantas veces como indique su frecuencia absoluta simple), dejando a su izquierda y derecha la misma cantidad de observaciones.

DAS:

- n impar: Med = dato que ocupa la posición $\frac{n+1}{2}$, esto es, si los datos son $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, Med $= X_{(\frac{n+1}{2})}$.
- n par: $Med = \frac{X_{(n/2)}+X_{(n/2+1)}}{2}$

Medidas de Tendencia Central - Mediana

DAIC

- Calcular $n/2$.
- Buscar en la tabla de frecuencias absolutas acumuladas el intervalo de clase que contenga a la frecuencia $n/2$. Llamaremos a dicho intervalo "intervalo Mediana", y denotaremos por $L_{Med_{inf}}$ al límite inferior de ese intervalo, por f_{Med} a la frecuencia absoluta simple del mismo y por F_{Med-1} a la frecuencia absoluta acumulada de la clase inmediata anterior. Sea A_{Med} la amplitud del intervalo mediana.
- $Med = L_{Med_{inf}} + \frac{n/2 - F_{Med-1}}{f_{Med}} A_{Med}$

Medidas de Tendencia Central - Mediana

Ventajas:

- No es afectada por los valores extremos, ya que no depende de los valores que toma la variable, sino del orden de las mismas. Por ello su uso es adecuado en distribuciones asimétricas.
- Es de cálculo rápido e interpretación sencilla.
- La mediana de una variable discreta es siempre un valor de la variable que estudiamos.

Desventajas:

- En su cálculo no interviene toda la información disponible.
- Hay que ordenar los datos antes de determinarla.
- Para poder calcularla, el nivel de medición debe ser al menos jerárquica.

Medidas de Tendencia Central - Moda

Datos en agrupación simple: Valor que aparece más frecuentemente que cualquier otro. Puede haber más de una moda (distribución bimodal, trimodal, multimodal).

Datos agrupados en intervalos:

1. Determinar la clase modal (la de mayor frecuencia absoluta).
2. Determinar valor de la moda dentro de la clase modal:

$$Mo = L_{Mod_{inf}} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) A_{Mod}$$

donde $L_{Mod_{inf}}$ es el límite inferior de la clase modal, $\Delta_1 = f_{Mod} - f_{preMod}$ y $\Delta_2 = f_{Mod} - f_{postMod}$. f_{Mod}, f_{preMod} y $f_{postMod}$ son las frecuencias absolutas simples de las clases Modal, pre Modal y post Modal, respectivamente. A_{Mod} es la amplitud del intervalo modal.

Medidas de Tendencia Central - Modo o Moda

Datos en agrupación simple: Ejemplo:

$$Mod = \text{Dato con mayor frecuencia} = 1170$$

Por lo que, la tasa más frecuente fue de 1170 libras/hora.

Datos agrupados en intervalos: Hay 2!

- Intervalo modal: [1100;1200);
- $L_{Mod_{inf}} = 1100$;
- $\Delta_1 = 9 - 0 = 9$; $\Delta_2 = 9 - 5 = 4$; $A_{Mod} = 100$;
- $Mod = 1100 + \frac{9}{13} \cdot 100 = 1169{,}23$

Medidas de Tendencia Central - Modo o Moda

Ventajas:

- No requiere cálculos, excepto en DAIC.
- Puede usarse para datos tanto cuantitativos como cualitativos.
- Fácil de interpretar.
- No se ve influenciado por valores extremos.

Medidas de Tendencia Central - Modo o Moda

Desventajas:

- Para conjuntos pequeños de datos su valor no tiene casi utilidad, si es que de hecho existe.
- No utiliza toda la información disponible.
- No siempre existe, si los datos no se repiten.
- Difícil de interpretar si la distribución de datos posee más de dos modas.
- Puede no ser una buena medida de tendencia central cuando se encuentra al comienzo o al final del campo de variabilidad de los valores de la variable en estudio.
- Está muy afectado por la manera en que se construyen los intervalos de clase; por lo que no es una medida estable.

Medidas de Posición

Cuantiles: Son ciertos valores del conjunto de observaciones que permiten dividirlo en partes iguales. Los cuantiles más usados son: los **Cuartiles (Q)**, los **Deciles (D)** y los **Percentiles (P)**.

- **Cuartiles (Q):** dividen el conjunto de observaciones en cuatro (4) partes iguales, cada una de las cuales contiene un cuarto (25 %) de la información. Se denotan Q_1, Q_2, Q_3, Q_4.
- **Deciles (D):** dividen el conjunto de observaciones en diez (10) partes iguales, son 10 Deciles denotados como: $D_1, \ldots, D_{10}$.
- **Percentiles (P):** dividen el conjunto de observaciones en cien (100) partes iguales cada una de las cuales contiene un 1 % de las observaciones, denotados por $P_1, P_2, \ldots, P_k, \ldots, P_{100}$, donde k denota el porcentaje de observaciones que quedan a la izquierda del percentil P_k.

Cálculo de Cuantiles

Datos en agrupación simple:

- Se ordenan los datos de menor a mayor.
- Se calcula la posiciÃ³*ndelcuantil*$\mathrm{l}_k = \frac{k \cdot n}{NoC}$, luego se busca el valor correspondiente del cuantil con la ayuda de la tabla de frecuencias acumuladas.
- $C_k :=$ Cuantil buscado; $1 \leq k \leq NoC$, $n =$ total de observaciones, $NoC = 4$ Para Cuartiles; 10 para Deciles y 100 para Percentiles.

Datos agrupados en intervalos:

- Se identifica el intervalo que contenga el cuantil buscado con la fórmula $I_{Ck} = \frac{k \cdot n}{NoC}$.
- Se calcula:

$$C_k = L_{inf} + \frac{\frac{k \cdot n}{NoC} - \sum f_{ant}}{f_{ICk}} \cdot A_{ICk}$$

Medidas de Dispersión o Variabilidad

Una vez que se han recogido los valores que toman las variables de nuestro estudio (datos), procedemos al análisis descriptivo de los mismos.
Para variables numéricas, en las que puede haber un gran número de valores observados distintos, se ha de optar por un método de análisis que responda:

1. ¿Alrededor de qué valor se agrupan los datos?
2. Supuesto que se agrupan alrededor de un número, ¿cómo lo hacen?¿muy concentrados? ¿muy dispersos?

Medidas de Dispersión o Variabilidad

Absolutas:

- Rango o Amplitud Máxima.
- Rango intercuartílico.
- Varianza.
- Desviación típica o Estándar.

Relativas:

- Coeficiente de variación.

Medidas de Dispersión o Variabilidad - Rango

Tambien llamado ancho o recorrido, es la diferencia entre el máximo y el mínimo valor del conjunto de datos:

$$R = x_{(n)} - x_{(1)}$$

Ventajas:

- Es fácil de calcular y es comúnmente usado como una medida burda, pero eficaz de variabilidad.
- Es comprensible para cualquier persona, aún cuando no conozca de Estadística.

Desventajas:

- No está basado en ninguna MTC.
- Está afectado por los valores OUTLIERS.
- En su cálculo sólo intervienen dos valores, por lo que se desaprovecha mucha información.

Medidas de Dispersión o Variabilidad - Desviación Intercuartílica

Indica la variación máxima que sufre el 50% central de los valores de la variable. Este desvío deja mucho a cada lado (el 25% de la información).
La mediana parte a la distribución en dos partes iguales, pero a veces es más significativo el 50% entre Q_3 y Q_1; porque es un 50% más puro, más homogéneo.

$$RIC = Q_3 - Q_1$$

El intervalo $\left[MTC - \frac{RIC}{2}\,;\, MTC + \frac{RIC}{2}\right]$ concentra, aproximadamente, el 50% de los datos centrales.
Problema: Sigue sin estar basado en una MTC.

Medidas de Dispersión o Variabilidad - Desvío Medio

Si calculamos los desvíos respecto de la media aritmética, por la propiedad vista anteriormente, resulta:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = 0$$

Para evitar este inconveniente se define el **Desvío Medio** como:

$$\text{D.S.A.: } DM = \frac{\sum_{i=1}^{n} |x_i - \bar{X}|}{n}$$

$$\text{D.A.S.: } DM = \frac{\sum_{i=1}^{k} |x_i - \bar{X}| f_i}{n}$$

$$\text{D.A.I.C.: } DM = \frac{\sum_{i=1}^{k} |M_{ci} - \bar{X}| f_i}{n}$$

Muchas veces, si el promedio aritmético no es una MTC confiable, se la suele reemplazar por la MEDIANA, definiendo así el **Desvío Mediana**.

Problema: El valor absoluto es muy complicado de trabajar algebraicamente, por lo que resulta poco práctico.

Medidas de Dispersión o Variabilidad - Varianza

Es el promedio de los cuadrados de las desviaciones de los valores muestrales respecto de la media aritmética $\bar{X}$. Se representa por S^2.

$$\text{D.S.A.: } S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}$$

$$\text{D.A.S.: } S^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{X})^2 f_i}{n}$$

$$\text{D.A.I.C.: } S^2 = \frac{\sum_{i=1}^{k}(M_{ci} - \bar{X})^2 f_i}{n}$$

Fórmula de Trabajo:

$$S^2 = \bar{X^2} - (\bar{X})^2$$

En la práctica, el denominador que se utiliza es $n-1$ en lugar de n. La medida que se utiliza es:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

Ventajas:

- En su cálculo intervienen todos los datos observados.
- Es una medida de variabilidad promedio respecto de una MTC.

Desventajas:

- Se pierde la unidad de medida original.

Propiedades:

- La varianza es un número real no negativo.
- La Varianza de una constante es nula.
- La Varianza de la suma de una variable más (o menos) una constante, es igual a la Varianza de la variable.
- La Varianza del producto (o cociente) de una variable por (o dividido) una constante no nula, es igual a la Varianza de la variable por (o dividido) la constante al cuadrado.

Medidas de Dispersión o Variabilidad - Desvío Estándar

Es la raíz cuadrada de la Varianza, y se representa por S. Expresa la dispersión de la distribución y se expresa en las mismas unidades de medida de la variable.
La Desviación Estándar es la medida de dispersión más utilizada en Estadística.

$$\text{D.S.A.: } S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}$$

$$\text{D.A.S.: } S = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{X})^2 f_i}{n-1}}$$

$$\text{D.A.I.C.: } S = \sqrt{\frac{\sum_{i=1}^{k}(M_{ci} - \bar{X})^2 f_i}{n-1}}$$

Medidas de Dispersión o Variabilidad - Desvío Estándar

Ventajas:

- En su cálculo intervienen todos los datos observados.
- Es una medida de variabilidad promedio respecto de una MTC.

Desventajas:

- Al estar basada en la media aritmética, que está fuertemente afectada por los valores OUTLIERS, ésta también se encuentra afectada por ellos.

Medidas de Dispersión o Variabilidad- Varianza - Desvío

Como medidas de variabilidad más importantes, conviene destacar algunas características de la Varianza y el Desvío Estándar.

- Son índices que describen la variabilidad o dispersión y por tanto cuando los datos están muy alejados de la media, el numerador de sus fórmulas será grande y la Varianza y la Desviación Estándar también lo serán.
- Al aumentar el tamaño de la muestra, disminuye la Varianza y el Desvío Estándar.
- Cuando todos los datos de la distribución son iguales, la Varianza y el Desvío Estándar son iguales a cero.
- Para su cálculo se utilizan todos los datos de la distribución; por tanto, cualquier cambio de valor será detectado.

Medidas de Dispersión o Variabilidad - Coeficiente de Variación

El Coeficiente de Variación es una medida de dispersión relativa que se expresa generalmente en porcentajes. Las medidas de dispersión que vimos anteriormente son absolutas; son útiles para describir la dispersión de un solo conjunto de datos. Si dos conjuntos van a ser comparados, los valores absolutos son convenientes para éste fin, únicamente si los promedios de dichos conjuntos son más o menos iguales y si se refieren a un mismo fenómeno. Por ejemplo, no tiene sentido comparar cuál entre dos compañías A y B presenta mayor dispersión en los salarios, si la primera paga en dólares y la segunda paga en pesos argentinos.

Medidas de Dispersión o Variabilidad - Coeficiente de Variación

Para estos casos, es necesario disponer de una medida que nos permita comparar qué tan pequeña o qué tan grande es una medida de dispersión absoluta como la Desviación Estándar.
El Coeficiente de Variación, simbolizado con CV, es una medida de dispersión relativa que resulta de comparar S con la $\bar{X}$ del conjunto, así:

$$CV = \frac{S}{\bar{X}} \cdot 100$$

Distribución Normal - Campana de Gauss

Es la distribución teórica más conocida y utilizada en Estadística. Fue creada por el matemático Gauss con el objeto de generalizar muchas distribuciones referidas a ciertos fenómenos de la naturaleza (por ejemplo: estatura y peso, por sexo) que presentaban características similares.

Características generales de una distribución Normal:

- Relaciona la Media con la Desviación Estándar, que son sus parámetros: μ y σ.
- Tiene forma de campana. Es una curva simétrica: tiene un pico máximo en el centro y decrece constantemente hacia los extremos.
- No corta el eje de abscisas.
- La Media Aritmética coincide con el Modo y la Mediana.

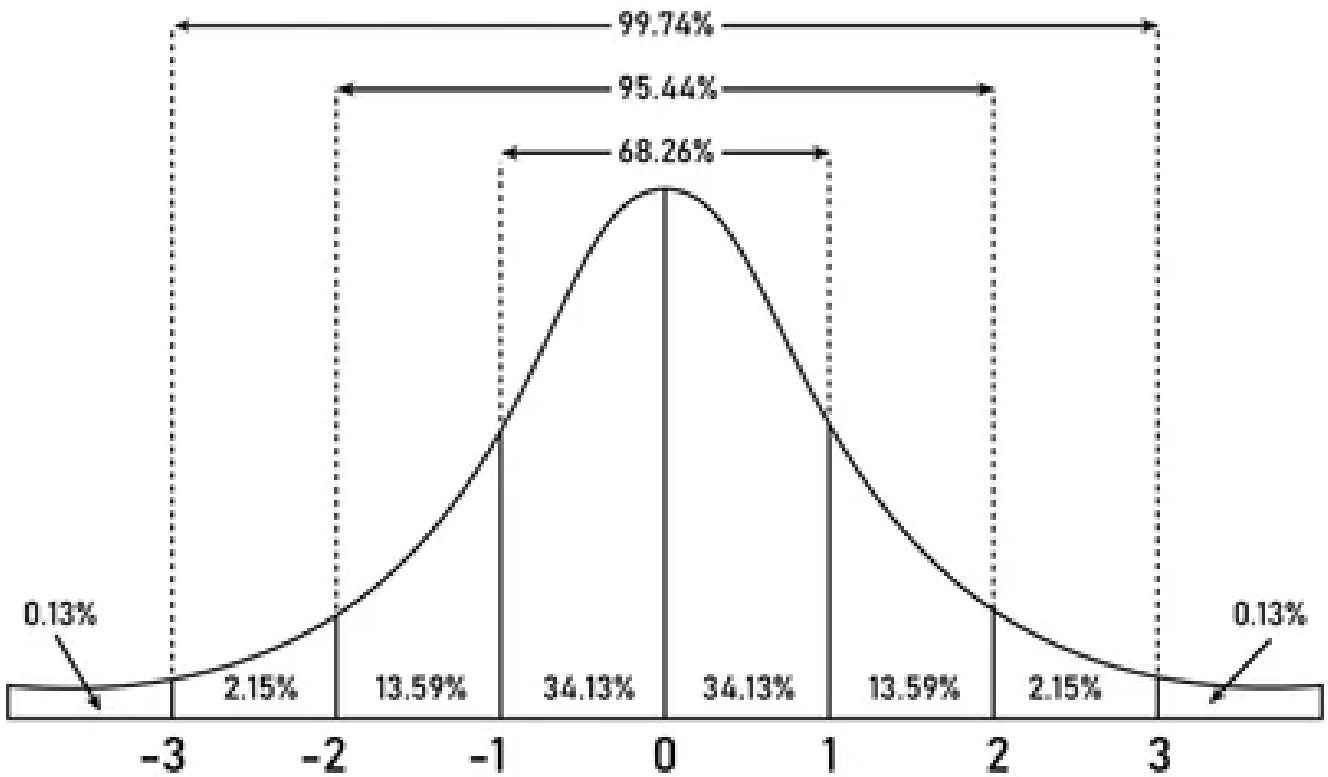
Es una distribución que se utiliza para describir otras características de una distribución en particular comparándola con ella (por ejemplo, asimetría y kurtosis). También para determinar valores de datos atípicos.

Distribución Normal - Campana de Gauss

Para distribuciones de datos que se aproximan a la distribución normal podemos también obtener fracciones de datos que caen dentro de ciertos límites. La más usada es la regla (68 - 95 - 99).

- Aproximadamente, el 68,27% de los casos están entre $\bar{X} - S$ y $\bar{X} + S$.
- Aproximadamente, el 95,45% de los casos están entre $\bar{X} - 2S$ y $\bar{X} + 2S$.
- Aproximadamente, el 99,73% de los casos están entre $\bar{X} - 3S$ y $\bar{X} + 3S$.

Distribución Normal - Campana de Gauss



Medidas de Forma - Coeficiente de Asimetría

Grado de simetría de una distribución respecto a su media. Una distribución puede ser:

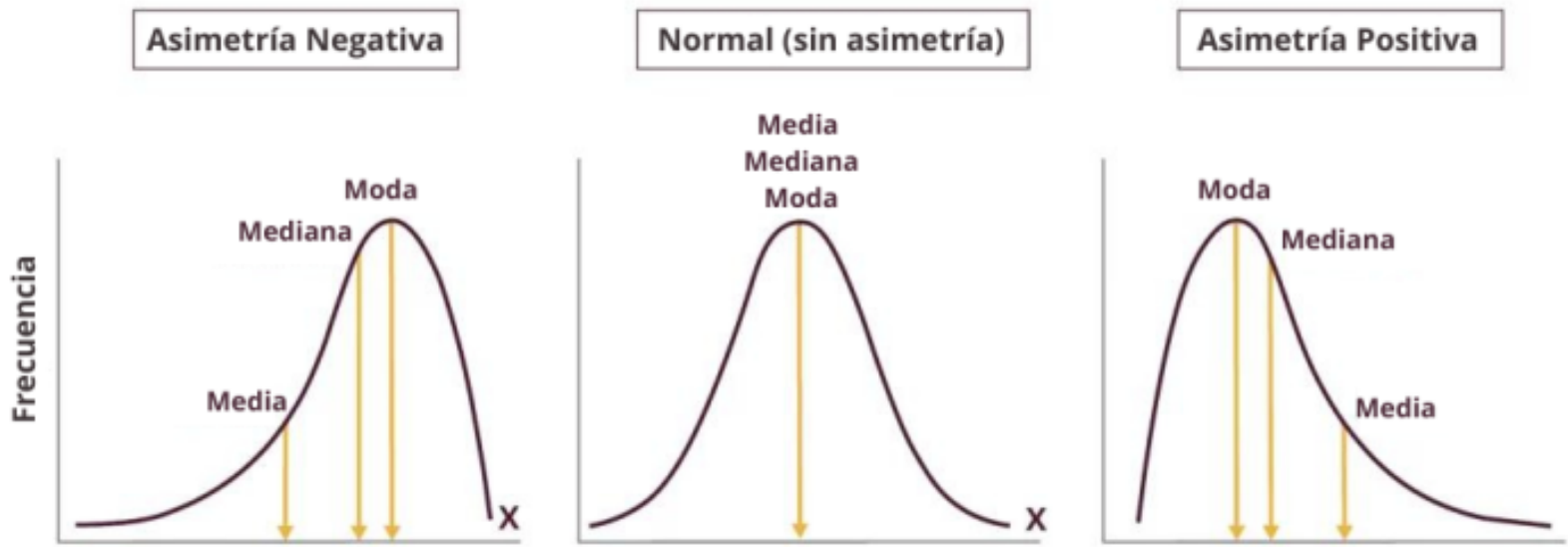
- **Simétrica:** los valores equidistantes de una posición central tienen la misma frecuencia.
- **Asimétrica positiva:** las frecuencias más altas corresponden a valores que se encuentran al lado izquierdo de esa posición central (cola a la derecha).
- **Asimétrica negativa:** distribuciones con cola a la izquierda.

Se define el coeficiente de asimetría de Pearson como:

$$a_s = \frac{3\,(\bar{X} - Med)}{S}$$

- $a_s = 0 \Rightarrow$ distribución simétrica $(\bar{X} = Med = Mo)$
- $a_s > 0 \Rightarrow$ distribución asimétrica positiva $(\bar{X} > Med > Mo)$
- $a_s < 0 \Rightarrow$ distribución asimétrica negativa $(\bar{X} < Med < Mo)$

Medidas de Forma - Coeficiente de Asimetría



Medidas de Forma - Kurtosis

Se aplica a distribuciones unimodales simétricas o ligeramente asimétricas, ya que representa la elevación o achatamiento de una distribución comparada con la distribución normal.

$$\kappa = \frac{m_4}{S^4} - 3$$

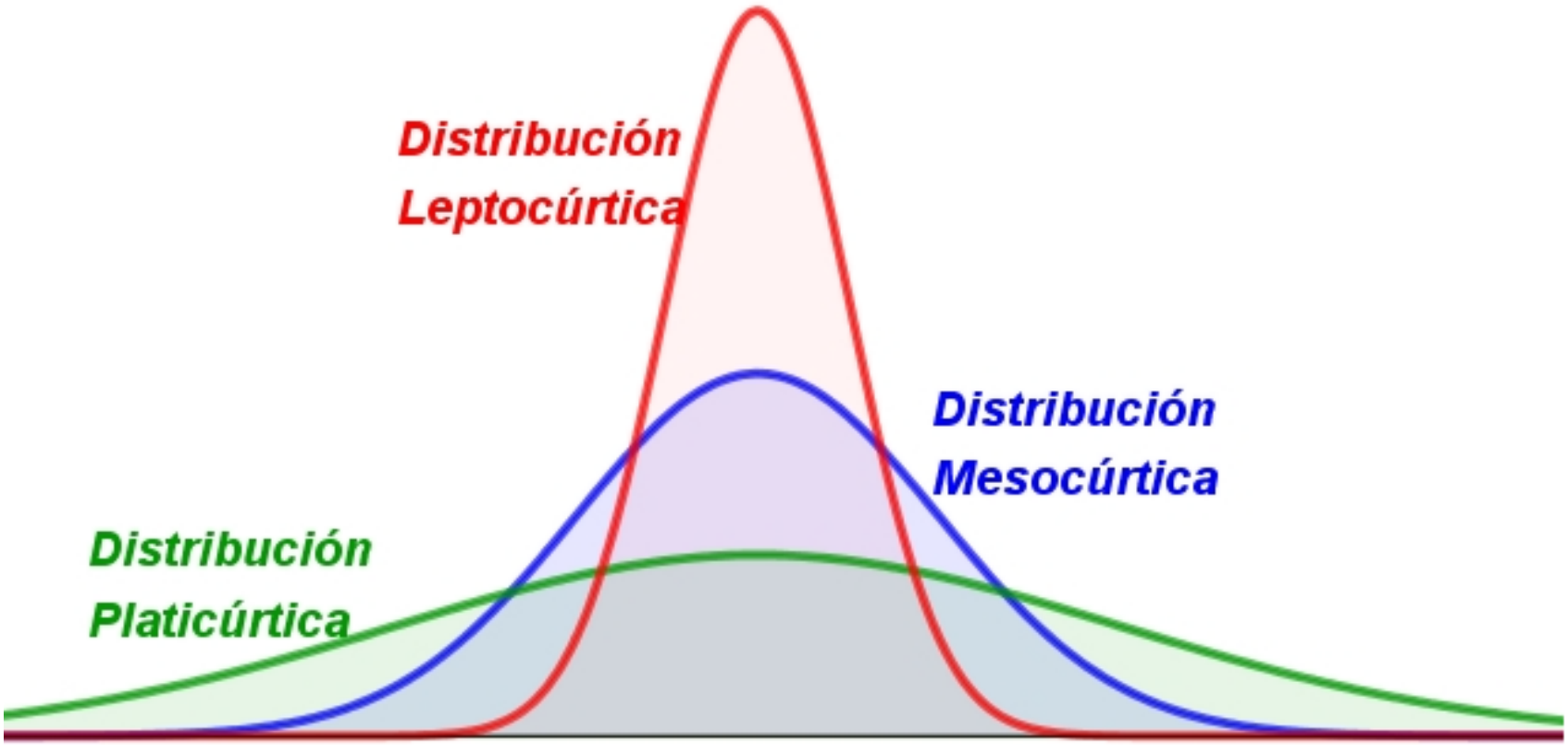
siendo

$$m_4 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^4}{n}$$

el momento central de orden 4.

- $\kappa = 0 \Rightarrow$ mismo grado de elevación que distribución normal (Mesocúrtica).
- $\kappa > 0 \Rightarrow$ más apuntamiento que distribución normal (Leptocúrtica).
- $\kappa < 0 \Rightarrow$ menor grado de elevación que distribución normal (Platicúrtica).

Medidas de Forma - Coeficiente de Kurtosis



Boxplot - Diagrama de Cajas y Bigotes

- Gráfico en forma de rectángulo (caja) construido en base a solamente cinco números que resumen los datos.
- La altura del rectángulo es el rango intercuartílico $Q_3 - Q_1$. La base inferior y superior del rectángulo son Q_1 y Q_3, el rectángulo se divide con una línea a la altura de la mediana (Q_2).
- Se calcula $1.5\times$ Rango intercuartílico, se dibuja una línea vertical desde la mitad de la parte superior (inferior) del rectángulo hasta la mayor (menor) observación que se encuentre entre ese extremo de la caja y $1.5\times$ Rango intercuartílico.
- Las observaciones que caen fuera de esos "bigotes" se representan con círculos rellenos si están a una distancia mayor a $3\times$ Rango intercuartílico, o por círculos sin rellenar en caso contrario.

Boxplot: Ejemplo, datos sin agrupar