Unidad 9:
ESTADÍSTICA DESCRIPTIVA.

Probabilidad y Estadística.
LSI FaCENA

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** respresentativa de la misma

Instroducción

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

Introducción

- Objetivo Básico: Describir lo más simplemente posible los resultados obtenidos en un experimento/encuesta/censo, etc.
- Organización y Presentación de resultados:
  - Representaciones tabulares (tablas): 1^{er} 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 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

- **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 susceptibles de tomar distintos estados entre unidades elementales, o que varían dentro de una misma unidad elemental a través del tiempo.

Las variables pueden clasificarse como cualitativas ó cuantitativas

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: género al nacer) ó **POLICOTÓMICAS**:más de dos categorías (ej: nivel de estudios alcanzado)

Cuantitativas

Asumen valores numéricos. Expresan una cantidad.

- **Discretas**: Surgen de contar. Son aquellas que sólo toman valores discretos dentro de su campo de variación (ej: cantidad de materias aprobadas)
- **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. Esto nos da una nueva información: ¿qué categoría tiene más elementos? ¿Qué frecuencia tiene cada categoría?
- Esta escala constituye el nivel de medición más bajo.

Escala de Medición

Variables Cualitativas

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).

Ejemplo: Nivel educacional.
No asistieron – Primaria Incompleta – Primaria Completa – Secundaria Incompleta – Secundaria Completa – Superior o Universitario Incompleto – Superior o Universitario Completo.

Las categorías pueden designarse con números, pero éstos deben guardar la misma relación de orden que las categorías a las que se asignaron.

No asistieron (0) – Primaria Incompleta (1) – Primaria Completa (2) – Secundaria Incompleta (3) – Secundaria Completa (4) – Superior o Universitario Incompleto (5) – Superior o Universitario Completo (6).

Escala de Intervalos

- Se utiliza cuando los datos son cuantitativos, por lo tanto los valores de cada categoría necesariamente deben ser números.
- 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.

Escala de Intervalos

Ejemplo: Temperatura, altura sobre el nivel del mar, etc.
Temperatura: **ºC** 0 – 10 – 20 –30 (0 es la temperatura de congelación del agua, 100 la de ebullición, el intervalo se divide en 100 partes y así se obtienen los grados centígrados).

ºF 32 – 50 – 68 –86
(0**ºC** =32**ºF**, 0**ºF** =-17,78**ºC**)
Notar que, por ejemplo, 32**ºC** NO indica el doble de calor que 16**ºC**
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 $aX + b; a > 0; b > 0$. En nuestro ejemplo, $^oF = \frac{9}{5}{}^oC + 32$.

Escala de Razón o Proporción

- Se utiliza para variables cuantitativas
- 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. (ej: una persona que pesa 50kg pesa el doble que una que pesa 25kg)
- Esta escala constituye el nivel de medición más alto.

Escalas de Medición

Escala de Razón o Proporción

Ejemplo: Peso, altura, distancia, etc.
Peso: **Kg** 0 – 1 – 2 – 3
Libra 0 – 2,205 – 4,41 – 6,615

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 $aX; a > 0$. En nuestro ejemplo, $Lb \approx 2Kg$.

EJEMPLO 1

Se desea estudiar el perfil de los clientes que compraron en cierto supermercado mayorista de la ciudad de Corrientes, durante el año 2022. Se decide elegir aleatoriamente a 300 clientes, a quienes se les consultó acerca de las siguientes características:

1. Barrio de residencia.
2. Estado Civil.
3. Cantidad de Hijos.
4. Peso.
5. Estatura.
6. Edad.
7. Grado de satisfacción con el supermercado.
8. Ocupación.
9. Ingreso (en pesos, sin centavos).

Se pide: Identificar en qué consiste el experimento, Unidad Elemental, Población, Muestra, Variables en estudio y Tipo de cada una de las Variables.

EJEMPLO 2

La siguiente tabla muestra el número de hermanos en edad escolar de una muestra de 3.000 estudiantes de la FaCENA, en el año 2002.

Nro. de Hermanos	Cantidad
0	740
1	1097
2	658
3	345
4	126
5	34

- Población: Estudiantes de la FaCENA, en el año 2002.
- Unidad elemental: Estudiante de la FaCENA, en el año 2002.
- Variable en Estudio: Nro. de hermanos.
- Tipo de Variable: Variable Cuantitativa Discreta.
- Escala de Medición:De razón o proporción.

EJEMPLO 2

La siguiente tabla muestra el número de hermanos en edad escolar de una muestra de 3.000 estudiantes de la FaCENA, en el año 2002.

Nro. de Hermanos	Cantidad
0	740
1	1097
2	658
3	345
4	126
5	34

- Población: Estudiantes de la FaCENA, en el año 2002.
- Unidad elemental: Estudiante de la FaCENA, en el año 2002.
- Variable en Estudio: Nro. de hermanos.
- Tipo de Variable: Variable Cuantitativa Discreta.
- Escala de Medición:De razón o proporción.

EJEMPLO 3

En una fábrica de bombillas eléctricas de la Ciudad de Rosario, en el año 2014, se observaron 200 de ellas para estudiar su duración y se obtuvieron los siguientes resultados:

Horas de Duración	f_i
(100–200]	2
(200–300]	7
(300–400]	16
(400–500]	49
(500–600]	62
(600–700]	41
(700–800]	23

- Población: Bombillas eléctricas producida por la fábrica de la ciudad de Rosario en estudio, durante el año 2014.
- Unidad elemental: Cada bombilla eléctricas producida por la fábrica de la ciudad de Rosario en estudio, durante el año 2014.
- Muestra: Las 200 bombillas eléctricas estudiadas.
- Variable en Estudio: Duración de la bombilla ó, Vida Útil.
- Tipo de Variable: Variable Cuantitativa Continua.
- Escala de Medición: De razón o proporción.

EJEMPLO 3

En una fábrica de bombillas eléctricas de la Ciudad de Rosario, en el año 2014, se observaron 200 de ellas para estudiar su duración y se obtuvieron los siguientes resultados:

Horas de Duración	f_i
(100–200]	2
(200–300]	7
(300–400]	16
(400–500]	49
(500–600]	62
(600–700]	41
(700–800]	23

- Población: Bombillas eléctricas producida por la fábrica de la ciudad de Rosario en estudio, durante el año 2014.
- Unidad elemental: Cada bombilla eléctricas producida por la fábrica de la ciudad de Rosario en estudio, durante el año 2014.
- Muestra: Las 200 bombillas eléctricas estudiadas.
- Variable en Estudio: Duración de la bombilla ó, Vida Útil.
- Tipo de Variable: Variable Cuantitativa Continua.
- Escala de Medición: De razón o proporción.

Datos en Agrupación Simple

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 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.

EJEMPLO: Ej. 2

Nro. de Hermanos	**Cantidad**	r_i	p_i	F_i	R_i	P_i
0	740	0,247	24,7	740	0,247	24,7
1	1097	0,366	36,7	1837	0,613	61,3
2	658	0,219	21,9	2495	0,832	83,2
3	345					
4	126					
5	34			3000	1	100
Total	3000	1	100			

EJEMPLO: Ej. 2

Nro. de Hermanos	**Cantidad**	r_i	p_i	F_i	R_i	P_i
0	740	0,247	24,7	740	0,247	24,7
1	1097	0,366	36,7	1837	0,613	61,3
2	658	0,219	21,9	2495	0,832	83,2
3	345	0,115	11,5	2840	0,947	94,7
4	126	0,042	4,2	2966	0,989	98,9
5	34	0,011	1,1	3000	1	100
Total	3000	1	100			

La tabla nos dice, por ejemplo, que:

- 1097 estudiantes tienen 1 sólo hermano.
- Aproximadamente el 4 % de los estudiantes tienen 4 hermanos.
- 2840 estudiantes tienen 3 hermanos o menos.
- Aproximadamente el 5 % de los estudiantes tienen más de 3 hermanos.

EJEMPLO: Ej. 4

Ppal Adicción	**Nro. de Per.**	r_i	p_i
Drogas Ilegales	54		
Alcohol	84		
Tabaco	162		
Total	300		

Ppal Adicción	**Nro. de Per.**	r_i	p_i
Drogas Ilegales	54	0,18	18
Alcohol	84	0,28	28
Tabaco	162	0,54	54
Total	300	1	100

EJEMPLO: Ej. 4

Ppal Adicción	**Nro. de Per.**	r_i	p_i
Drogas Ilegales	54		
Alcohol	84		
Tabaco	162		
Total	300		

Ppal Adicción	**Nro. de Per.**	r_i	p_i
Drogas Ilegales	54	0,18	18
Alcohol	84	0,28	28
Tabaco	162	0,54	54
Total	300	1	100

EJEMPLO 5

A continuación se muestran las mediciones sobre la tasa de flujo (libras/hora) de una torre de destilación (“A Self-Scalilng 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, 1170, 1620, 1800, 1170,1440, 1800, 1260, 1170, 1260, 1710, 1710,1350, 1530, 1440, 1530, 1170, 1350, 1620, 1495, 1440, 1260, 1540, 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**.

Introducción

Ordenando las mediciones ...

1170, 1170, 1170, 1170, 1170, 1170, 1170, 1170, 1170, 1260,
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 ....

EJEMPLO 5 - Agrupación simple

Tasa	f_i	f_{r_i}	p_i	F_i	F_{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

La tabla nos dice, por ejemplo, que:

- 7 observaciones dieron una tasa de 1350 libras/hora
- Aproximadamente un 14 % de las observaciones midieron 1350 libras/hora
- 21 observaciones midieron una tasa de 1350 libras/hora o menos.
- Un 12 % de las observaciones midieron más de 1710 libras/hora.

Datos Agrupados en Intervalos de Clase

X variable cuantitativa continua, o cuantitativa discreta pero:

- Demasiados datos;
- Pocos datos, pero muy dispersos (muy diferentes entre sí).
- Interesa una clasificación particular de los resultados.

Se agrupan los valores en intervalos de clase:

- Intervalos deben ser disjuntos: cada observación debe estar contenida en un, y sólo un intervalo de clase.
- Se pierde información sobre las observaciones, pero la variable es más “manejable”.

Intervalos de Clases

¿Cuántos intervalos?

- Algunos consideran $\sqrt{n}$ como primera aproximación.
- Recomendación: no inferior a 5, no superior a 20.
- Puede definirse previamente de acuerdo al criterio de los investigadores.

Intervalos de Clase

Amplitud

- Observaciones ordenadas: $x_1 \leq x_2 \leq \ldots \leq x_n$. Rango de la información: $R = x_n - x_1$.
- I: cantidad de intervalos. $\rightarrow$ Amplitud de los intervalos: $A = R/I$ (redondeando, si es necesario, al entero inmediato superior).
- El primer intervalo de clase debe contener el valor mínimo (x_1) y el último al valor máximo (x_n).

Tener en cuenta:

- NO pueden existir intervalos con frecuencia cero (0). En tal caso, definir intervalos de distinta amplitud.
- La agrupación en intervalos no debe modificar la forma de la distribución original de los datos.

Intervalos de Clase

Sean $L_{i\,inf}$ y $L_{i\,sup}$ los límites inferior y superior del intervalo $i-$ésimo, $i = 1; \ldots; k, k$: número de intervalos.

- Los intervalos, que deben ser disjuntos, pueden ser del tipo:
  - $L_{i\,inf} \leq x < L_{i\,sup}$ (incluyen al límite inferior),
  - $L_{i\,inf} < x \leq L_{i\,sup}$ (incluyen al límite superior).
- Se define la Marca de clase del intervalo i a

$$M_{ci} = \frac{L_{i\,inf} + L_{i\,sup}}{2}$$

- M_{ci} es el punto medio del intervalo i y es el "representante"de dicho intervalo.

Tablas de Frecuencias para datos agrupados en intervalos de clase

- 1^{er} columna: Intervalos de clase.
- 2^{da} columna: Marca de clase.
- El resto de las columnas se construye de igual manera que la tabla de frecuencias para datos en agrupación simple:
  - f_i : frecuencia absoluta simple del intervalo i , cantidad de observaciones que caen en dicho intervalo.
  - $r_i = \frac{f_i}{n}$: frecuencia relativa simple del intervalo i.
  - $p_i = 100 \times r_i$: frecuencia porcentual simple del intervalo i.
  - $F_i = \sum_{j=1}^{i} f_j$: frecuencia absoluta acumulada, suma de las frecuencias absolutas anteriores hasta el intervalo i.
  - $R_i = \frac{F_i}{n}$: frecuencia relativa acumulada.
  - $P_i = 100 \times R_i$: frecuencia porcentual acumulada.

Ejemplo: Ej 5

Ejemplo 5, agrupando los datos en intervalos de amplitud 100, comenzando en 1100

Tasa	M_{C_i}	f_i	f_{r_i}	p_i	F_i	F_{r_i}	P_i
[1100,1200)	1150	9	0.19	19	9	0.19	19
[1200,1300)	1250	5	0.10	10	14	0.29	29
[1300,1400)	1350	7	0.14	14	21	0.43	43
[1400,1500)	1450	9	0.19	19	30	0.62	62
[1500,1600)	1550	6	0.12	12	36	0.74	74
[1600,1700)	1650	3	0.06	6	39	0.80	80
[1700,1800)	1750	5	0.10	10	44	0.90	90
[1800,1900)	1850	5	0.10	10	49	1.00	100

La tabla nos dice, por ejemplo, que:

- 7 observaciones midieron entre 1300 y 1400 libras/hora
- 36 observaciones midieron una tasa menor a 1600 libras/hora
- Un 20 % de las observaciones midió 1700 libras/hora o más.

Representación Gráfica

Descripción

- Es un complemento importante de la presentación tabular.
- En las gráficas, los datos estadísticos se presentan en términos de magnitudes interpretadas visualmente.
- Los hechos y las relaciones esenciales que son difíciles de reconocer en masas de datos estadísticos, se observan con mayor claridad en la gráfica.

Ventajas

- Son más eficaces para llamar la atención que cualquier otro sistema.
- Una gráfica sencilla, atractiva y bien trazada, que represente un número limitado de datos, es más fácil de comprender que un cuadro.

Representación Gráfica

Desventajas

- No pueden representar tantos grupos de datos como un cuadro. Sólo presenta a la vez una cantidad limitada de información.
- Sólo se pueden presentar valores aproximados.
- Son útiles para dar una rápida idea de la situación general, pero no de los detalles.

Métodos Gráficos

Permiten obtener en forma rápida una primera idea del comportamiento de los datos.
Debe tenerse en cuenta:

- El gráfico más efectivo es el que alcance su objetivo de la manera más simple posible.
- Una “mirada”debería bastar para tener una idea de como están distribuidos los datos.
- Deben indicarse: título, origen, escala, variable que deben responder a las preguntas: ¿qué? (la variable), ¿cómo? (cifras absolutas, %, unidad de medida), ¿cuándo? (tiempo: año, mes, etc.), y ¿dónde?(lugar).
- Si los datos provienen de una base de datos debe indicarse la fuente: autor, título, volumen, etc.

Métodos Gráficos

Según el tipo de la variable en estudio se utilizará un determinado gráfico, los más comunes son:

- Variables Cualitativas Nominales:
  - Diagrama de barras.
  - Diagrama de sectores.
- Datos en Agrupación Simple (variables cuantitativas discretas, cualitativas ordinales):
  - Diagramas de barras.
  - Polígonos de frecuencias.
- Datos Agrupados en intervalos de clase:
  - Histogramas.
  - Polígonos de frecuencias

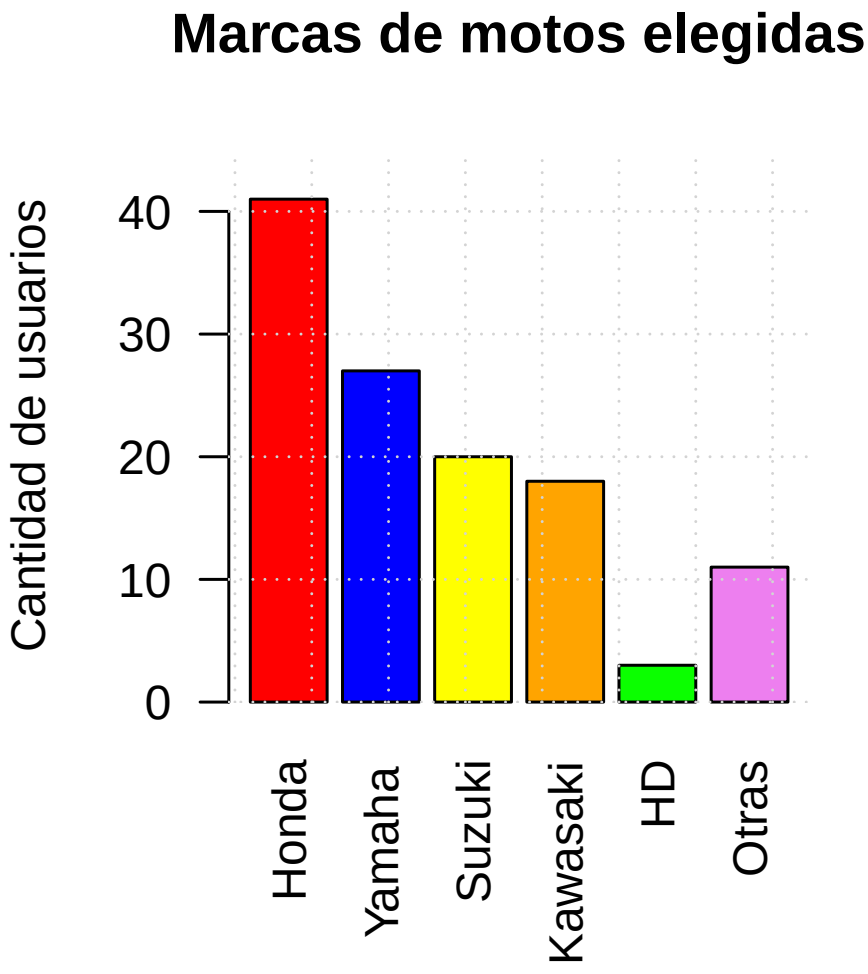
Diagramas o Gráficos de Barras

Las “barras”pueden ser horizontales o verticales. Caso vertical:

- Eje de las abscisas (x): categorías, (o valores de las variables caso cuantitativo discreta). Sobre ellos se levantan barras de igual base que no se solapen, (o líneas en el caso de la variable cuantitativa discreta).
- Eje de las ordenadas (y): altura de las barras o líneas proporcional a la frecuencia simple que representan.

Se le preguntó la marca de sus motos a 120 individuos poseedores de motos y se confeccionaron la siguiente tabla y diagrama de barras:

Marca	f_i	f_{r_i}
Honda	41	0.33
Yamaha	27	0.23
Suzuki	20	0.17
Kawasaki	18	0.15
Harley-Davidson	3	0.03
Otra	11	0.09



Ejemplo

Supongamos ahora que además tenemos información sobre la edad de los propietarios de las motos:

Marca	$\geq$ **50**	$<$ **50**	**Total**
Honda	11	30	41
Yamaha	20	7	27
Suzuki	15	5	20
Kawasaki	9	9	18
Harley-Davidson	2	1	3
Otra	7	4	11

Gráficos de Barras

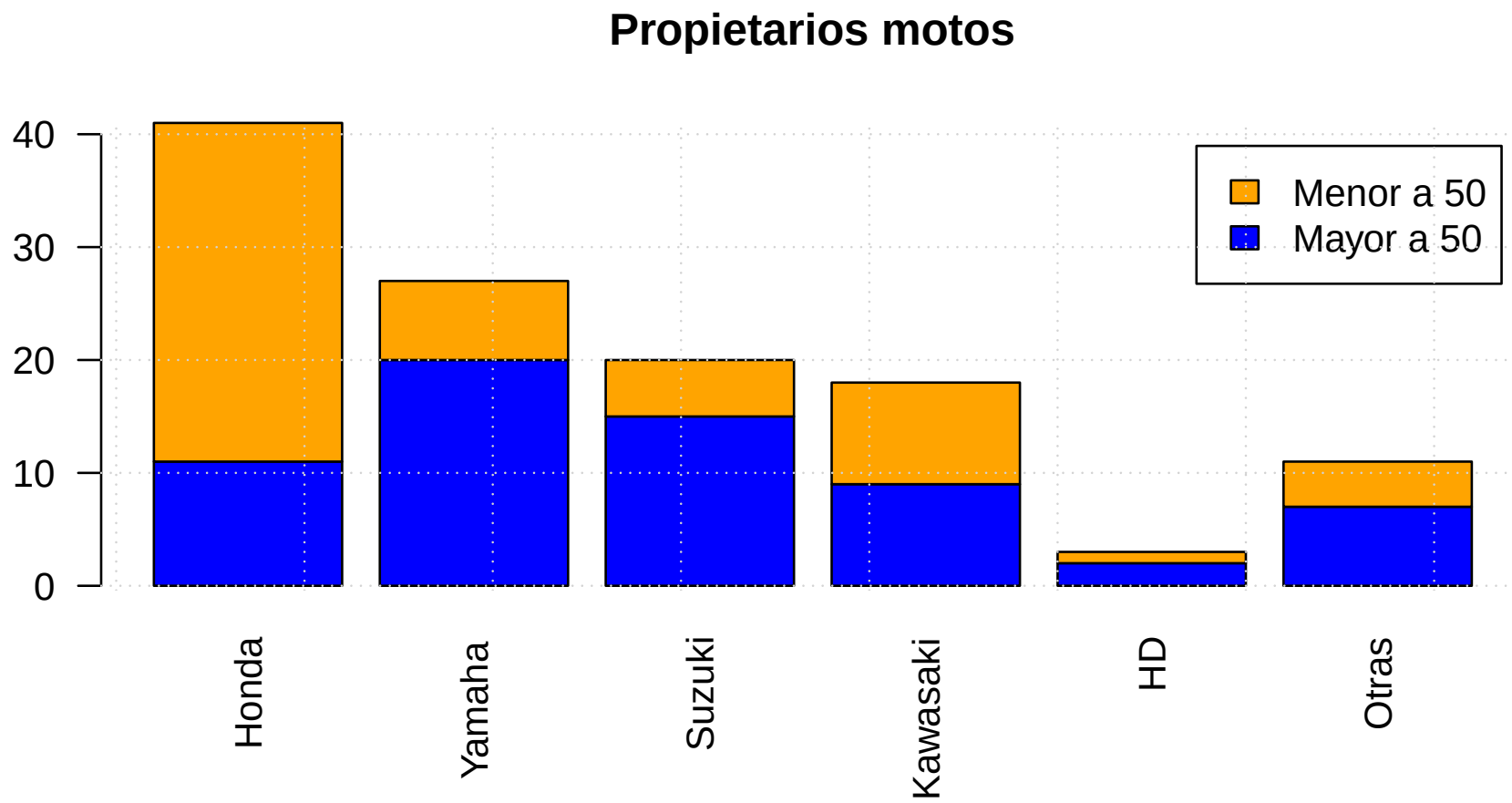
Barras apiladas, subdivididas o segmentadas

- Cada barra se segmenta en sus partes componentes.
- Los gráficos de columnas apiladas muestran la relación de cada elemento con el todo.

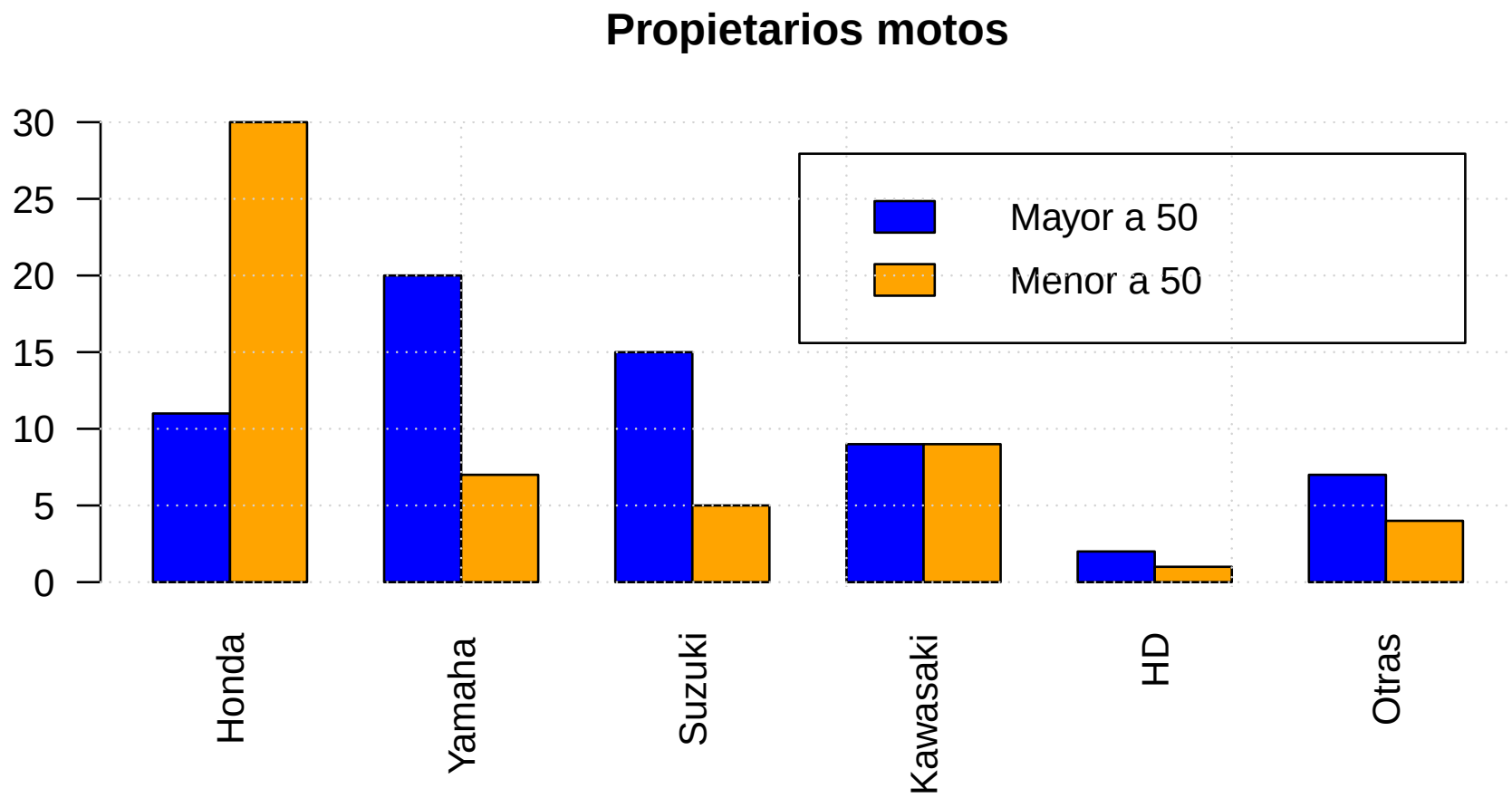
Barras Agrupadas

- Son de gran utilidad cuando deseamos comparar varias poblaciones entre sí.
- Se usan para comparar los componentes de un fenómenos o para comparar el mismo fenómenos en momentos, lugares, grupos diferentes.

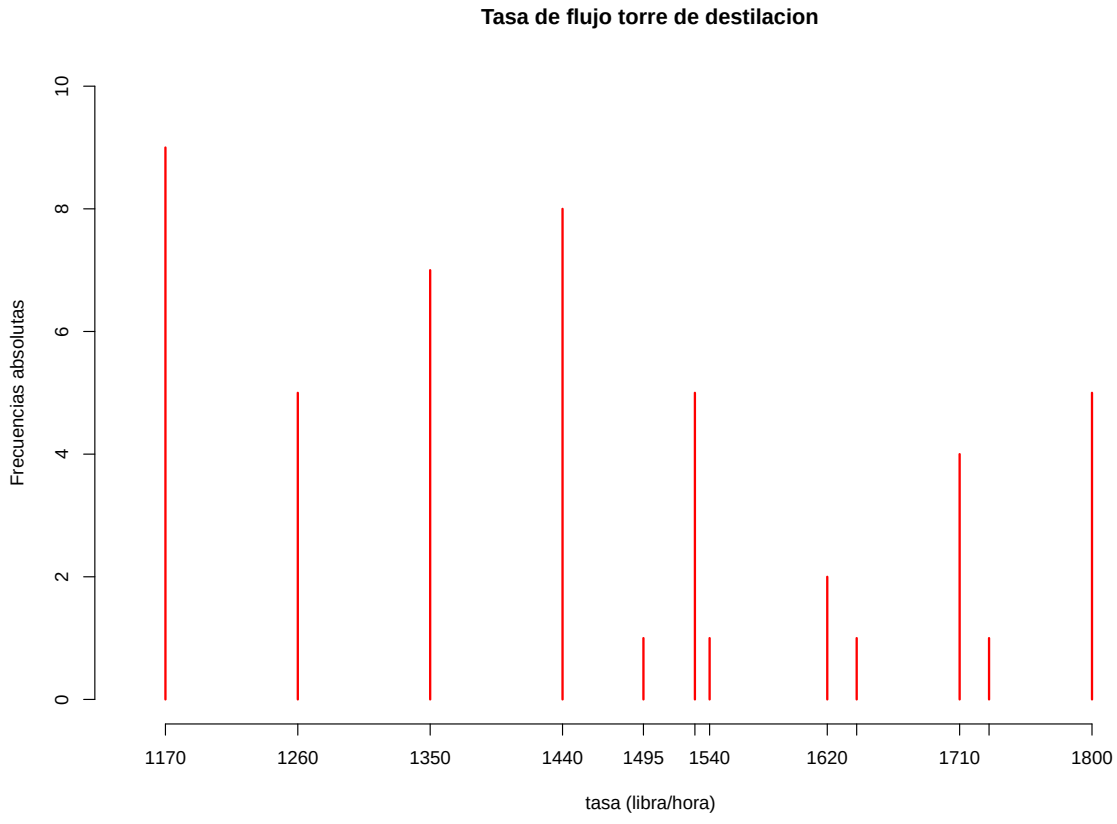
Barras apiladas



Barras agrupadas



Ejemplo: Mediciones de la torre de destilación, datos en agrupación simple



Ventajas y Desventajas

- **Ventajas:**
  - Representación de los datos de forma sencilla.
  - Los gráficos de barras representan una recopilación de datos que pueden ser comparados con otros.
- **Desventajas:**
  - A pesar de ser fáciles de crear y de entender, no son exactamente precisos.
  - No pueden utilizarse si los valores son muy diferentes entre si, por ejemplo: 10, 10000, 30, 1, 500

Gráficos Circulares o de Torta

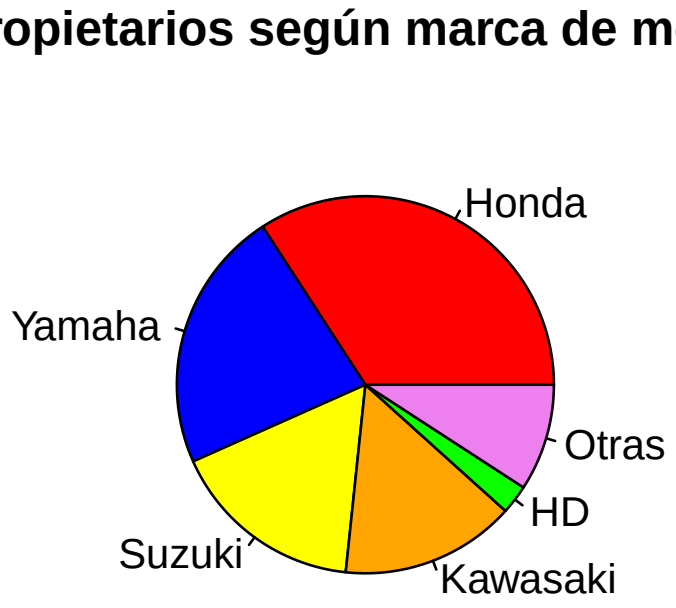
Caraterísticas

- Representan de forma significativa las comparaciones de varias medidas simultáneas en el tiempo y su importancia relativa, es decir los tantos por ciento que representan respecto al conjunto.
- Muestran el tamaño proporcional de los elementos que conforman una serie de datos en función de la suma de elementos. Siempre mostrará una única serie de datos, y resulta de utilidad cuando se desea destacar un elemento significativo.
- Se divide un círculo en tantas porciones como clases existan, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa.
- El arco de cada porción se calcula usando la regla de tres simple:

$$n \longrightarrow 360^\circ$$

$$n_i \longrightarrow x_i = \frac{360 \cdot n_i}{n}$$

Diagramas de Sectores - Variables Cualitativas

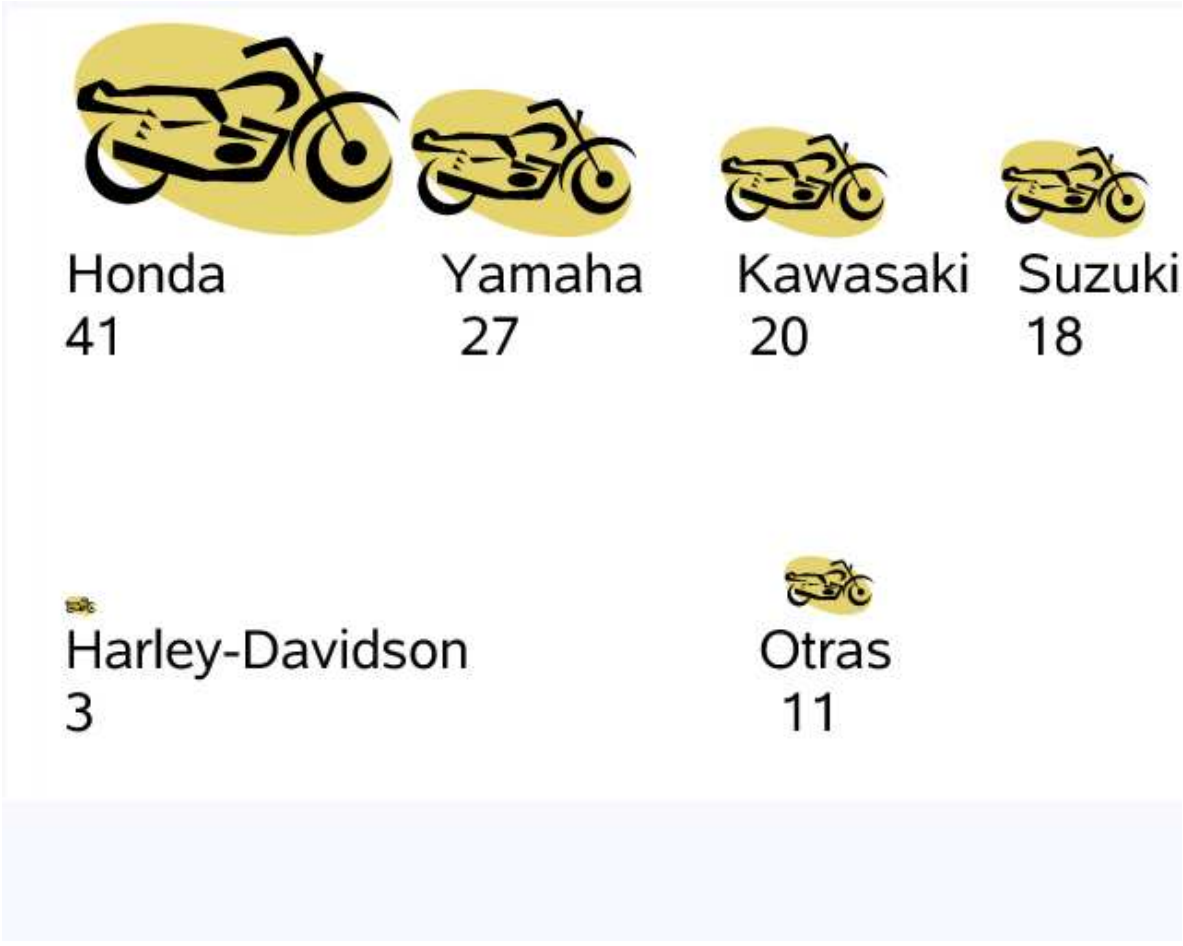


Pictogramas

Caraterísticas

- Son gráficos que expresan con dibujos alusivos al tema de estudio las frecuencias de las modalidades de la variable.
- Si usamos símbolos para indicar las magnitudes, éstos deben ser uniformes en tamaño y estar ordenados en forma de barras pictóricas.
- Se conserva el valor pictórico, usando varios dibujos pequeños, pero todos del mismo tamaño, y arreglándolos de manera que se forme una gráfica de barras.

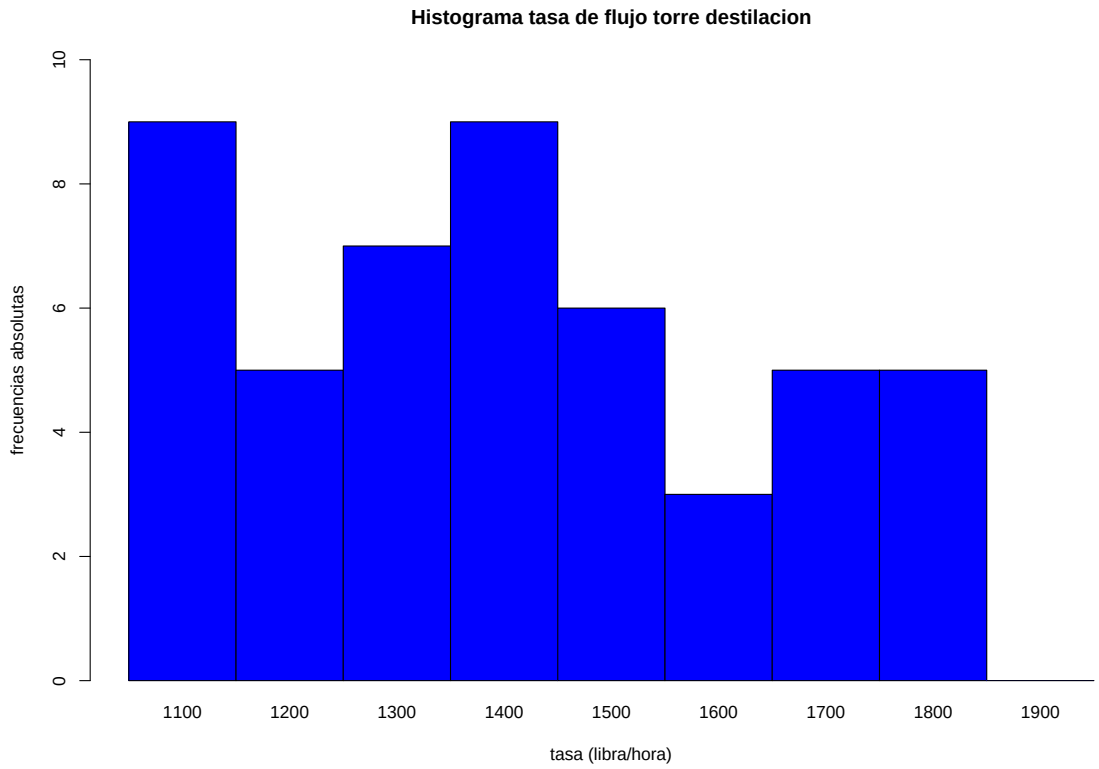
Pictogramas



Histogramas

- Sólo para datos agrupados en intervalos de clase.
- Eje de las abscisas: intervalos de clase.
- Eje de las ordenadas: frecuencias (simples).
- Sobre cada intervalo se levantan rectángulos cuya base es la longitud del intervalo de clase y su altura es tal que el área del rectángulo sea proporcional a la frecuencia del intervalo.
- Si los intervalos son de la misma amplitud, el ancho del rectángulo es fijo y la altura del mismo corresponde a la frecuencia que se esté considerando.

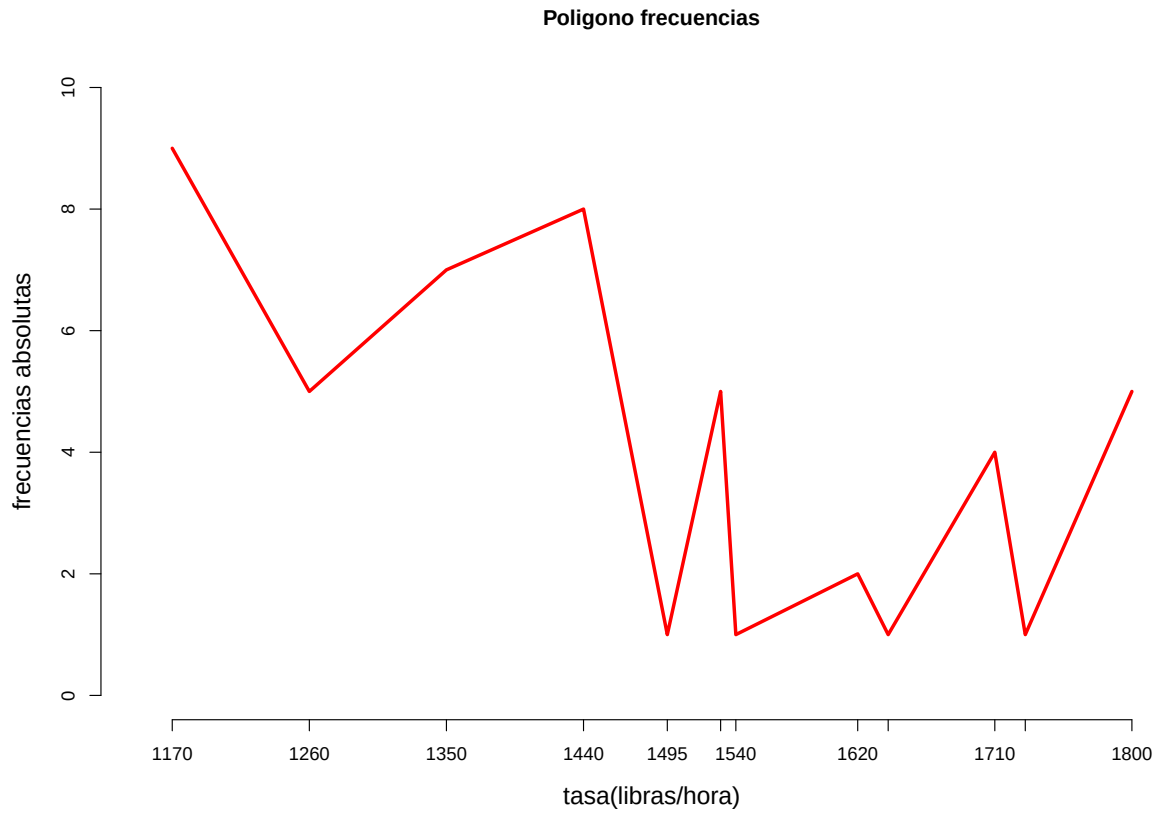
Ejemplo - Continuación Torre Destilación



Polígonos de Frecuencias - Datos en Agrupación Simple

Datos Agrupación Simple

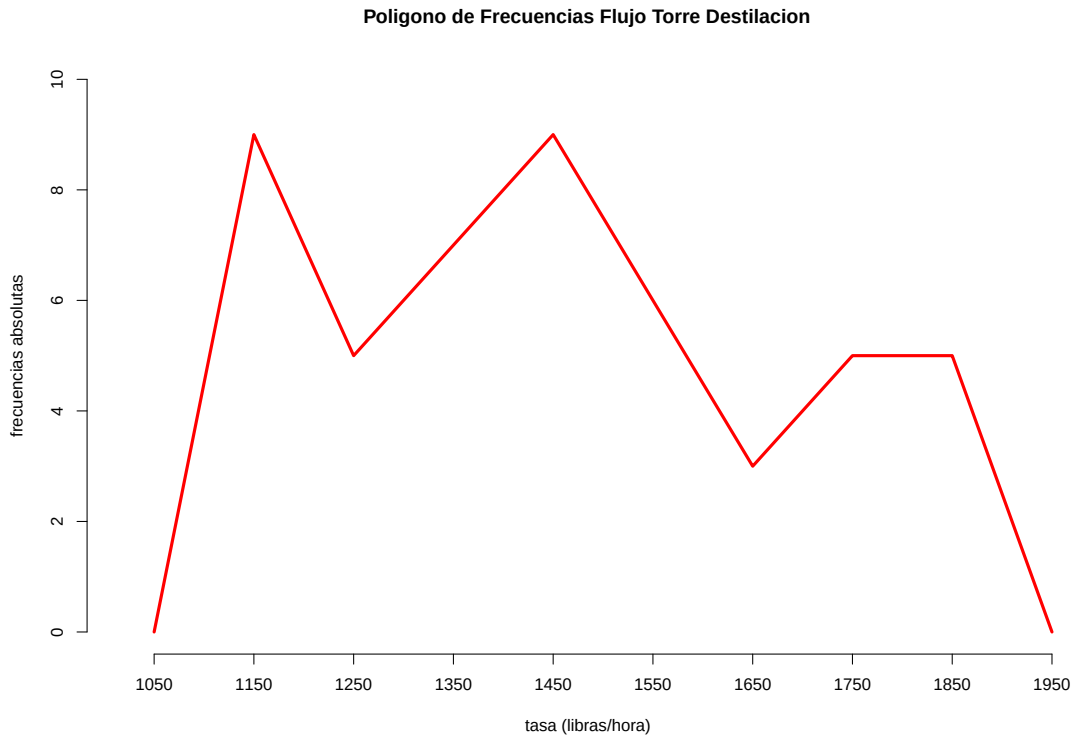
1. Eje de las abscisas: x_i
2. Eje de las ordenadas: frecuencia simple
3. Se unen esos puntos con segmentos de recta



Polígonos de Frecuencias-Datos Agrupados en Intervalos

Datos Agrupados en Intervalos de Clase

1. Eje de las abscisas: marcas de clase de los intervalos.
2. Eje de las ordenadas: frecuencia simple
3. Se unen esos puntos con segmentos de recta

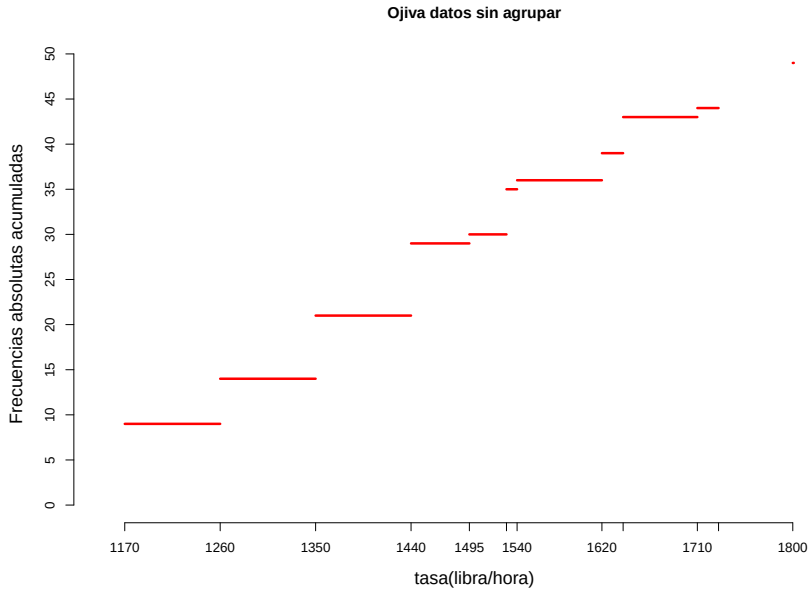


Polígonos de Frecuencias Acumuladas - Ojiva

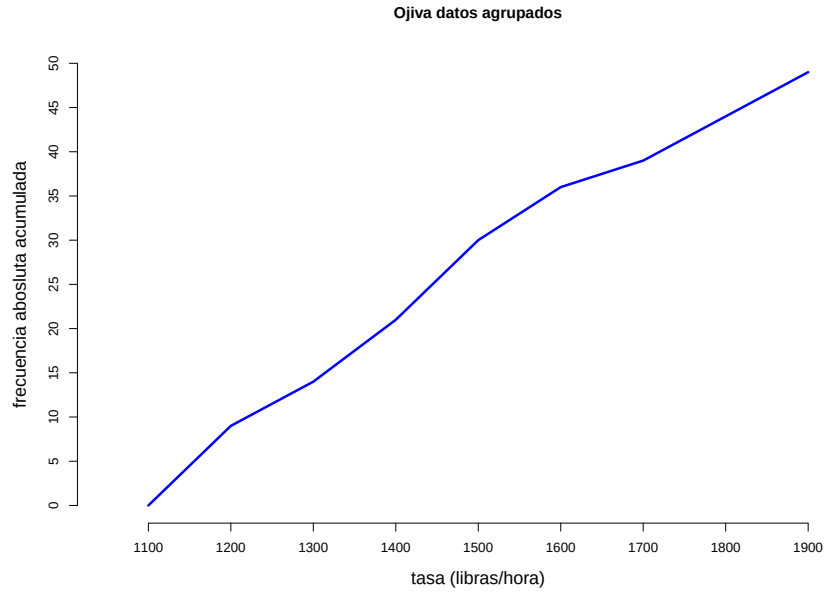
1. Datos en Agrupación Simple
   - Función escalonada.
   - Eje de las abscisas: x_i.
   - Eje de las ordenadas: Frecuencia acumulada.
2. Datos Agrupados en Intervalos
   - Eje de las abscisas: límites de los intervalos.
   - Eje de las ordenadas: frecuencias acumuladas sobre el límite superior del intervalo.
   - Al límite inferior del primer intervalo se le asigna el valor 0.
   - Se unen esos puntos con segmentos de recta.

Ejemplo: Datos torre de destilación

Datos en Agrupación simple



Datos Agrupados en Intervalos de clase



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 cuan esparcidos o concentrados están los datos correspondientes a una variable. Ej: Rango, Rango intercuartílico, varianza, desviación estandar, 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 $x_1 < \ldots < x_k$. Sean $f_1; \ldots; f_k$ sus respectivas frecuencias absolutas.
- Datos agrupados en intervalos:
  Supongamos datos agrupados en intervalos $\left(L_{inf\,1}; L_{sup_1}\right]; \ldots; \left(L_{inf\,k}; L_{sup_k}\right]$, cuyas marcas de clase son $M_{C1}; \ldots; M_{Ck}$, y f_i frecuencias absoluta 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

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

Ejemplo 5 (torres destilación - datos en agrupación simple):

$$\bar{x} = \frac{1170 \cdot 9 + 1260 \cdot 5 + 1350 \cdot 7 + 1440 \cdot 8 + 1495 \cdot 1 + 1530 \cdot 5 + 1540 \cdot 1 + 1620 \cdot 2 + 1640 \cdot 1 + 1710 \cdot 4 + 1730 \cdot 1 + 1800 \cdot 5}{49} = 1447,65$$

Esto nos dice que la tasa de flujo promedio de esa torre de medición es de 1447,65 libras/hora.

Medidas de Tendencia Central - Media

Datos sin agrupar

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

Datos en agrupación simple

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

Ejemplo 5 (torres destilación - datos en agrupación simple):

$$\bar{x} = \frac{1170 \cdot 9 + 1260 \cdot 5 + 1350 \cdot 7 + 1440 \cdot 8 + 1495 \cdot 1 + 1530 \cdot 5 + 1540 \cdot 1 + 1620 \cdot 2 + 1640 \cdot 1 + 1710 \cdot 4 + 1730 \cdot 1 + 1800 \cdot 5}{49} = 1447,65$$

Esto nos dice que la tasa de flujo promedio de esa torre de medición es de 1447,65 libras/hora.

Datos Agrupados en Intervalos

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

Ejemplo 5, datos agrupados en intervalos:

$$\bar{x} = \frac{1150 \cdot 9 + 1250 \cdot 5 + 1350 \cdot 7 + 1450 \cdot 9 + 1550 \cdot 6 + 1650 \cdot 3 + 1750 \cdot 5 + 1850 \cdot 5}{49} = 1456,12$$

Esto nos dice que la tasa de flujo promedio de esa torre de medición es de 1456,12 libras/hora.

Datos Agrupados en Intervalos

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

Ejemplo 5, datos agrupados en intervalos:

$$\bar{x} = \frac{1150 \cdot 9 + 1250 \cdot 5 + 1350 \cdot 7 + 1450 \cdot 9 + 1550 \cdot 6 + 1650 \cdot 3 + 1750 \cdot 5 + 1850 \cdot 5}{49} =$$

$1456,12$

Esto nos dice que la tasa de flujo promedio de esa torre de medición es de 1456,12 libras/hora.

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 $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, $Med = x_{(\frac{n+1}{2})}$.
- n par: $Med = \frac{x_{(n/2)}+x_{(n/2+1)}}{2}$

Medidas de Tendencia Central-Mediana

Ejemplo:

1170, 1170, 1170, 1170, 1170, 1170, 1170, 1170, 1170, 1260,
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.

n**=49** $\Rightarrow Med = x_{25} = 1440$

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.
- $Med = L_{Med\,inf} + \dfrac{n/2 - F_{Med-1}}{f_{Med}} A_{Med}.$

Medidas de Tendencia Central-Mediana

Tasa	M_{C_i}	f_i	f_{r_i}	p_i	F_i	F_{r_i}	P_i
[1100,1200)	1150	9	0.19	19	9	0.19	19
[1200,1300)	1250	5	0.10	10	14	0.29	29
[1300,1400)	1350	7	0.14	14	21	0.43	43
[1400,1500)	1450	9	0.19	19	30	0.62	62
[1500,1600)	1550	6	0.12	12	36	0.74	74
[1600,1700)	1650	3	0.06	6	39	0.80	80
[1700,1800)	1750	5	0.10	10	44	0.90	90
[1800,1900)	1850	5	0.10	10	49	1.00	100

$n/2 = 49/2 = 24{,}5$;
$L_{Med\,inf} = 1400$; $F_{Med-1} = 21$;
$f_{med} = 9$; $A_{med} = 100$; por lo que:

$$Med = 1440 + \frac{24{,}5 - 21}{9} \cdot 100 = 14$$

Esto nos dice que el 50 % de las mediciones dieron una tasa de 1478,89 libras/hora o menos; o equivalentemente, el 50 % de las mediciones dieron una tasa de 1478,89 libras/hora o más.

Medidas de Tendencia Central-Mediana

Datos en agrupación simple

Ejercicio 2

N. de Herm.	Cantidad
0	740
1	1097
2	658
3	345
4	126
5	34

$$Med = \frac{x_{(1500)} + x_{(1501)}}{2} = \frac{1+1}{2} = 1.$$

Esto nos dice que el 50 % de los alumnos de la FaCENA del año 2002 tiene 1 hermano o menos.

Medidas de Tendencia Central-Mediana

Datos en agrupación simple

Ejercicio 2

N. de Herm.	Cantidad
0	740
1	1097
2	658
3	345
4	126
5	34

$$Med = \frac{x_{(1500)} + x_{(1501)}}{2} = \frac{1+1}{2} = 1.$$

Esto nos dice que el 50 % de los alumnos de la FaCENA del año 2002 tiene 1 hermano o menos.

Medidas de Tendencia Central-Mediana

Datos en agrupación simple

Ejercicio 2

N. de Herm.	Cantidad
0	740
1	1097
2	658
3	345
4	126
5	34

$$Med = \frac{x_{(1500)} + x_{(1501)}}{2} = \frac{1+1}{2} = 1.$$

Esto nos dice que el 50 % de los alumnos de la FaCENA del año 2002 tiene 1 hermano o menos.

Medidas de Tendencia Central-Mediana

Datos en agrupación simple

Ejercicio 2

N. de Herm.	Cantidad
0	740
1	1097
2	658
3	345
4	126
5	34

$$Med = \frac{x_{(1500)} + x_{(1501)}}{2} = \frac{1+1}{2} = 1.$$

Esto nos dice que el 50 % de los alumnos de la FaCENA del año 2002 tiene 1 hermano o menos.

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.
- Dificil 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 posición del cuantil $l_k = \frac{k \cdot n}{NoC}$, 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 $l_{Ck} = \frac{k \cdot n}{NoC}$.
- Se calcula:

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

Cálculo de Cuantiles

Datos en agrupación simple:

- Se ordenan los datos de menor a mayor.
- Se calcula la posición del cuantil $l_k = \frac{k \cdot n}{NoC}$, 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 $l_{Ck} = \frac{k \cdot n}{NoC}$.
- Se calcula:

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

Cuantiles- Ejemplos

Datos en agrupación simple:
Para primer cuartil Q_1, $C_1 = (1 \times 49)/4 = 12{,}25$, y en la tabla de frecuencias acumuladas vemos que ese dato corresponde a $Q_1 = 1260$.
Para tercer cuartil, Q_3, $C_3 = (3 \times 49)/4 = 36{,}75$, esto es, $Q_3 = 1520$.
P_{90} es el dato que deja a su izquierda el 90 % de las observaciones. $(90 \times 49)/100 = 44{,}1$, esto es, $P_{90} = 1800$.
Datos agrupados en intervalos:
Para el primer cuartil, $C_1 = 49/4 = 12{,}25$, entonces:

$$Q_1 = 1200 + \frac{12{,}25 - 9}{5} \cdot 100 = 1265$$

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 reponda:

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

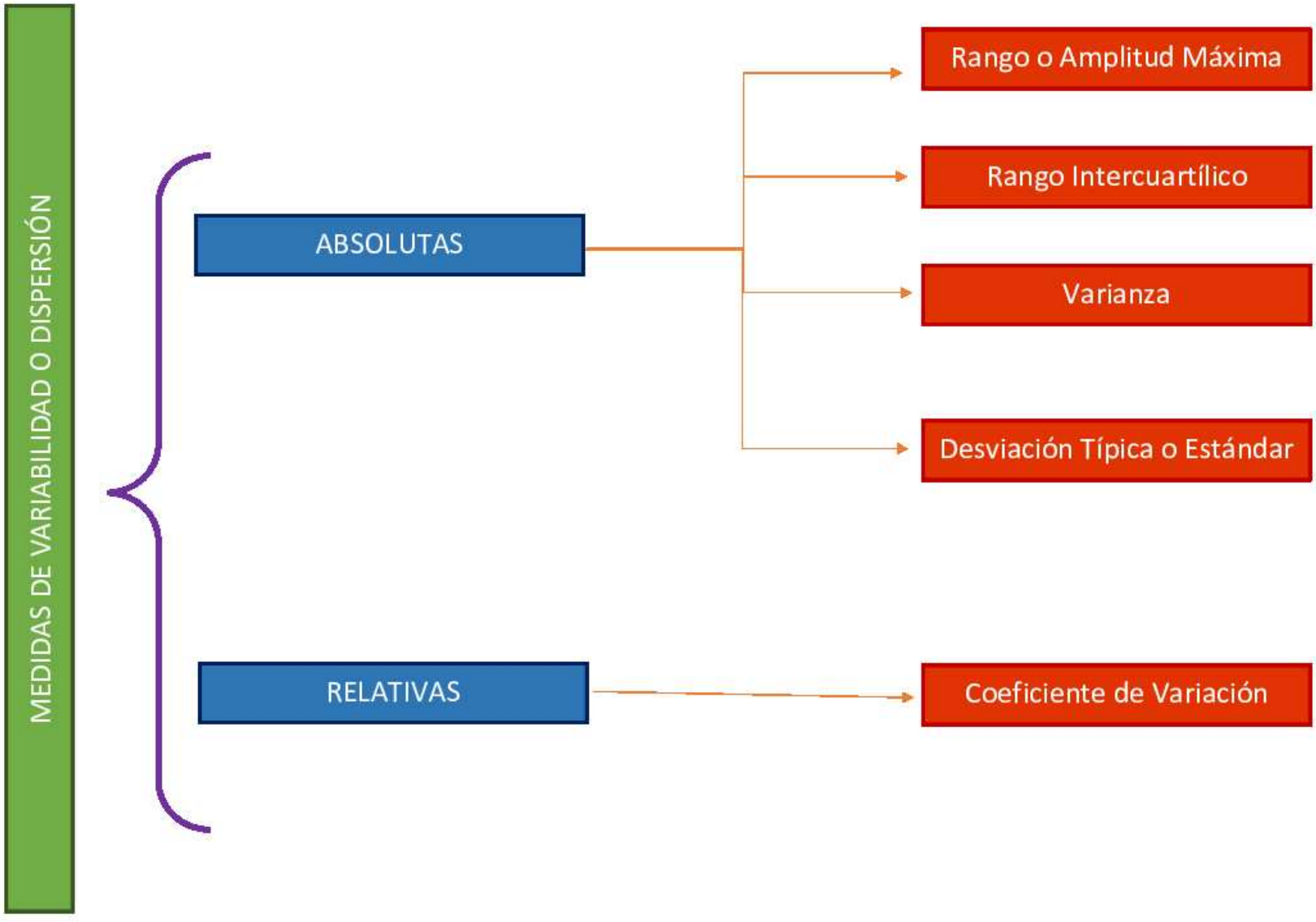
La medidas de tendencia central (MTC) vienen a reponder a la primera pregunta. Estas medidas de centralización, sirven para describir un aspecto de los datos, pero no nos dicen nada acerca de otro aspecto de igual importancia: la dispersión de los valores observados.

Un promedio, por ejemplo, no nos dice nada acerca de la dispersión de los datos, para esto utilizaremos las medidas de dispersión o variabilidad.

Si el valor de estas medidas de dispersión es pequeño, nos indica que los datos están estrechamente agrupados alrededor de la MTC, entonces ésta se considera representativa de los datos. Inversamente, una medida de dispersión grande indica que la MTC no es confiable.

Por lo tanto toda MTC, para que brinde una información eficaz, debe ir acompañada de alguna medida de variabilidad.

Medidas de Dispersión o Variabilidad



Medidas de Dispersión o Variabilidad- Rango

También 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 facil 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.

Medidas de Dispersión o Variabilidad- Rango

Desventajas:

- No está basado en ninguna MTC.
- Está afectado por los valores OUTLIERS.
- Está afectado por el tamaño de la muestra.
- En su cálculo sólo intervienen dos valores, por lo quer 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 $[MTC - RIC/2; MTC + RIC/2]$ 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

$$\bar{d} = \frac{\sum_i^n d_i}{n} = 0$$

Para evitar este inconveniente se define el **Desvío Medio** como: D.S.A: $DM = \frac{\sum_i^n \left| x_i - \bar{X} \right|}{n}$

D.A.S.: $DM = \frac{\sum_i^k \left| x_i - \bar{X} \right| f_i}{n}$

D.A.I.C.: $DM = \frac{\sum_i^k \left| M_{ci} - \bar{X} \right| f_i}{n}$

Muchas veces, si el promedio aritmético no es una MTC confiable, se la suele reemplazar por la MEDIANA, definiendo se 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. D.S.A: $S^2 = \dfrac{\sum_i^n \left(x_i - \bar{X}\right)^2}{n}$

D.A.S.: $S^2 = \dfrac{\sum_i^k \left(x_i - \bar{X}\right)^2 f_i}{n}$

D.A.I.C.: $S^2 = \dfrac{\sum_i^k \left(M_{Ci} - \bar{X}\right)^2 f_i}{n}$

Fórmula de Trabajo: $S^2 = \bar{X^2} - \left(\bar{X}\right)^2$.

Aunque esta fórmula de la varianza es correcta, en la práctica, el denominador que se utiliza es $n - 1$ en lugar de n. Por lo tanto, la medida que se utiliza es

$$S^2 = \frac{\sum_i^n \left(x_i - \bar{X}\right)^2}{n - 1}$$

El hecho de dividir por $n - 1$ en lugar de n es apenas apreciable cuando n es grande.

Medidas de Dispersión o Variabilidad- Varianza

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 (queda afectada por el cuadrado).

Medidas de Dispersión o Variabilidad- Varianza

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.

D.S.A: $S = \sqrt{\frac{\sum_i^n \left(x_i - \bar{X}\right)^2}{n-1}}$

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

D.A.I.C.: $S = \sqrt{\frac{\sum_i^k \left(M_{Ci} - \bar{X}\right)^2 f_i}{n-1}}$

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 catacterísticas 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 absolutasz 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 companías 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:

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

Medidas de Dispersión o Variabilidad - Coeficiente de Variación

Ejemplo:
Si tenemos dos conjuntos de estudiantes A y B, cuyo peso presenta la misma dispersión $S = 12$ kilos, pero el conjunto A tiene un peso promedio de 72 kilos, mientras que el conjunto B tiene un peso promedio de 62 kilos.
Es claro que desde el punto de vista de la dispersión absoluta, la variabilidad en ambos conjuntos es idéntica. No obstante, también es claro que *relativamente, el conjunto A presenta mayor homogeneidad en sus pesos*, ya que 12 respecto a 72 es **relativamente** menor que 12 respecto a 61.
Como es de esperarse, $CV_A < CV_B$.

Distribución Normal - Campana de Gauss

Es la distribución teórica más conocida y utilizada en Estadítica. 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 máxmo en el centro y decrece constantemente hacia los extremos.
- No corta al eje de abscisas.
- La Media Aritmética coincide con el Modo y la Mediana.

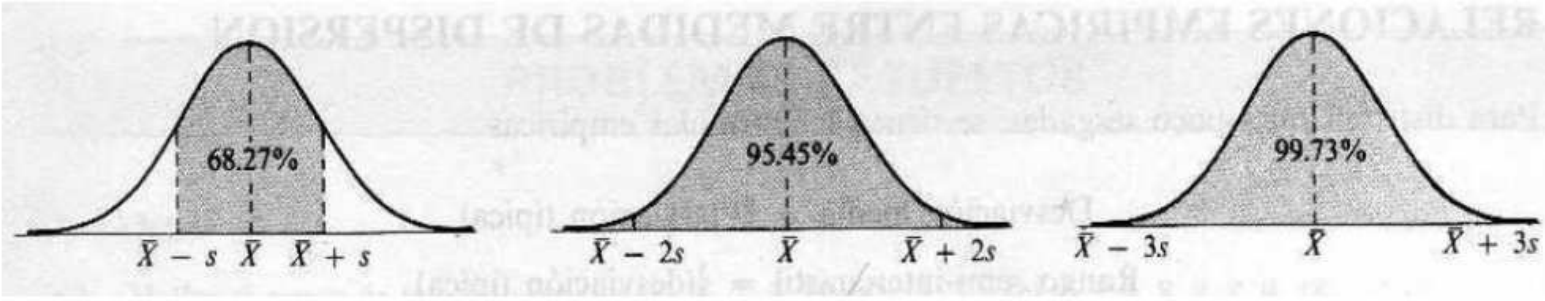
Es una distribución que se utiliza para describir otras características de una distribución en particular comparándola con ella (por ejemplo, asimetría y kurtosis). También para determinar valores de datos atípicos.

Distribución Normal - Campana de Gauss

Para distribuciones de datos que se aproxima 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:

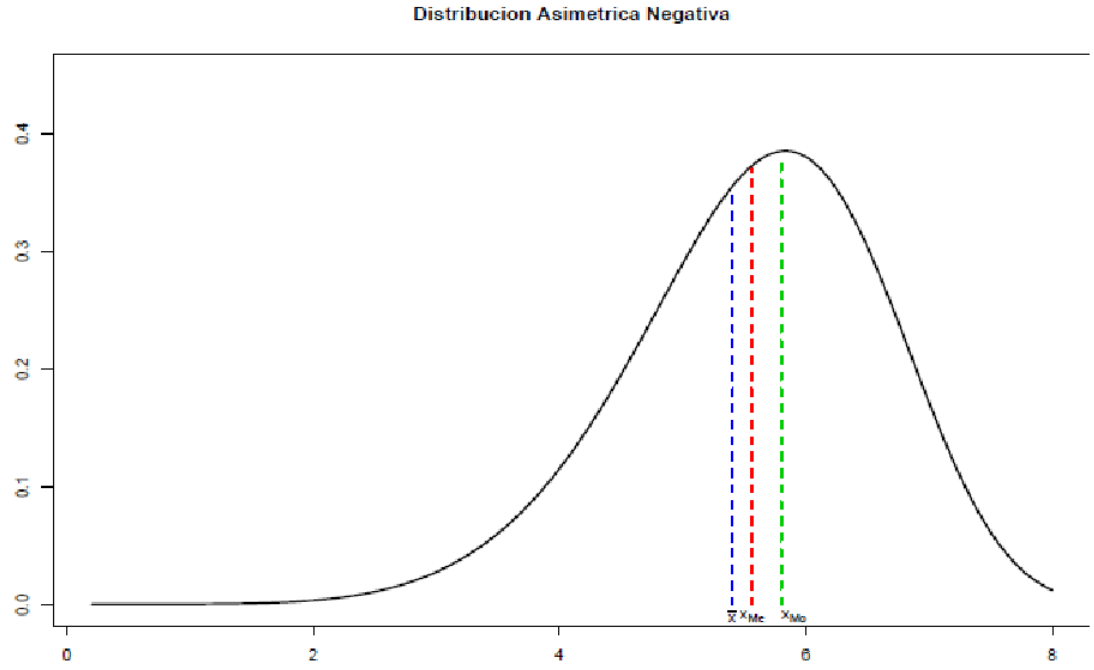
- **simétrica**: los valores equidistiantes de una posición central tienen la misma frecuencia,
- **asimétrica positiva**: las frecuencias más altas corresponden a valores que se encuentran al lado izquierdo de esa posición central (cola a la derecha),
- **asimétrica negativa**: distribuciones con cola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

Se define el coeficiente de asimetría de Pearson como

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

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

Coeficiente de Asimetría- Asimetría Negativa



Medidas de forma - Kurtosis

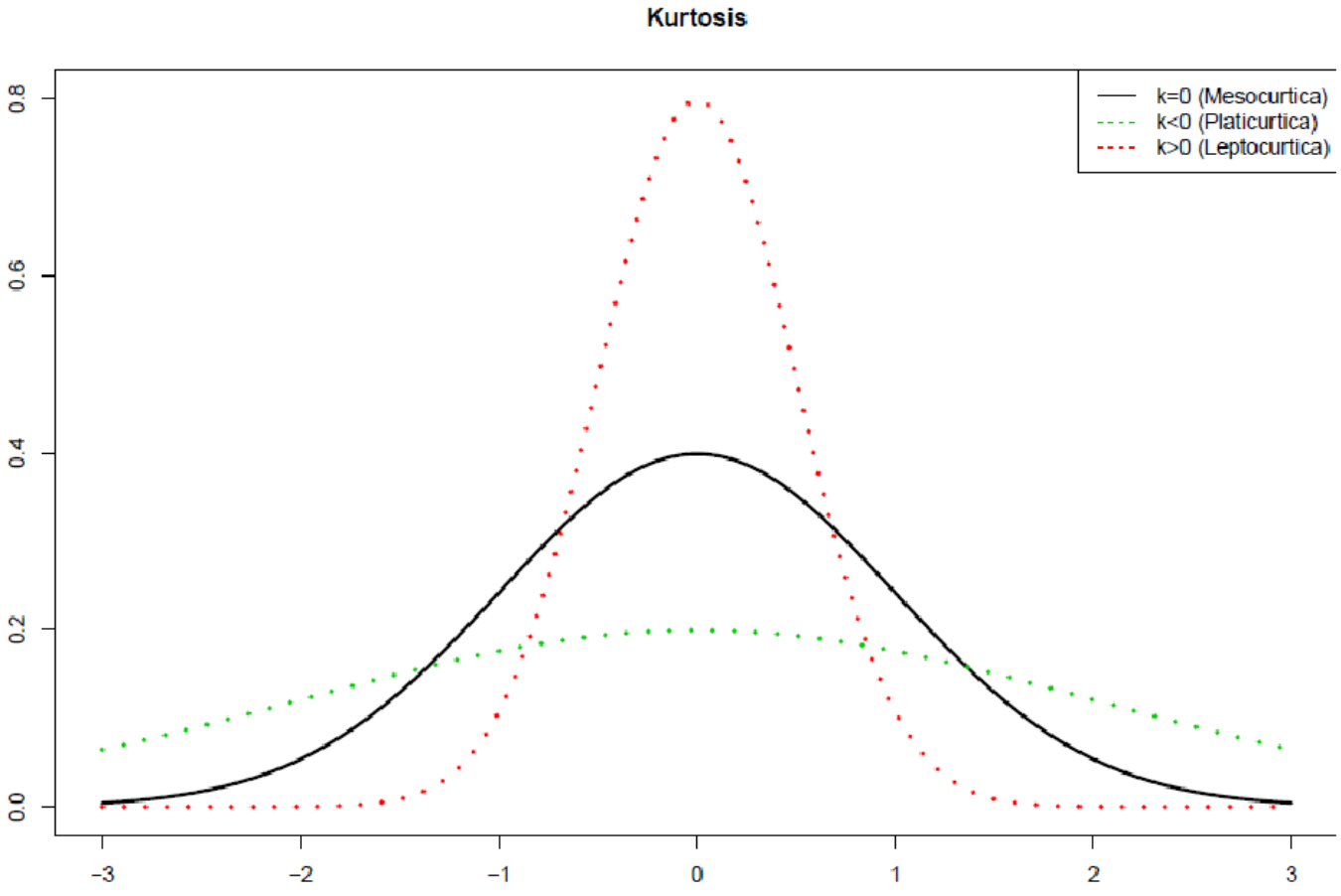
Se aplica a distribuciones unimodales simétricas o ligeramente asimétricas, ya que representa la elevación o achatamiento de una distribución comparada con la distribución normal.

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

siendo $m_4 = \dfrac{\sum_{i=1}^{n} \left(x_i - \bar{X}\right)^4}{n}$ 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)

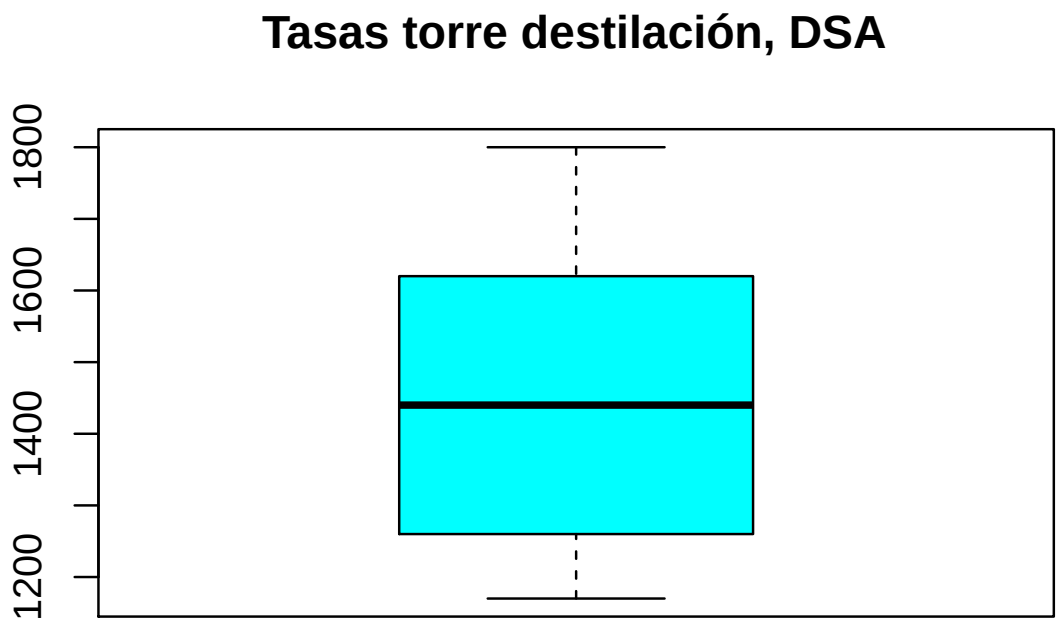
Coeficiente de Kurtosis



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 * 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 *Rango intercuartílico.
- Las observaciones que caen fuera de esos “bigotes”se representan con círculos rellenos si están a una distancia mayor a 3* Rango intercuartílico, o por círculos sin rellenar en caso contrario.

Boxplot: Ejemplo, datos sin agrupar



Diagramas de Tallo y Hoja

Es un diagrama de gran utilidad para representar un conjunto de datos cuantitativos, este tipo de representación presenta similitudes con el histograma en cuanto que proporciona información del recorrido de la distribución de datos en estudio, muestra la ubicación de la mayor concentración de mediciones y revela la presencia o ausencia de simetría.
Este diagrama tiene ventajas sobre el histograma, porque conserva la información que puede arrojar las mediciones individuales, situación que se pierde en los intervalos del histograma.

Diagramas de Tallo y Hoja

¿Cómo construir el diagrama de tallo y hojas?

- Se debe dividir cada medición en dos partes, la primera se llama tallo y la segunda hojas.
- El tallo se forma con uno o más dígitos iniciales de la medición, y las hojas se forman con uno o más de los dígitos restantes.
- La cantidad de tallos preferiblemente deben ser mayores o iguales a 5 y menores o iguales a 20.
- Los tallos forman una columna ordenada de menor a mayor del lado izquierdo del diagrama.
- Registrar las hojas por cada observación junto al valor correspondiente del tallo.

Diagramas de Tallo y Hoja

Los siguientes datos representan la evaluación de los latidos cardíacos de un grupo de 30 personas después de cierta actividad física. 82 – 95 – 92 – 62 – 85 – 92 – 82 – 95 – 70 – 85 – 84 – 95 – 91 – 82 – 94 – 76 – 88 – 91 – 87 – 80 – 68 – 58 – 76 – 85 – 110 – 60 – 75 – 88 – 64 – 74

Ordenamos los datos: 58 – 60 – 62 –64 – 68 –70 – 74 –75 – 76 –76 – 80 – 82 –82 – 82 – 84 – 85 – 85 – 85 – 87 – 88 – 88 –91 – 91 – 92 – 92 –94 –95 –95 –95 –110

Tallo	Hoja
5	8
6	0 2 4 8
7	0 4 5 6 6
8	0 2 2 2 4 5 5 5 7 8 8
9	1 1 2 2 4 5 5 5
11	0

Diagramas de Tallo y Hoja

Recomendaciones:

- Es importante tomar en cuenta que este tipo de diagramas, no es aconsejable en informes anuales o en algún tipo de medios de difusión para un público en general.
- Algunas veces, la utilización del primero o de los dos primeros dígitos de los datos puntuales como tallos no proporcionan suficientes tallos como para permitir detectar la forma de su distribución. Una manera de solucionar esto es utilizar tallos dobles. Es decir, utilizar cada tallo dos veces: una vez para trazar las hojas inferiores y otra vez para trazar las hojas superiores.
- Los diagramas de tallos y hojas dan una idea de la localización de los datos y de la forma de la distribución. Esta técnica funciona bien para los conjuntos de datos que no tienen una dispersión muy grande.