Pontificia Universidad JAVERIANA Bogotá

ESTADÍSTICAS DESCRIPTIVAS: MEDIDAS Y GRÁFICOS

ING. HERIBERTO FELIZZOLA JIMENEZ

RUTA DE CIENCIA DE DATOS

Educación Continua

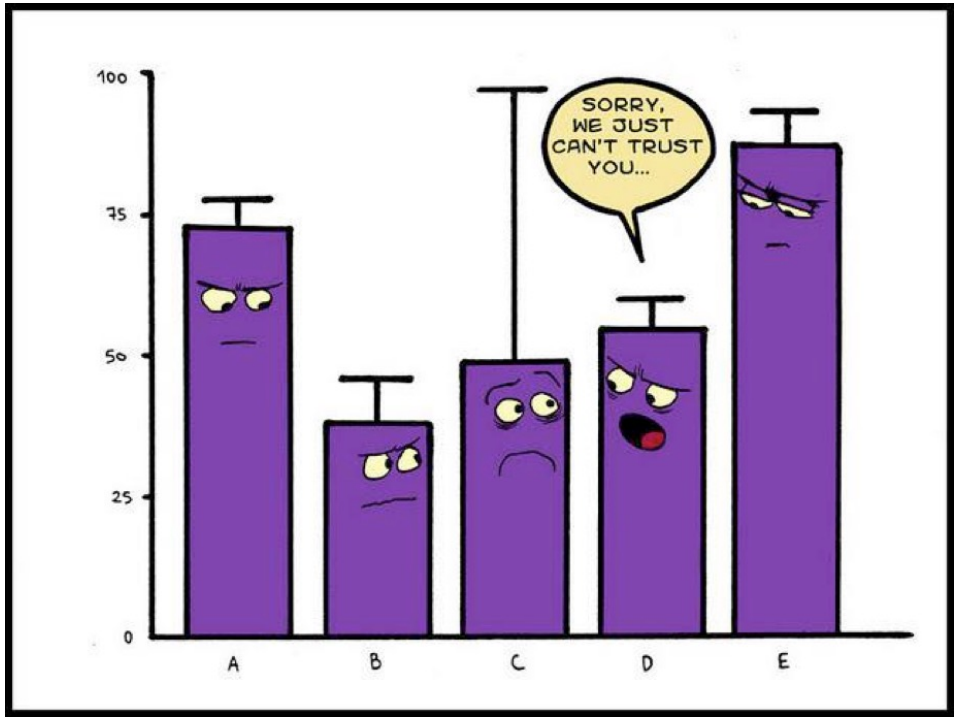
Generamos experiencias educativas

CONTENIDO

1. El rol de la estadística
2. Fuentes y tipos de datos
3. Tablas de frecuencia e histogramas
4. Medidas de tendencia central
5. Medidas de dispersión
6. Medidas de forma
7. Medidas de posición
8. Gráficos de caja y valores extremos
9. Análisis de variables categóricas

El ROL DE LA ESTADÍSTICA

- Los métodos estadísticos nos ayudan a describir y comprender la variabilidad.
- Por variabilidad entendemos que las observaciones sucesivas de un sistema o fenómeno no producen exactamente el mismo resultado.
- La variabilidad conduce a la incertidumbre sobre los resultados de un sistema o problema.
- La variabilidad hace que los eventos o fenómenos sean de cierta manera impredecibles con 100% de confianza.
- Todos nos encontramos con la variabilidad en nuestra vida cotidiana, y el pensamiento estadístico puede ofrecernos una forma útil de incorporar esta variabilidad a nuestros procesos de toma de decisiones.



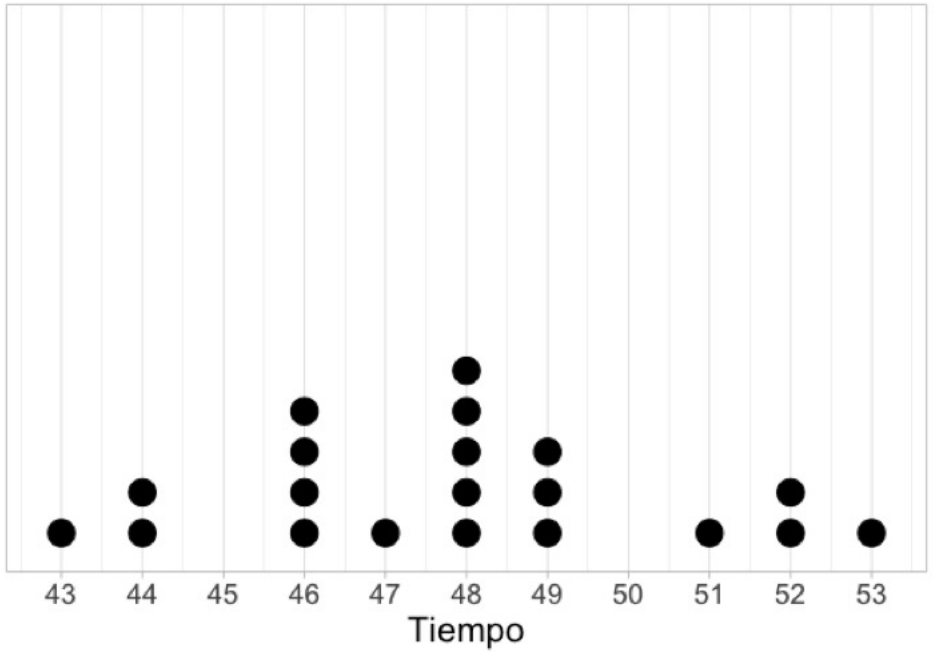
Fuente. https://cheezburger.com/5551937792

El ROL DE LA ESTADÍSTICA

Pontificia Universidad JAVERIANA Bogotá

Consideremos el siguiente ejemplo:

- Suponga que se quiere analizar el nivel de servicio del área de soporte técnico de una organización.
- Se ha establecido 48 horas como el tiempo estándar para la solución de un requerimiento.
- *¿Cree usted que todas las solicitudes se resuelven en este tiempo?*
- Piense y comente otras situaciones de la vida cotidiana/industria/sociedad donde observe el fenómeno de la variabilidad.
- Suponga que la empresa toma una muestra aleatoria de 20 servicios en una semana y registra el tiempo, en horas, de resolución de la orden.

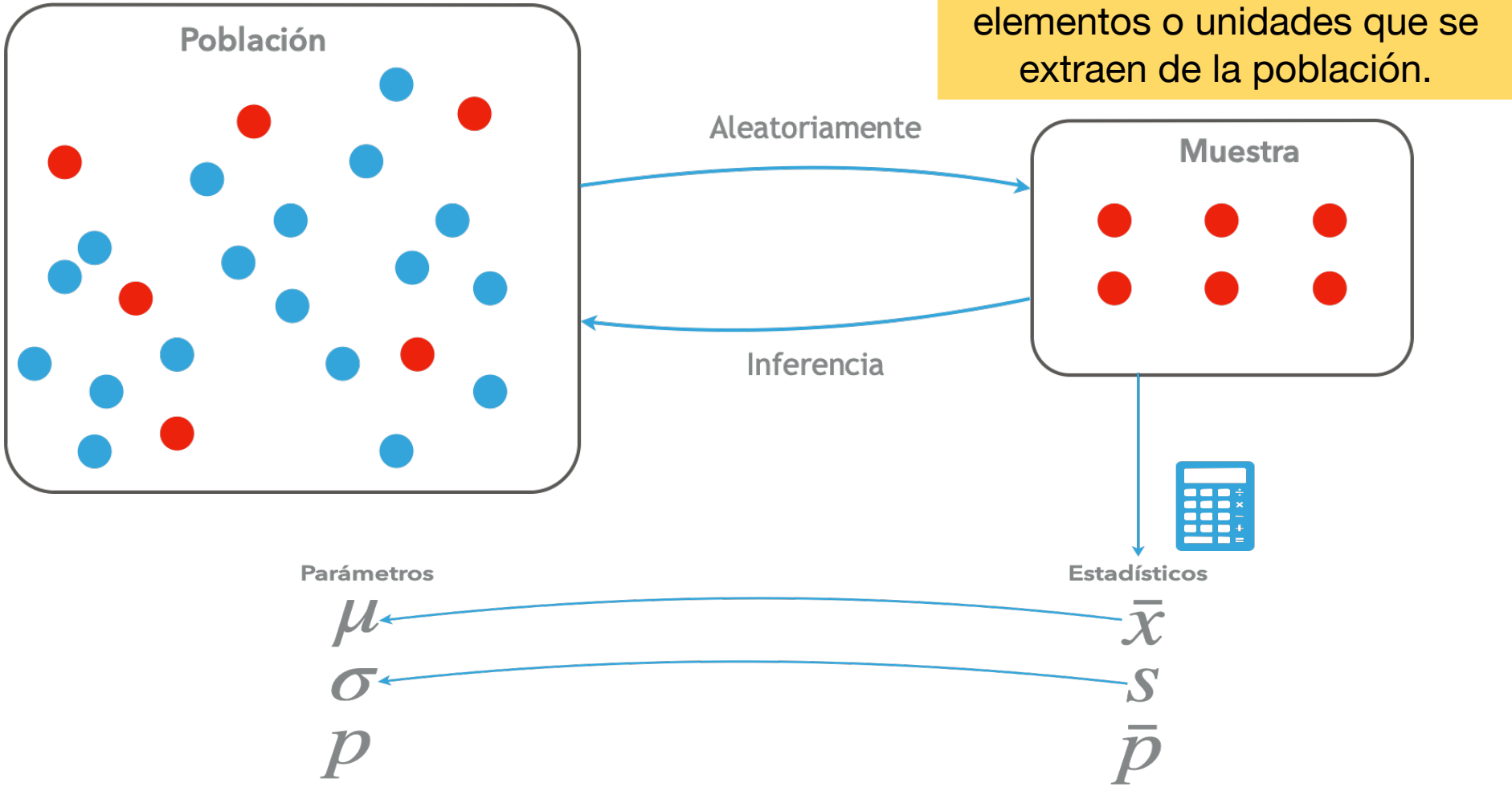


¿Qué efecto tiene esta variabilidad en el contexto del problema?

POBLACION VS MUESTRA

Población: Conjunto de elementos o unidades con al menos un atributo en común.

Muestra: es un subconjunto de elementos o unidades que se extraen de la población.



EJEMPLOS

Pontificia Universidad JAVERIANA Bogotá

Población	Muestra	Propósito
Pacientes atendidos en un mes	Seleccionar aleatoriamente 100 pacientes por mes	• Estimar el tiempo promedio de atención. • Estimar el tiempo de espera. • Verificar fallas en el servicio.
Clientes atendidos por un call center por día	Seleccionar aleatoriamente 5 clientes por hora	• Estimar el tiempo en espera. • Estimar y controlar el tiempo de atención. • Medir la calidad del servicio.
Paquetes transportados en un semana	Seleccionar aleatoriamente 10 paquetes por día	• Estimar el tiempo de envío. • Estimar el nivel de servicio. • Verificar fallas en el proceso.

Defina un ejemplo de población, muestra y propósito de acuerdo a su área de trabajo (Coloque la respuesta en el chat)…

FUENTES DE DATOS

Pontificia Universidad JAVERIANA Bogotá

Encuestas

Formatos

Bases de Datos de Sistemas Transaccionales

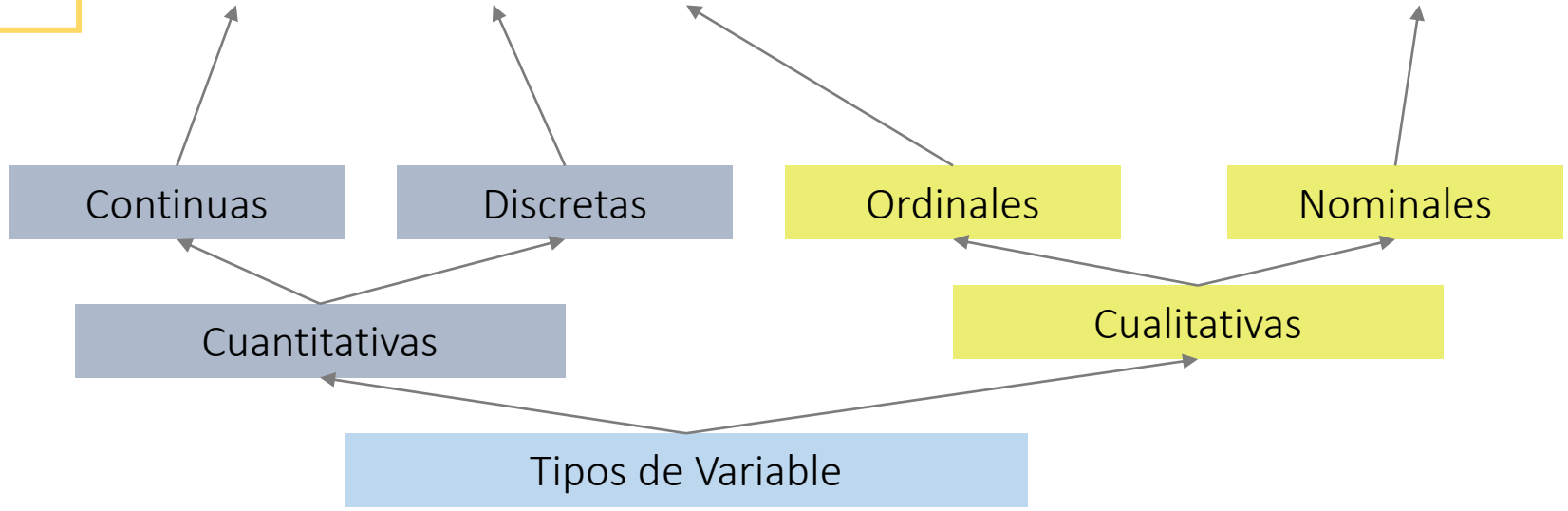
TIPOS DE DATOS

Las columnas representan las características de cada uno de los préstamos, también denominadas variables.

Cada fila representa un único préstamo. Esto puede ser un caso/instancia/observación.

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Cada celda es un valor único de la variable para una observación particular.



ESTADISTICAS Y GRAFICOS

Pontificia Universidad JAVERIANA Bogotá

		Tipo de variable	
		Numérica	Categórica
Método	Estadísticas	Media, mediana, moda, rango, varianza, desviación estándar, cuártiles, percentiles, sesgo, curtosis, tablas de frecuencia	Conteo, proporción, tablas de frecuencia, tablas de contingencia
	Gráfico	Histograma, gráfico de caja, valores individuales	Gráfico de barras, Pareto, donas, torta

RESUMEN DE DATOS

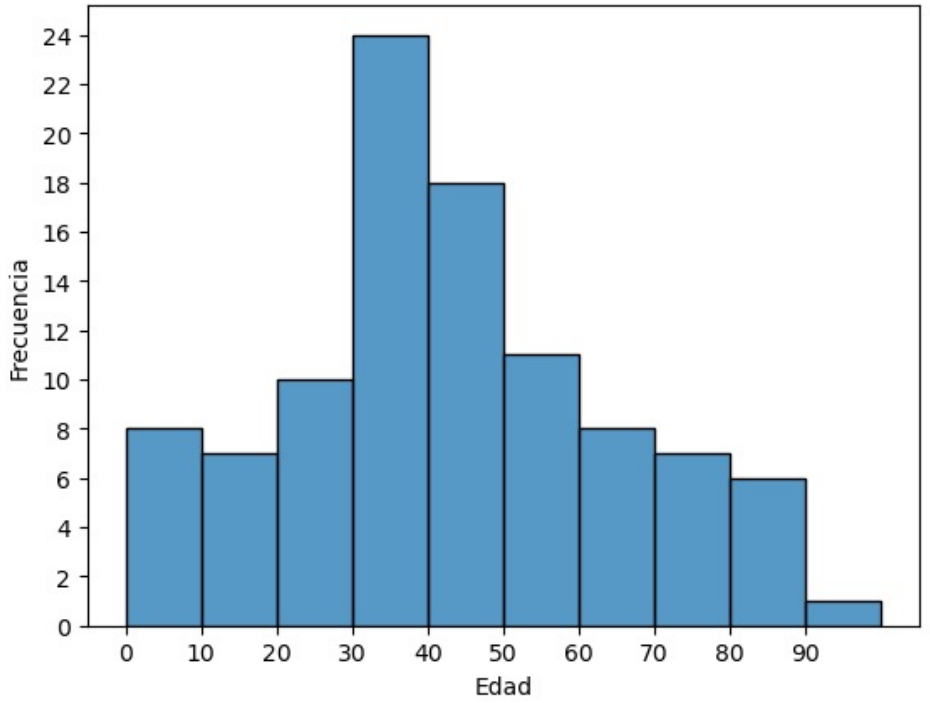
- Imagine que tenemos una muestra de 100 de casos de Covid-19 en Colombia y queremos resumir los datos de edad: [38, 86, 13, 26, 44, 62, 24, 37, 36, 62, 63, 42, 62, 5, 27, 46, 26, 30, 81, 38, 47, 1, 19, 38, 44, 75, 22, 56, etc…]
- Una forma de resumir los datos es construir tablas de frecuencia e histogramas.
- Esto nos permite analizar ciertas características en los datos y el fenómeno subyacente.

Tabla de frecuencia

Rango Edad	Frecuencia	Porcentaje
(0, 10]	8	8,0%
(10, 20]	7	7,0%
(20, 30]	13	13,0%
(30, 40]	21	21,0%
(40, 50]	19	19,0%
(50, 60]	10	10,0%
(60, 70]	10	10,0%
(70, 80]	6	6,0%
(80, 90]	5	5,0%
(90, 100]	1	1,0%
Total	100	

La tabla de frecuencia nos permite resumir pero *“una imagen vale mas que mil palabras”.*

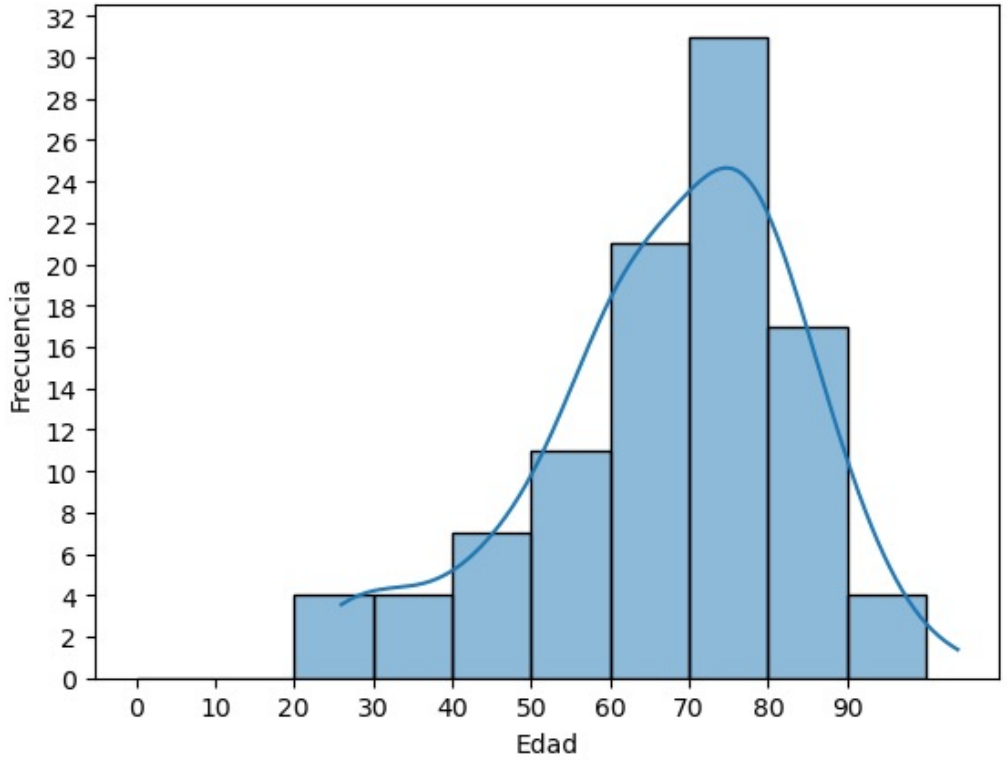
Histograma



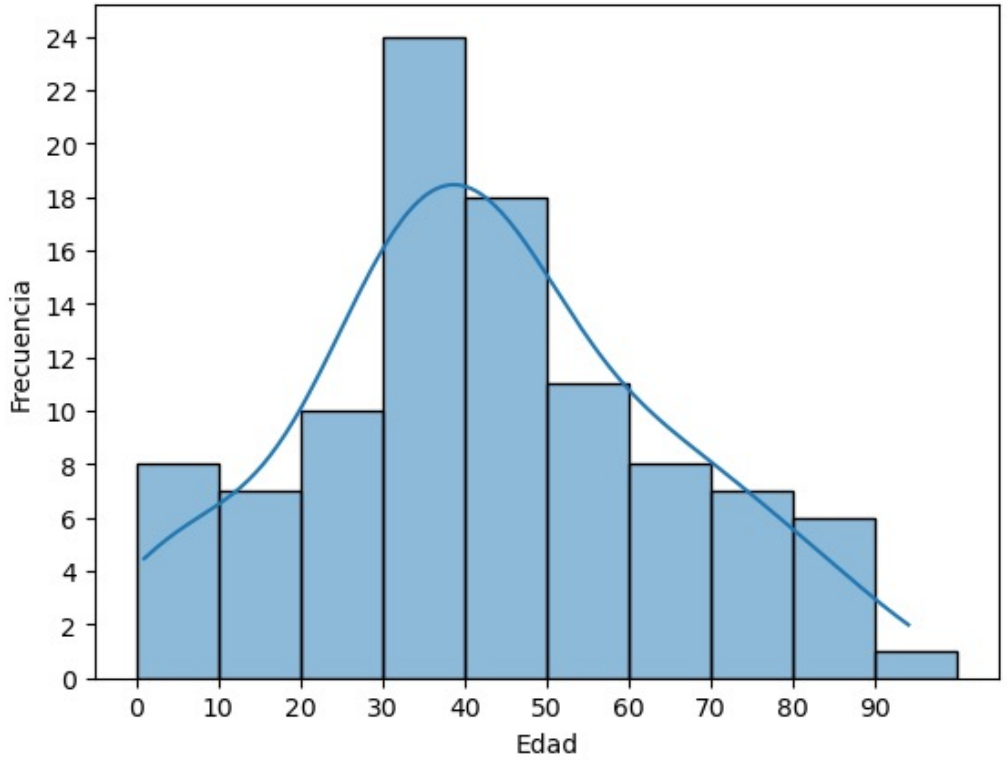
RESUMEN DE DATOS

Imagine que ahora tomamos una muestra de 100 de casos de muertes por covid-19 y representamos la distribución de casos por edad.

Muertes por rango de edad



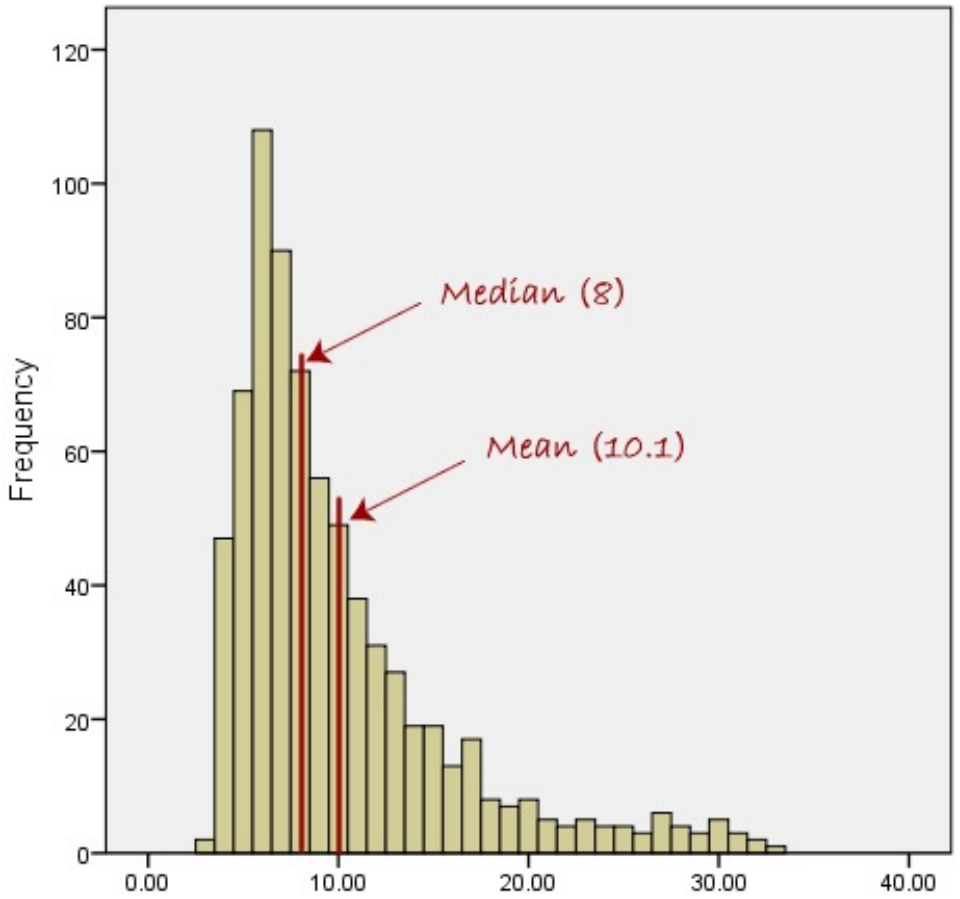
Casos por rango de edad



¿Qué nos muestran los datos con respecto a la dinámica de la pandemia de acuerdo a los rango de edad?

MEDIDAS DE TENDENCIA CENTRAL

- Las medidas de tendencia central también son conocidas como medidas de localización.
- El objetivo principal de las medidas de tendencia central es dar una idea de cuál es el valor típico o común de una variable determinada.
- Las tres medidas de tendencia central más comunes son la media aritmética, la mediana y la moda.



MEDIDAS DE DISPERSION

Pontificia Universidad JAVERIANA Bogotá

Media Aritmética

Es la suma de todos los datos dividida entre la cantidad de datos. Calculemos el promedio de la nota de 7 estudiantes: 2 3 3 3 4 4 5

$$\overline{X} = \frac{\sum x_i}{n} = \frac{2+3+3+3+4+4+5}{7} = \frac{24}{7} = 3.43$$

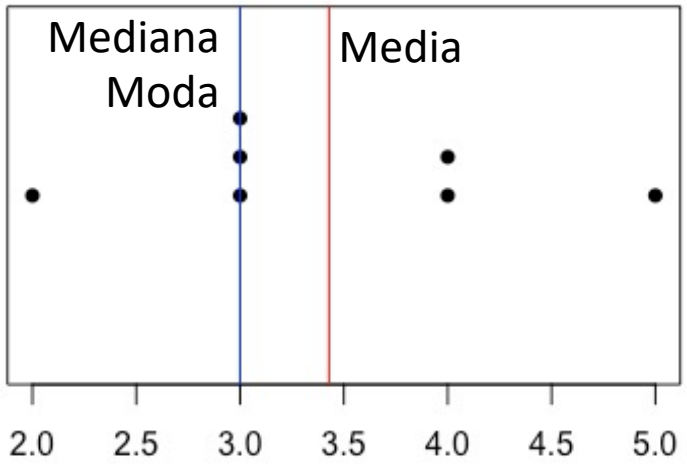
Mediana

Es el dato que divide la muestra en dos partes iguales: 2 3 3 3 4 4 5.

$$\tilde{X} = 3.0$$

La regla general para calcular la mediana:

$$\tilde{X} = \begin{cases} x_{(n+1)/2}, & \text{si } n \text{ es impar}, \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right), & \text{si } n \text{ es par}. \end{cases}$$



Moda

Es el dato de la muestra que más se repite: 2 3 3 3 4 4 5

Moda = 3.0

MEDIA VS MEDIANA

- Existen algunos casos donde la media aritmética “falla” como medida de tendencia central.
- La media es sensible a los sesgos fuertes y los valores extremos.
- Considere los siguientes datos:
  - X = [90, 95, 95, 100, 105]
  - Y = [90, 95, 95, 100, 500]
- Al calcular los promedio de X y Y

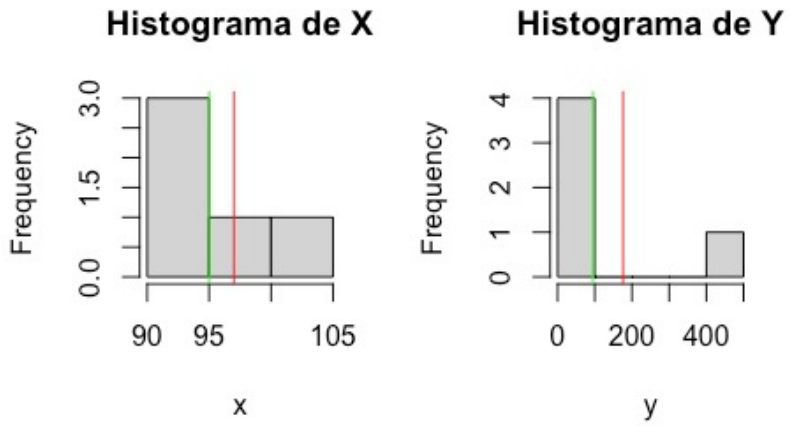
$$\overline{X} = \frac{90 + 95 + 95 + 100 + 105}{5} = \frac{485}{5} = 97$$

$$\overline{Y} = \frac{90 + 95 + 95 + 100 + 500}{5} = \frac{880}{5} = 176$$

- Al calcular la mediana de X y Y

X = [90, 95, 95, 100, 105] -> $\tilde{X} = 95$

Y = [90, 95, 95, 100, 500] -> $\tilde{Y} = 95$

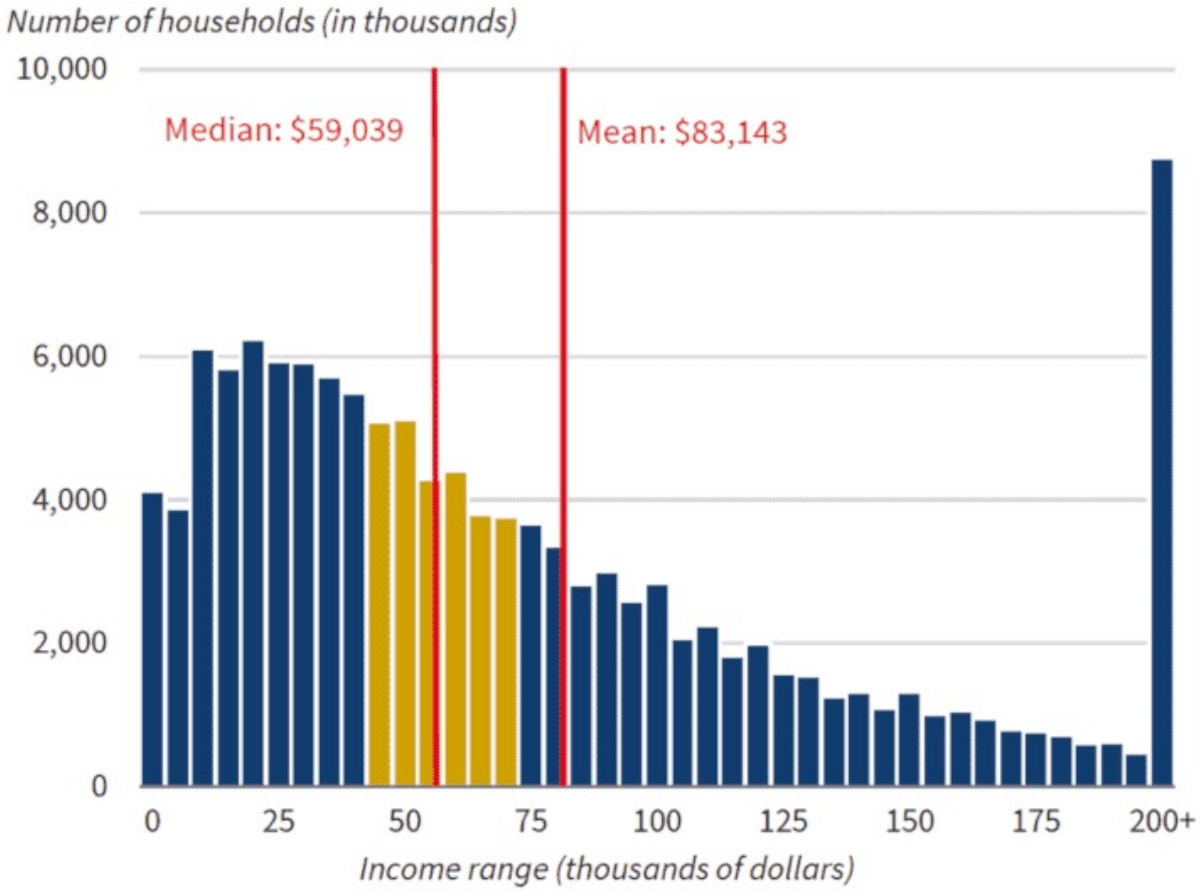


Piense en casos reales donde se pueden presentar sesgo o valores extremos

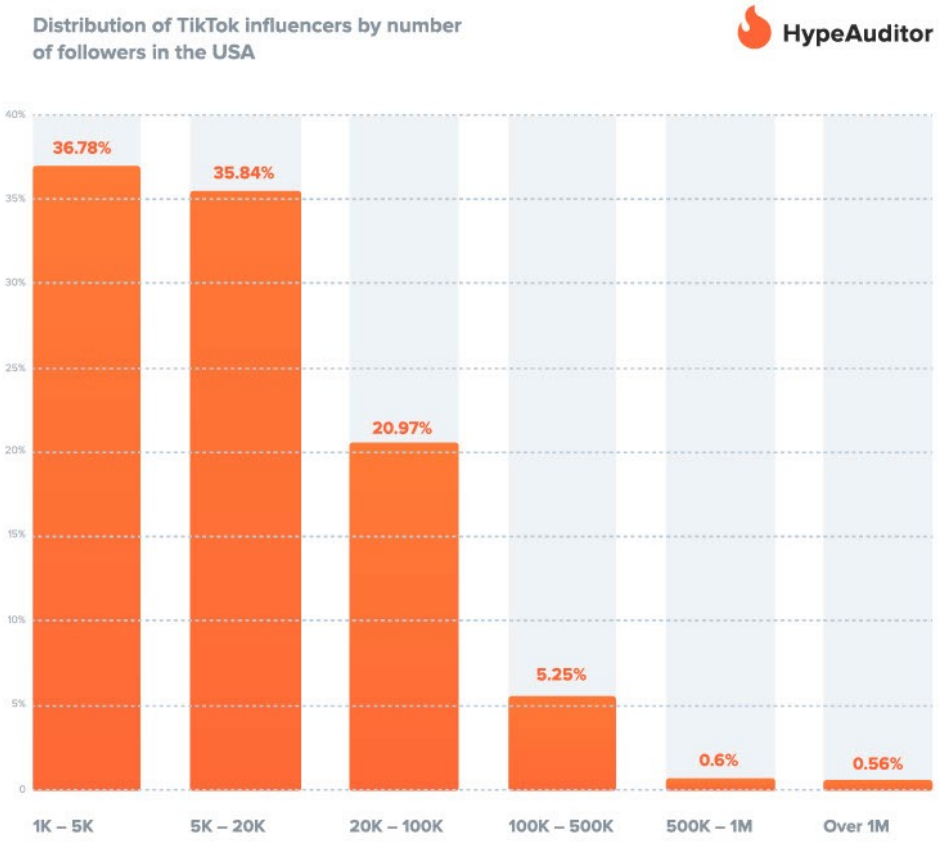
CASOS DE SESGO

Pontificia Universidad JAVERIANA Bogotá

Figure 3-i. U.S. Household Income Distribution in 2016



Source: U.S. Census Bureau.
Note: Middle quintile range (shaded yellow) is an estimate. Data are right-censored at $200,000.



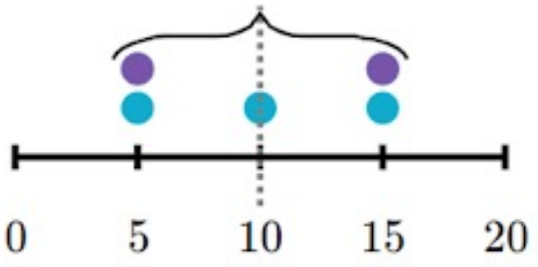
MEDIDAS DE DISPERSIÓN

- La dispersión se refiere a la discrepancia o diferencia en los valores o resultados de una muestra.
- Por esta razón, las medidas de dispersión se denominan medidas de variabilidad.
- Las medidas de dispersión son:
  - Rango: R = mayor valor – menor valor
  - Varianza: $S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$
  - Desviación estándar: $S = \sqrt{S^2}$

MEDIDAS DE DISPERSIÓN

Pontificia Universidad JAVERIANA Bogotá

Muestra A



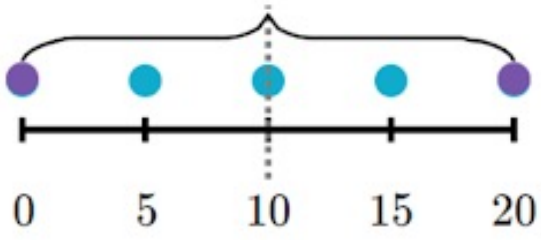
$$\text{Media}: \overline{X}_A = \frac{5+5+10+15+15}{5} = 10$$

$$\text{Rango}: R_A = 15 - 5 = 10$$

$$\text{Varianza}: S_A^2 = \frac{(5-10)^2+(5-10)^2+(10-10)^2+(15-10)^2+(15-10)^2}{5-1} = \frac{100}{4} = 25$$

$$\text{Desv. Estándar}: S_A = \sqrt{25} = 5.0$$

Muestra B



$$\text{Media}: \overline{X}_B = \frac{0+5+10+15+20}{5} = 10$$

$$\text{Rango}: R_B = 20 - 0 = 20$$

$$\text{Varianza}: S_B^2 = \frac{(0-10)^2+(5-10)^2+(10-10)^2+(15-10)^2+(20-10)^2}{5-1} = \frac{250}{4} = 62.5$$

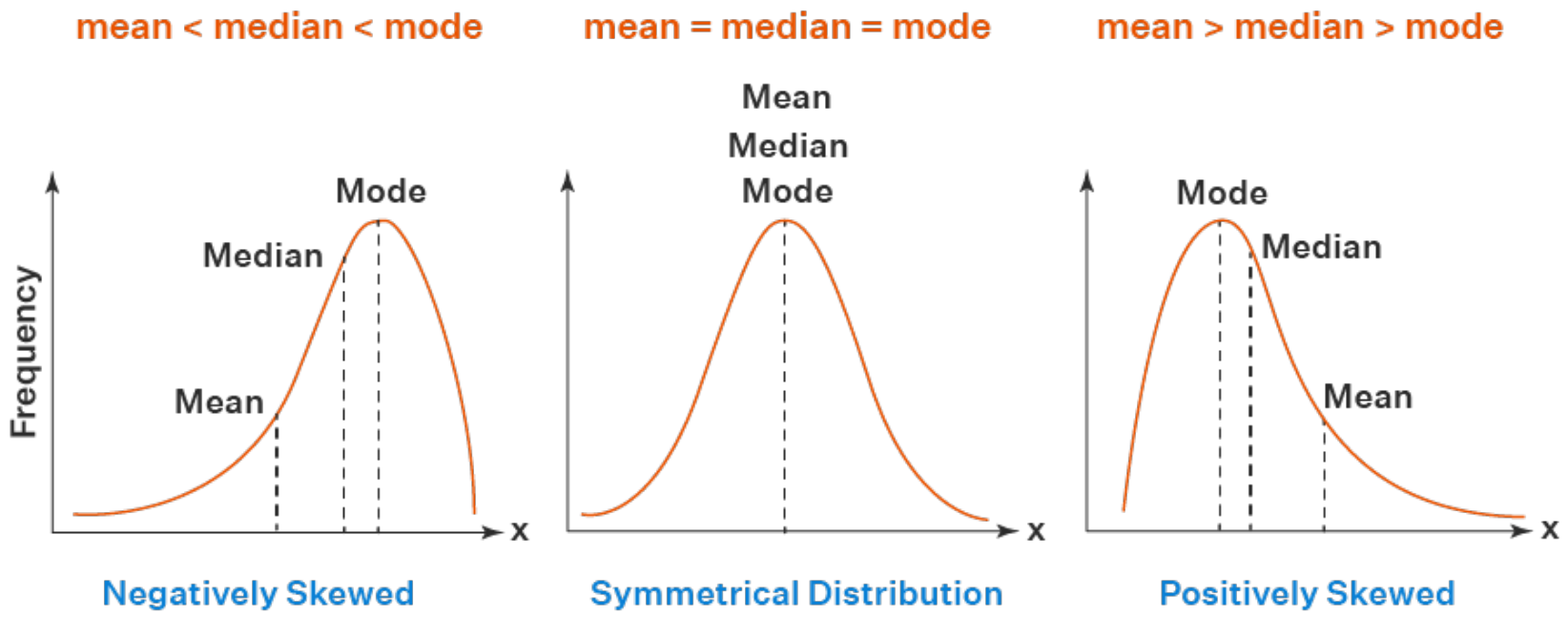
$$\text{Desv. Estándar}: S_B = \sqrt{62.5} = 7.9$$

Observe que en la muestra B los datos están más dispersos en comparación a los datos de la muestra. Esto a su vez se refleja en las medidas de dispersión.

Sesgo - Skewness

- Como ya se analizó con el histograma y las medidas de tendencia central un aspecto relevante en el análisis de un conjunto de datos o una variable es estudiar la forma de su distribución.
- Por ello, el histograma se complementa con las mediciones de sesgo y curtosis.
- Sesgo: es una media del sesgo que puede existir en la muestra.
  - El signo del sesgo indica el lado donde la cola de la distribución es más larga.
  - Hacia la derecha (signo −)
  - Hacia la izquierda (signo +)

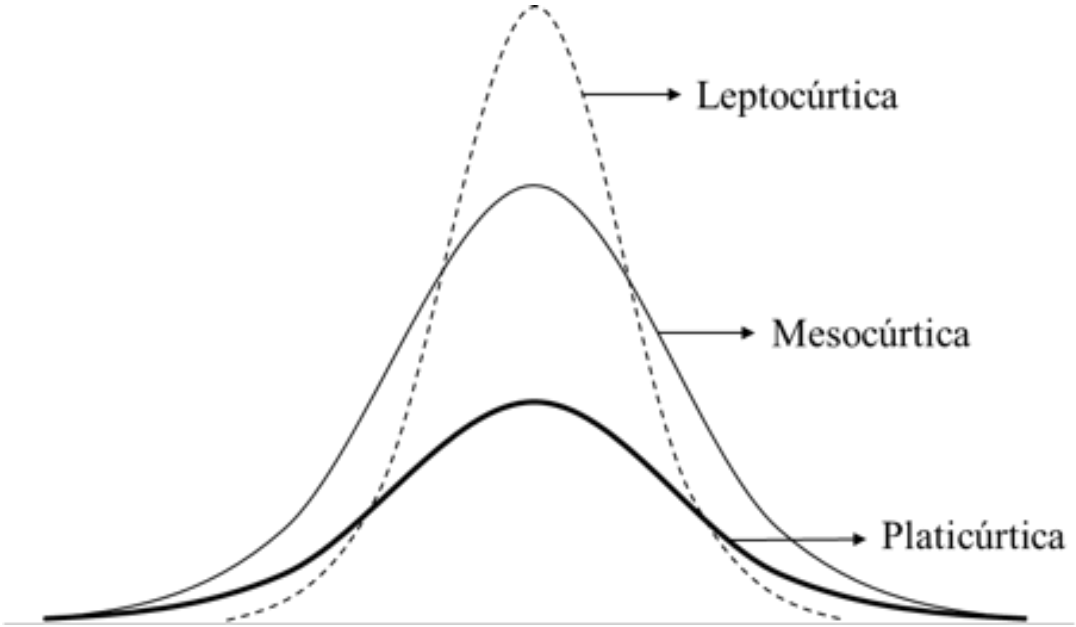
$$\text{Sesgo} = \frac{1}{n}\frac{\sum_i^n (x_i - \bar{x})^3}{S^3}$$



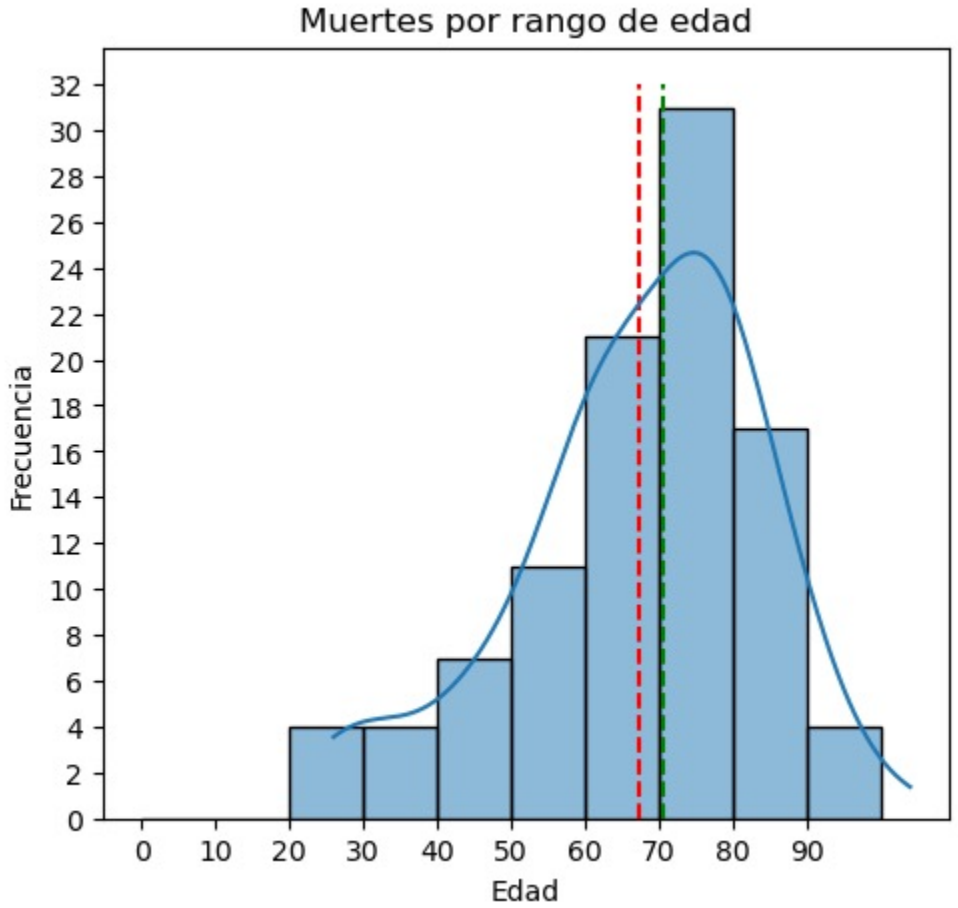
Curtosis

- Curtosis: Es una medida para determinar qué tan elevada o plana es la distribución de la muestra.
- Las distribuciones con una curtosis inferior a 3 se denominan platicúrticas, aunque esto no implica que la distribución sea "plana".
- Más bien significa que la distribución produce menos valores atípicos y menos extremos que la distribución normal.
- Las distribuciones con una curtosis superior a 3 se denominan leptocúricas. Lo que indica que se valores se agrupan mayormente hacia la tendencia central.

$$\text{Curtosis} = \frac{1}{n}\frac{\sum_i^n (x_i - \bar{x})^4}{S^4}$$



EJEMPLO COVID-19

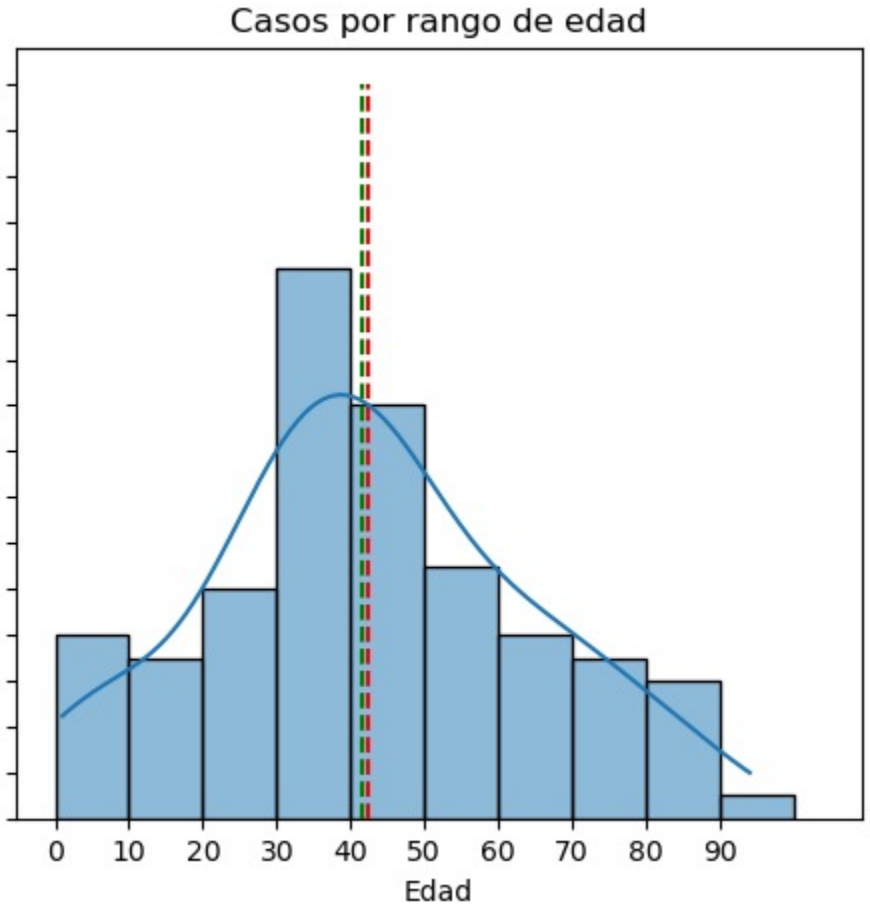


Media = 67.4

Mediana = 70.5

Sesgo = -0.665

Curtosis = 3.24



Media = 42.5

Mediana = 41.5

Sesgo = 0.154

Curtosis = 2.63

MEDIDAS DE POSICIÓN

- Las medidas de posición se utilizan para determinar puntos de referencias que dividen la muestra ordenada en partes iguales.
- Los más utilizados son los:
  - Cuartiles (4 partes iguales)
  - Quintiles (5 partes iguales)
  - Deciles (10 partes iguales)
  - Percentiles (100 partes iguales).
- En estadística, un cuantil es una puntuación por debajo de la cual cae un determinado porcentaje de los valores de la muestra.
- Existen diferentes formas de calcular estas medidas, a continuación ilustramos una de estas.

Suponga que se tiene una muestra de 10 salarios anuales (millones) en una organización:

28, 30, 35, 38, 39, 46, 50, 52, 53, 70

Los cuartiles de la muestra:

Q2(50%) = (39 + 46)/2 = 42.5

Q1(25%) = 35

Q3(75%) = 52

Los quintiles:

Q1(20%) = (30 + 35)/2 = 32.5

Q2(40%) = (38 + 39)/2 = 38.5

Q3(60%) = (46 + 50)/2 = 48

Q4(80%) = (52 + 53)/2 = 52.5

USO DE LOS CUANTILES

- Los cuantiles se utilizan para posicionar o rankear las entidades o individuos dentro de una muestra o población.

- Esto permite tener una idea de la posición relativa dentro de la muestra.

GOBIERNO DE COLOMBIA · MINEDUCACIÓN · icfes mejor saber

REPORTE DE RESULTADOS DE ESTUDIANTES
•SABER PRO•

Fecha de aplicación
Fecha de publicación de resultados
Número de registro
Nombres y apellidos
Identificación
Institución: CORPORACION UNIVERSITARIA MINUTO DE DIOS -UNIMINUTO- BOGOTÁ D.C.
Código SNIES: 91237
Programa

PUNTAJE GLOBAL

De 300 puntos posibles, su puntaje global es: **116**

GRUPO REFERENCIA

ADMINISTRACIÓN Y AFINES

¿EN QUÉ PERCENTIL ME ENCUENTRO?

Con respecto a los estudiantes del país, usted está aquí.

MÓDULOS COMPETENCIAS GENÉRICAS

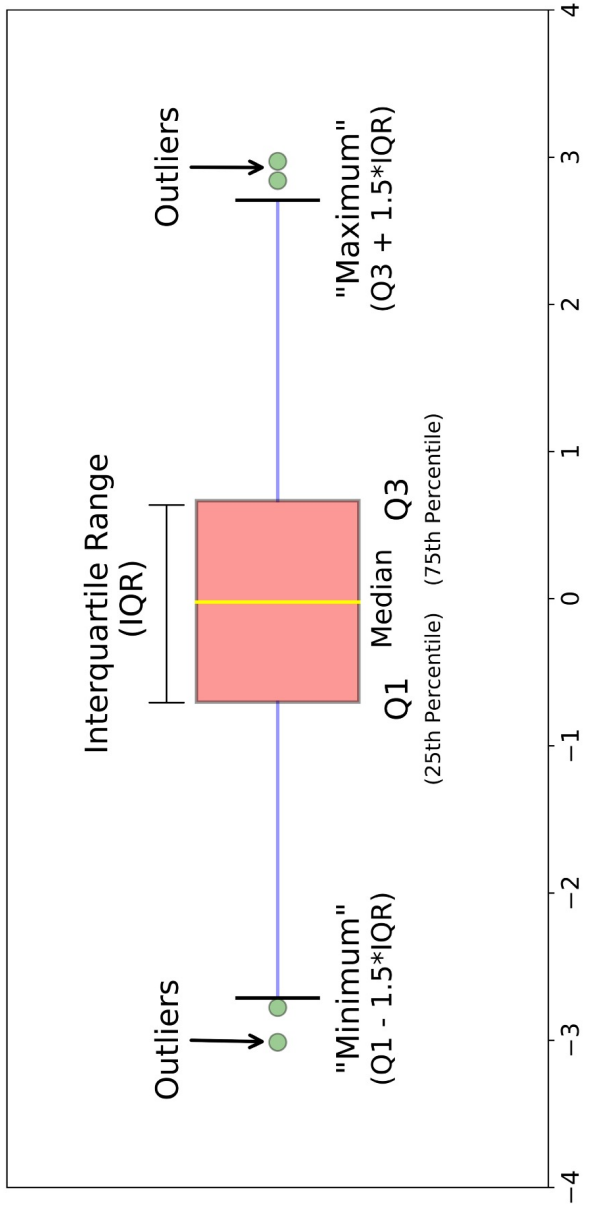
Módulos	De 300 puntos posibles, su puntaje global es	¿EN QUE PERCENTIL ME ENCUENTRO?
Comunicación Escrita[1]	--	
Razonamiento Cuantitativo	**123**	24 / 22
Lectura Crítica	**176**	86 / 78
Competencias Ciudadanas	**132**	46 / 40
Inglés	**150**	61 / 55

MÓDULOS COMPETENCIAS ESPECÍFICAS

Módulos	De 300 puntos posibles, su puntaje global es	¿EN QUE PERCENTIL ME ENCUENTRO?
Gestión De Organizaciones	**159**	73
Formulación, Evaluación Y Gestión De Proyectos	**173**	81
Gestión Financiera	**98**	6

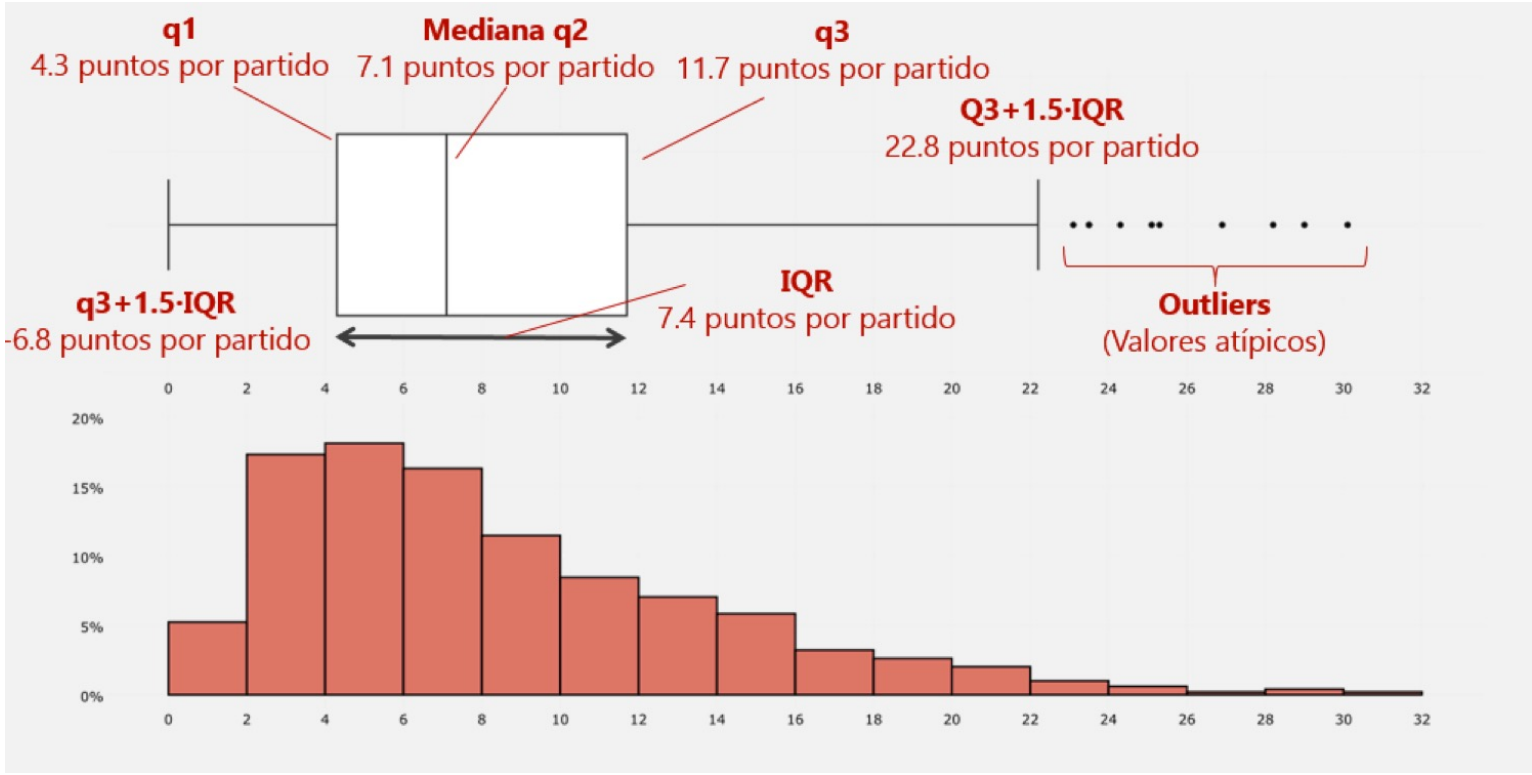
GRÁFICOS DE CAJA

- El diagrama de caja se basa en los cuartiles y divide los datos ordenados en cuatro grupos, cada uno con el 25%.
- De esta forma es posible visualizar dónde termina de acumularse 25% de los datos menores, y a partir de donde se localiza 25% de los datos mayores.
- Entre estos dos cuartiles se ubica el 50% de los datos que están al centro.
- Además de los cuartiles están involucrados los siguientes conceptos:
  - Rango intercuartílico: IQR = Q3 – Q1
  - Barrera interior izquierda: Q1 – 1.5IQR
  - Barrera interior derecha: Q1 + 1.5IQR

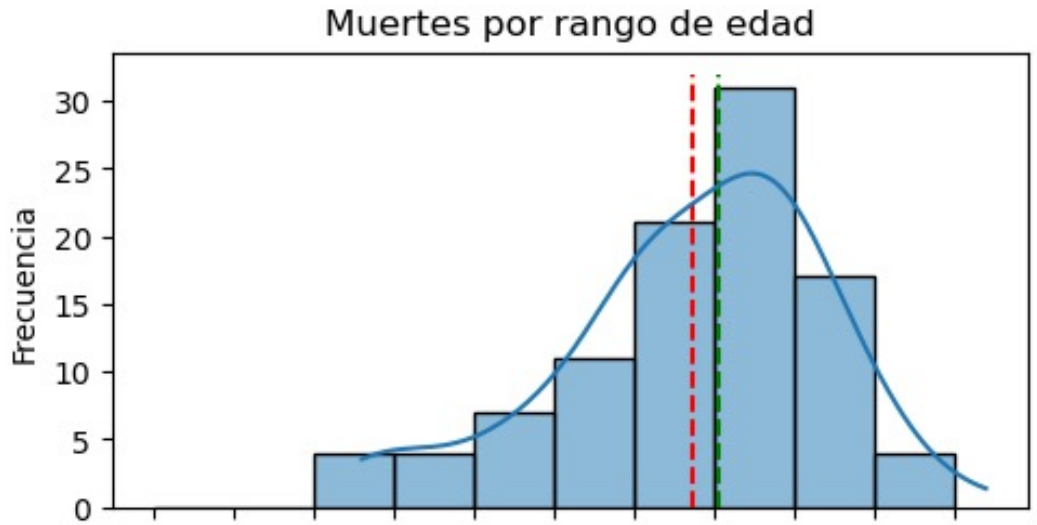


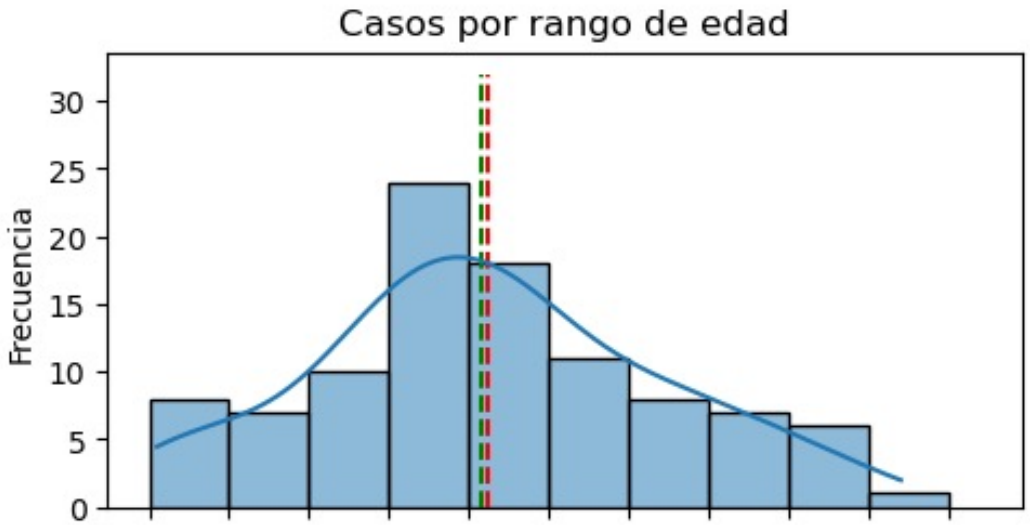
VALORES EXTREMOS

- Usualmente los valores atípicos se localizan de la siguiente forma:
  - Puntos por fuera de los bigotes en el gráfico de cajas.
  - Valores abajo y arriba de los percentiles 1% y 99%.
  - Aunque estos criterios pueden variar de acuerdo al contexto y la distribución de los datos.

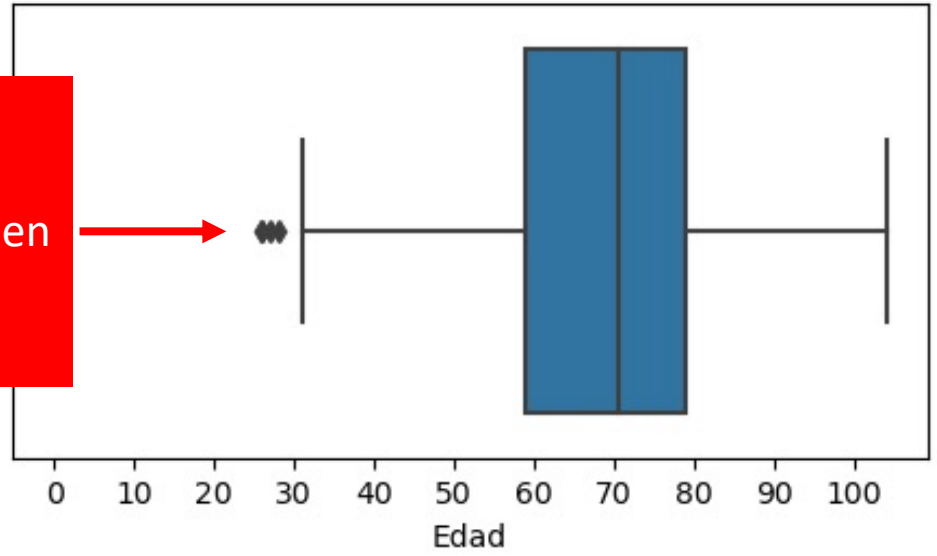


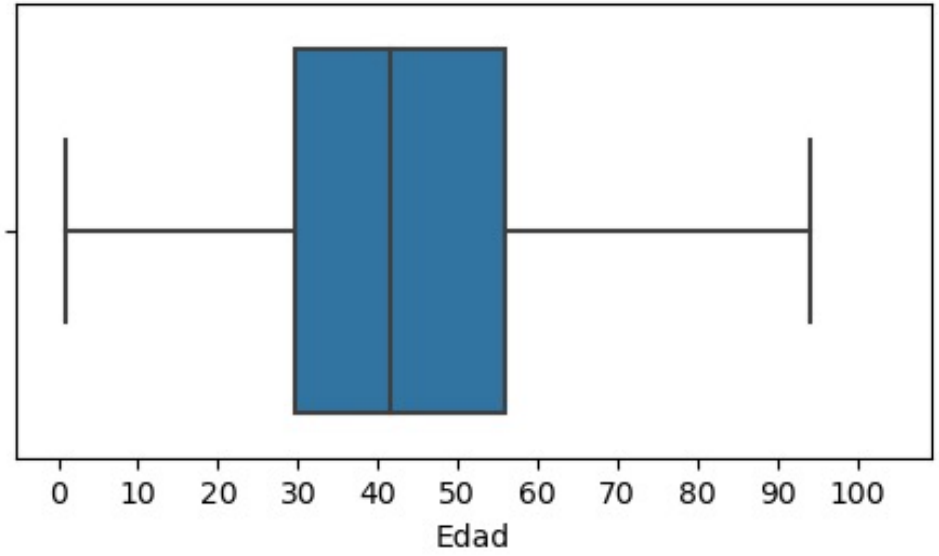
EJEMPLO COVID-19





Desde el contexto del problema tienen sentido estos datos atípicos



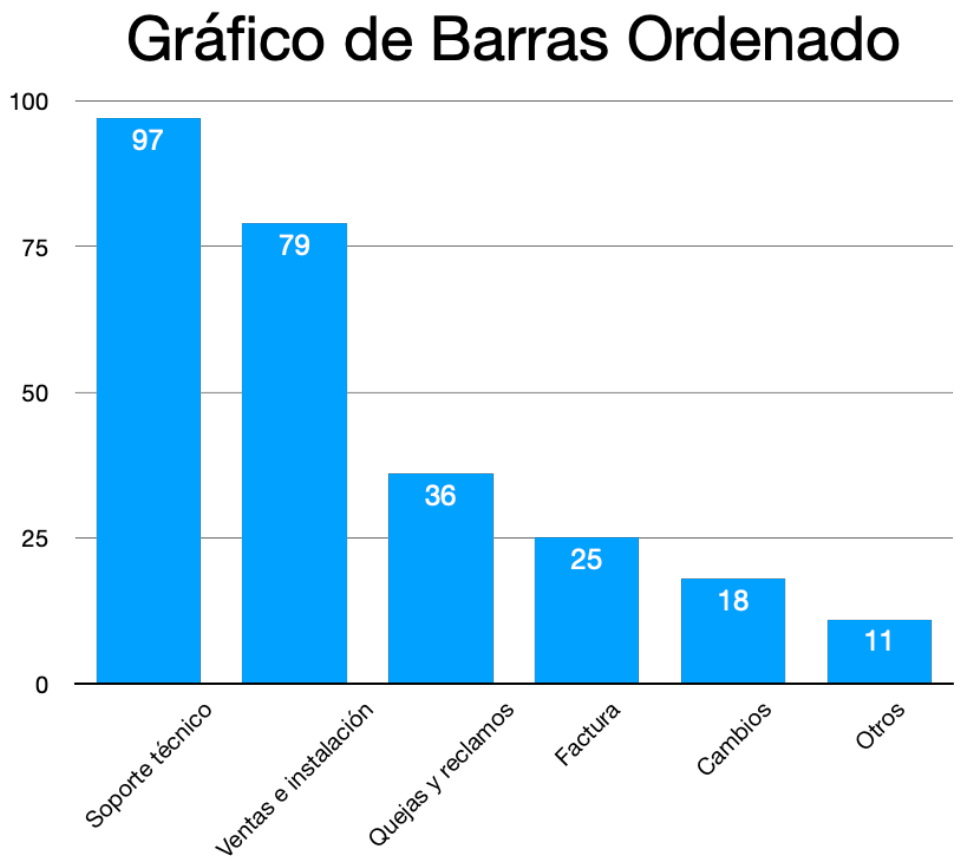


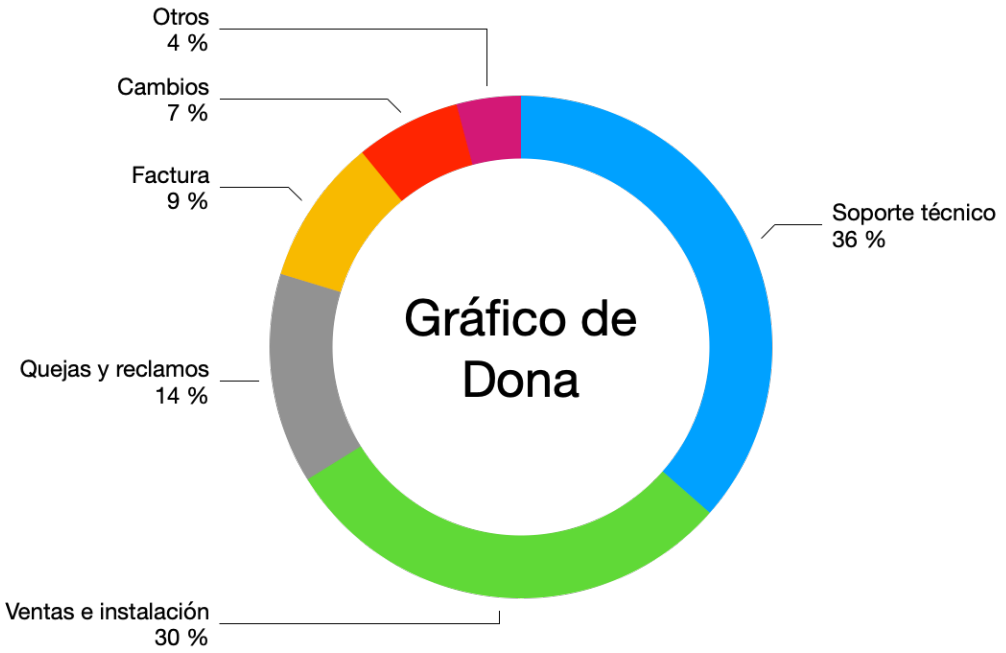
VARIABLES CUALITATIVAS

- Cuando se tienen variables categóricas se utilizan las tablas de frecuencia para generar estadísticas, tales como:
  - Sumas
  - Sumas acumuladas
  - Porcentajes
  - Porcentajes acumulados
- Las tablas de frecuencia suministran información clave para:
  - Analizar las categorías
  - Su importancia relativa
  - Generar gráficos
- Cuando la variable categórica es nominal, es recomendable generar un ordenamiento según su frecuencia.
- Este tipo de análisis se conoce como Pareto, el cual permite identificar las categorías principales bajo la regla 80%/20%

Requerimiento	Total	Porcentaje	Porcentaje Acumulado
Soporte técnico	97	36,47 %	36,47 %
Ventas e instalación	79	29,70 %	66,17 %
Quejas y reclamos	36	13,53 %	79,70 %
Factura	25	9,40 %	89,10 %
Cambios	18	6,77 %	95,86 %
Otros	11	4,14 %	100,00 %
Total	266	100,00 %	

GRÁFICOS





Los gráficos de barras se utilizan cuando se quiere presentar las categorías como un ranking de acuerdo a su importancia o peso

Los gráficos de dona o torta se utiliza cuando se quiere representar el peso o porcentaje de las categorías con respecto al todo.

¿PREGUNTAS?