ESTADISTICA DESCRIPTIVA
II PARTE

METODOS GRAFICOS PARA DESCRIBIR VARIABLES CUALITATIVAS

Gráficos para variables cualitativas.

Una vez que conocemos la distribución de la variable, nos interesa presentarla de alguna manera gráfica, uno de los gráficos o diagramas más usados en variables cualitativas son los diagramas sectoriales o de torta y los gráficos de barra.

METODOS GRAFICOS PARA DESCRIBIR VARIABLES CUALITATIVAS

Un **gráfico sectorial (o de torta)** muestra la distribución de una variable cualitativa dividiendo un círculo en partes que corresponden a las categorías de la variable, tal que el tamaño (ángulo) de cada pedazo es proporcional al porcentaje de ítems en cada categoría.

Un **gráfico de barras** muestra la distribución de una variable cualitativa listando las categorías o valores de la variable en el eje X y dibujando una barra sobre cada categoría. La altura de la barra es igual al porcentaje de ítems en esa categoría. Las barras deben tener el mismo ancho.

METODOS GRAFICOS PARA DESCRIBIR VARIABLES CUALITATIVAS

Gráfico sectorial.

Figura 1 (a):
Diagrama sectorial con 1/4 de los ítems que comparten alguna propiedad.

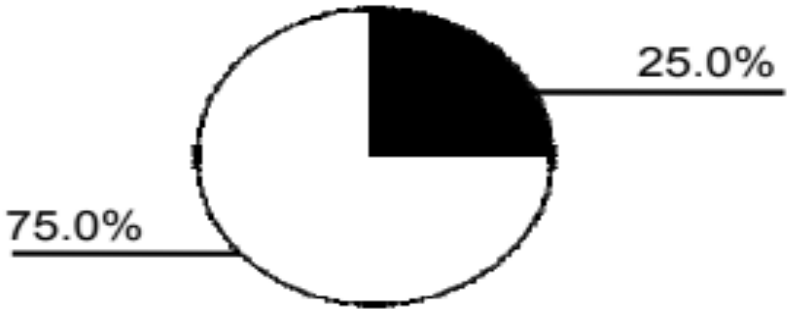


Figura 1 (b):
Diagrama sectorial con 7/8 de los ítems que comparten alguna propiedad

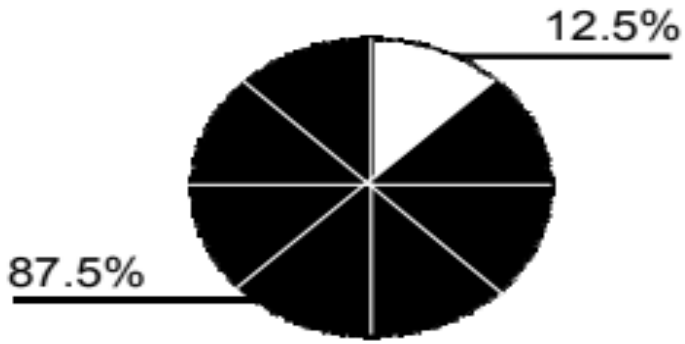
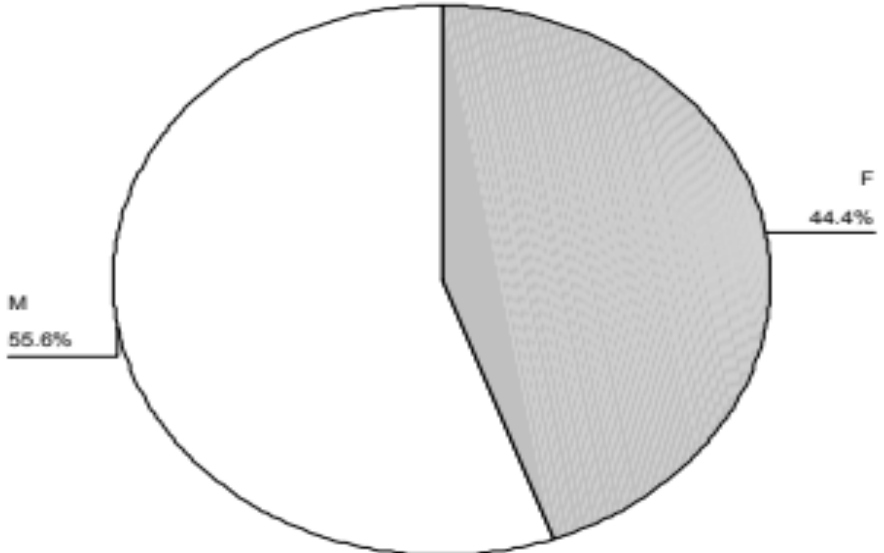


Diagrama sectorial para la variable SEXO de base de datos 1

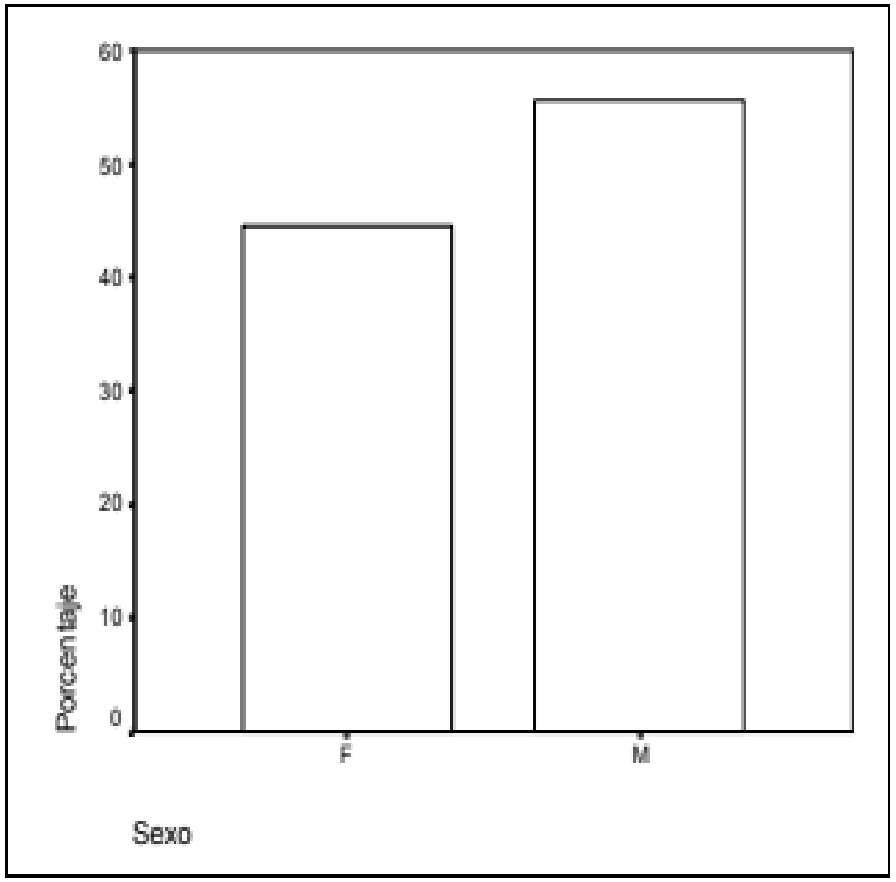


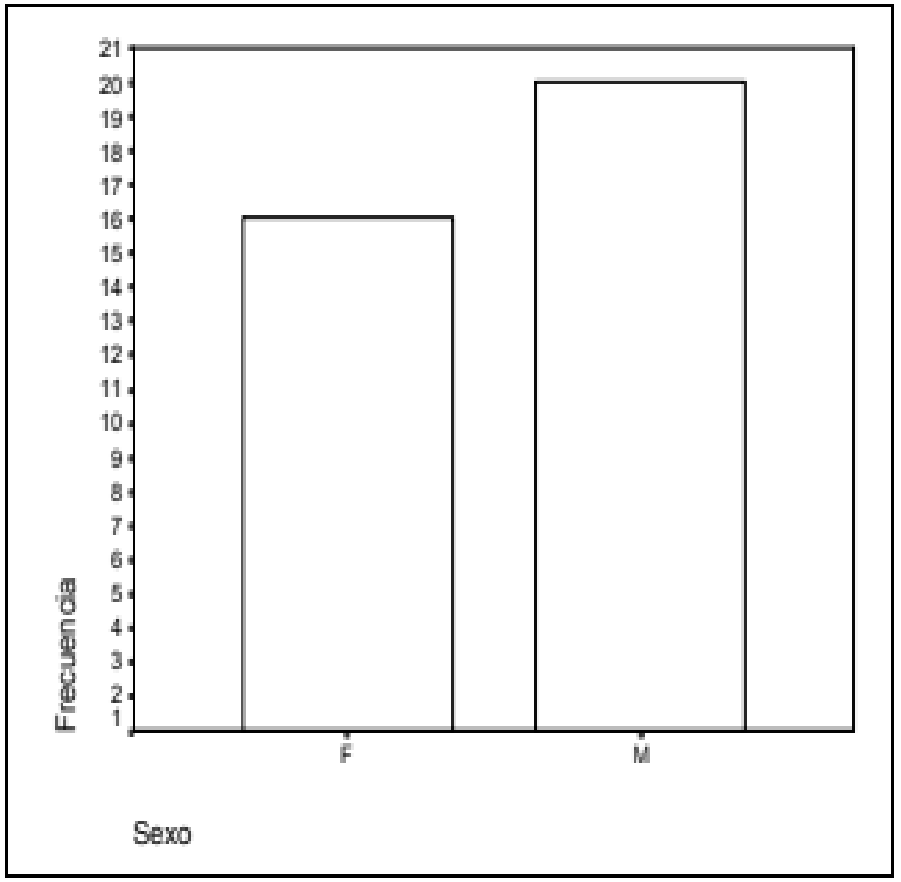
METODOS GRAFICOS PARA DESCRIBIR VARIABLES CUALITATIVAS

Gráfico de barras

Compare los siguientes gráficos. ¿Cuáles son las diferencias?

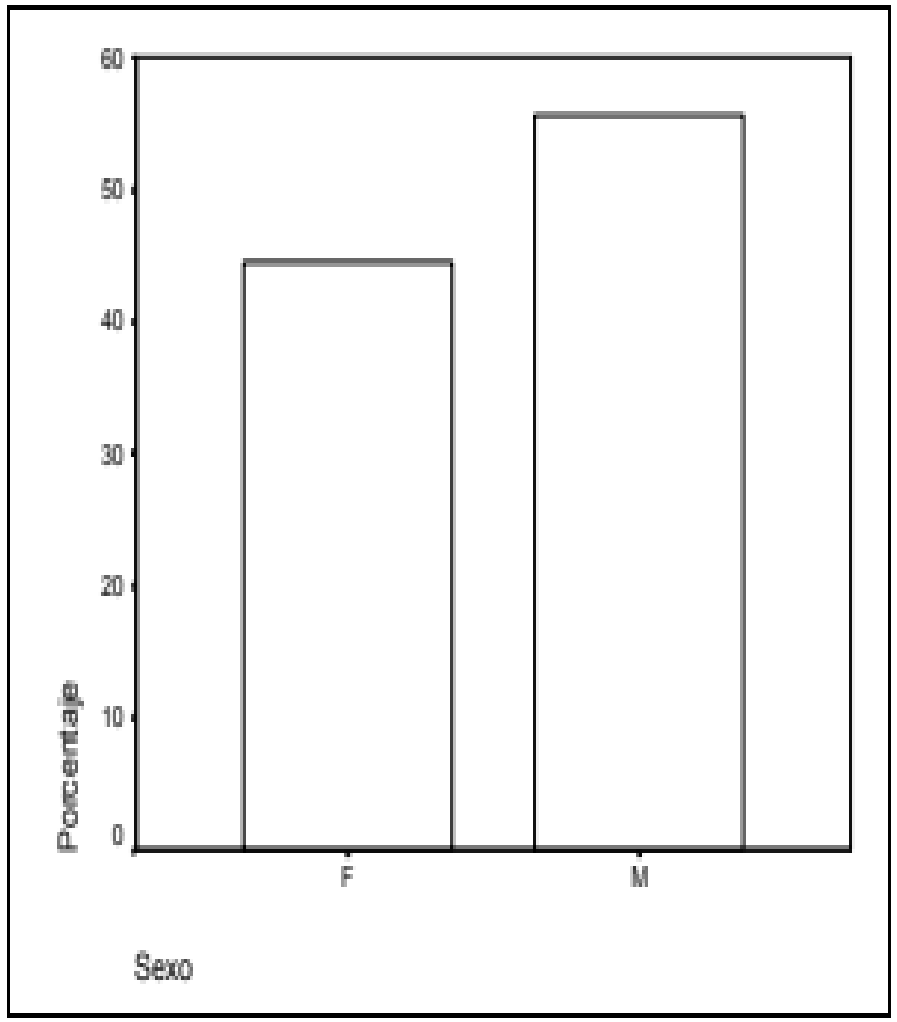
Gráfico de barras: Sexo en la base de datos 1.

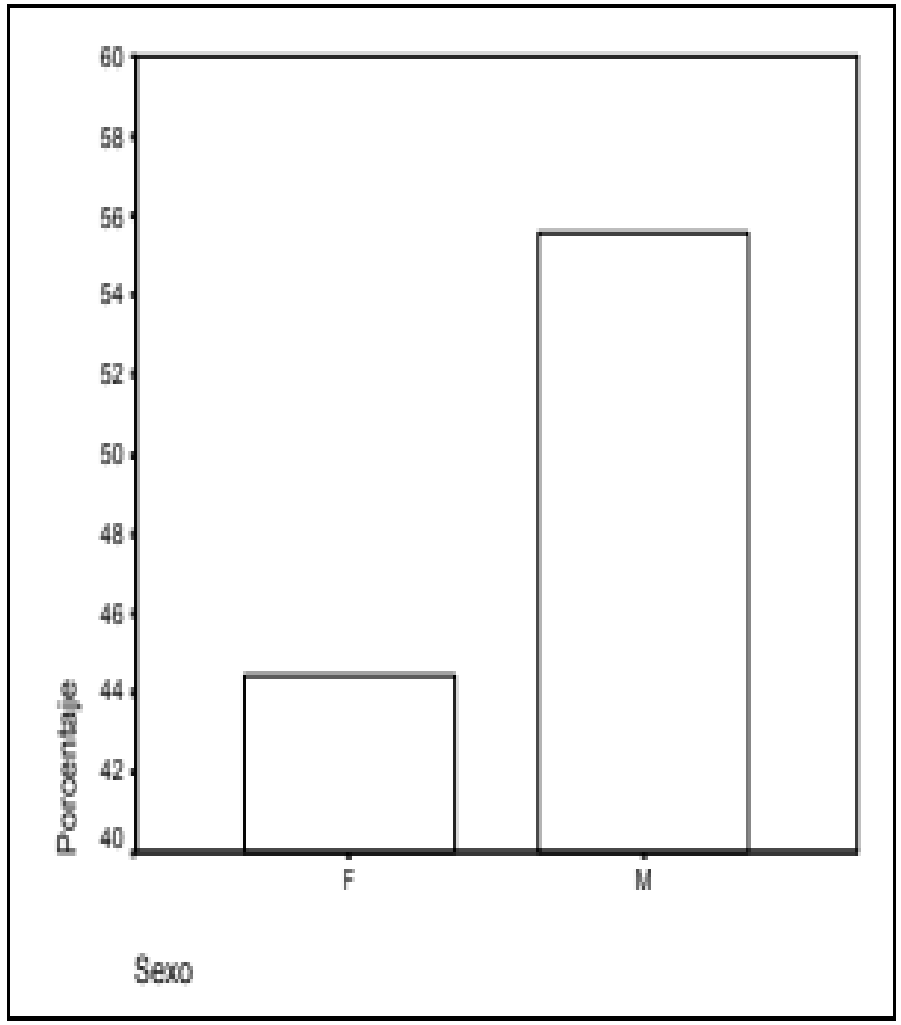




METODOS GRAFICOS PARA DESCRIBIR VARIABLES CUALITATIVAS

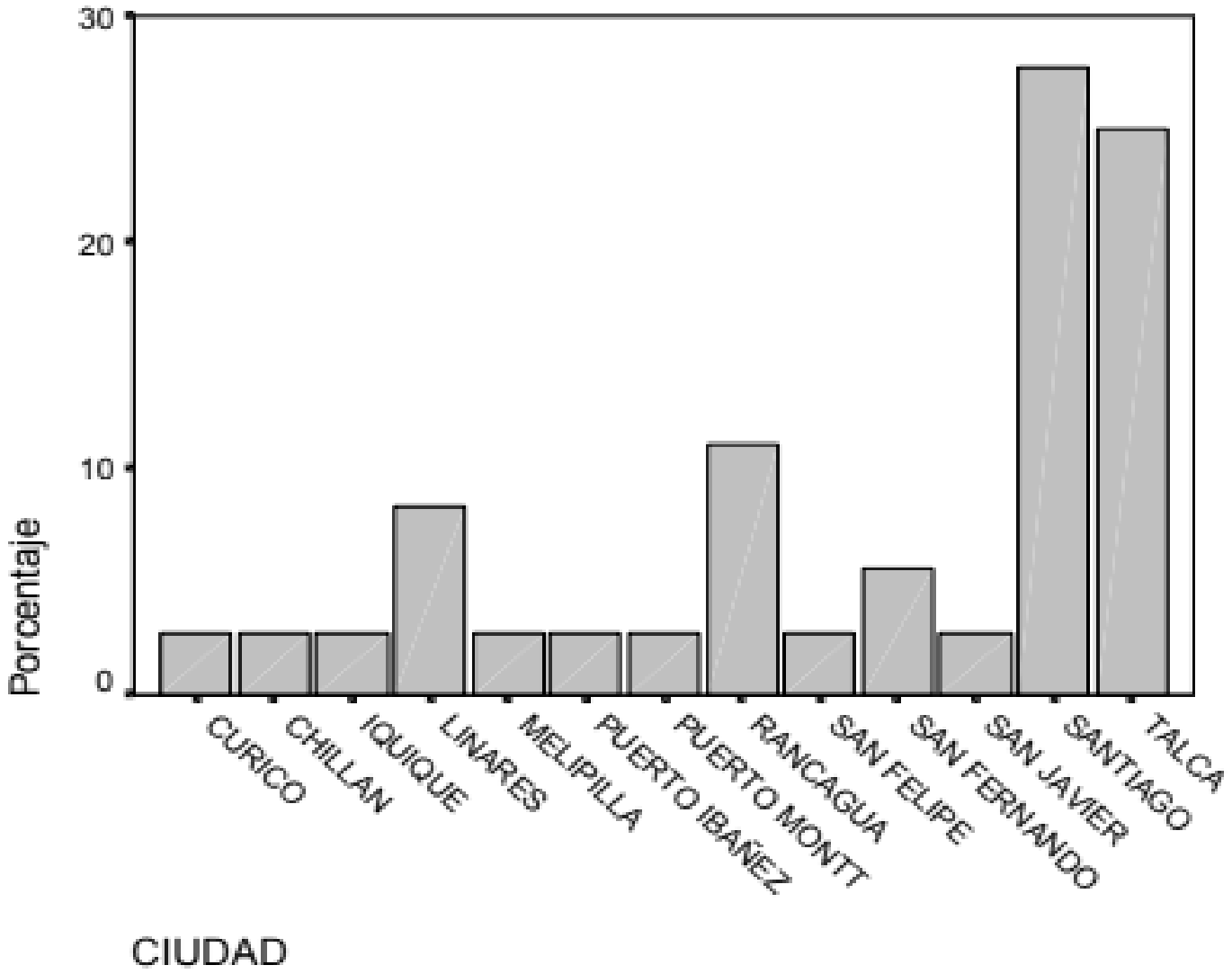
Compare los siguientes gráficos. ¿Cuáles son las diferencias?





METODOS GRAFICOS Y NUMERICOS PARA DESCRIBIR VARIABLES CUALITATIVAS

Gráfico de Barras: Ciudad de procedencia de alumnos de base de datos 1.



Métodos gráficos para describir variables cuantitativas.

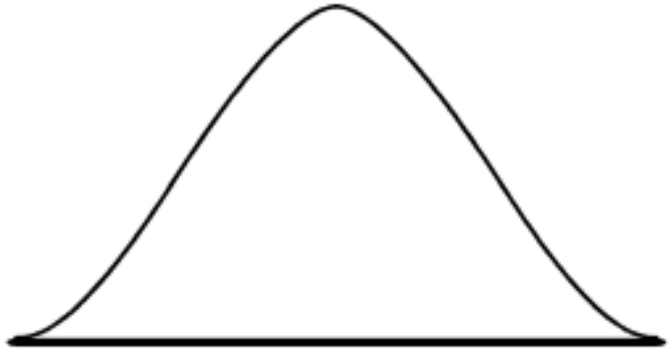
En esta sección veremos de qué manera podemos describir gráficamente las variables cuantitativas.

Veremos 2 tipos de gráficos:

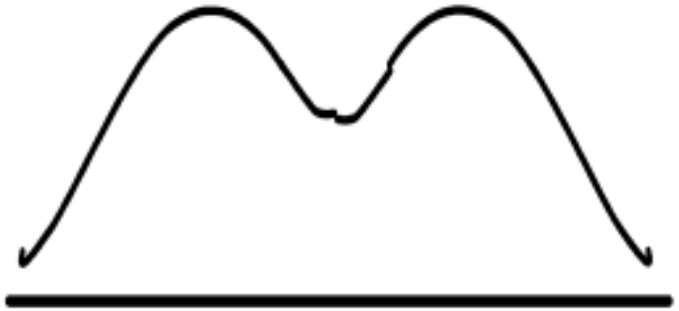
1. Gráfico de puntos.
2. Histograma.

1. Gráfico de Puntos

Formas de Distribuciones



Simétrica, acampanada, unimodal



Bimodal

Sesgada a la derecha (sesgo positivo)

Sesgada a la izquierda (sesgo negativo)

Uniforme

1. Gráfico de Puntos

Los términos usados para describir la forma de una distribución son:

• Simétrica: La distribución puede ser dividida en dos partes alrededor de un valor central y cada parte es el reflejo de la otra.

• Sesgada: Un lado de la distribución se alarga más que el otro. La dirección del sesgo es la dirección del lado más largo.

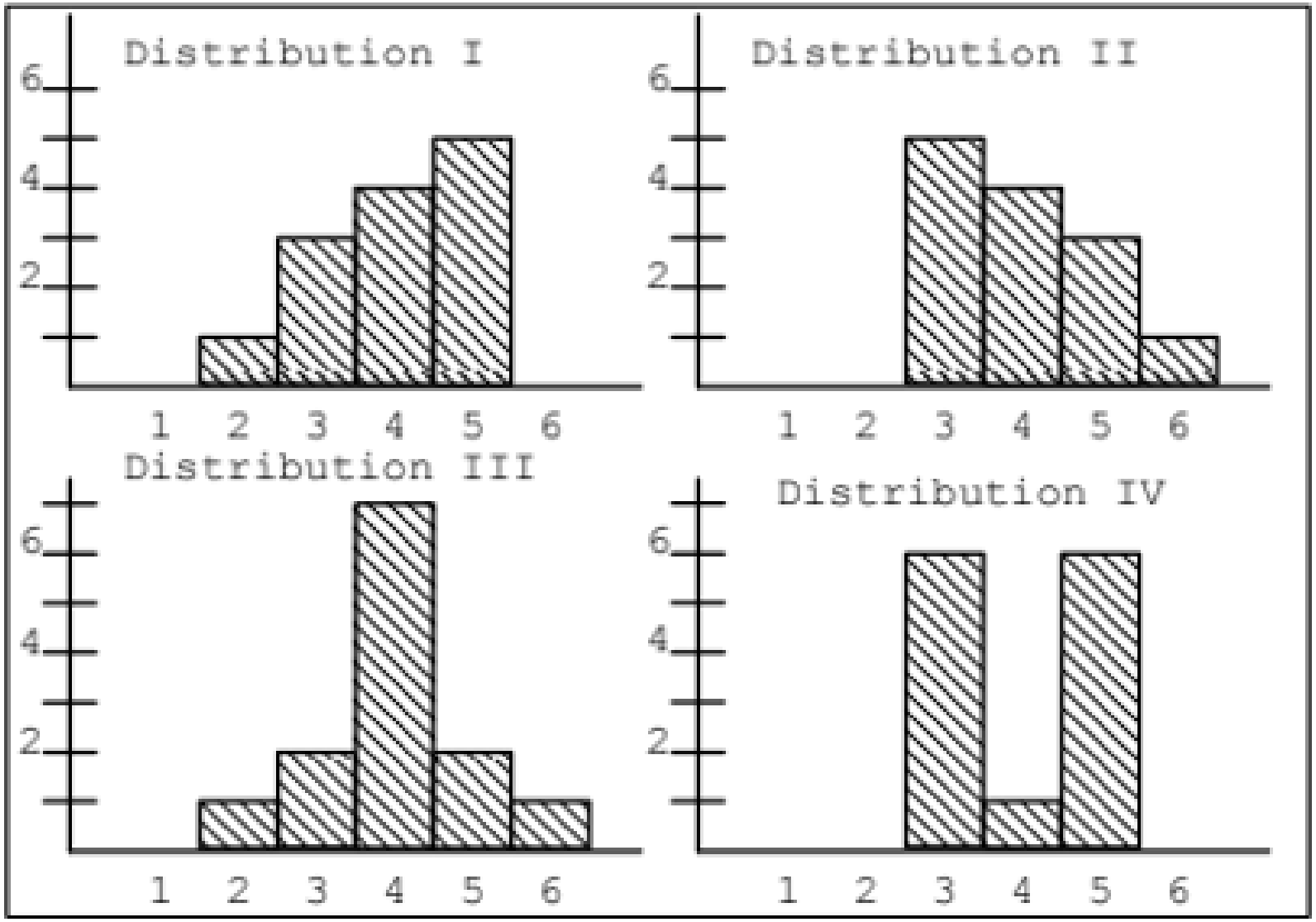
• Unimodal: La distribución tiene un único máximo que muestra el o los valores más comunes en los datos.

• Bimodal: La distribución tiene dos máximos. Esto resulta a menudo cuando la muestra proviene de dos poblaciones.

• Uniforme: Los valores posibles tienen la misma frecuencia.

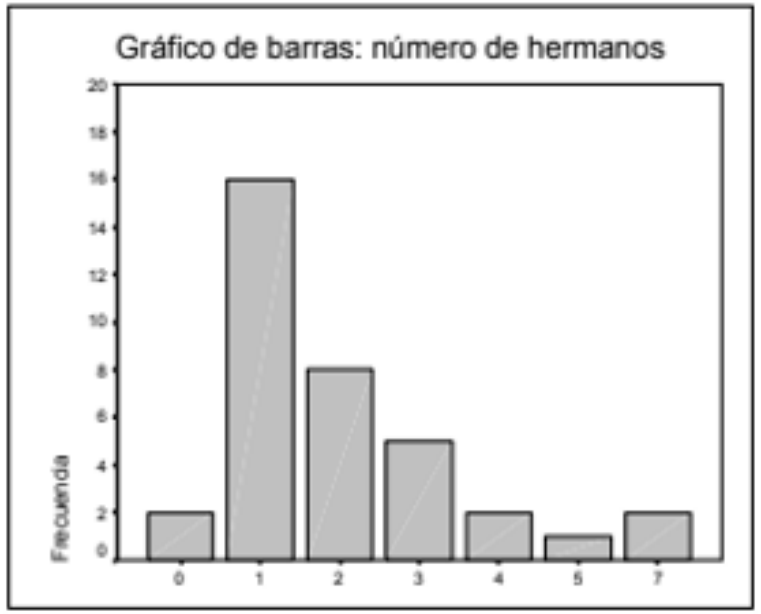
2. Histograma

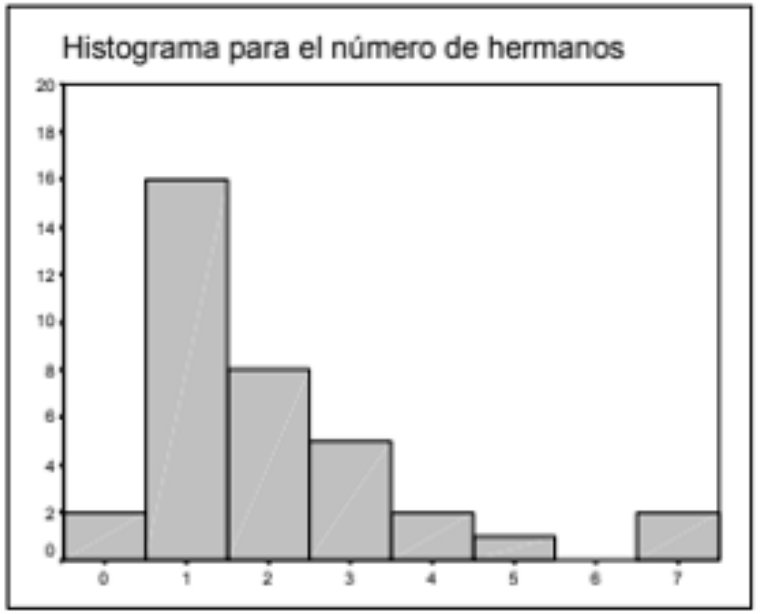
Los histogramas son otra manera de mostrar la distribución de una variable cuantitativa.



2. Histograma

Cuidado con usar gráficos de barras para variables cuantitativas:





Guía para gráficos, figuras o diagramas:

Hay ciertos errores comunes que aparecen en gráficos que pueden hacer que se mal interprete la información. Cuando construya gráficos:

- Póngale un título apropiado.
- Incluya la fuente de los datos o cualquier información relevante.
- Escriba el nombre de la variable que se describe en los ejes.
- Incluya las unidades de medida de las variables.
- Verifique si el eje de la frecuencia, proporción o porcentaje comienza en cero.
- Verifique si los ejes mantienen una escala constante

Métodos Numéricos para describir Variables Cuantitativas

Métodos numéricos para describir variables cuantitativas

En este capítulo, empezamos a organizar y resumir los datos, primero tratamos las variables cualitativas, luego la descripción gráfica de variables cuantitativas, ahora estudiaremos cómo obtener buen resumen numérico de los datos. Específicamente estudiaremos medidas de resumen o medidas descriptivas numéricas que son de tres tipos:

- las que ayudan a encontrar el **centro** de la distribución, llamadas medidas de tendencia central.
- las que miden la **dispersión**, llamadas medidas de dispersión.
- las que describen la **posición relativa** de una observación dentro del conjunto de datos, llamadas medidas de posición relativa.

1. **Medidas de Tendencia Central**.

Las medidas de tendencia central son valores numéricos que quieren mostrar el centro de un conjunto de datos, nos interesan especialmente dos medidas: la ***media*** y la ***mediana***.
Si los datos son una muestra, el promedio y la mediana se llamarán *estadísticas*. Si los datos son una población entonces estas medidas de tendencia central se llamarán *parámetros*.

Una **Estadística** es una medida descriptiva numérica calculada a partir de datos de una muestra.

Un **Parámetro** es una medida descriptiva numérica que usa la totalidad de las unidades de una población.

Métodos Numéricos para describir Variables Cuantitativas

a) **Promedio**.

El **promedio** de un conjunto de *n* observaciones es simplemente la suma de las observaciones dividida por el número de observaciones, *n*.

Promedio de edad de los 20 sujetos en el estudio médico:
Sume las 20 edades y divida por 20:

$$\frac{45+41+51+46+47+\cdots+45+37}{20} = 43{,}35 \text{ años}$$

Notación: Si $x_1, x_2, \ldots, x_n$ denota una muestra de *n* observaciones, entonces el *promedio de la muestra* se llama "x-barra" y se denota por:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad 1$$

Si se tiene TODOS los valores de una población, el promedio de la población es la suma de todos los valores dividida por cuántos son.

Métodos Numéricos para describir Variables Cuantitativas

El *promedio de la población* se denota por la letra Griega μ (mu): $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$.

☑ Ejemplo

Número promedio de niños por hogar.

Los datos siguientes son el número de niños en una muestra aleatoria de 10 casas en un vecindario: 2, 3, 0, 2, 1, 0, 3, 0, 1, 4.

El promedio de estas 10 observaciones es: 1,6

El resultado es 1,6 aunque no sea posible observar 1,6 niños en una casa. El promedio es 1,6

Supongamos que una observación en la última casa se anotó como 40 en vez de 4, ¿Qué le pasará al promedio?

Notar que 9 de las 10 observaciones son menores que el promedio. El promedio es *sensible a las observaciones extremas*.

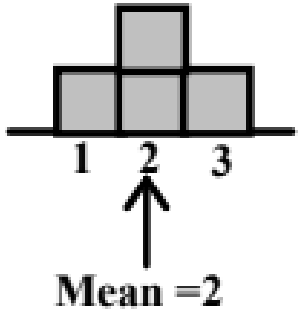
La mayoría de los métodos gráficos nos ayudarán de detectar observaciones extremas.

☑ Ejemplo

Un promedio NO es siempre representativo.

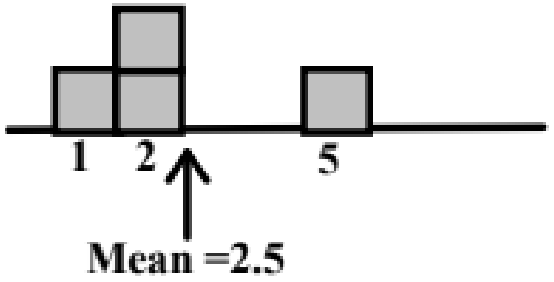
Las notas en varias pruebas de Juanita son 1,0 6,9 2,0 1,8 1,3, calcule el promedio de Juanita.

Métodos Numéricos para describir Variables Cuantitativas

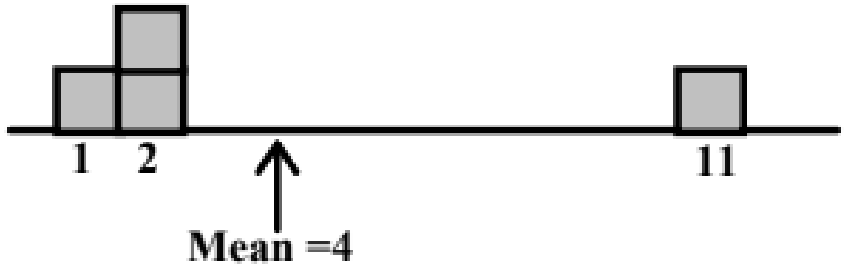


El promedio también se define como el ***punto de equilibrio,*** el punto donde distribución se balancea.

Si la distribución es ***simétrica***, el promedio será exactamente el centro de la distribución.



Si la observación más grande se mueve a la derecha, el ***promedio se mueve con la observación extrema***.



Si la distribución es sesgada, vamos a querer usar una medida que sea más ***resistente*** para mostrar el centro. La medida de tendencia central que es más resistente a los valores extremos es la ***mediana***.

Métodos Numéricos para describir Variables Cuantitativas

b) **Mediana**.

Definición:

La **mediana** de un conjunto de *n* observaciones, ordenadas de menor a mayor, es un valor tal que la mitad de las observaciones son menores o iguales que tal valor y la mitad de las observaciones son mayores o iguales que ese valor.

Pasos para encontrar la mediana:

1. Ordenar los datos de menor a mayor;
2. Calcular la posición de la mediana: $\boldsymbol{(n+1)/2}$, donde n es el número de observaciones
3. a) Si el número de observaciones es **impar**, la mediana es un único término central.

   b) Si el número de observaciones es **par**, la mediana es el promedio de los dos términos centrales.

Métodos Numéricos para describir Variables Cuantitativas

Nota: La mediana es resistente (robusta), es decir, no cambia o cambia muy poco con observaciones extremas.

Calcule la mediana de los siguientes conjuntos de datos:

Datos I: 2 3 3 3 4 4 4 4 5 5 5 5 5

Datos II: 5 6 6 6 7 7 8 8

Métodos Numéricos para describir Variables Cuantitativas

c) **Moda**.

Definición:

La **moda** de un conjunto de observaciones es el valor más frecuente.

- La moda de los valores: { 0, 0, 0, 0, 1, 1, 2, 2, 3, 4 } es 0.
- { 0, 0, 0, 1, 1, 2, 2, 2, 3, 4 } dos modas, 0 y 2 (*bimodal*).
- ¿Cuál sería la moda del siguiente conjunto de valores? { 0, 1, 2, 4, 5, 8 }.
- {0, 0, 0, 0, 0, 1, 2, 3, 4, 4, 4, 4, 5} ...

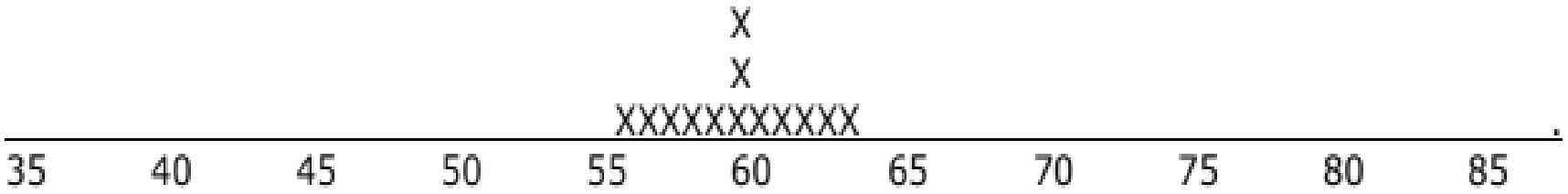
La Moda no se usa a menudo como medida de tendencia central para datos cuantitativos. Sin embargo la Moda es LA medida de tendencia central que puede ser calculada en datos ***cualitativos***.

Métodos Numéricos para describir Variables Cuantitativas

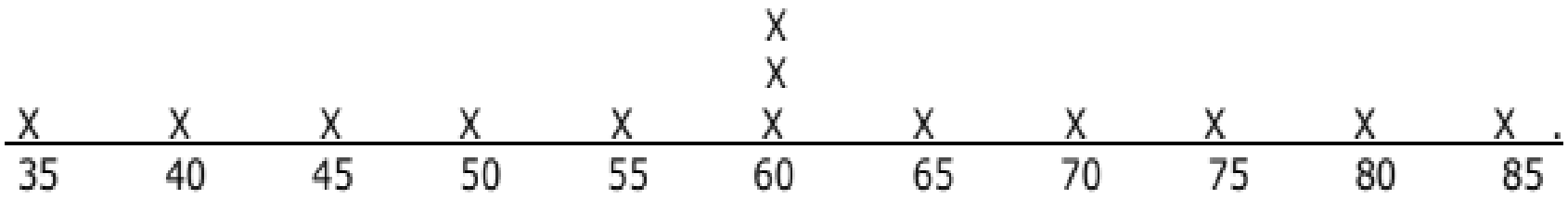
2. **Medidas de Dispersión.**

Las medidas de tendencia central son útiles pero nos dan una interpretación parcial de los datos. Considere los dos siguientes conjuntos de datos:

Datos 1: 55, 56, 57, 58, 59, 60, 60, 60, 61, 62, 63, 64, 65



Datos 2: 35, 40, 45, 50, 55, 60, 60, 60, 65, 70, 75, 80, 85



Métodos Numéricos para describir Variables Cuantitativas

a) Rango.

Es la medida de variabilidad o dispersión más simple. Se calcula tomando la diferencia entre el valor máximo y el mínimo observado.

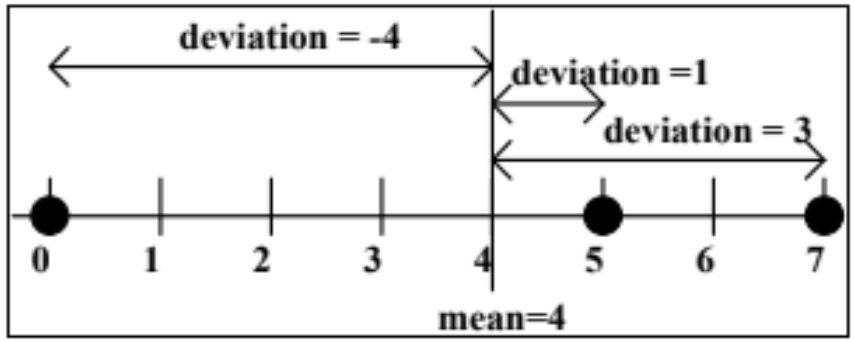
Rango = Máximo – Mínimo.

Métodos Numéricos para describir Variables Cuantitativas

b) **Desviación Estándar**.

Es una medida *de la dispersión de las observaciones a la media*. Es un "promedio de la distancia de las observaciones a la media".

☑ Ejemplo



Observación x	Desviación $x - \bar{x}$	Desviación al cuadrado $(x - \bar{x})^2$
0	0 – 4 = - 4	16
5	5 – 4 = 1	1
7	7 – 4 = 3	9
Promedio = 4	Suma = 0	Suma = 26

La **varianza muestral** está definida como la suma de las desviaciones al cuadrado divididas por el tamaño muestral menos 1, es decir, dividas por $n-1$.

$$\text{varianza muestral} = \frac{(-4)^2 + (1)^2 + (3)^2}{3-1} = \frac{16+1+9}{2} = \frac{26}{2} = 13$$

$$\text{desviación estándar muestral} = \sqrt{13} \approx 3{,}6$$

Métodos Numéricos para describir Variables Cuantitativas

☑ Ejemplo

Desviación estándar para el número de niños por hogar.

Recordemos los datos del número de niños por hogar en una muestra de 10 casas de un barrio:
2, 3, 0, 2, 1, 0, 3, 0, 1, 4

Use su calculadora científica y compruebe es siguiente resultado:

*"Los hogares tienen, **en promedio** 1,6 niños con una **variación** de alrededor de 1,43 niños".*

📎 En Resumen

Pensemos la desviación estándar como aproximadamente un *promedio de las distancias* de las observaciones a la media.

Si todas las observaciones son iguales, entonces la desviación estándar es cero.

La desviación estándar es positiva y mientras más alejados están los valores del promedio, mayor será la desviación estándar.

Métodos Numéricos para describir Variables Cuantitativas

Si $x_1, x_2, \ldots, x_n$ denota una muestra de n observaciones, la **varianza muestral** se denota por:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

La desviación estándar muestral, denotada por s, es la raíz cuadrada de la varianza: $s = \sqrt{s^2}$.

La **desviación estándar poblacional**, se denota por la letra Griega σ (sigma), es la raíz cuadrada de la varianza poblacional y se calcula como:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} .$$

Notas:

- La varianza y la desviación estándar no son medidas de variabilidad distintas, debido a que la última no puede determinarse a menos que se conozca la primera.

- A menudo se prefiere la desviación estándar en relación con la varianza, porque se expresa en las mismas unidades físicas de las observaciones.

- Así como el promedio es una medida de tendencia central que no es resistente a las observaciones extremas, la desviación estándar, que usa el promedio en su definición, tampoco es una medida de dispersión resistente a valores extremos.

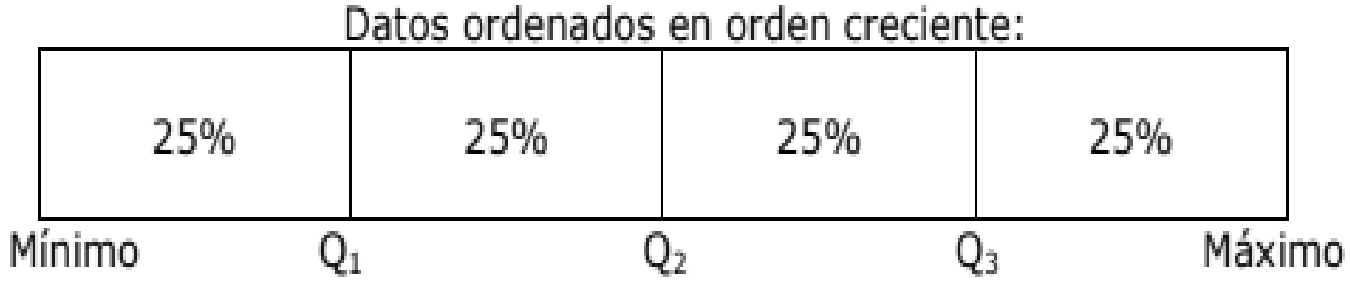
- Tenemos argumentos estadísticos para demostrar por qué dividimos por $n-1$ en vez de n en el denominador de la varianza muestral.

Métodos Numéricos para describir Variables Cuantitativas

Cuartiles

La mediana de una distribución divide los datos en dos partes iguales:

También es posible dividir los datos en más de dos partes. Cuando se dividen un conjunto ordenado de datos en cuatro partes iguales, los puntos de división se conocen como **cuartiles** y los representamos por Q_1, Q_2 y Q_3.



Métodos Numéricos para describir Variables Cuantitativas

c) **Rango entre cuartiles**.

La diferencia entre el tercer cuartil y el primer cuartil se llama ***rango entre cuartiles***, denotado por $\mathbf{RQ=Q_3-Q_1}$. El rango entre cuartiles mide la variabilidad de la mitad central de los datos.

Pasos para calcular cuartiles:

1. Encontrar la mediana de todas las observaciones.
2. Encontrar el primer cuartil = Q_1 = mediana de las observaciones que son menores a la mediana.
3. Encontrar el tercer cuartil = Q_3 = mediana de las observaciones que son mayores a la mediana.

Notas:

- Cuando el número de observaciones es impar, la observación del medio es la mediana. Esta observación no se incluye luego en los cálculos de Q_1 y Q_3.
- Pueden encontrar diferentes fórmulas en libros, calculadoras o computadores, pero todas estas fórmulas se basan en el mismo concepto.
- Si la distribución es simétrica, los cuartiles deben estar a la misma distancia de la mediana.

Métodos Numéricos para describir Variables Cuantitativas

☑ Ejemplo

¿Qué es Variabilidad?

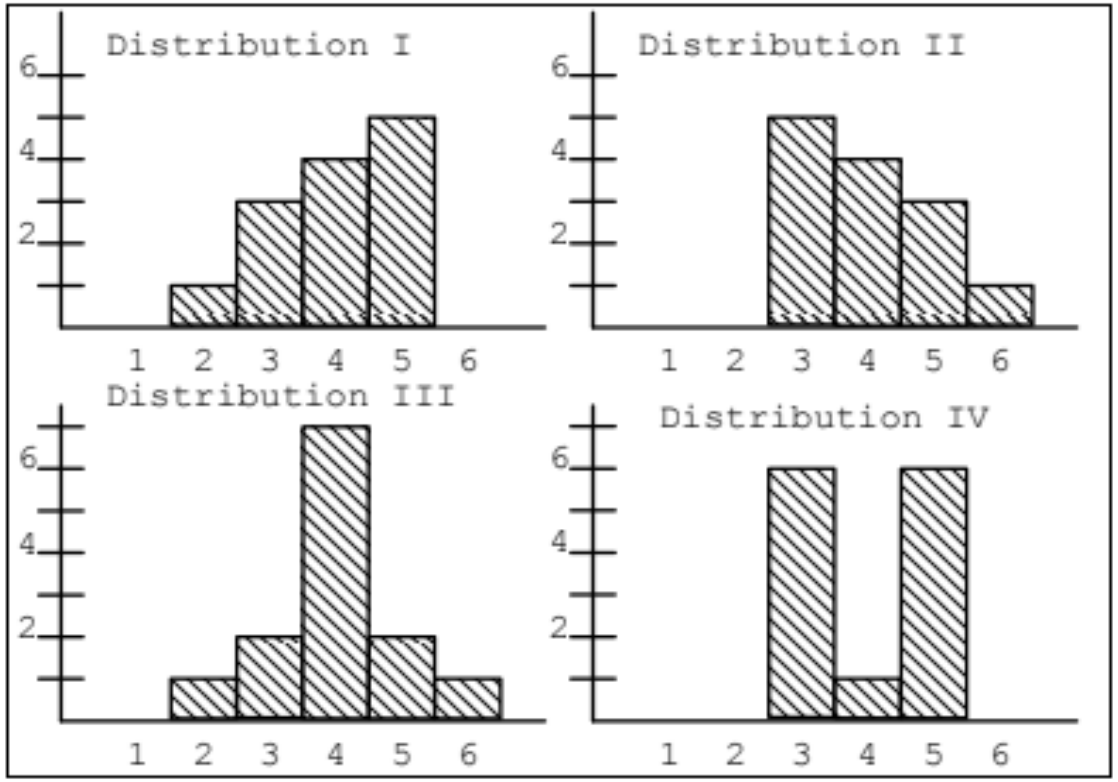
Considere los 4 conjuntos de datos siguientes y sus histogramas:

Datos I:
2 3 3 3 4 4 4 4 5 5 5
5 5

Datos II:
3 3 3 3 3 4 4 4 4 5 5
5 6

Datos III:
2 3 3 4 4 4 4 4 4 4 5
5 6

Datos IV:
3 3 3 3 3 3 4 5 5 5 5
5 5



Medidas de variabilidad	I	II	III	IV
Rango				
Rango entre cuartiles				
Desviación Estándar				

Métodos Numéricos para describir Variables Cuantitativas

En Resumen

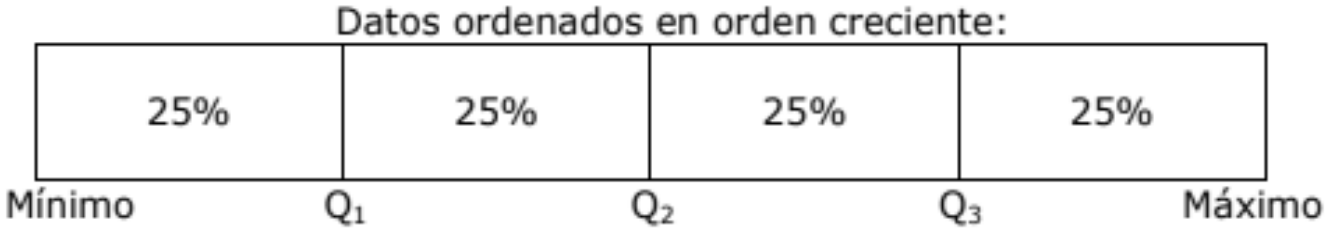
Cuando queremos describir una variable usamos alguna medida de posición central y una medida de dispersión. El par de medidas más comúnmente usado es el promedio y la desviación estándar. Pero vimos que cuando la distribución de las observaciones es sesgada, el promedio no es una buena medida de posición central y preferimos la mediana. La mediana en general va acompañada del rango como medida de dispersión. Pero cuando observamos valores extraños (extremos) el rango se ve muy afectado, por lo que preferimos usar el rango entre cuartiles.

Medida de tendencia central	**Medida de dispersión**	**Uso en Distribuciones**	**Ventajas**	**Desventajas**
Promedio	Desviación estándar	Simétricas	Buenas propiedades, muy usados.	Sensible a valores extremos.
Mediana	Rango	Sesgadas, sin valores extremos	Mediana robusta a valores extremos. Rango muy conocido, fácil de entender.	Rango sensible a valores extremos.
Mediana	Rango entre cuartiles	Sesgadas con valores extremos	Medidas robustas a valores extremos.	El rango entre cuartiles no es muy conocido.

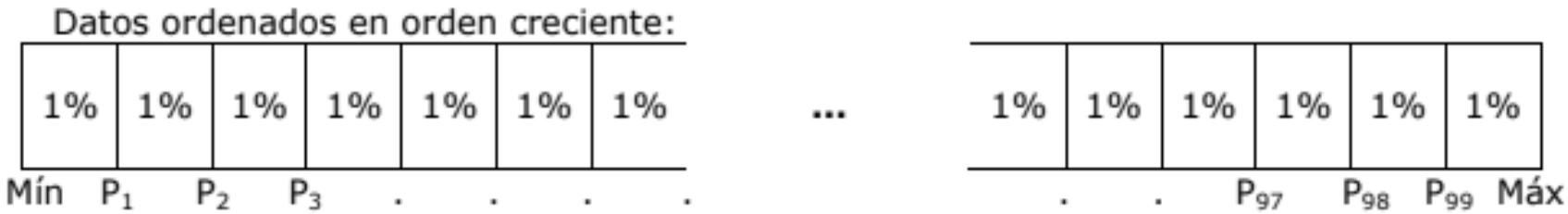
Métodos Numéricos para describir Variables Cuantitativas

3. **Medidas de posición relativa**.

Los ***cuartiles*** dividen un conjunto ordenado de datos, en cuatro partes iguales:

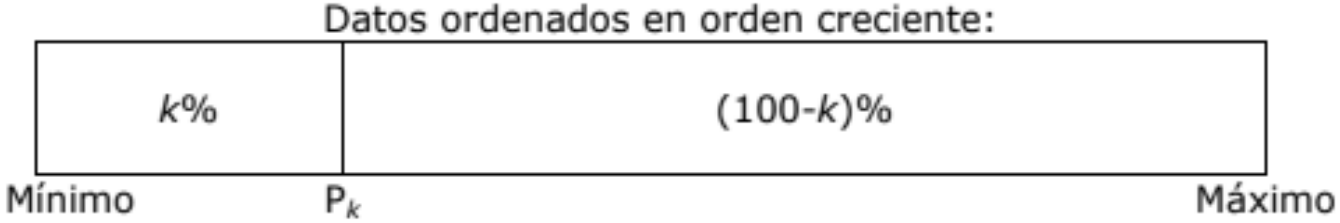


También podemos dividir conjuntos de datos en 100 partes iguales y los puntos de división se conocen como ***percentiles*.**



Es así como los cuartiles son en realidad los ***percentiles*** 25, 50 y 75, respectivamente.

En general, el ***k*-ésimo percentil** es un valor tal que el ***k*%** de los datos son menores o iguales que él, y el **(100-*k*)%** restante son mayores o iguales que él.



Métodos Numéricos para describir Variables Cuantitativas

Por ejemplo, el 25-ésimo percentil o **percentil 25** (P_{25}) es un valor tal que el ***25%*** de los datos son menores o iguales que él, y el **(100-25) = 75%** restante son mayores o iguales que él.

Definición:

Las **medidas de posición relativa** son medidas que describen la posición que tiene un valor específico en relación con el resto de los datos.

☑ Ejemplo

Si su nota estuvo en el percentil 84, entonces el 84% de las notas fueron inferiores a la suya y el 16% superiores.

Usos de medidas de posición relativa en:

- Calificaciones de exámenes.
- Puntajes en tests Psicológicos.
- Curvas de crecimiento en salud (http://www.cdc.gov/growthcharts/)

Método para elaborar un diagrama de caja y bigote

Definición

Valores extremos (outliers): son valores que se alejan del conjunto de datos.

Regla para identificar valores o datos extremos:

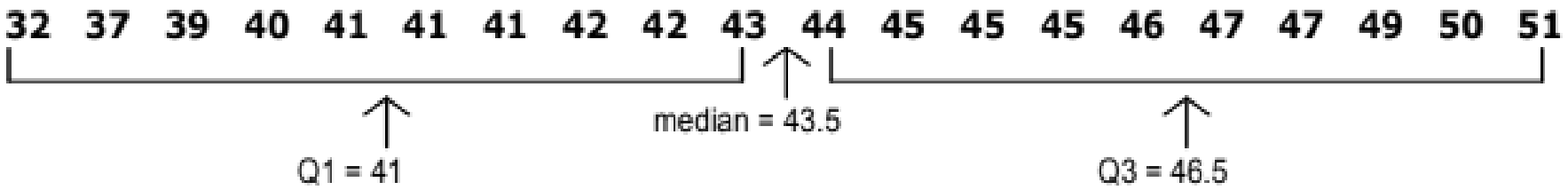
Vamos a definir una observación x_i como ***extrema*** si:

$$x_i < Q1 - 1{,}5 * (Q3-Q1) \quad \text{o} \quad x_i > Q3 + 1{,}5 * (Q3-Q1)$$

donde x_i serán las primeras y últimas observaciones en la serie ordenada de los datos.

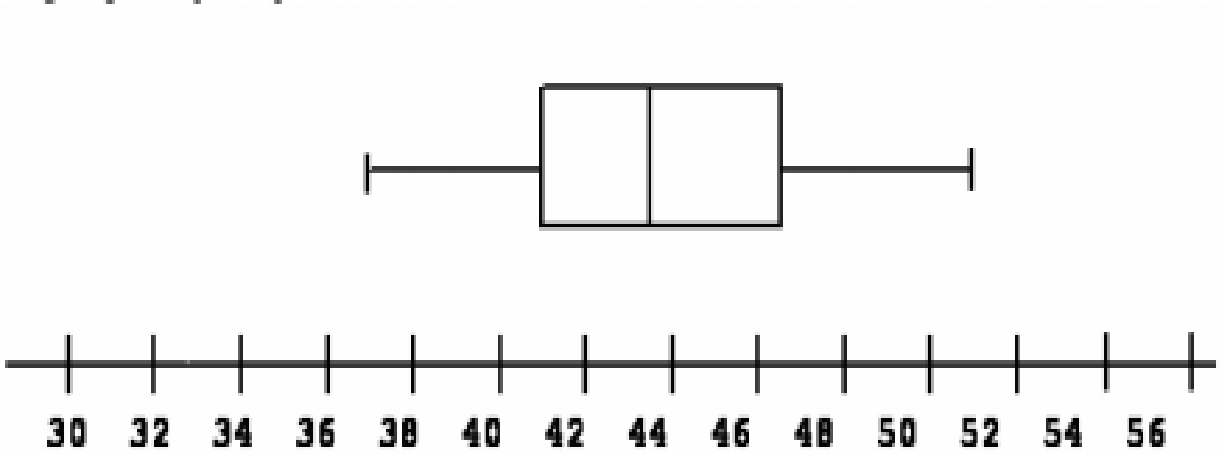
☑ Ejemplo

¿Tiene valores extremos, la variable edad de los 20 sujetos en el estudio médico?



Método para elaborar un diagrama de caja y bigote

Diagramas de caja (boxplot):



El diagrama de **caja** se construye de la siguiente manera:

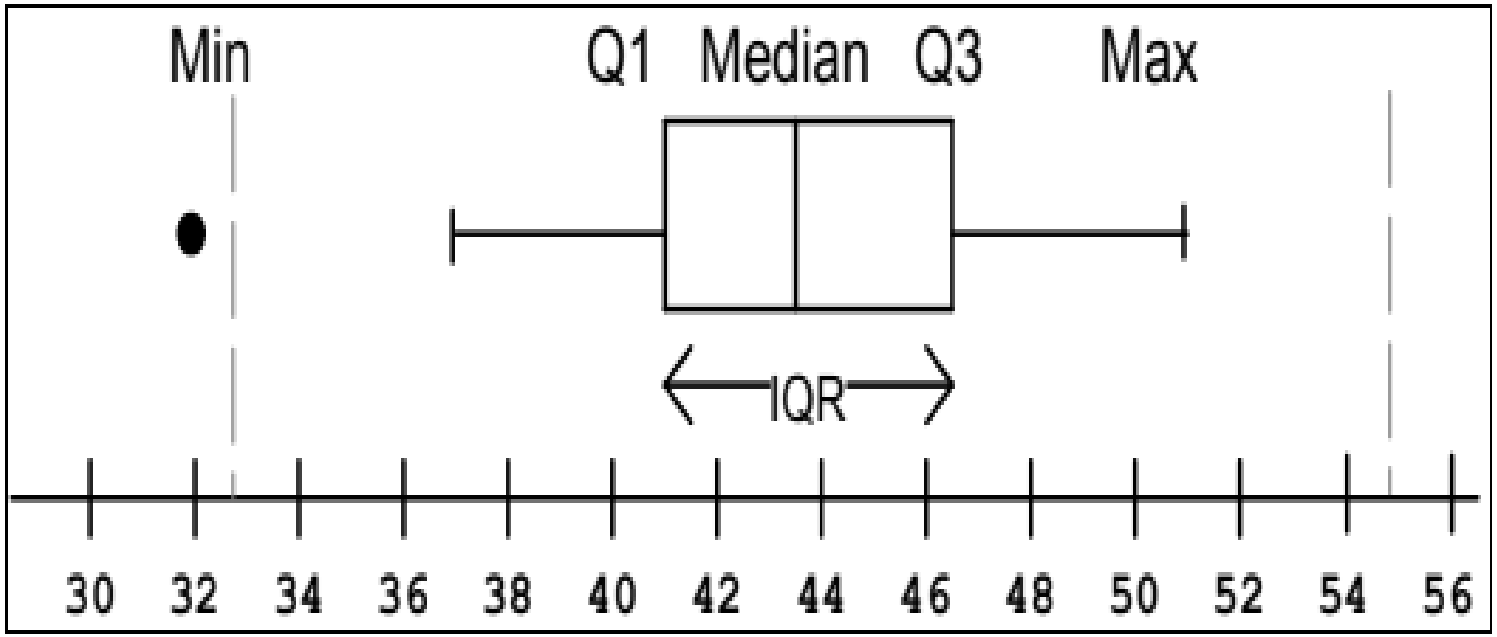
1. Dibujar la caja que empieza en el primer cuartil y termina en el tercer cuartil.
2. Dibujar la mediana con una línea dentro de la caja.
3. Por último se extienden las líneas, llamadas bigotes, saliendo de la caja hasta el mínimo y el máximo (salvo en la presencia de valores extremos).

Método para elaborar un diagrama de caja y bigote

☑ Ejemplo

Gráfico de caja para la EDAD

min = 32 Q1 = 41 mediana = 43,5 Q3= 46,5 max = 51



En la presencia de valores extremos, los "bigotes" se extienden hasta el valor observado anterior al valor extremo.

EJEMPLOS RCOMMANDER

Realice en Rcommander un histograma y gráfico de caja y bigote para las variables de base de datos 1. Interprete cada uno de los gráficos.

Encuentre las medidas de tendencia central y de dispersión de base datos 1. Interprete.

Encuentre todas las medidas de tendencia central y de dispersión de la base de datos de los alumnos de ingeniería civil generación 2015. Interprete.

EJEMPLOS RCOMMANDER

Analice mediante un gráfico de caja y bigote los valores extremos en las variables de la base de datos de los alumnos de ingeniería civil generación 2015.

Realice un histograma para describir variables cuantitativas de la base de datos de los alumnos de ingeniería civil generación 2015. Interprete cada uno de los gráficos.

EJEMPLOS RCOMMANDER

Realice un resumen descriptivo numérico calculando las medidas de tendencia central y de dispersión y gráfico mediante un histograma y un diagrama de caja y bigote los ejemplos que se detallan a continuación. Interprete los resultados obtenidos.

1.

Ejemplo:

En la construcción de una edificación de vivienda se estudia la estatura de un conjunto de 30 trabajadores con el objetivo de analizar las tallas de la ropa de trabajo. Los datos obtenidos para la estatura en metros luego de una medición cuidadosa son los siguientes.

Estatura [m]									
1.85	1.49	1.70	1.79	1.69	1.79	1.63	1.73	1.61	1.68
1.68	1.65	1.60	1.65	1.72	1.72	1.60	1.91	1.78	1.58
1.68	1.60	1.78	1.83	1.74	1.73	1.69	1.75	1.67	1.55

2.

El crecimiento económico en nuestro país en los últimos años, han evidenciado un creciente desarrollo en el rubro de la Construcción, situación que se puede apreciar en el medio, pues hoy en día existe una creciente oferta de casas, departamentos, dúplex, etc, en diferentes zonas urbanas.

Este hecho ha propiciado que el estado norme sobre las medidas de seguridad que deben ofrecer las empresas dedicadas a la actividad, a sus diferentes trabajadores.

La Empresa Liberteña "Constructora FM" dedicada exclusivamente a la edificación de Torres Habitacionales, ha sido comunicada acerca del cumplimiento de los estándares de seguridad que debe cumplir respecto a la integridad de los trabajadores y como es sabido, las edificaciones en nuestro medio permiten a los trabajadores a desplazarse en diferentes niveles y muchas veces lo hacen sobre rampas de madera y adicionalmente muchas veces deben transportar cargas y/o herramientas, es por ello que se desea determinar cómo varían los pesos en kilogramos de las trabajadores con la finalidad de reforzar o replantear dichas rampas, para ello se ha tomado una muestra de 50 trabajadores al azar sin tener en cuenta el rango operativo de función a fin de obtener sus pesos.

Los datos obtenidos fueron los siguientes:

65	63	65	63	69	67	83	88	60	61
64	65	74	72	68	77	85	87	70	68
84	85	64	71	68	66	86	89	61	72
63	65	63	70	67	76	87	79	71	62
64	64	83	69	67	66	88	60	61	62