Estadística Descriptiva

Hugo S. Salinas

Fuente: http://dta.utalca.cl/estadistica/

Tipo de Variables

Número	Sexo	Edad	Estatura	Peso	Ciudad de residencia	Número de hermanos
1	M	22	180	74	SAN FERNANDO	7
2	M	20	175	95	CHILLAN	2
3	M	20	178	68	TALCA	2
4	M	22	183	75	TALCA	7
5	M	25	180	76	LINARES	3
6	M	22	180	78	SANTIAGO	1
7	M	21	180	.	TALCA	1
8	M	24	182	85	TALCA	1
9	M	21	177	78	CURICO	1
10	M	21	184	85	SANTIAGO	0
11	M	20	172	70	SAN FERNANDO	3
12	M	21	173	59	IQUIQUE	4
13	F	20	162	56	SANTIAGO	0
14	M	22	194	105	LINARES	4
15	M	20	174	79	SANTIAGO	1
16	F	20	165	50	SAN JAVIER	1
17	F	22	167	58	TALCA	1
18	F	20	155	52	PUERTO MONTT	2
19	M	20	174	65	LINARES	2
20	F	20	160	48	SANTIAGO	2
21	F	22	155	58	SANTIAGO	1
22	M	19	174	80	SAN FELIPE	1
23	F	19	162	60	MELIPILLA	1
24	M	19	180	82	TALCA	3
25	F	20	160	57	TALCA	1
26	F	21	170	70	SANTIAGO	2
27	F	20	155	50	SANTIAGO	1
28	F	21	160	60	TALCA	1
29	F	22	166	61	PUERTO IBAÑEZ	1
30	M	19	170	68	RANCAGUA	3
31	F	22	160	60	SANTIAGO	1
32	M	20	182	72	TALCA	1
33	F	19	162	55	RANCAGUA	2
34	F	20	154	46	SANTIAGO	3
35	F	19	155	50	RANCAGUA	2
36	M	20	184	85	RANCAGUA	5

Tipo de variables

La base de datos anterior contiene la información de 36 alumnos de un curso de Estadística de la Universidad de Talca. En esta base de datos podemos notar que los alumnos tienen distintas características, por ejemplo, no todos vienen de la misma ciudad.

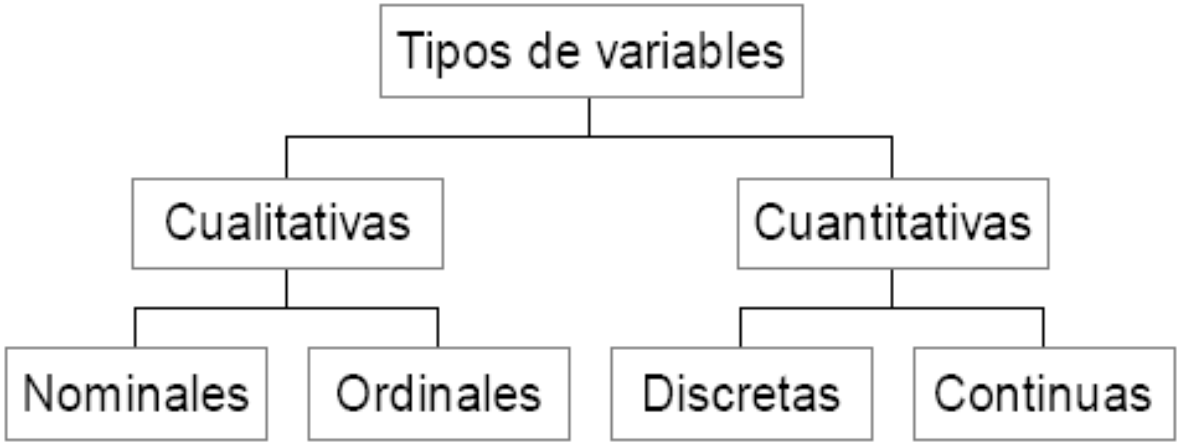
Definiciones:

Unidad es el objeto que observamos. Cuando el objeto es una persona, lo referimos como ***sujeto***.

Observación es la información o característica que registramos de cada unidad.

Una característica que puede variar de unidad en unidad es llamada **variable**.

Una colección de observaciones con una o más variables se llama **base de datos**.



Tipo de variables cont.

Ejemplo:
Determinar el tipo de variable. Si son variables **cualitativas** (nominal u ordinal) o **cuantitativas** (discretas o continuas).
a) Marca de automóvil.
b) Duración de un disco compacto (segundos).
c) Número de temas de un disco compacto.
d) Nivel educacional (básica, media, universitaria).
e) Temperatura al mediodía en Copiapó (grados Celsius).
f) Estado civil (soltero, casado, divorciado, viudo).
g) Cantidad de lluvia en un año en Copiapó (mm^3).

Métodos gráficos y numéricos para describir variables cualitativas

Definición:

La **distribución** de una variable nos da los valores posibles de la variable y cuantas veces ocurren. La distribución de una variable nos muestra la forma en que varía la variable.

Tablas de distribución de frecuencias.

Lo primero que hacemos al querer describir variables cualitativas es contar cuántas unidades caen en cada categoría de la variable. Esto lo presentamos en una tabla de distribución de frecuencias de la forma:

Valor o categoría de la variable	Frecuencia	Porcentaje
...		
Total	**n**	**100**

Tabla de distribución de frecuencias del sexo de la base de datos 1

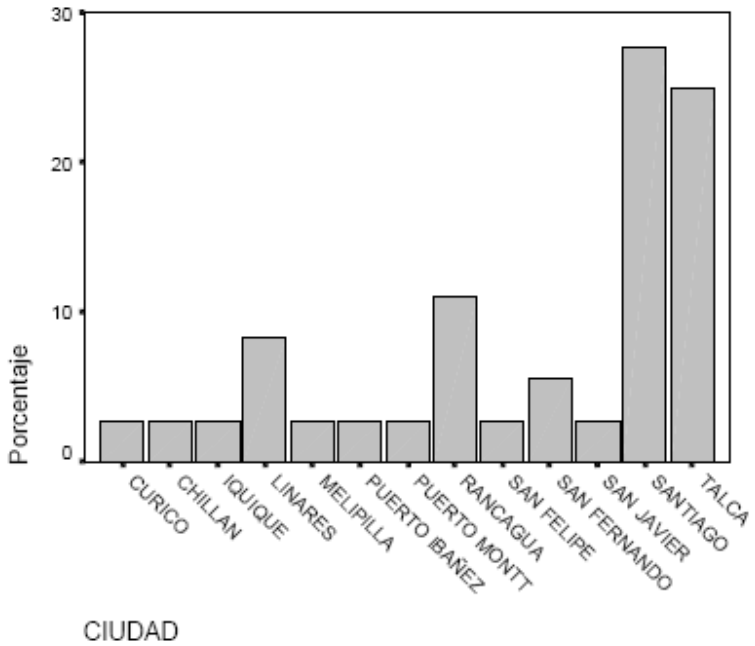
Sexo	Número de alumnos	Porcentaje de alumnos
Femenino	16	44,4
Masculino	20	55,6
Total	36	100,0

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**.

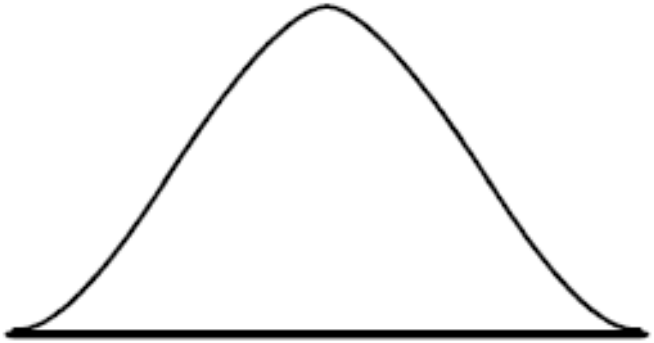
Un **gráfico sectorial (o de torta)** muestra la distribución de una variable cualitativa dividiendo un círculo en partes que corresponden a las categorías de la variable, tal que el tamaño (ángulo) de cada pedazo es proporcional al porcentaje de ítems en cada categoría.

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.

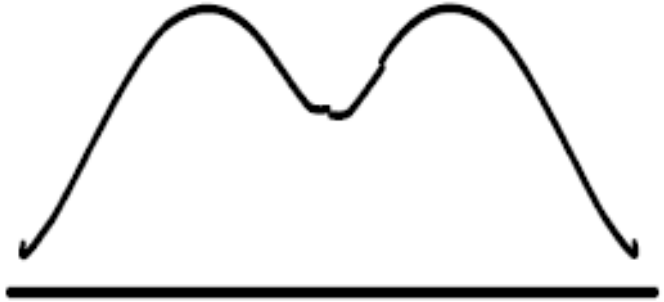


Métodos gráficos para describir variables cuantitativas
1. Gráficos de puntos

Formas de Distribuciones



Simétrica, acampanada, unimodal



Bimodal

Sesgada a la derecha (sesgo positivo)

Sesgada a la izquierda (sesgo negativo)

Uniforme

Gráficos de puntos cont.

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.

Ejemplo:

¿Cuántas llaves tiene en su bolsillo?

Hacer un gráfico de frecuencias (de puntos) con el número de llaves que tienen los estudiantes que asisten hoy a clases. Describir la forma del gráfico.

2. Diagrama de Tallo y Hojas

Pasos para hacer un Tallo y Hoja:

1. Separar cada medida en un tallo y una hoja.
   Generalmente la hoja consiste en exactamente un dígito (el último) y el tallo consiste en uno o más dígitos.

   Ejemplo: 734 => tallo=73, hoja=4 2,345 => tallo=2,34, hoja=5.

   A veces se deja fuera el decimal pero se agrega una nota de cómo leer el valor.
   Para 2,345 por ejemplo podremos decir que 234 | 5 se debe leer como 2,345.

2. Escribir los tallos en orden creciente de arriba abajo y dibujar una línea a la derecha de los tallos.

3. Agregar las hojas a su respectivo tallo en orden creciente.

Ejemplo:
Diagrama básico de Tallo y Hoja para la Edad de base de datos de un estudio médico.
45 41 51 46 47 42 43 50 39 32
41 44 47 49 45 42 41 40 45 37

2. Diagrama de Tallo y Hojas cont.

Una modificación útil es que podemos ***dividir los tallos:***

```
3 | 2
3 | 7 9
4 | 0 1 1 1 2 2 3 4
4 | 5 5 5 6 7 7 9
5 | 0 1
```

Note que el menor valor representado por 3 | 2 se lee 32 años.

Así podemos visualizar mejor que la distribución de las edades de los sujetos es aproximadamente simétrica, centrada en aproximadamente 43-44, sin valores extremos evidentes (observaciones que caen fuera del patrón general de datos).

3. Histograma

Pasos para hacer un histograma:

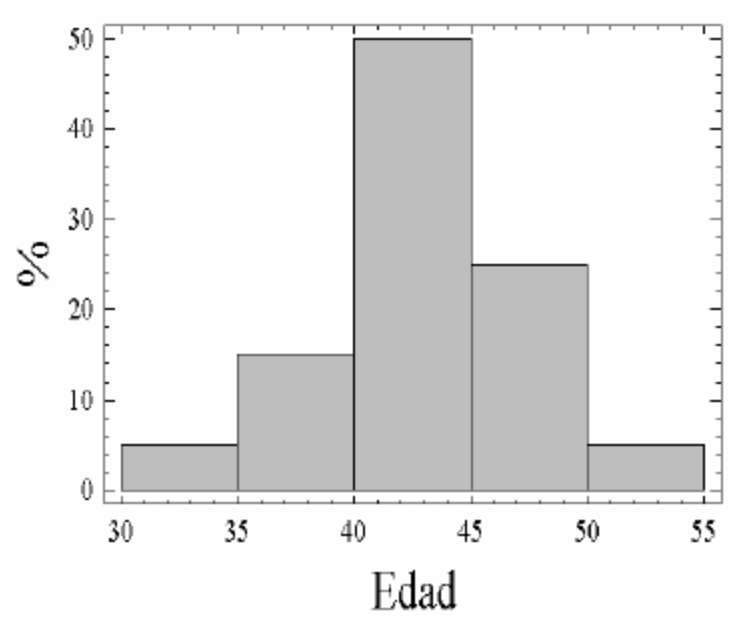
1. Dividir el rango de los datos (menor a mayor) en clases del mismo ancho. Las clases deben contener el rango posible de datos y no se deben superponer. Ej. Si los datos van de 0 a 29, comience en 0 hasta 30 de ancho 5.
2. Contar el número de observaciones (frecuencias) que caen en cada clase.
3. Dibujar en el eje horizontal y marcar las clases.
4. El eje vertical puede contener la frecuencia, la proporción, o el porcentaje.
5. Dibujar un rectángulo (una barra vertical) en cada clase con la altura igual a la frecuencia, la proporción, o el porcentaje.

Histograma de Edad

Veamos nuevamente las edades de la base de datos médica. El rango va de 32 a 51, entonces podemos crear clases que comiencen en 30 con incrementos de 5 hasta 55. Puede intentar diferentes clases con distinto ancho hasta obtener una buena representación.

3. Histograma cont.

Clase	Cuenta	Número de observaciones	Porcentaje
(30,35]	/	1	1/20 = 0.05 => 5%
(35,40]	///	3	3/20 = 0.15 => 15%
(40,45]	//////////	10	10/20 = 0.50 => 50%
(45,50]	/////	5	5/20 = 0.25 => 25%
(50,55]	/	1	1/20 = 0.05 => 5%



Métodos numéricos para describir variables cuantitativas

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: la **media y la mediana**.
Si los datos son una **muestra**, la media (o promedio) y la mediana se llamarán **estadísticas**. Si los datos son una **población** entonces estas medidas de tendencia central se llamarán **parámetros**.

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:

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

Medidas de Tendencia Central

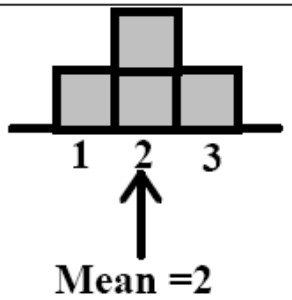
Notación: Si $X_1, X_2, \ldots, X_n$ denota una muestra de *n* observaciones, entonces el *promedio de la muestra* se llama "x-barra" y se denota por:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$ [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.

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

El promedio de 3 estudiantes es 5,4 y el promedio de otros 4 estudiantes es 6,7, ¿Cuál es el promedio de los 7 estudiantes?

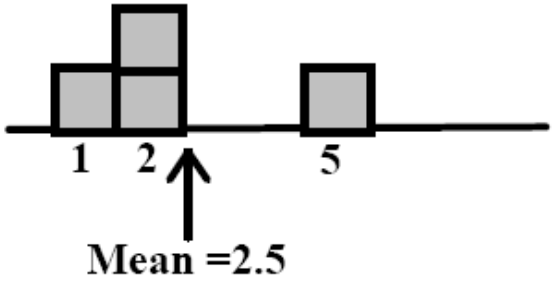


El promedio también se define como el ***punto de equilibrio,*** el punto donde distribución se balancea.

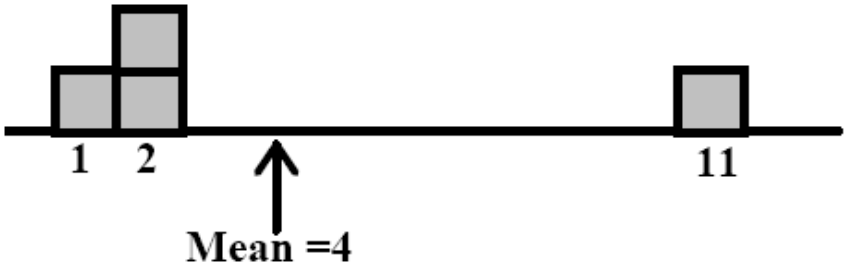
Si la distribución es ***simétrica***, el promedio será exactamente el centro de la distribución.

1 Pueden revisar la notación de sumatorias en Hopkins, K. Hopkins, B. Glass, G. (1997) Estadística básica para las ciencias sociales y del comportamiento. Tercera edición. Prentice Hall.

Medidas de Tendencia Central cont.



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**.

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.

Medidas de Tendencia Central cont.

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.

Ejemplo:
Encuentre la mediana del número de niños por hogar en la muestra de 10 hogares.
Número de Niños: 2, 3, 0, 1, 4, 0, 3, 0, 1, 2.
a) Ordenar las observaciones de menor a mayor:
b) Calcular (n+1)/2 =
c) Mediana =
d) ¿Qué le pasa a la mediana si la quinta observación en la lista se anota incorrectamente como 40 en vez de 4?
e) ¿Qué le pasa a la mediana si la tercera observación en la lista se anota incorrectamente como –20 en vez de 0?
La mediana es resistente (robusta), es decir, no cambia o cambia muy poco con observaciones extremas.

Medidas de Tendencia Central cont.

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.

Medidas de Tendencia Central cont.

Diferentes medidas pueden dar diferentes impresiones

El famoso trío - **media**, **mediana** y **moda** – representan tres métodos diferentes para encontrar el valor del **centro**. Estos tres valores pueden ser un mismo valor pero a menudo son distintos. Cuando son distintos, pueden servir para diferentes interpretaciones de los datos que queremos resumir. Considere el ingreso mensual de cinco familias en un barrio:

$120 000 $120 000 $300 000 $900 000 $1 000 000

¿Cuál es el **ingreso típico** de este grupo?

El ingreso mensual promedio es:

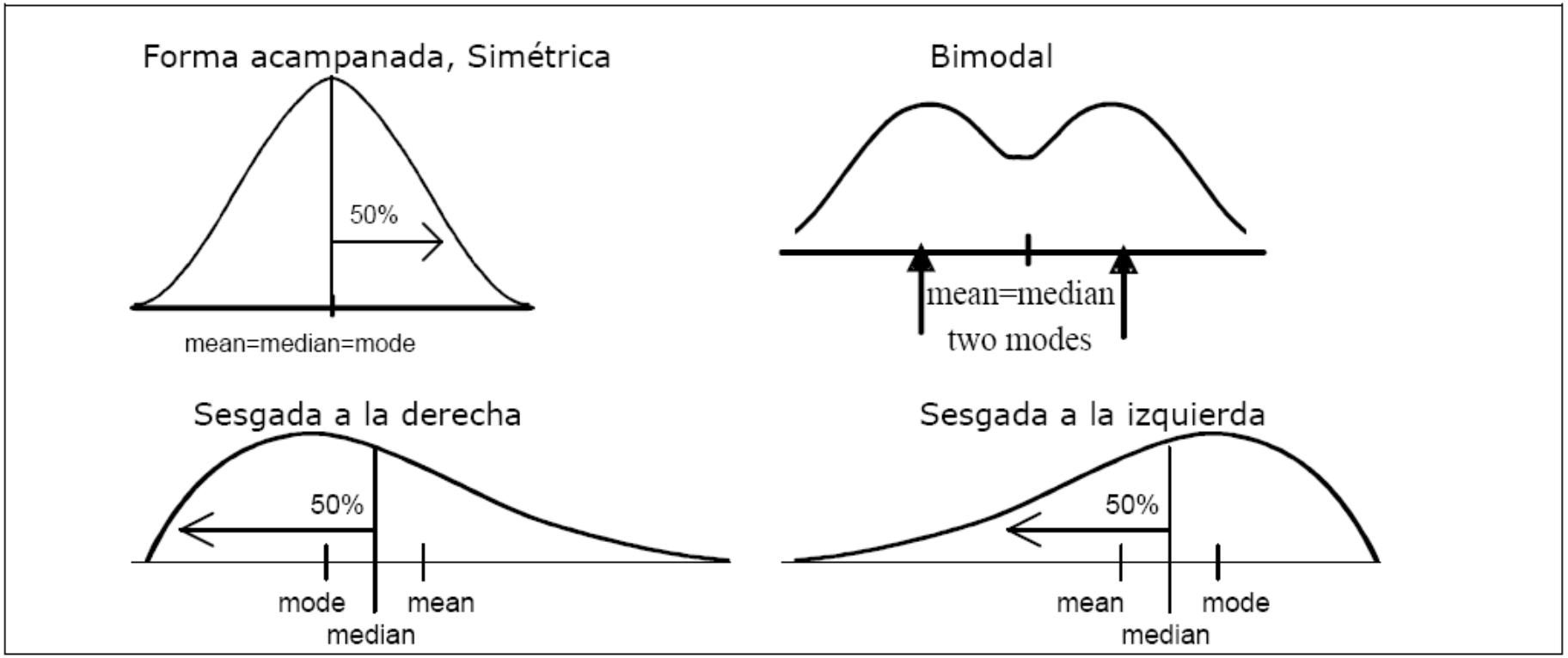
La mediana del ingreso mensual es:

La moda del ingreso mensual es:

Si tú estás tratando de promover el barrio, ¿Qué medida usarías?

Si tú estás tratando que bajen las contribuciones, ¿Qué medida usarías?

¿Qué medida de tendencia utilizar?



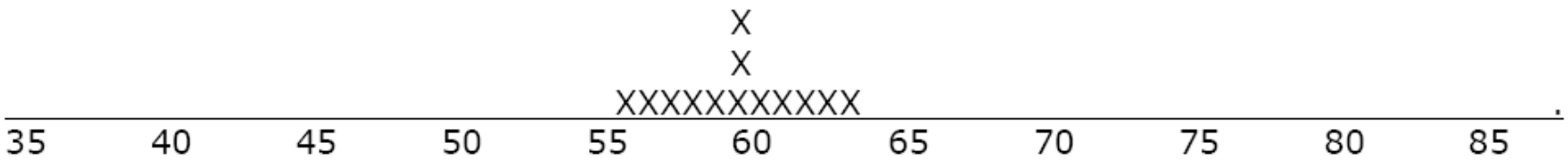
Responder:

1. Supongamos que calculamos la media, mediana y moda de una lista de números, ¿Qué medida es siempre un número en la lista?
2. Si la distribución es simétrica, ¿Qué medida de tendencia central calcularías: la media o la mediana?, ¿Por qué?

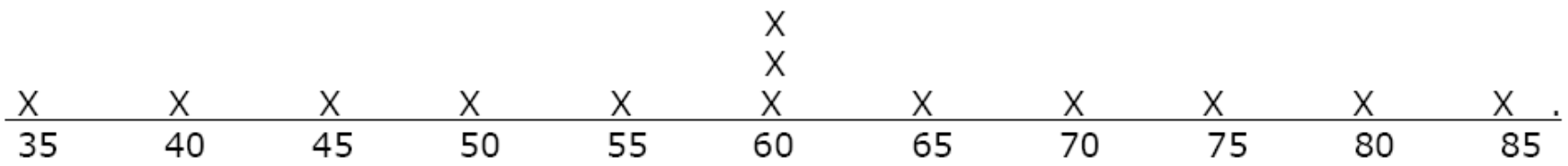
Medidas de Dispersión

Las medidas de tendencia central son útiles pero nos dan una interpretación parcial de los datos. Consideremos 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

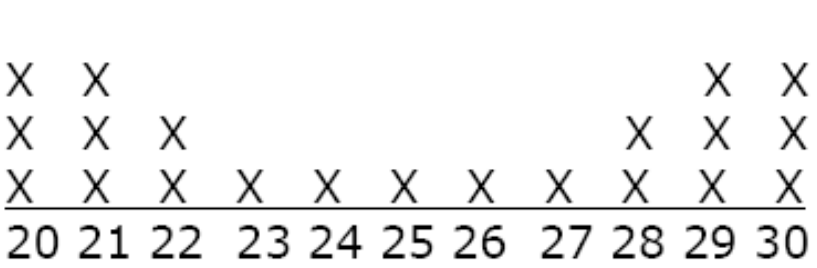


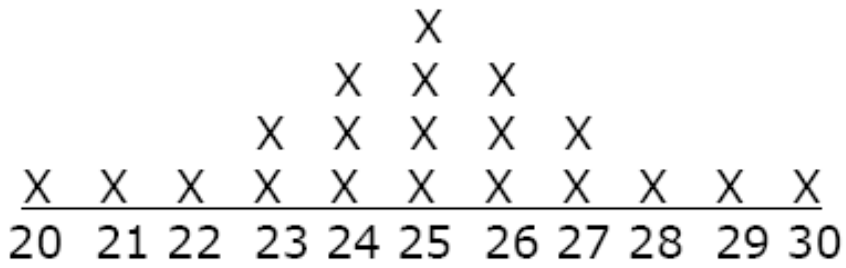
Rango:

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.

Medidas de Dispersión cont.

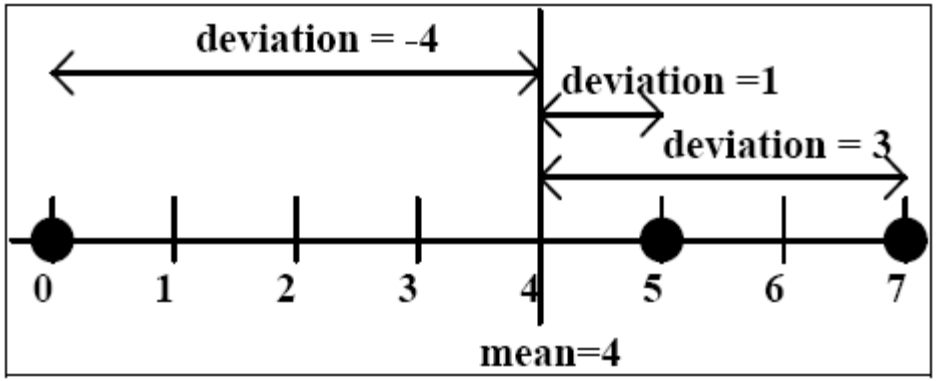




Analizar cuáles podrían ser las ventajas y desventajas del rango como medida de variabilidad.

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**.



Medidas de Dispersión cont.

Observación x	Desviación $x - \overline{x}$	Desviación al cuadrado $(x - \overline{x})^2$
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 .

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

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

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.

Medidas de Dispersión cont.

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

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

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:

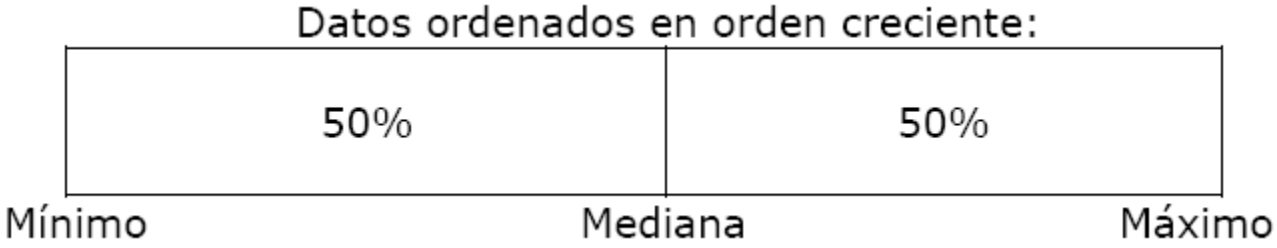
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} .$$

- 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**.

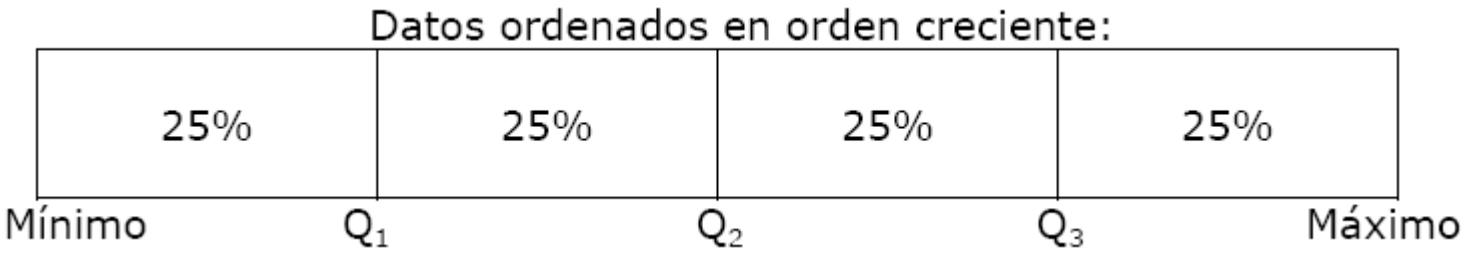
Medidas de Dispersión cont.

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 **Q1, Q2 y Q3**.



Rango entre cuartiles

La diferencia entre el tercer cuartil y el primer cuartil se llama **rango entre cuartiles**, denotado por **RQ=Q3-Q1**. El rango entre cuartiles mide la variabilidad de la mitad central de los datos.

Medidas de Dispersión cont.

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.

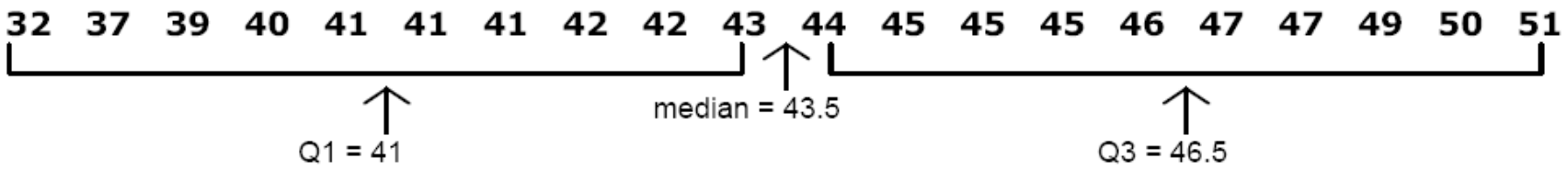
Notas:

- Cuando el número de observaciones es impar, la observación del medio es la mediana. Esta observación no se incluye luego en los cálculos de **Q1 y Q3**.
- 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.

Ejemplo

Lista ordenada de las edades de los 20 sujetos en el estudio médico:



Medidas de Dispersión cont.

¿Qué es Variabilidad?

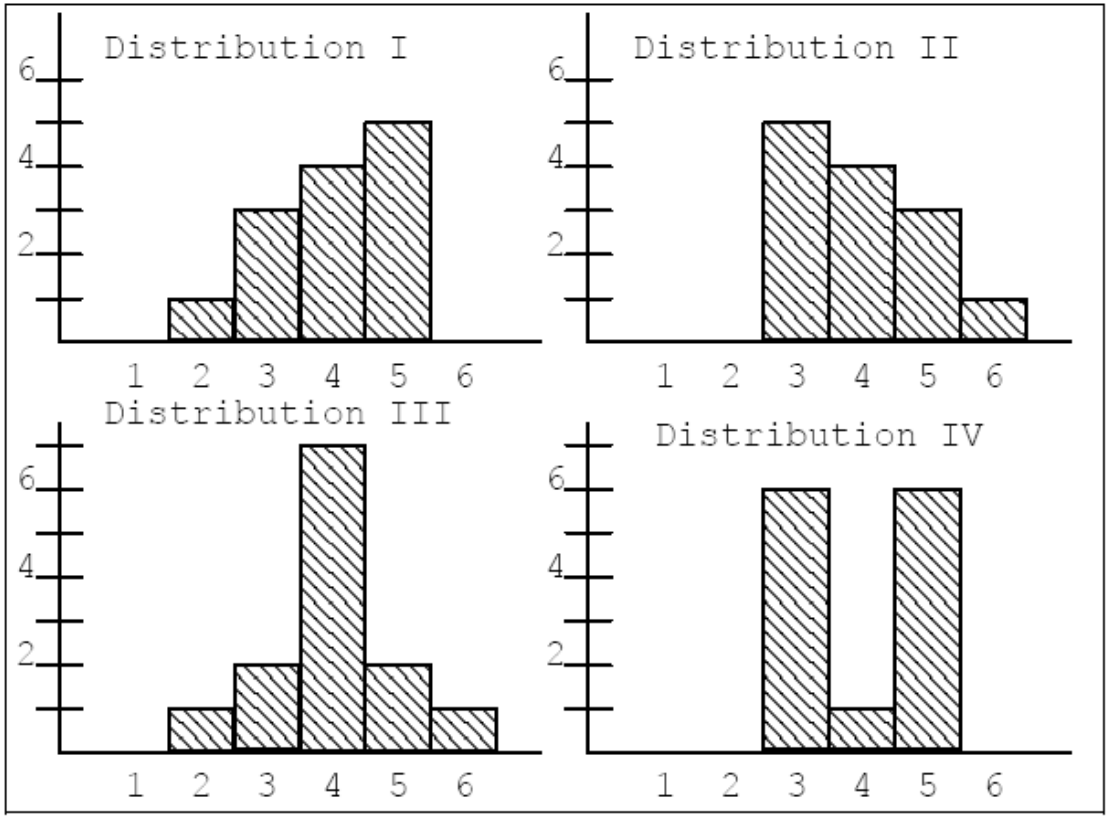
Considere los 4 conjuntos de datos siguientes y sus histogramas:

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				

Medidas de Dispersión cont.

Algunas personas asocian variabilidad con rango mientras que otras asocian variabilidad con cómo difieren los valores de la media. Hay muchas medidas de variabilidad, y la **desviación estándar** es la más usada. Pero recuerden que una distribución con la menor desviación estándar no es necesariamente la distribución que es menos variable con respecto a otras definiciones de variabilidad.

Resumen:

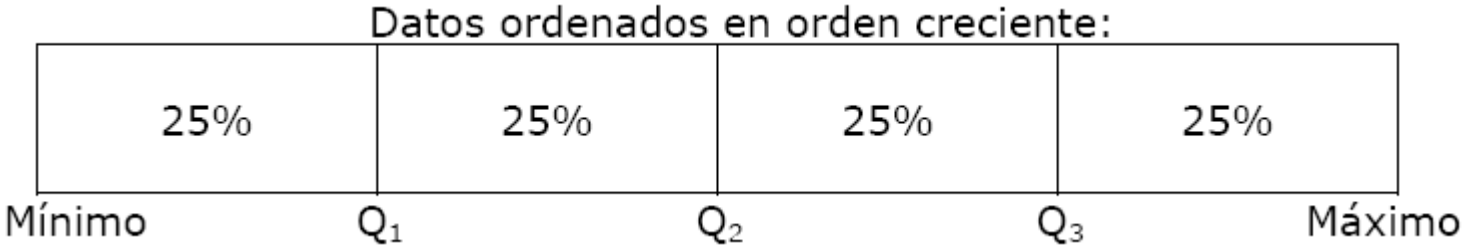
Cuando queremos describir una variable usamos alguna **medida de posición central** y una **medida de dispersión**. El par de medidas más comúnmente usado es el **promedio** y la **desviación estándar**. Pero vimos que cuando la distribución de las observaciones es sesgada, el promedio no es una buena 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.

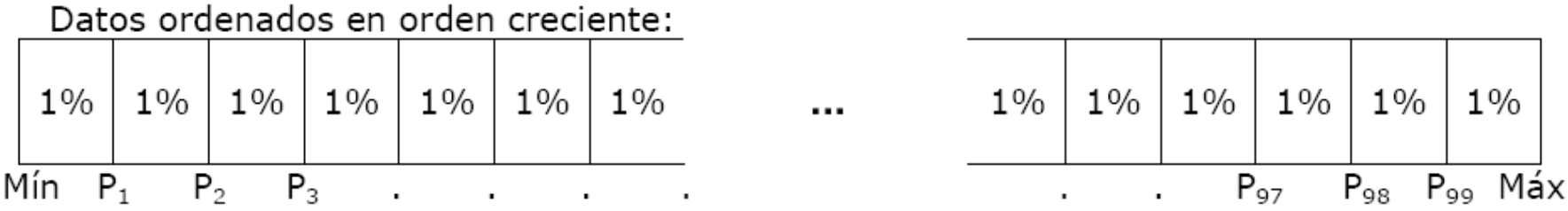
Medidas de Dispersión cont.

Medidas de posición relativa.

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



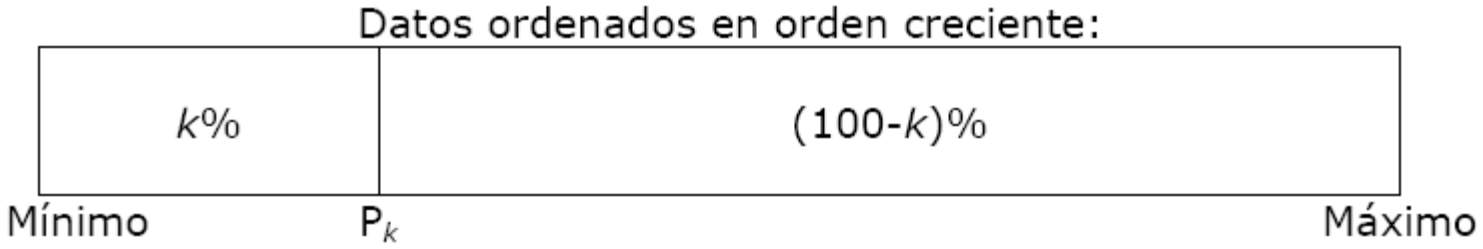
También podemos dividir conjuntos de datos en 100 partes iguales y los puntos de división se conocen como **percentiles**.



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.

Medidas de Dispersión cont.



Por ejemplo, el **25-ésimo percentil** o **percentil 25 (P25)** 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.

Definición

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

Medidas de Dispersión cont.

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.

Los valores extremos por lo general son atribuibles a una de las siguientes causas:

- La observación se registra incorrectamente.
- La observación proviene de una población distinta.
- La observación es correcta pero representa un suceso poco común (fortuito).

Volvamos al ejemplo de las edades.

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

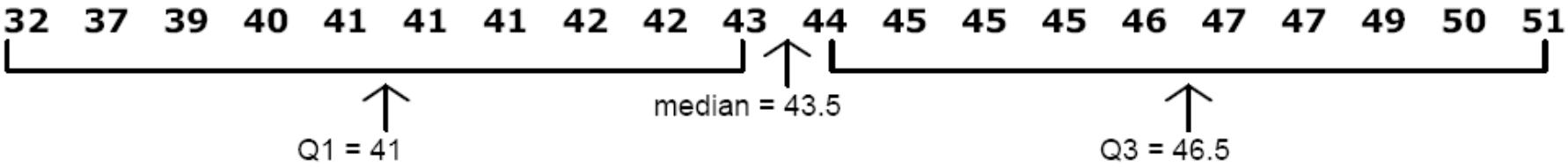
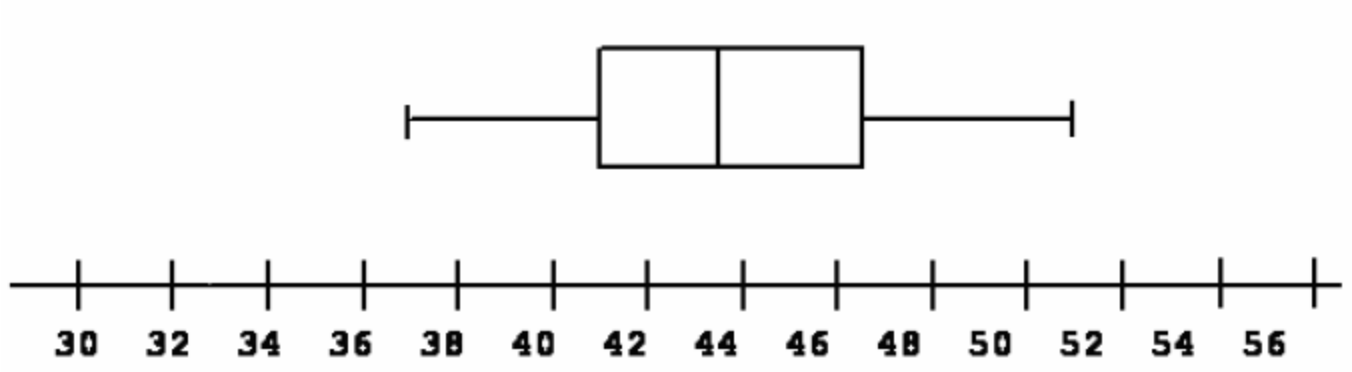


Diagrama de Cajas (Blox-plot)



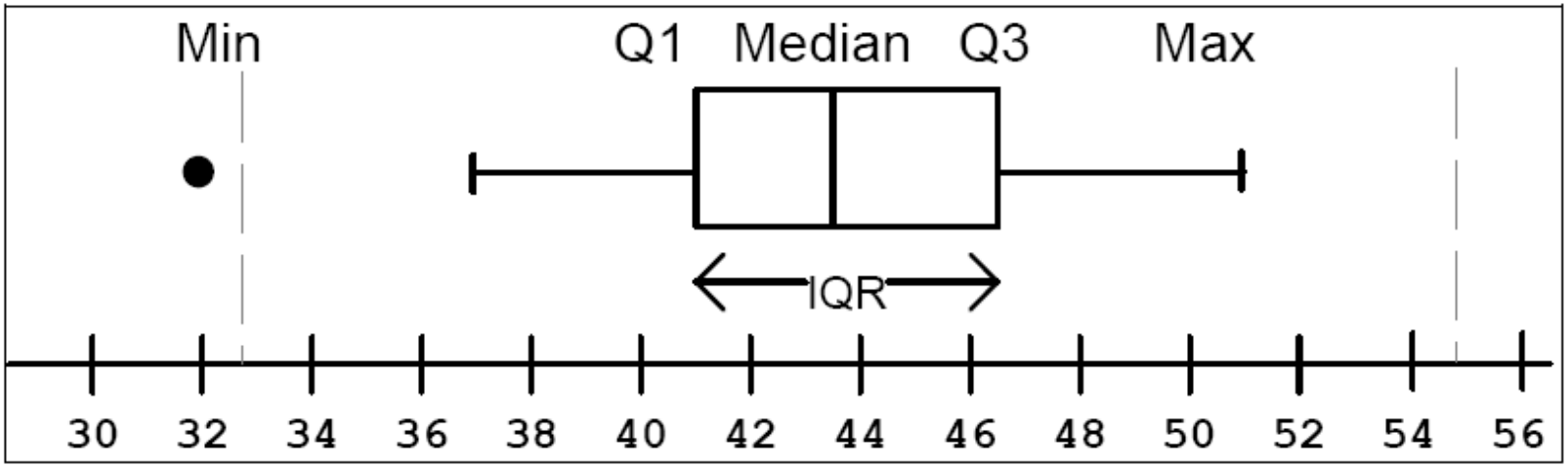
El diagrama de **caja** se construye de la siguiente manera:

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).

Diagrama de Cajas (Blox-plot) cont.

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.

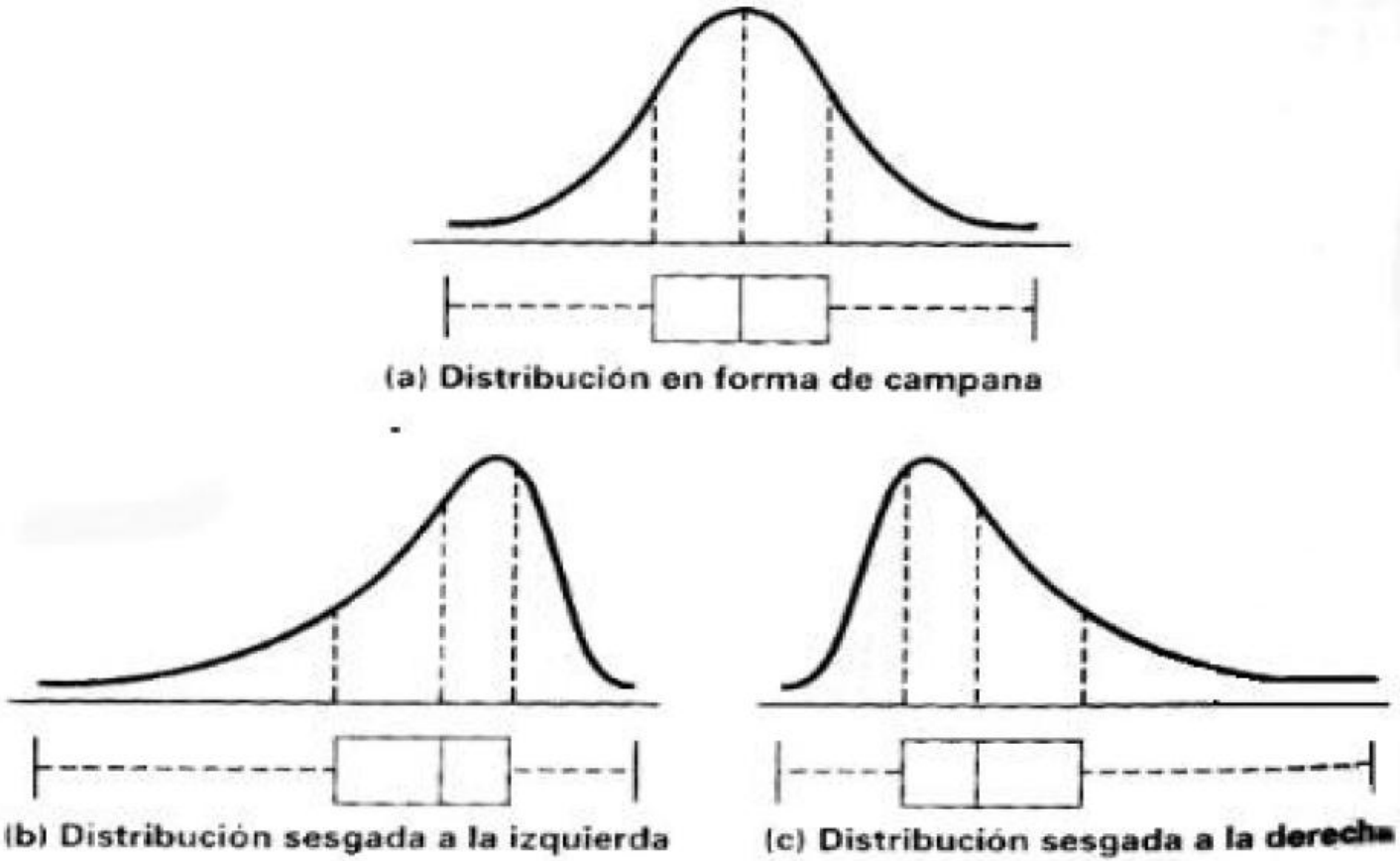
La distancia entre la mediana y los cuartiles es aproximadamente la misma, lo que nos hace pensar que la distribución de los datos es más o menos simétrica como vimos antes en el histograma y en el tallo y hoja.

Medidas de Dispersión cont.

La distancia entre la mediana y los cuartiles es aproximadamente la misma, lo que nos hace pensar que la distribución de los datos es más o menos simétrica como vimos antes en el histograma y en el tallo y hoja.

Los gráficos de caja son muy útiles para comparar distribuciones de dos o más grupos.

Si el gráfico de caja es simétrico, ¿Podemos concluir que la distribución de los datos es simétrica?



Coeficiente de Variación

Es una medida de **variación relativa**. Se simboliza c.v. y es igual a:

$$c.v. = \frac{s}{\bar{x}} . 100$$

Es la desviación estándar expresado como porcentaje de la media (promedio), por lo tanto no viene expresado en unidades.
Es útil para la **comparación de la variabilidad relativa entre distribuciones** que no están expresadas en la misma unidad de medida o bien, entre distribuciones que si bien están expresadas en la misma unidad, poseen promedios muy dispares.

Ejemplo:
En marzo del año pasado, los datos de préstamos personales de un Banco mostraron un promedio de $6500000 y una desviación estándar de $3000000. Recientemente se calculó la media y la desviación estándar correspondiente a los préstamos personales de marzo del presente año resultando las mismas $ 9000000 y $ 3500000 respectivamente.
¿En cuál de los dos años los préstamos personales presentaron menor dispersión relativa?.
c.v. año pasado=(30/65)x100=45%, c.v. presente año=(35/90)x100=39%
La menor dispersión relativa se presenta en los préstamos personales otorgados este año.

Regla Empírica

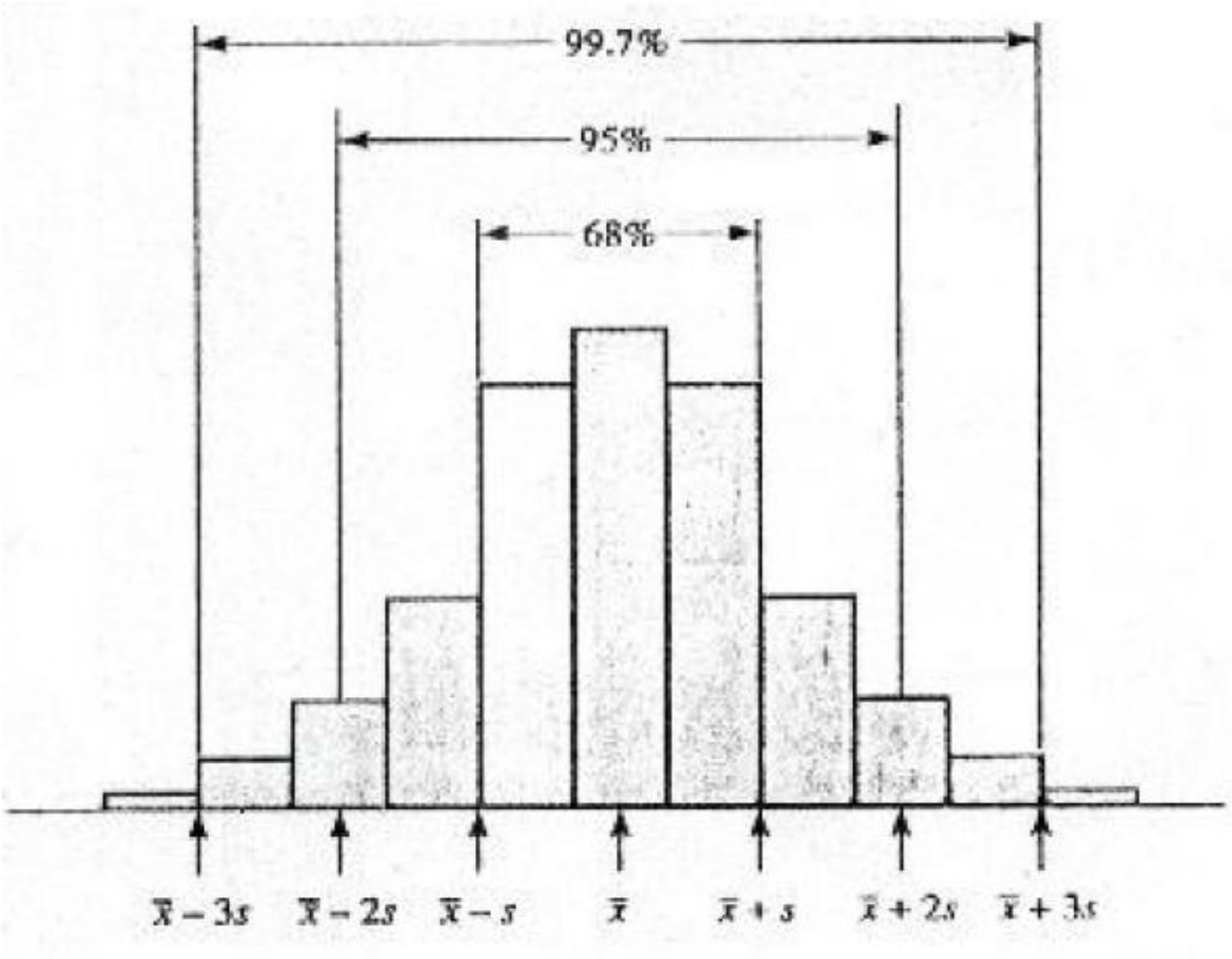
Es posible que dos conjuntos de datos distintos tengan el mismo rango pero difieran considerablemente en el grado de variación de los datos. En consecuencia, el rango es una medida relativamente insensible de la variación de los datos. La varianza tiene importancia teórica, pero es difícil de interpretar porque las unidades de medición de la variable de interés están elevadas al cuadrado. En cambio, las unidades de medición de la desviación estándar son las unidades de la variable. Si la desviación estándar se combina con la media del conjunto de datos, resulta fácil interpretarla.

Si un conjunto de datos tiene una **distribución aproximadamente simétrica se pueden utilizar** las siguientes reglas prácticas para describir el conjunto de datos:

- Aproximadamente el 68 % de las observaciones quedan a **una desviación estándar** de su media (es decir, dentro del intervalo $\bar{x} \pm s$)
- Aproximadamente el 95 % de las observaciones quedan **a dos desviaciones estándar** de su media (es decir, dentro del intervalo $\bar{x} \pm 2s$)
- Casi todas las observaciones quedan a **tres desviaciones estándar** de su media (es decir, dentro del intervalo $\bar{x} \pm 3s$)

La **regla empírica** es el resultado de la experiencia práctica de investigadores en muchas disciplinas, que han observado muy diferentes tipos de conjuntos de datos de la vida real.

Regla Empírica cont.



Fuente : Estadística Elemental. Johnson – Kuby pag 82

Transformaciones Lineales y Estandarización

Una transformación:

Se tiene datos del número de niños por hogar de 10 viviendas de un barrio:

2, 3, 2, 2, 1, 0, 3, 2, 1, 4

El **promedio** es 2,0 y **desviación estándar** =es 1,1547 niños

a) Queremos describir el número de personas en cada vivienda y supongamos que en cada vivienda hay 2 adultos, entonces: 4, 5, 4, 4, 3, 2, 5, 4, 3, 6

• Encontrar el promedio y la desviación estándar de esta nueva variable y comparar con las observaciones originales.

• ¿Cómo cambia el promedio? ¿Cómo cambia la desviación estándar?

• **Describir cómo afecta al promedio y la desviación estándar el sumar una constante a cada observación**.

b) Supongamos que cada niño recibe una mesada semanal de $500. Describir ahora el gasto en mesadas de cada vivienda.

• Encontrar el promedio y la desviación estándar y comparar con los obtenidos de las observaciones originales.

• ¿Cómo cambia el promedio?, ¿Cómo cambia la desviación estándar?

• **Describir cómo afecta al promedio y la desviación estándar el multiplicar una constante a cada observación**.

Transformaciones Lineales y Estandarización cont.

Si X representa una variable, $\overline{x}$ su promedio y s_X su desviación estándar. Sea $\boldsymbol{Y=aX+b}$, una **transformación lineal** de X, entonces:

El promedio de Y es: $\overline{y} = a\overline{x} + b$

y la desviación estándar: $s_Y = |a| s_X$

NOTA: |a| es el valor absoluto o módulo de la constante a, donde a es cualquier valor positivo o negativo y su módulo es siempre positivo.

Si X representa una variable, $\overline{x}$ su promedio y s_X su desviación estándar. Llamaremos z a la variable estandarizada:

$$z = \frac{x - \overline{x}}{s_X}$$

Una variable está **estandarizada** si la variable tiene media cero y desviación estándar uno.

Note que la **variable estandarizada** $\frac{x - x}{s_X}$ se puede expresar de la forma de una **transformación lineal**:

$$\frac{x - \overline{x}}{s_X} = \left(\frac{1}{s_X}\right)x + \left(-\frac{\overline{x}}{s_X}\right) \text{ con } a = \left(\frac{1}{s_X}\right), \text{ y } b = \left(-\frac{\overline{x}}{s_X}\right).$$