CIENCIA
DE DATOS

Estadísticas Descriptivas

ESTADÍSTICA PARA CIENCIA DE DATOS

PROF. ESTEBAN BALLESTERO

Parámetro vs población

- Es muy importante distinguir entre un parámetro de población y un estadístico de muestra.
- Un parámetro es un resumen numérico de una población. Como una población tiene muchos miembros o elementos, el valor del parámetro es difícil de conocer, al menos que se realice un censo.
- Cualquier medida calculada de un subconjunto de la población (muestra) es un estadístico.

Parámetro vs población

Situación	Parámetro Poblacional	Estadístico de muestra
Encuesta política a favor de candidatos presidenciales	Proporción de votantes a favor de un candidato	Proporción a favor de un candidato de 1000 encuestados
Nivel de DDT (pesticida) en huevos de una especie de ave rapaz que habita en Guatemala	Nivel medio de DDT en todos los huevos de la especie en Guatemala	Nivel medio de DDT en 150 huevos de la especie colectados en Guatemala

- Para distinguir entre parámetros poblacionales y estadísticos de muestra se usan diferentes símbolos. Generalmente se usan letras Griegas para referirse a parámetros y letras Romanas para referirse a estadísticos.
- Se debe recordar que los estadísticos de muestra estiman el parámetro de población y la Estadística tiene como objetivo evaluar que tan bien el estadístico estima el verdadero parámetro de la población.

Medidas de tendencia central

Cuando se trabaja con datos, el uso de los datos crudos para presentar información, no son una buena estrategia, por lo que se hace necesario utilizar algunas medidas de resumen de estos.

La medidas de tendencia central son valores que resumen los datos o los representan y reciben su nombre porque usualmente sus valores se ubican en centro de la distribución. También se les conoce como medidas de posición central

A saber:

- Media, media aritmética o promedio
- Mediana
- Moda

Media

Poblacional	Muestral
$\mu = \frac{\sum X}{N}$	$\bar{x} = \frac{\sum X}{n}$

N: Población
n: muestra tomada de una población
X: conjunto de datos

Mediana

- Suponga que se tiene n datos. Primeramente se ordenan los datos ascendentemente (de menor a mayor)
- Si n es impar, la mediana corresponde al dato que se encuentra en la posición $\frac{n+1}{2}$, es decir: $\boldsymbol{x_{\frac{n+1}{2}}}$.
- Si n es par, la mediana corresponde al promedio de los datos que queden en las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$, es decir: $\frac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2}$.

Moda

- Es el dato que más se repite o el dato de mayor frecuencia
- Las series de datos pueden ser:
  1. Amodal (ninguna moda).
  2. Unimodal o Monomodal (una moda).
  3. Bimodal (dos modas).
  4. Trimodal (tres modas).
  5. Tetramodal (cuatro modas).
  6. Polimodal (más de cuatro modas).

Ejemplo1:

Valores de DDT (ppm) en una muestra de 20 huevos de una especie de ave rapaz en México.

18 18 21 21 21 21 21 21 21 21 25 25 25 25 25 25 35 35 50 70

- En lugar de reportar los 20 valores se quiere reportar un valor “típico” que caracterice a la muestra.
- Una medida apropiada puede ser un valor en el medio reconociendo que algunos huevos van a tener menos DDT y otros van a tener más DDT debido a variación en el tipo y cantidad de alimento consumido por las hembras.
- 3 medidas de tendencia central son la media, la mediana y la moda.

Ejemplo1:

Datos ordenados de forma ascendente

18 18 21 21 21 21 21 21 21 21 25 25 25 25 25 25 35 35 50 70

21 + 25/2
Mediana

Tamaño de muestra	Media muestral	Mediana muestral	Moda muestral
20	544/20 = 27,2 ppm	(21 + 25)/2 = 23 ppm	21 ppm

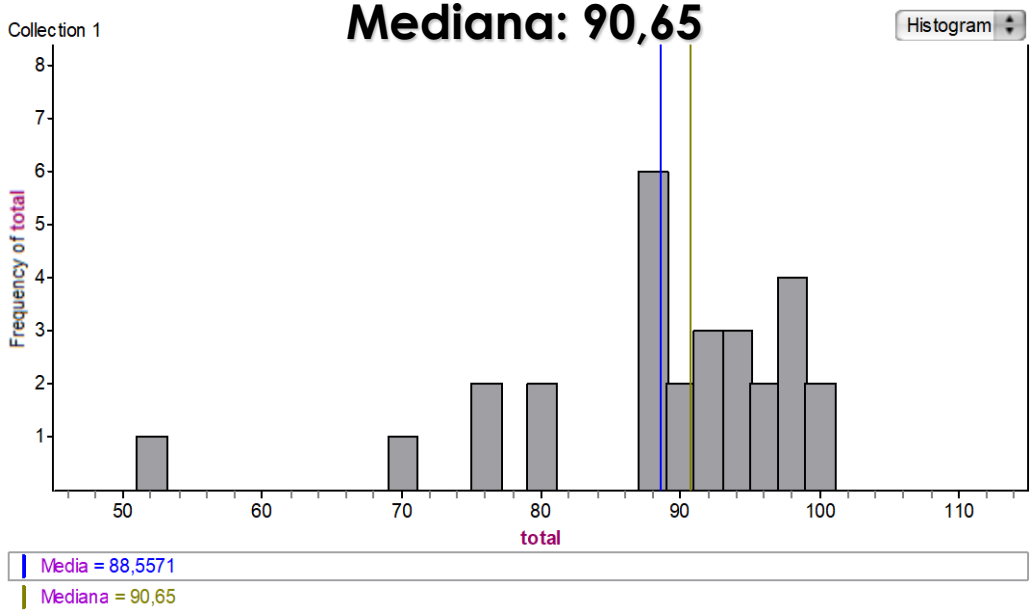
Ejemplo 2: MTC y valores extremos

El siguiente gráfico de barras muestra los porcentaje de alfabetización de los países de América Latina.

Los valores extremos de los datos afectan significativamente a la media, no así a mediana y a la moda

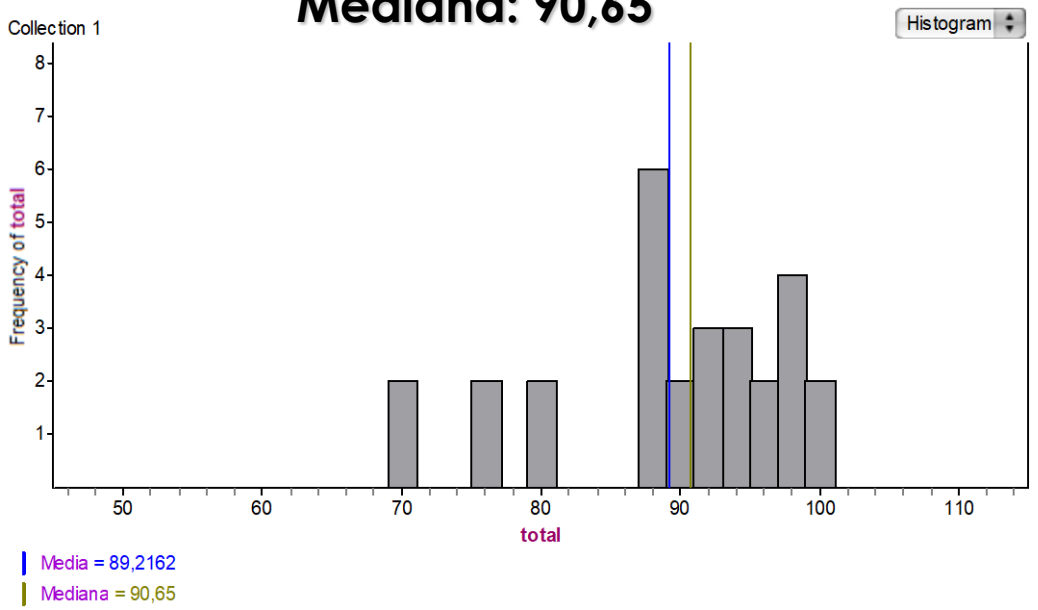
Media: 88.5%

Mediana: 90,65



Media: 89.17%

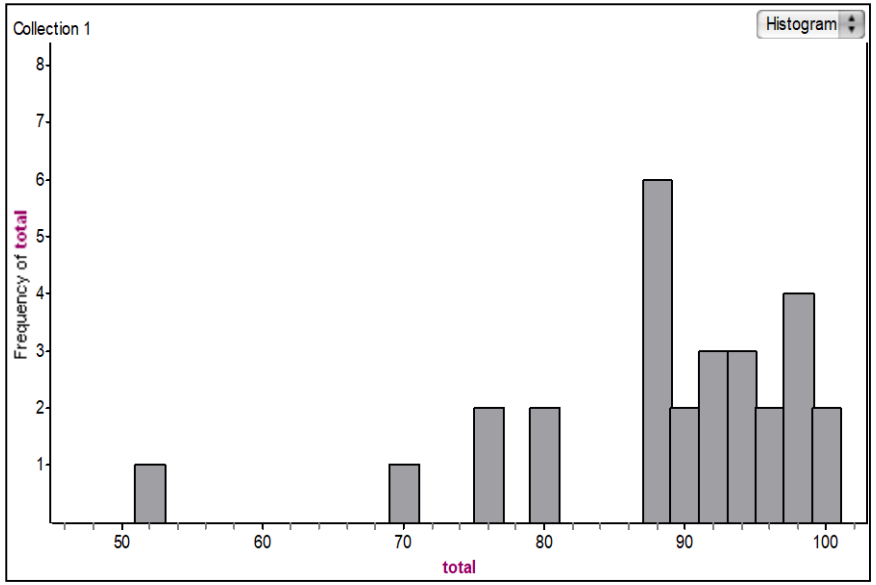
Mediana: 90,65



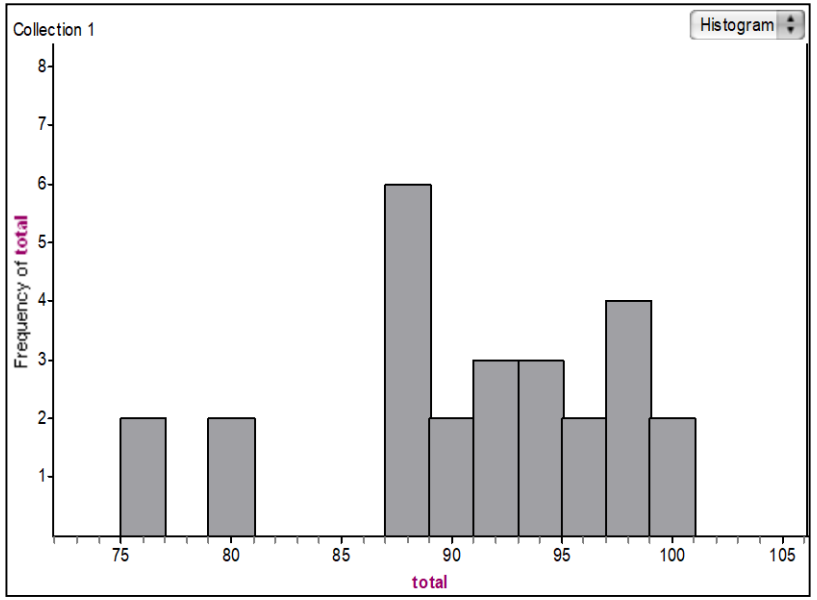
Ejemplo 2: MTC y número de datos

El número de datos de una muestra afecta significativamente a la media, la mediana, no a la moda

Media: 88.5%
Mediana: 90.6%



Media: 90.7%
Mediana: 91.3%



Conclusiones: MTC

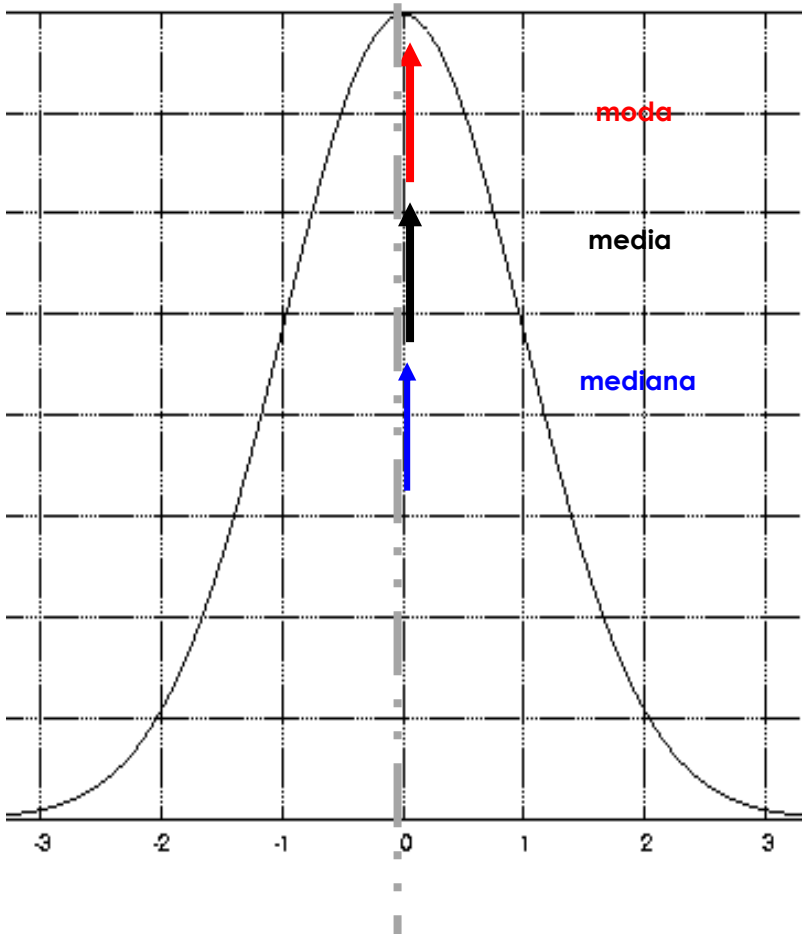
Media Símbolo: $\bar{x}$	Es válida para datos de Intervalo o cociente Es sensible a valores atípicos o extremos, así como al número de observaciones
Mediana Símbolo:$\tilde{x}$	Divide al conjunto de datos en dos parte con igual número de observaciones. Es útil para datos ordinales, intervalos o cocientes. No es sensible a valores atípicos, pero sí se ve afectada si se modifica el número de datos
Moda Símbolos: $\hat{x}$	Es válido para todas las escalas

Conclusiones: MTC

- Es muy raro que las 3 medidas sean exactamente iguales.
- Si los datos tiene una distribución perfectamente simétrica entonces las medidas serán exactamente iguales (media = mediana = moda).
- Como la distribución se torna asimétrica, la media muestral se aleja de la mediana en la dirección de la cola más larga.
- La media es afectada por valores extremos, pero la mediana no.

MTC y distribuciones sesgadas

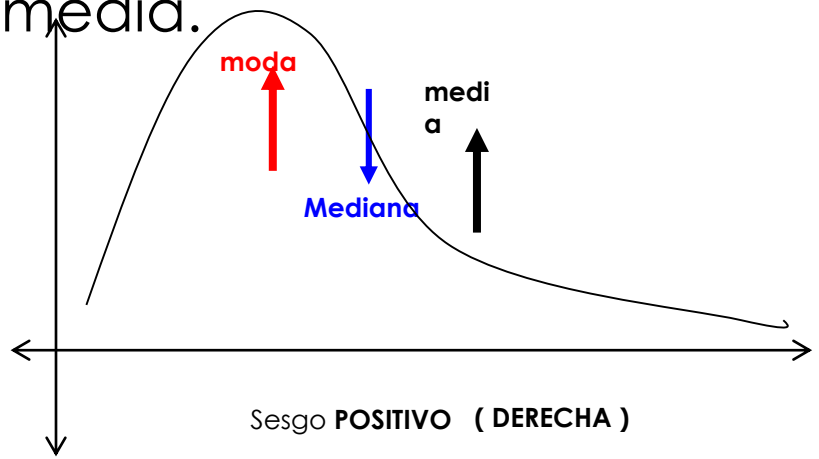
En las distribuciones simétricas las medidas de tendencia central (moda, media y mediana) coinciden con el eje de simetría



Distribución sesgada: se representa por una curva asimétrica.

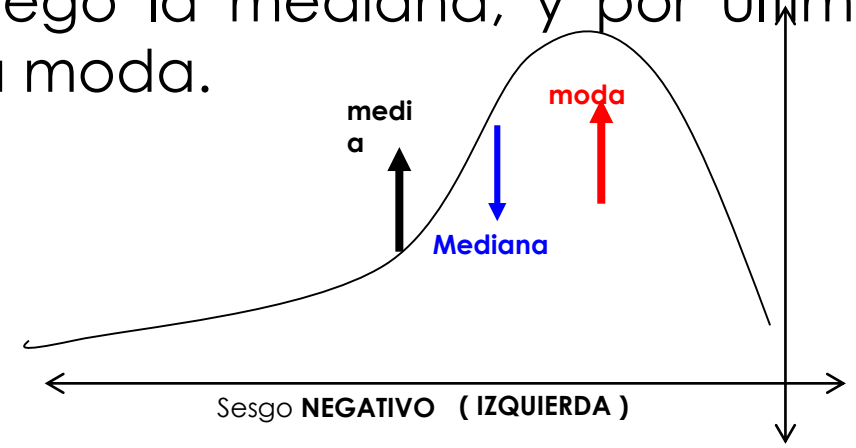
Sesgo **POSITIVO**

En una distribución sesgada a la **DERECHA**, la cola es más larga a la derecha que a la izquierda, y aparece primero la moda, luego la mediana, y por último la media.

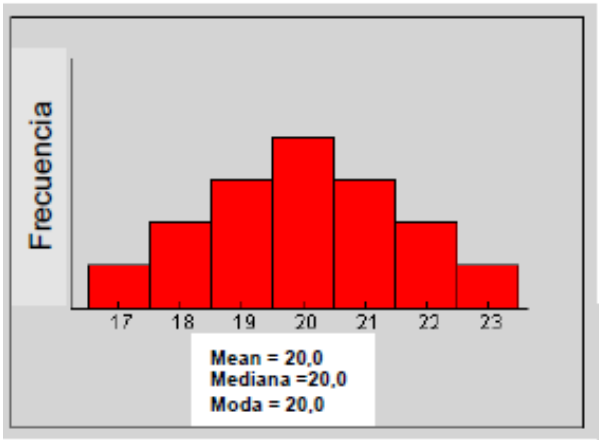


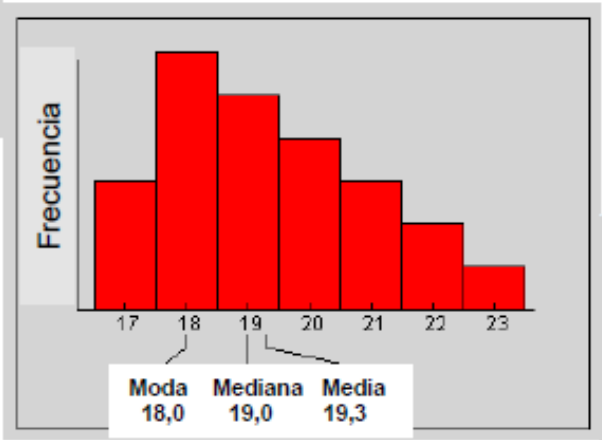
Sesgo **NEGATIVO**

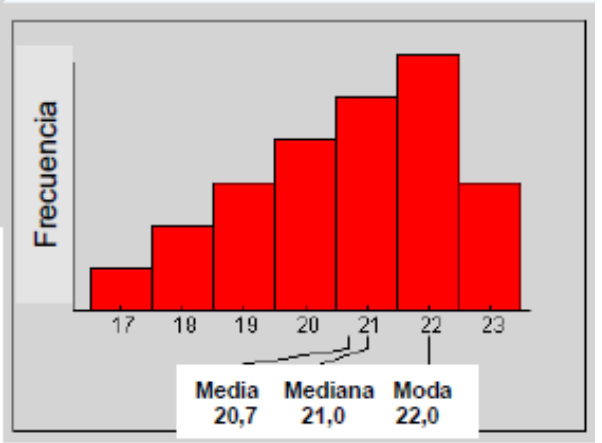
En una distribución sesgada a la **IZQUIERDA**, la cola es más larga a la izquierda que a la derecha, y aparece primero la media, luego la mediana, y por último la moda.



Conclusiones: MTC







Ejemplo 3:

Si tuviese que reportar la medida de tendencia central o localización para describir los salarios de los empleados de su empresa, ¿qué medida escogería?

Asimetría:

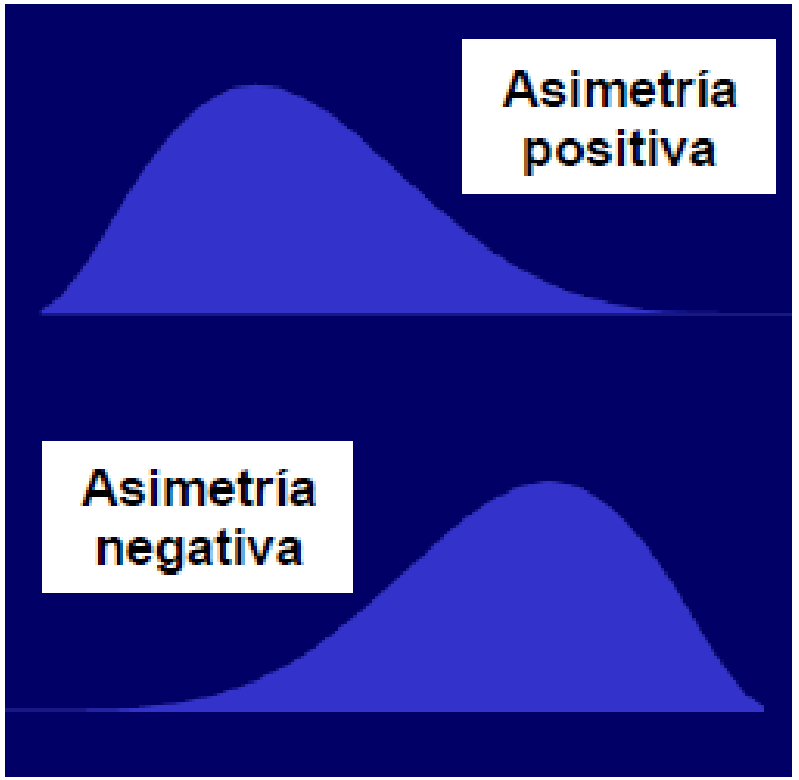
- La asimetría describe como la muestra difiere en su forma de una distribución simétrica.
- La asimetría es el 3er momento dividido por el desvío estándar al cubo (s^3) (es el 3er momento de una distribución estandarizado).
- Una distribución normal es simétrica y tiene asimetría = 0.
- Una asimetría positiva indica una cola larga hacia la derecha.
- Una asimetría negativa indica un acola larga hacia la izquierda.

3er Momento →

$$m_3 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^3}{n}$$

$$s_3 = s_y^{\,3} = (\sqrt{s^2})^3$$

$$Asimetría = g_1 = \frac{m_3}{s_3}$$



Ejemplo 3:

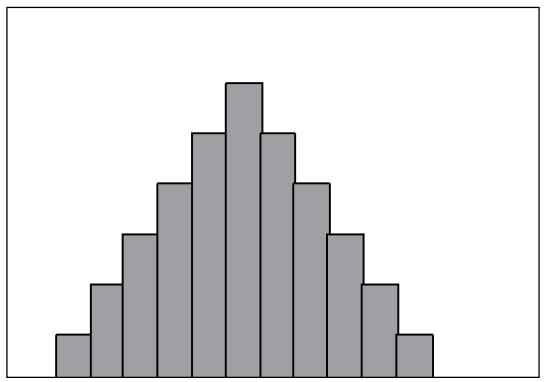
Un comité escolar de una pequeña ciudad quería determinar el promedio del número de niños por familia en esa ciudad. Ellos dividieron el número total de niños por 50, que es el número de familias. ¿Cuál de las siguientes afirmaciones es verdadera si sabemos que el promedio de niños por familia es de 2,2?

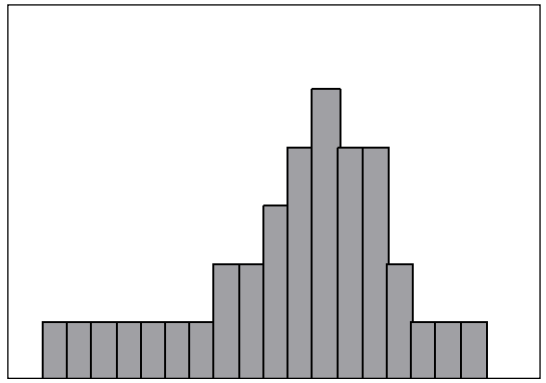
1. La mitad de las familias de esa ciudad tiene más de dos niños
2. La mayor parte de las familias de esa ciudad tienen 3 niños en lugar de 2
3. Existe un total de 110 niños en esa ciudad
4. Hay 2,2 niños en esa ciudad por cada adulto
5. El número de niños por familia más común es de 2
6. Ninguna de las anteriores

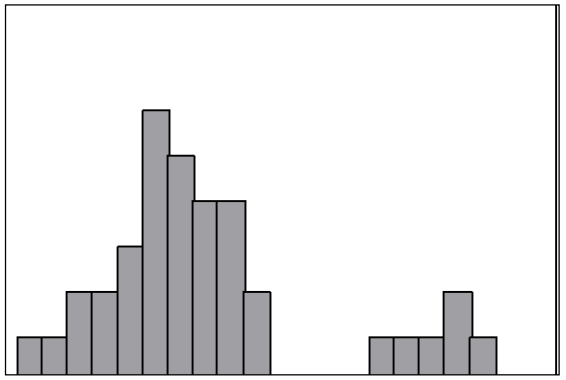
Ejemplo 4:

La siguiente tabla posee los valores de la media, la moda y la mediana de tres distribuciones diferentes. Además, bajo la tabla aparece el gráfico de cada una de estas distribuciones, sin ningún orden. Usted debe anotar el nombre de la distribución sobre la línea que aparece bajo el gráfico, de acuerdo con la distribución que considere sea la que ese gráfico representa

Medida	Distribución 1	Distribución 2	Distribución 3
Moda	5	6	12
Mediana	6	6	11.5
media	7.17	6	10.69







Ejemplo 5:

Hay 10 personas en un ascensor: 4 mujeres y 6 hombres. El peso medio de las mujeres es de 60 kilos y el de los hombres 80. ¿Cuál es el peso medio de las 10 personas del ascensor?

Variabilidad y medidas de dispersión

La variabilidad es el corazón de la estadística y es el componente fundamental del pensamiento estadístico (Pfannkuch, 1997; Pfannkuh & Wild, 2004; Shaughnessy, 1997; Garfield & Ben-Zvi, 2005; Hammerman &Rubin, 2004; Watson & Kelly, 2002).

Variabilidad y medidas de dispersión

La variabilidad es omnipresente en nuestro mundo y fuera de él, por lo que asumir homogeneidad de resultado en un estudio de un fenómeno cualquiera, sería pecar de ingenuo.

La esencia de la estadística se centra en tratar de explicar la variabilidad de los datos para entender el entorno

Percentiles y mediana

- Un percentil es el valor de una variable por debajo del cual se encuentra cierto porcentaje de las observaciones.

  Por ejemplo: P_{20} (percentil de orden 20) es el valor de la variable, tal que, el 20% de los datos son iguales o inferiores a P_{20}.

  Puede o no coincidir con un valor observado

- El 25 percentil es el 1er cuartil, el 50 percentil es la mediana, y el 75 percentil es el 3er cuartil.

- Por ejemplo, si un estudiante mide 185 cm y constituye el 90 percentil, entonces 90% de los estudiantes tienen alturas menos de 185 cm y 10% de los estudiantes tienen pesos mayores a 185 cm.

Percentiles y mediana

La fórmula de la derecha no nos da información sobre cual es el percentil, sino la posición que ocupa el valor que representa al percentil, de ahí que los datos deben estar ordenamos.

$$P_m = \frac{m}{100}(n+1)$$

donde

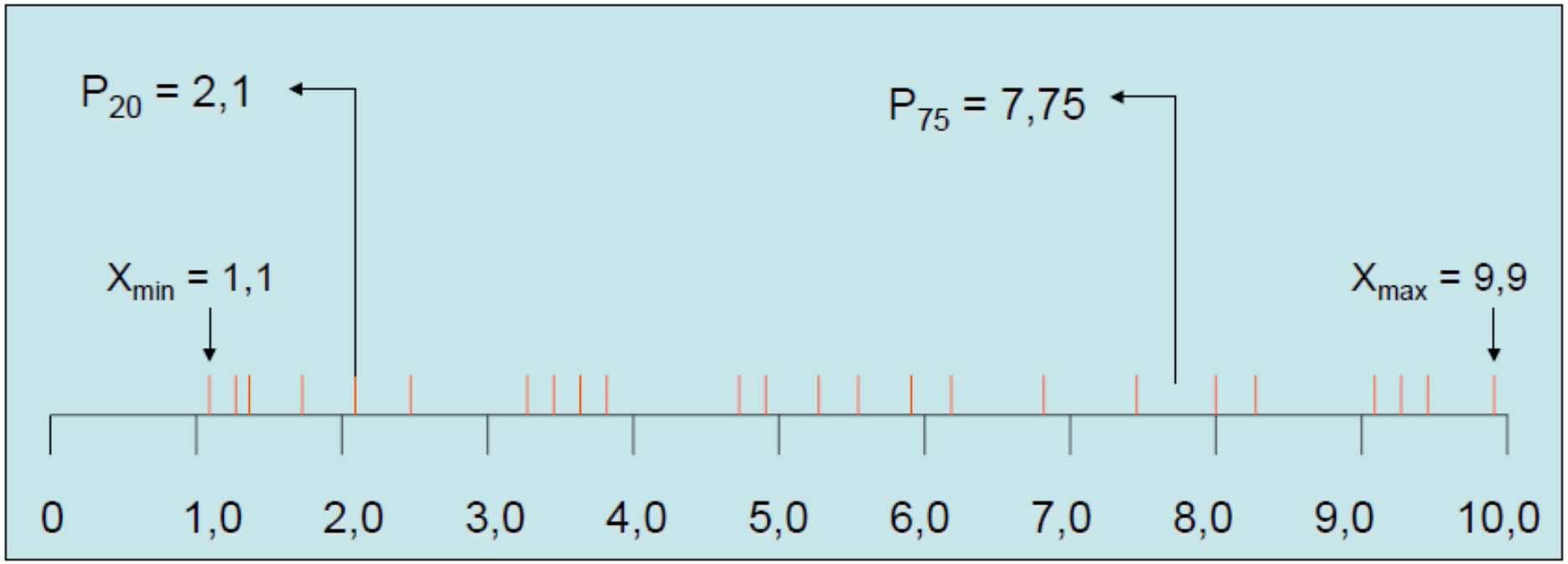
m : Número que indica el percentil deseado.
n : Número total de datos.
Q_1 (cuartil 1)= P_{25} (Percentil 25)
Q_2 (cuartil 2)= P_{50} (percentil 50)= D_5 (decil 5)= mediana

Percentiles y mediana

$P_{20} = X_{(0,20 \times 24)} = 4,8 \cong 5$ (la 5ta observación) = 2,1

$P_{75} = X_{(0,75 \times 24)} = X_{18} + X_{19}/2$ =(el promedio de las observaciones 18 y 19)= (7,5 + 8,0)/2 =7,75

$P_{50} = X_{(0,50 \times 24)} = X_{12} + X_{13} = 4,9 + 5,2/2 = 5,05$



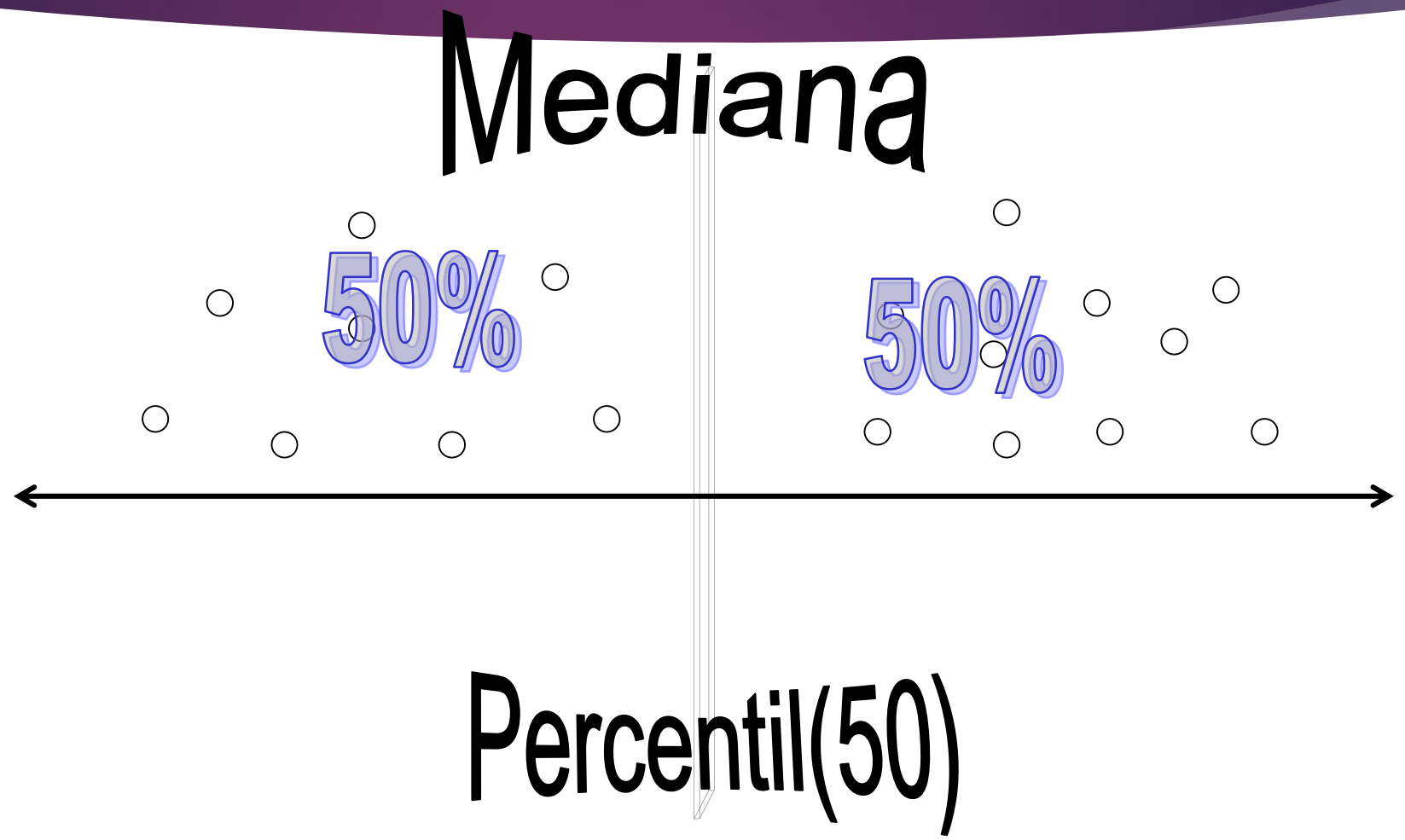
Percentiles y mediana

- Los datos deben de estar ordenados
- El percentil es una medida de dispersión
- Los percentiles que podemos calcular oscilan entre 0 y 100
- El ***percentil(n)*** nos indica que para un conjunto de datos determinado, el ***n%*** de los datos se encuentran a la izquierda de este valor

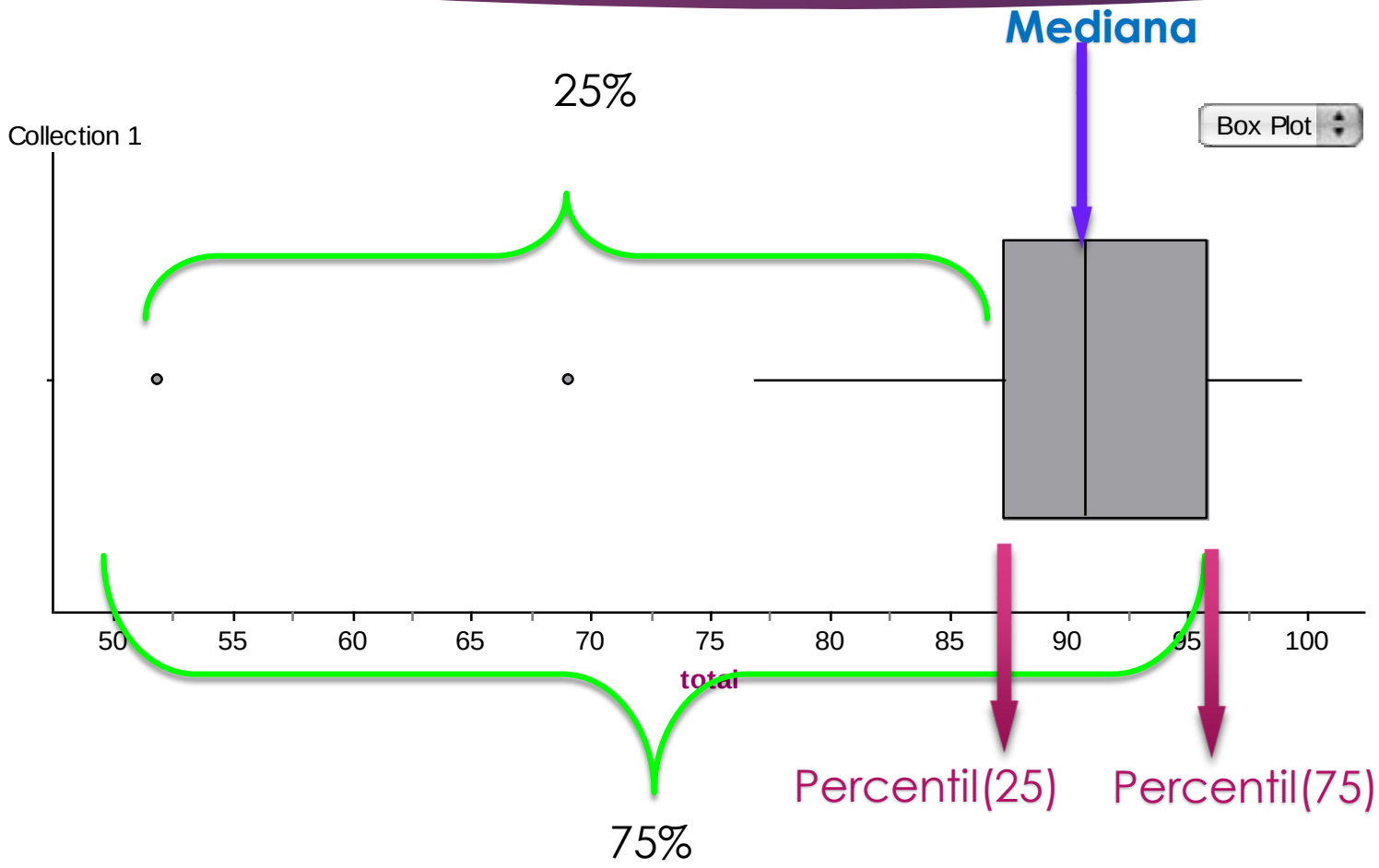
Percentiles y mediana

- La mediana, divide al conjunto de datos en dos subconjuntos donde cada uno de ellos reúne el 50% de los datos
- Estos subconjuntos son mutuamente excluyentes
- La mediana es equivalente al percentil(50)

Percentiles y mediana



Percentiles y mediana

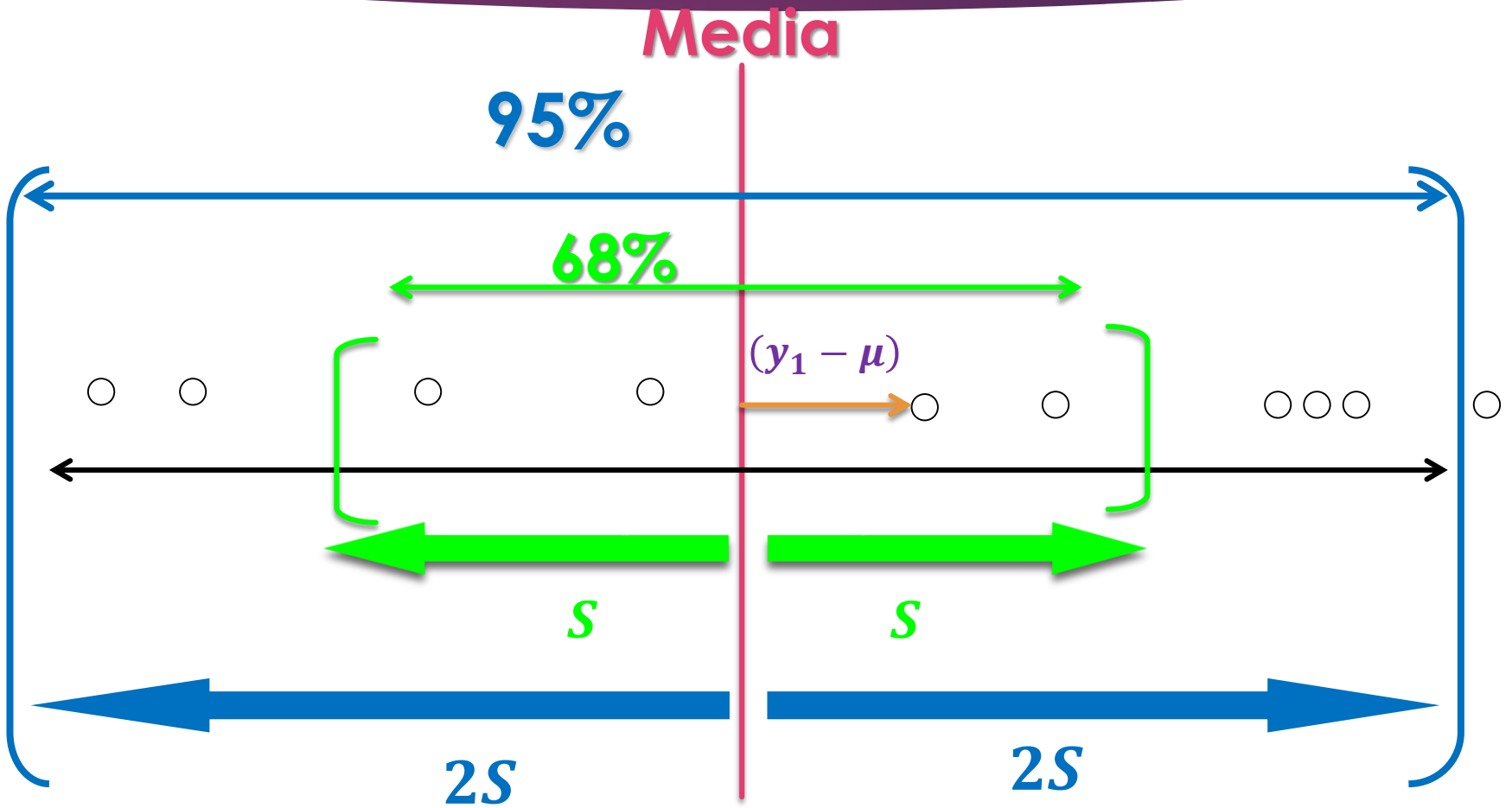


Desviación estándar y media

La desviación estándar es una medida de dispersión

Poblacional	Muestral
$s = \sqrt{\frac{\sum(x_i - \overline{x})^2}{N}}$	$s = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}}$

Desviación estándar y media



Conclusión

- La desviación estándar es la medida de dispersión de los datos con respecto a la media, mientras que los percentiles es la medida de dispersión de los datos con respecto a la mediana.
- La mediana constituye el valor central de los percentiles
- La media representa el valor central a partir del cual se comparan los datos para obtener la desviación estándar

Coeficiente de variación

- El coeficiente de variación es una medida de variabilidad.
- Se calcula dividiendo el desvío estándar por la media (y se lo multiplica por 100 para expresarlo como porcentaje).
- Por ejemplo, si un grupo de datos tiene un CV = 16,5% y otro grupo de datos tiene un CV = 25%, se puede concluir que el primer grupo de datos es menos variable que el segundo.

$$CV = \frac{s}{\bar{y}} \times 100$$

Desvío estándar

Media

Ejemplo 6:

a. Los siguientes datos reflejan las estaturas en centímetros de niñas de 12 y 16 años. ¿Cuál grupos es más variables, el de niñas de 16 o 12 años?

Edad	Estatura media	Desviación estándar
12	84	3
16	160	5

b. En un estudio sobre el peso de conejos adultos, se determinó que el peso promedio es de 4,3kg con una desviación estándar de 0,7 kg, mientras que las mediciones de vacas adultas de la raza Simmental dio como resultado una media de 680 kg con una desviación estándar de 92 kg. ¿Cuál muestra presenta mayor variabilidad?

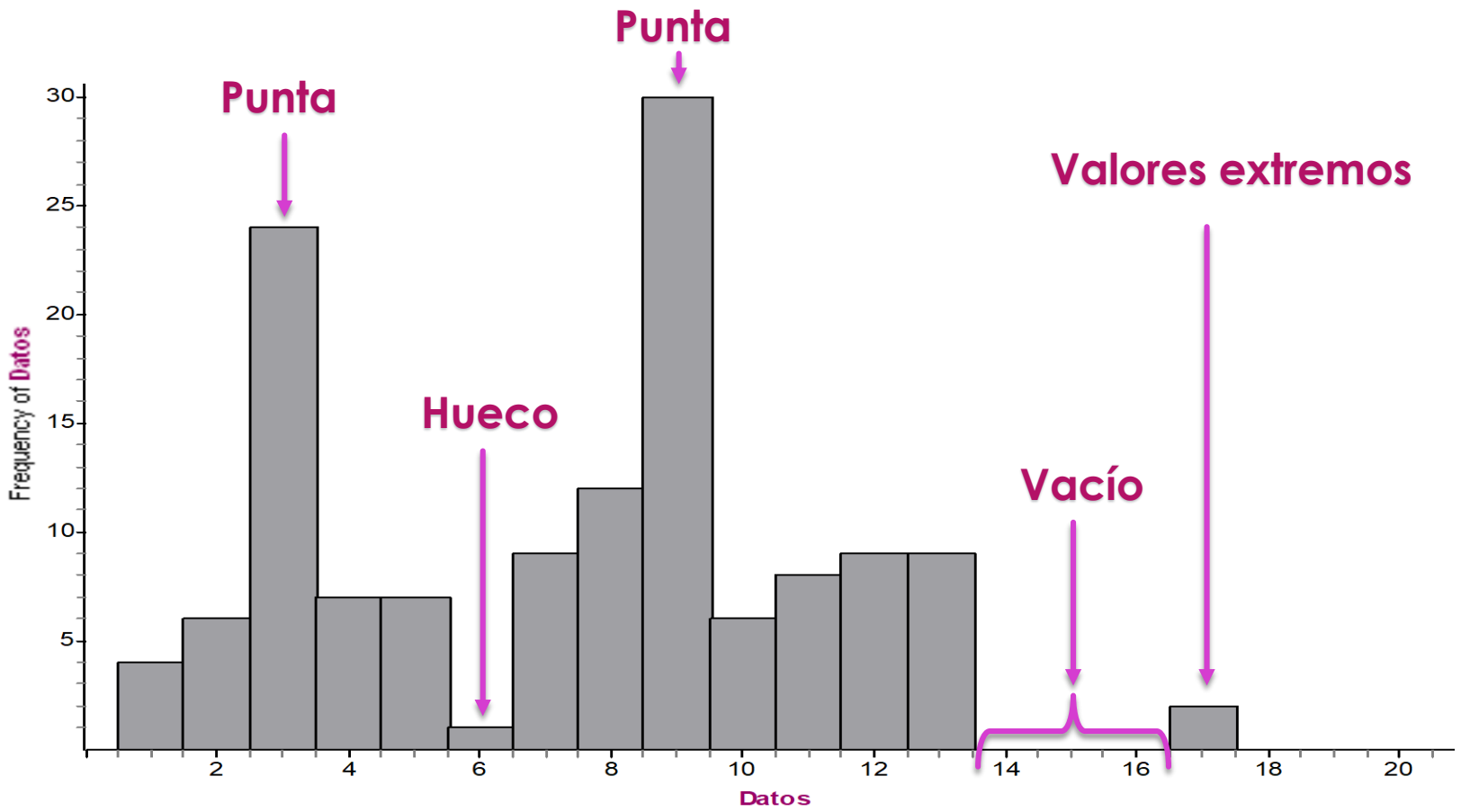
Análisis gráfico

¿Qué debemos observar en histograma?

F.Mosteller, S.E. Fienberg y R.E.K. Rourke en su libro : *Beginning statistics with data analysis,* consideran que cuando analizamos un histograma debemos tomar en cuenta las siguientes irregularidades:

Puntas, huecos, crestas, valles, vacíos, eje de simetría y valores extremos

Análisis gráfico



Análisis gráfico