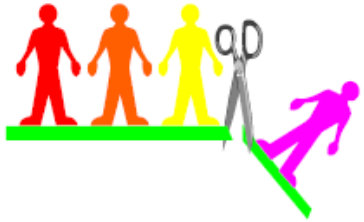
Estadística: conceptos básicos y definiciones.

Conceptos básicos

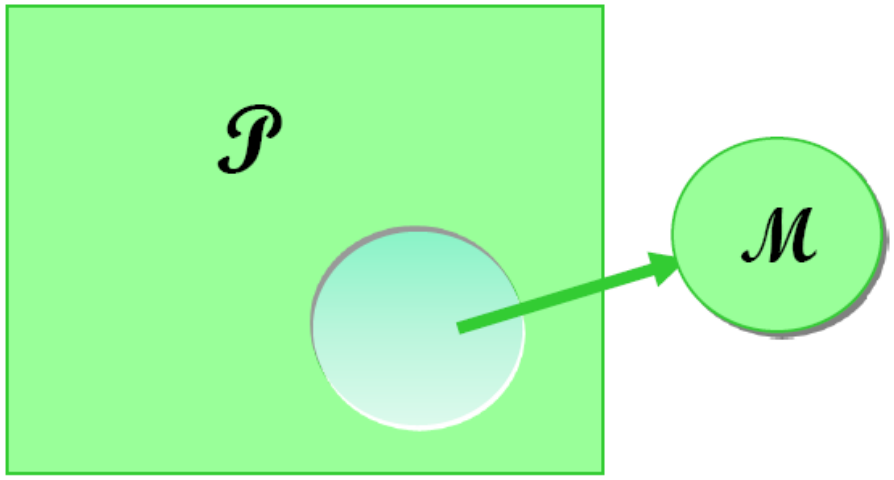
- **Población:** es el conjunto sobre el que estamos interesados en obtener conclusiones (hacer inferencia).
  - Normalmente es demasiado grande para poder abarcarlo.

- **Muestra:** es un subconjunto de la población al que tenemos acceso y sobre el que realmente hacemos las observaciones (mediciones)
  - Debería ser "representativo"
  - Esta formado por miembros "seleccionados" de la población (individuos, unidades experimentales).



Conceptos básicos cont.

- **Muestra Aleatoria:** es una muestra bien representativa de la población. Se considera que cada elemento de la población ha tenido la misma oportunidad de formar parte de la muestra. Las conclusiones basadas en una muestra aleatoria son confiables.



$\mathcal{P}$: población

$\mathcal{M}$: muestra

Conceptos básicos cont.

- **Variable:** una ***variable*** es una característica observable *que varía entre los diferentes individuos* de una población. La información que disponemos de cada individuo es resumida en **variables**.
- **Dato:** es un valor particular de la variable

- En los individuos de la *población* chilena, de uno a otro ***es variable***:
  - El grupo sanguíneo
    - {A, B, AB, O}
  - Su nivel de felicidad "declarado"
    - {Deprimido, ……, Muy Feliz}
  - El número de hijos
    - {0,1,2,3,...}
  - La altura
    - {1.62 , 1.74, ...}

Conceptos básicos cont.

- **Parámetro:** Es una cantidad numérica calculada sobre una población.
  - La altura media de los individuos de un país.
  - La idea es resumir toda la información que hay en la población en unos pocos números (parámetros).

- **Estadístico:** Ídem (cambiar población por muestra).
  - La altura media de los que estamos en este aula.
    - Somos una muestra (¿representativa?) de la población.
  - Si un estadístico se usa para aproximar un parámetro también se le suele llamar estimador.

Conceptos básicos cont.

- **Censo:** es un listado de una o más características de todos los elementos de una población. Los censos poblacionales se hacen cada 10 años a nivel mundial.

- **Encuesta:** Es un listado de una o más características de todos los elementos de una muestra.

Definición de Estadística

La estadística es la Ciencia de la

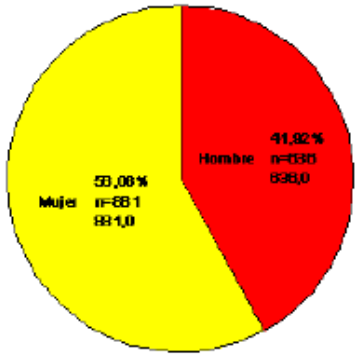
Descriptiva
- **sistematización, recogida, ordenación y presentación** de los datos referentes a un fenómeno que presenta variabilidad o incertidumbre para su estudio metódico, con objeto de

Probabilidad
- **deducir las leyes** que rigen esos fenómenos

Inferencia
- y poder hacer previsiones sobre los mismos, tomar **decisiones** u obtener **conclusiones**.

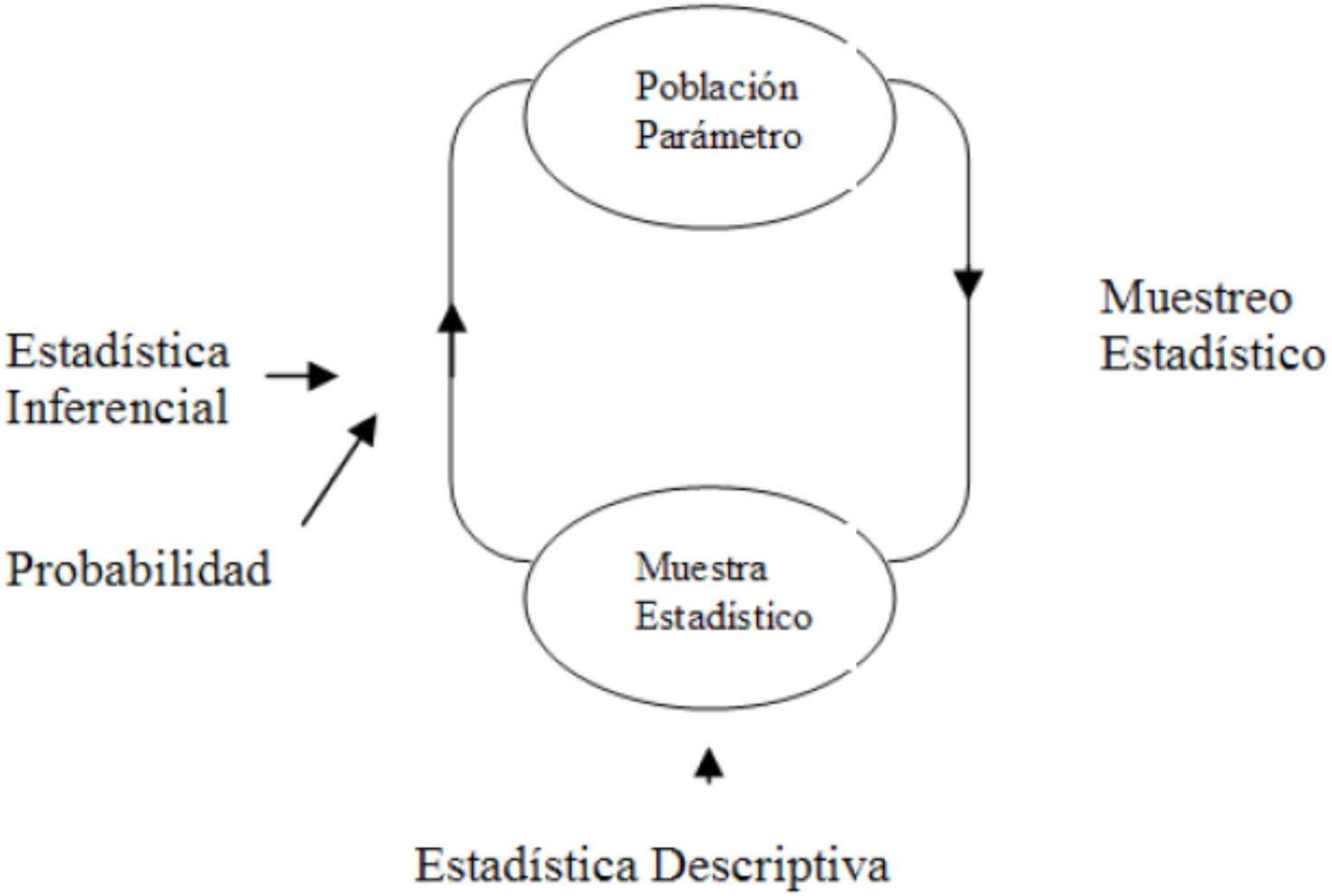
División de la Estadística

- **Estadistica Descriptiva:** Conjunto de técnicas y métodos que son usados para recolectar, organizar, y presentar en forma de tablas y gráficas información numérica. También se incluyen aquí el cálculo de medidas estadísticas de centralidad y de variabilidad.



- **Estadistica Inferencial:** Conjunto de técnicas y métodos que son usados para sacar conclusiones generales acerca de una población usando datos de una muestra tomada de ella.

Gráfica del Análisis Estadístico



Pasos en un estudio estadístico

- **Plantear hipótesis sobre una *población:***
  - **Los fumadores tienen *"más ausencias"* laborales que los no fumadores.**
  - ¿En qué sentido? ¿Mayor número? ¿Tiempo medio?

- **Decidir qué datos recoger (diseño de experimentos)**
  - Qué individuos pertenecerán al estudio (*muestras*).
    - Fumadores y no fumadores en edad laboral.
    - Criterios de exclusión: ¿Cómo se eligen? ¿Descartamos los que padecen enfermedades crónicas?
  - Qué datos recoger de los mismos (*variables*).
    - Número de ausencias.
    - Tiempo de duración de cada ausencia.
    - ¿Sexo? ¿Sector laboral? ¿Otros factores?

Pasos en un estudio estadístico cont.

- **Recoger los datos (*muestreo*):**
  - De qué forma recolecto la información.

- **Describir (resumir) los datos obtenidos:**
  - Tiempo medio de ausencia en fumadores y no fumadores (*estadísticos*)
  - % de ausencias por fumadores y sexo (*frecuencias*), gráficos,...

- **Realizar una inferencia sobre la población:**
  - Los fumadores están de ausencia al menos 10 días/año más *de media* que los no fumadores.

- **Cuantificar la confianza en la inferencia:**
  - *Nivel de confianza del 95%*
  - *Significación del contraste: valor-p = 2% ¿?*

Técnicas de Muestreo

a) **Muestreo Aleatorio**. Se usa cuando a cada elemento de la población se le quiere dar la misma oportunidad de ser elegido en la muestra.

b) **Muestreo Estratificado**. Se usa cuando se conoce de antemano que la población está dividida en estratos, que son equivalentes a categorías y los cuales por lo general no son de igual tamaño. Luego, de cada estrato se saca una muestra aleatoria, usualmente proporcional al tamaño del estrato.

c) **Muestreo por conglomerados ("Clusters").** En este caso la población se divide en grupos llamados conglomerados. Luego se elige al azar un cierto número de ellos y todos los elementos de los conglomerados elegidos forman la muestra.

d) **Muestreo Sistemático.** Se usa cuando los datos de la población están ordenados en forma numérica. La primera observación es elegida al azar de entre los primeros elementos de la población y las siguientes observaciones son elegidas guardando la misma distancia entre si.

Tipo de Variables

- **Cualitativas**
  Si sus valores (*modalidades*) no se pueden asociar naturalmente a un número (no se pueden hacer operaciones algebraicas con ellos)
  - Nominales: Si sus valores no se pueden ordenar
    - Sexo, Grupo Sanguíneo, Religión, Nacionalidad, Fumar (Sí/No)
  - Ordinales: Si sus valores se pueden ordenar
    - Mejoría a un tratamiento, Grado de satisfacción, Intensidad del dolor
- **Cuantitativas o Numéricas**
  Si sus valores son numéricos (tiene sentido hacer operaciones algebraicas con ellos)
  - Discretas: Si toma valores enteros
    - Número de hijos, Número de cigarrillos, Num. de "cumpleaños"
  - Continuas: Si entre dos valores, son posibles infinitos valores intermedios.
    - Altura, ingreso familiar, Dosis de medicamento administrado, edad

Tipo de variables cont.

Ejemplos:

- Es buena idea codificarlas variables como números para poder procesarlas con facilidad en un computador.
- Es conveniente asignar “etiquetas” a los valores de las variables para recordar qué significan los códigos numéricos.

–**Género** (Cualitativa : Códigos arbitrarios)

1 : Hombre

2 : Mujer

–**Raza** (Cualitativa: Códigos arbitrarios)

1 : Blanca

2 : Negra, ...

–**Felicidad** Ordinal: Respetar un orden al codificar.

1 : Muy feliz

2 : Bastante feliz

3 : No demasiado feliz

- Se pueden asignar códigos a respuestas especiales como

0 : No sabe

99 : No contesta...

Ejemplo: Tipo de variables cont.

En un programa para la detección de hipertensión en una muestra de 30 hombres en edades entre 30 y 40 años, la distribución de la presión diastólica (mínima) en mm Hg fue la siguiente:

70	85	85	75	65	90	110	95	90	70
60	75	80	120	85	95	90	70	100	65
80	90	95	90	95	110	100	85	80	75

La variable en estudio es :

Presión diastólica (medida en mm de Hg)

una variable numérica continua.

Tabla de Frecuencias

- Exponen la información recogida en la muestra de manera inteligente:
  - Frecuencias absolutas: Contabilizan el número de individuos de cada modalidad.
  - Frecuencias relativas (porcentajes unitarios): Ídem, pero dividido por el total, normalizadas.
  - Frecuencias acumuladas absolutas y relativas: Acumulan las frecuencias absolutas y relativas. Son especialmente útiles para calcular cuantiles (como veremos más adelante).

Tabla de Frecuencias cont.

Ordenamos los datos en forma creciente:

60	65	65	70	70	70	75	75	75	80
80	80	85	85	85	85	90	90	90	90
90	95	95	95	95	100	100	110	110	120

La **amplitud** total **A = 120 –60**

Número de clases: **K = $30^{1/2}$ = 5.48**. Aprox. 6 clases

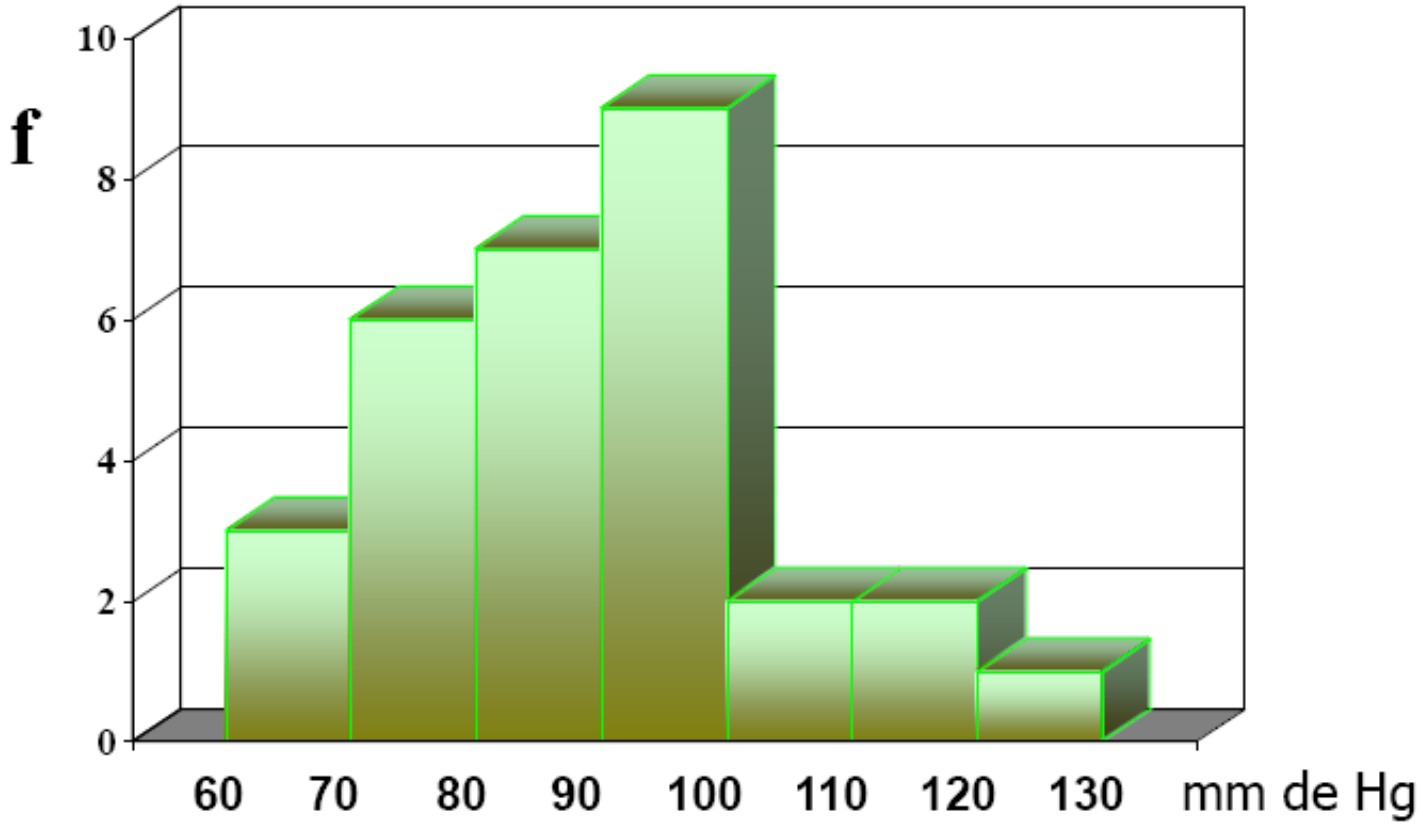
Extensión del intervalo: **H = A/ K = 60/6 = 10**

En este caso, entonces, la tabla de frecuencias tendrá aproximadamente 6 clases de amplitud 10 unidades en cada clase.

Tabla de Frecuencias cont.

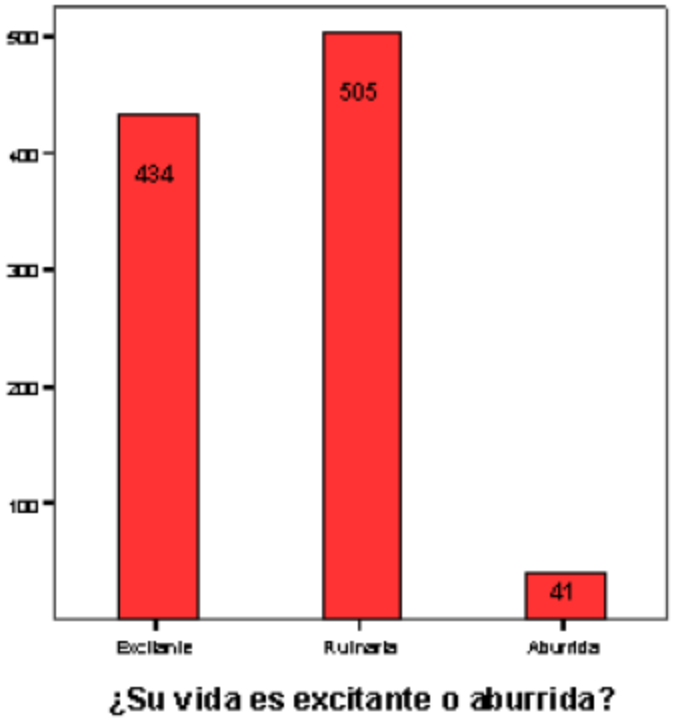
Variable X_i	Frecuencia Absoluta n_i	Frecuencia Absoluta Acumulada N_i	Frecuencia Relativa h_i	Frecuencia Relativa Acumulada H_i
60 - 70	3	3	0.1	0.1
70 - 80	6	9	0.2	0.3
80 - 90	7	16	0.23	0.53
90 - 100	9	25	0.3	0.83
100 - 110	2	27	0.07	0.90
110 - 120	2	29	0.07	0.97
120 - 130	1	30	0.03	1.00
total	30		1.0	

Histograma de la distribución de presión diastólica en mm de Hg según las frecuencias absolutas:

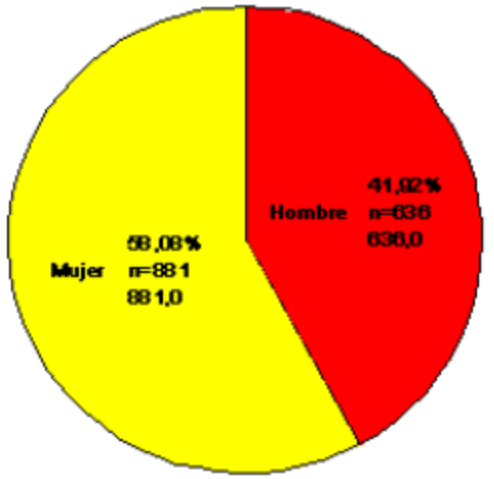


Gráficos para variables cualitativas

- Diagramas de barras
  - Alturas proporcionales a las frecuencias (abs. o rel.)
  - Se pueden aplicar también a variables discretas



- Diagramas de sectores (tartas, polares)
  - El área de cada sector es proporcional a su frecuencia (abs. o rel.)



Gráficos para variables cualitativas cont.

- Pictogramas
  - Fáciles de entender.
  - Cada modalidad debe ser proporcional a la frecuencia.

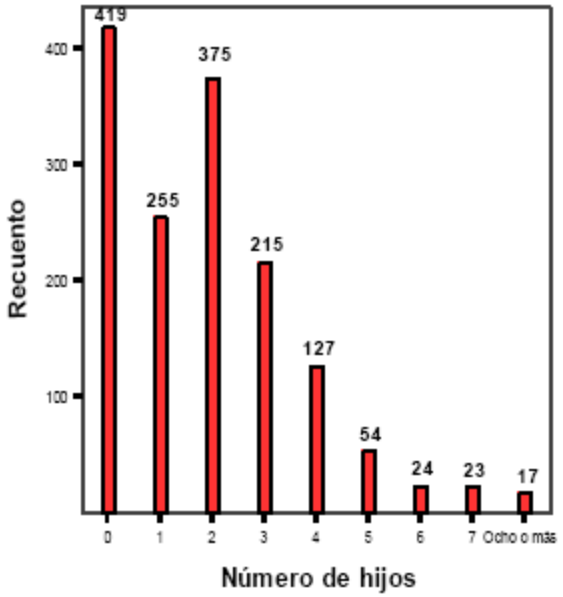
Gráficos diferenciales para variables numéricas

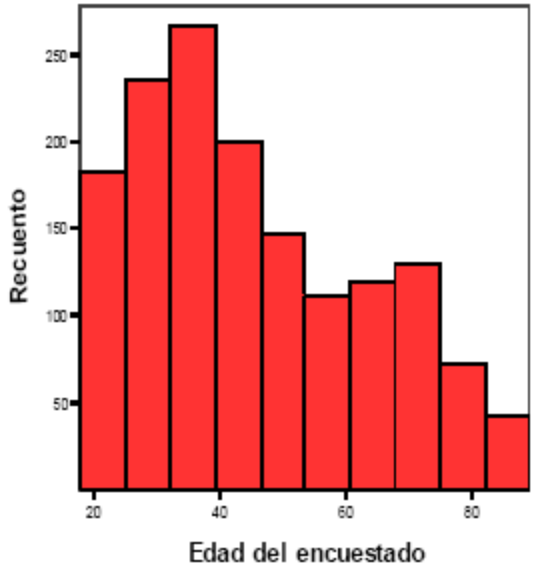
Son diferentes en función de que las variables sean discretas o continuas.

Valen con frec. absolutas o relativas.

- **Diagramas barras para variables discretas**
  - Se deja un espacio entre barras para indicar los valores que no son posibles

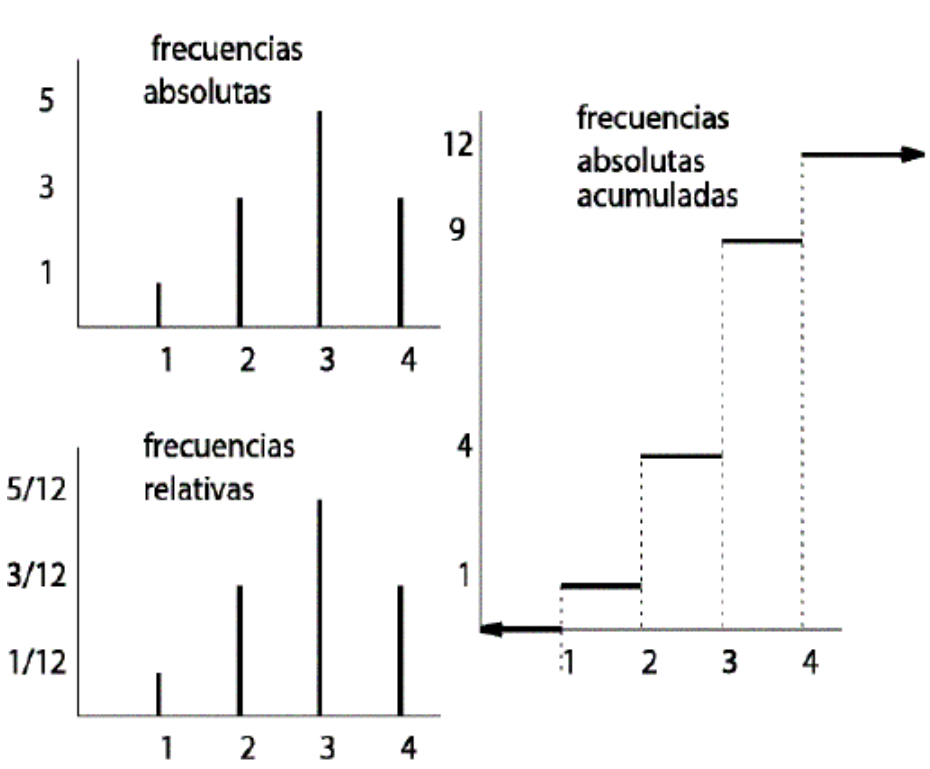
- **Histogramas para v. continuas**
  - El área que hay bajo del histograma entre dos puntos cualesquiera indica la cantidad (porcentaje o frecuencia) de individuos en el intervalo.

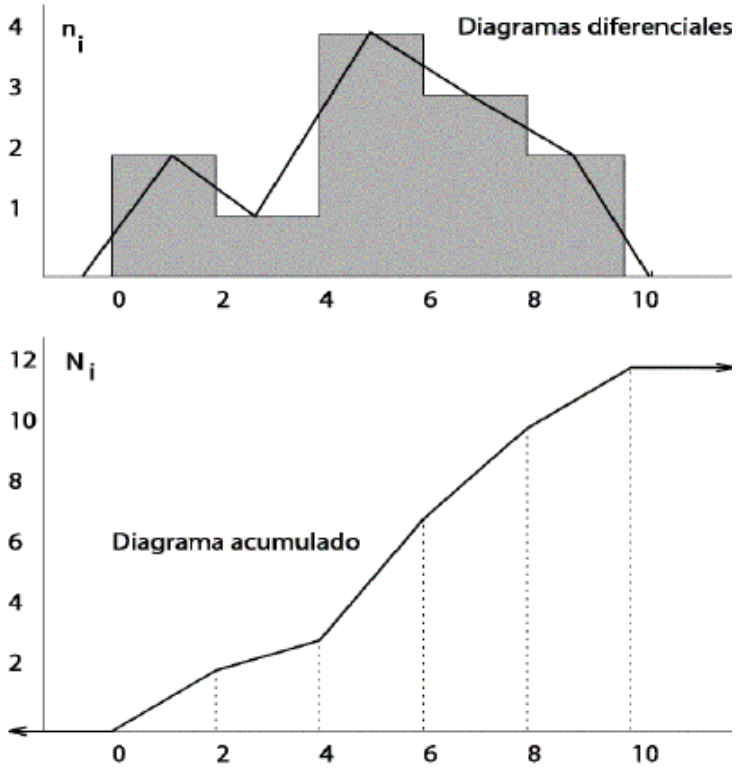




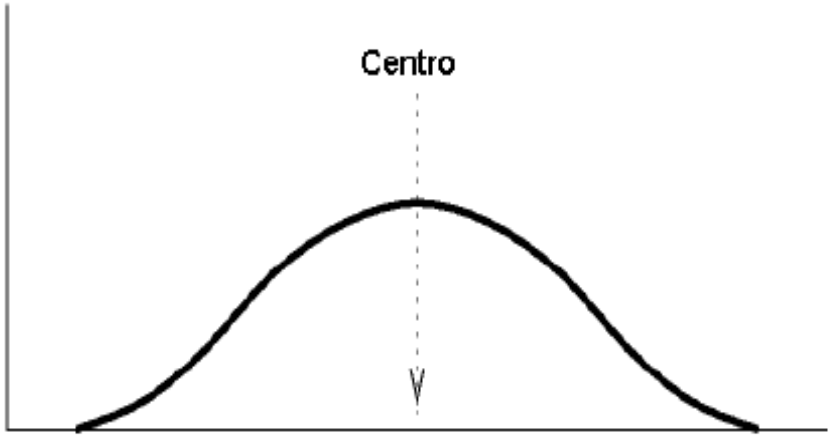
Diagramas Integrales

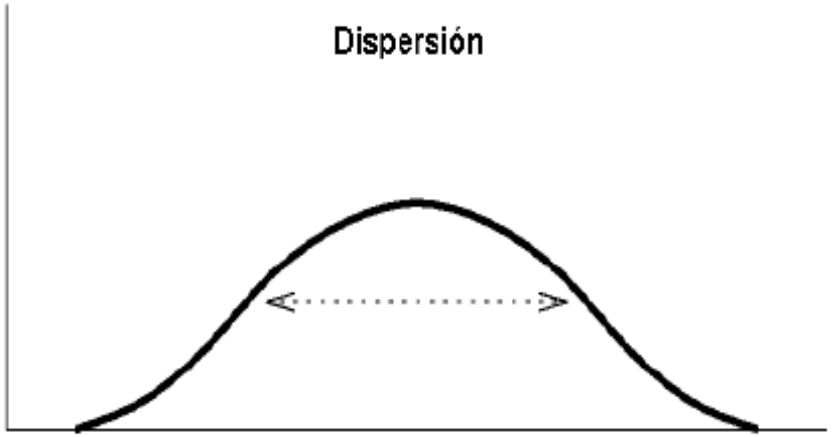
- Cada uno de los anteriores diagramas tiene su correspondiente diagrama integral. Se realizan a partir de las frecuencias acumuladas. Indican, para cada valor de la variable, la cantidad (frecuencia) de individuos que poseen un valor inferior o igual al mismo.

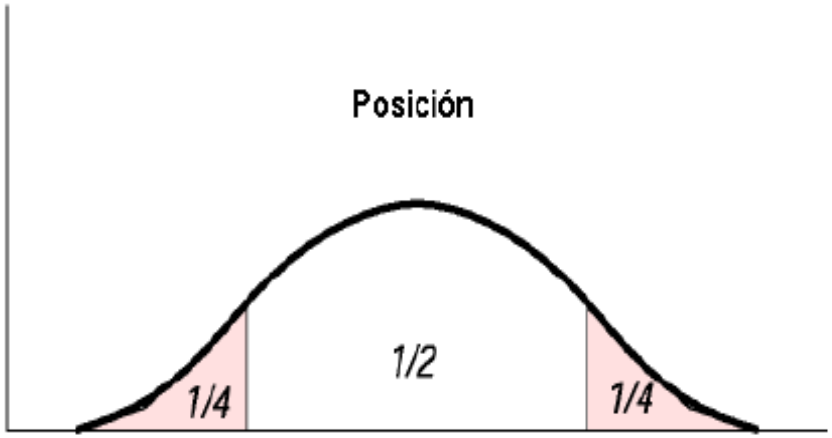


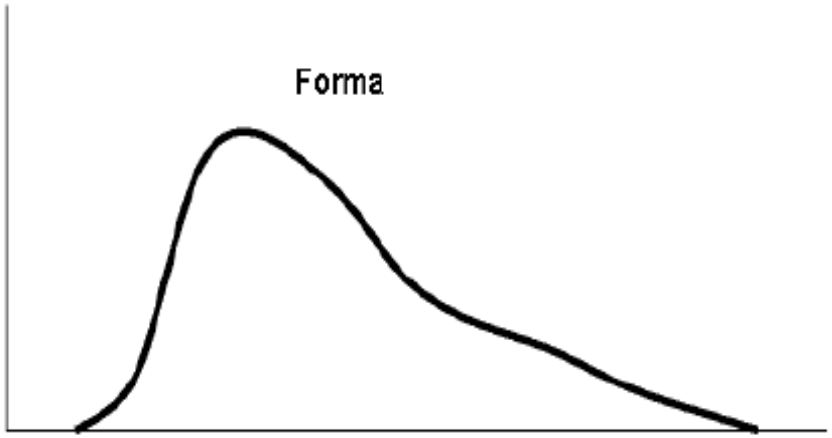


Estadísticos de forma intuitiva









Estadísticos

- Posición **(Basados en el orden)**
  - Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos.
    - Cuantiles, percentiles, cuartiles, deciles,...
- Centralización
  - Indican valores con respecto a los que los datos parecen agruparse.
    - Media, mediana y moda
- Dispersión
  - Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización.
    - Desviación estándar, coeficiente de variación, rango, varianza
- Forma
  - Asimetría
  - Apuntamiento o curtosis

Centralización

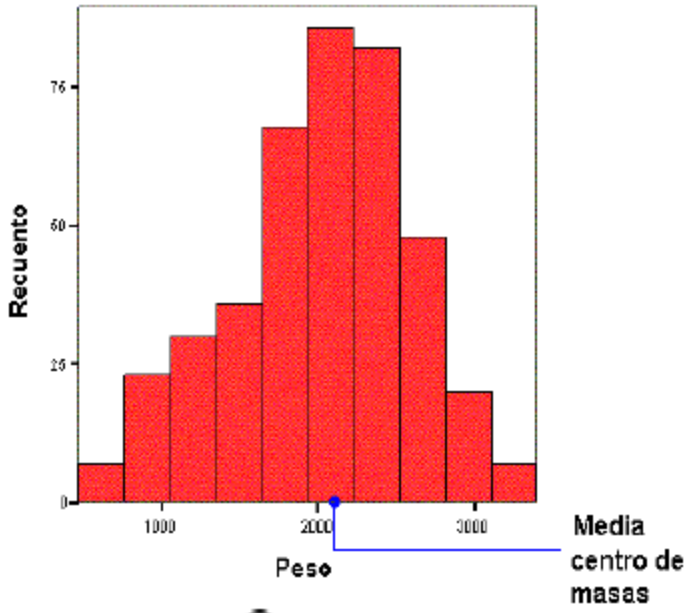
- Añaden unos cuantos casos particulares a las medidas de posición. Son medidas que buscan posiciones (valores) con respecto a los que los datos muestran tendencia a agruparse.

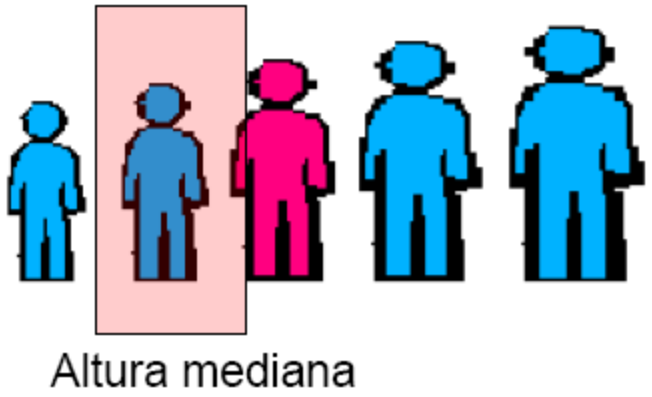
- Media: es la media aritmética (promedio) de los valores de una variable. Suma de los valores dividido por el tamaño muestral.
  - Media de {2, 2, 3, 7} es (2+2+3+7)/4 = 3.5
  - Conveniente cuando los datos se concentran simétricamente con respecto a ese valor. Muy sensible a valores extremos.
  - Centro de gravedad de los datos.

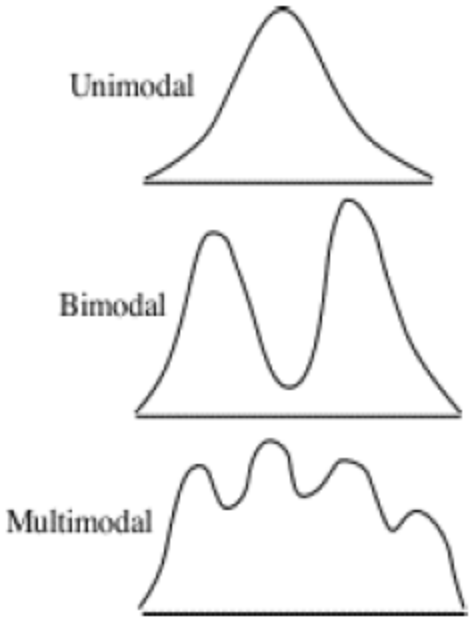
Centralización

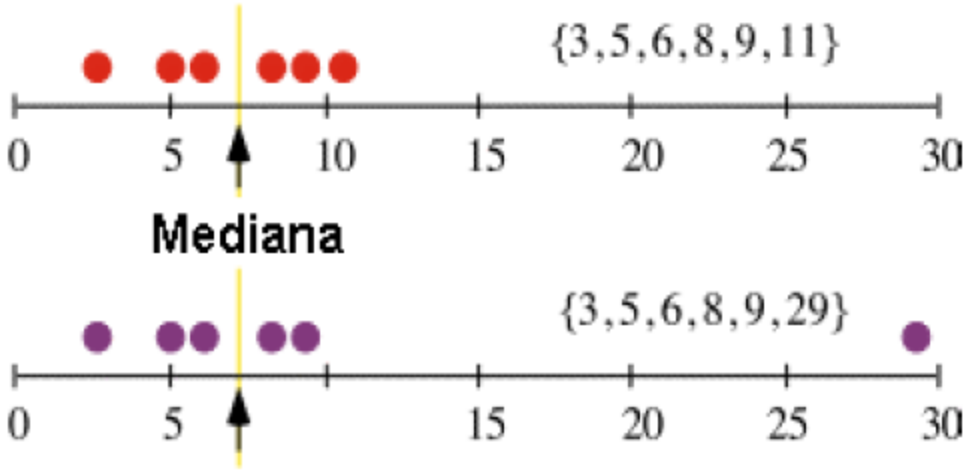
- Mediana: es un valor que divide a las observaciones en dos grupos con el mismo número de individuos (percentil 50). Si el número de datos es par, se elige la media de los dos datos centrales.
  - Mediana de 1, 2, 4, **5**, 6, 6, 8 es 5
  - Mediana de 1, 2, 4, **5**, **6**, 6, 8, 9 es (5+6)/2 = 5.5
  - Es conveniente cuando los datos son asimétricos. No es sensible a valores extremos.
    - Mediana de 1, 2, 4, **5**, 6, 6, 800 es 5. ¡La media es 117.7!

- Moda: es el/los valor/es donde la distribución de frecuencia alcanza un máximo.


Recuento
75
50
25
0
1000
2000
3000
Peso
Media
centro de
masas

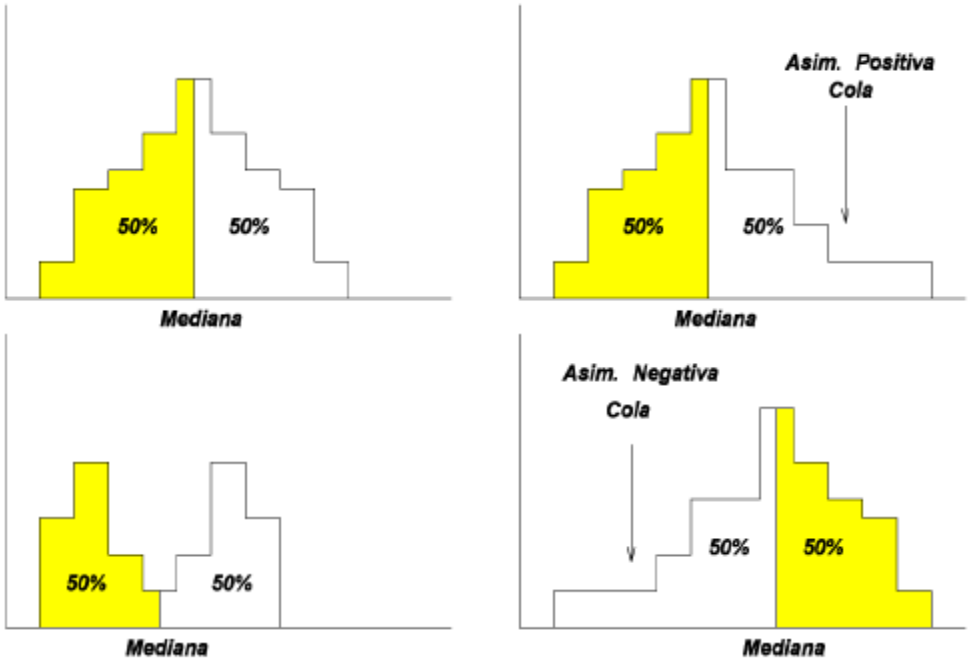

Altura mediana

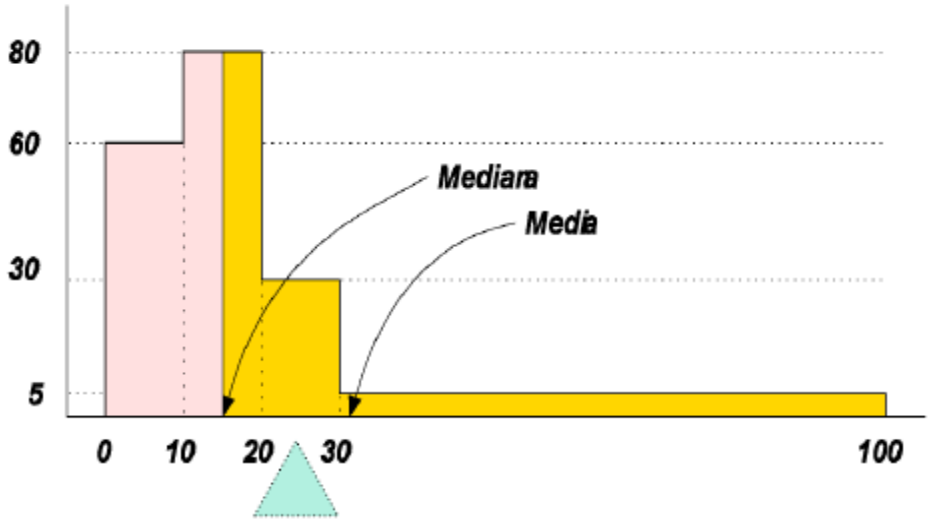

Unimodal
Bimodal
Multimodal


{3,5,6,8,9,11}
0
5
10
15
20
25
30
Mediana
{3,5,6,8,9,29}
0
5
15
20
25
30

Asimetría o sesgo

- Una distribución es simétrica si la mitad izquierda de su distribución es la imagen especular de su mitad derecha.
- En las distribuciones simétricas media y mediana coinciden. Si sólo hay una moda también coincide.
- La asimetría es positiva o negativa en función de a qué lado se encuentra la cola de la distribución.
- La media tiende a desplazarse hacia las valores extremos (colas).
- Las discrepancias entre las medidas de centralización son indicación de asimetría.

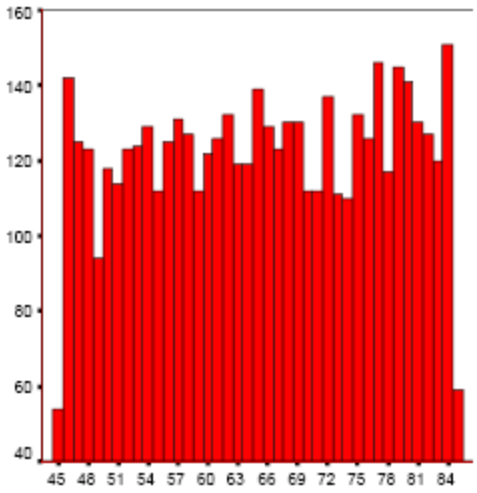




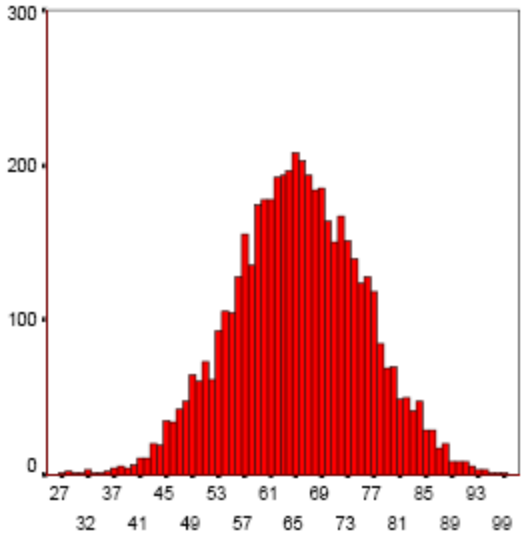
Apuntamiento o curtosis (kurtosis)

- La curtosis nos indica el grado de apuntamiento (aplastamiento) de una distribución con respecto a la distribución normal o gaussiana. Es adimensional.
- Platicúrtica: curtosis < 0
- Mesocúrtica: curtosis = 0
- Leptocúrtica: curtosis > 0

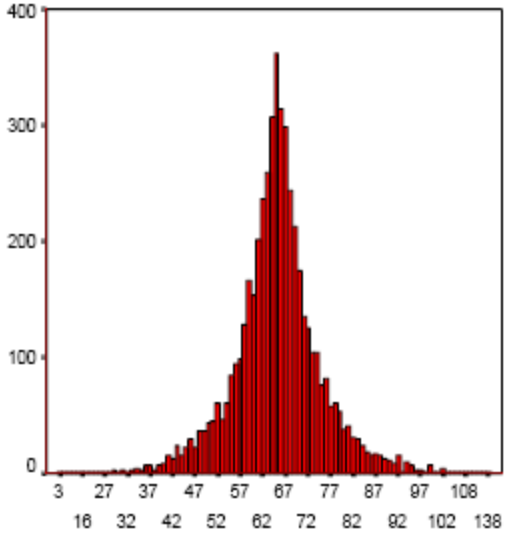
Los gráficos poseen la misma media y desviación típica, pero diferente grado de apuntamiento o curtosis.



Platicúrtica



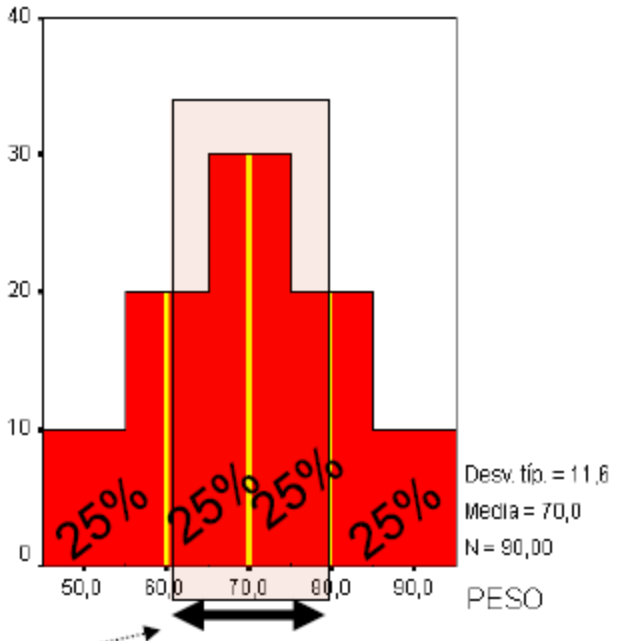
Mesocúrtica



Leptocúrtica

Medidas de dispersión

•Miden el grado de dispersión (variabilidad) de los datos, independientemente de su causa.

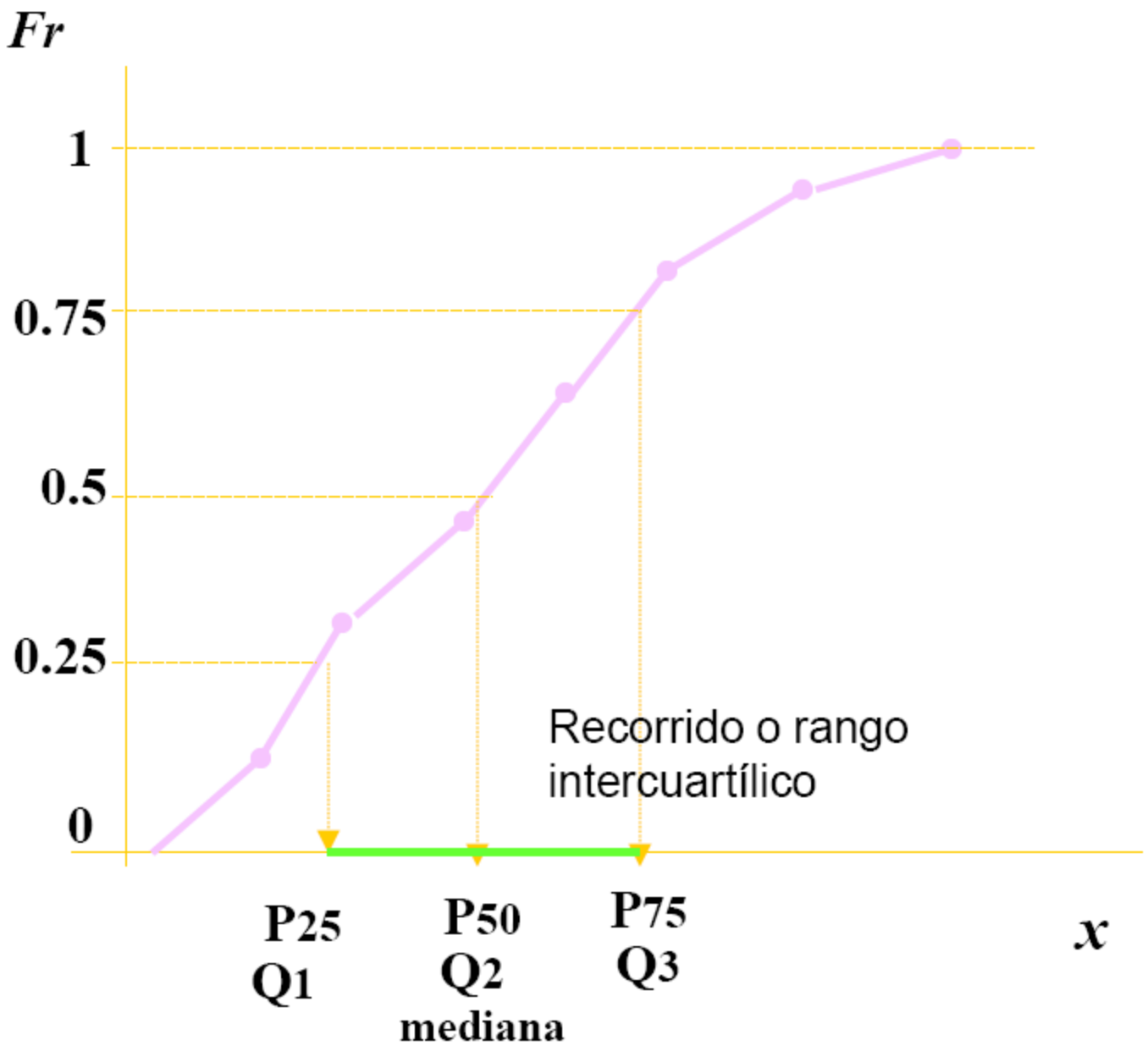


•**Amplitud o Rango** ('range'): La diferencia entre las observaciones extremas.

- 2,1,4,3,8,4. El rango es 8-1=7
- Es muy sensible a los valores extremos.

•**Rango intercuartílico** ('interquartile range'):

- Es la distancia entre el primer y tercer cuartil.
  - Rango intercuartílico = $P_{75} - P_{25}$
- Parecida al rango, pero eliminando las observaciones más extremas inferiores y superiores.
- No es tan sensible a valores extremos.


Fr
1
0.75
0.5
0.25
0
Recorrido o rango intercuartílico
P25
Q1
P50
Q2
mediana
P75
Q3
x

Concepto de Variabilidad

- El concepto de variabilidad está instalado en el centro de la estadística como disciplina, ya que a través de sus herramientas podemos cuantificar, entender, y explicar las diferentes fuentes de variabilidad en el problema que nos hemos propuesto estudiar.
- Nada que no tenga variabilidad podría ser de interés en este contexto, ya que el estudiar un solo objeto o un solo individuo sería suficiente para dar respuesta a todas nuestras preguntas.
- Variabilidad Entre-Sujetos:
  - Ej: Los clientes tienen comportamientos, hábitos de compras, características y gustos distintos. Una forma de visualizar la variabilidad es observando como se distribuyen los clientes en cuanto a comportamiento, hábitos de compras, características y gustos distintos.

Conceptos de Variabilidad cont.

- Variabilidad Intra-Sujetos:
  - Ej: Los mismos clientes pueden cambiar a través del tiempo. Los ciudadanos con derecho a voto en un país, pueden cambiar sus preferencias, especialmente a días de una elección. Esto puede ser más intenso en aquellos individuos más indecisos. También es relevante mencionar que el instrumento mismo y cómo fue diseñado, puede alterar las respuestas de los individuos. También es llamada variabilidad por error de medición según sea el caso.
- Variabilidad Muestral:
  - Se introduce al estudiar una muestra de la población. En muchas investigaciones de mercado al estudiar una población objetivo mediante una encuesta, debemos hacerlo a través de una muestra.

Conceptos de Variabilidad cont.

- Supongamos que hemos tomado una muestra representativa de la población, es decir, que ha sido elegida aleatoriamente de ella. En ese escenario, los resultados del análisis de la encuesta en nuestra muestra arrojarán valores distintos, cuantificablemente distintos, a los valores que arrojarían en otra muestra de la misma población. Incluso usando el mismo mecanismo o esquema de muestreo. A esta fuente de variabilidad la llamamos variabilidad muestral, y es el tema central de la inferencia estadística.

Conceptos de Variabilidad cont.

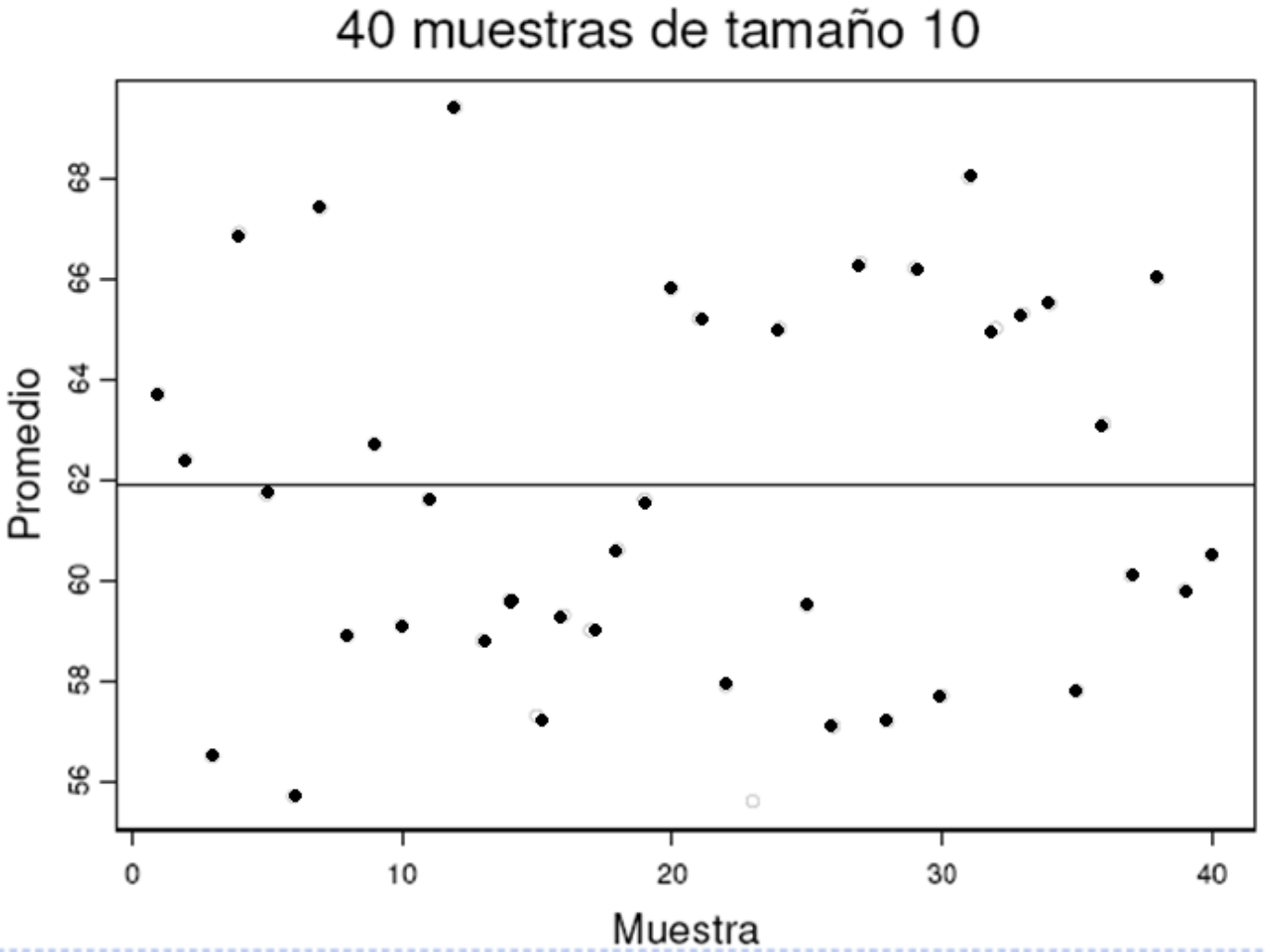
	A	B	C	D	E	F	G
1	64	66	46	71	65	73	61
2	75	58	90	73	85	75	44
3	64	76	73	50	59	54	74
4	84	65	41	73	57	73	69
5	73	59	63	66	48	60	55
6	79	75	93	45	72	60	78
7	63	73	75	49	61	41	70
8	71	42	45	71	62	38	79
9	76	44	72	65	64	49	60
10	51	50	73	78	58	76	53
11	49	63	68	62	71	67	60
12	51	63	59	67	33	62	61
13	65	38	40	80	63	57	67
14	68	76	81	65	50	79	42
15	49	63	72	62	62	53	86
16	84	59	40	57	67	48	54
17	60	67	70	44	52	68	76
18	68	47	59	73	63	61	59
19	63	63	72	95	61	61	86
20	33	52	63	69	51	53	54

Conceptos de Variabilidad cont.

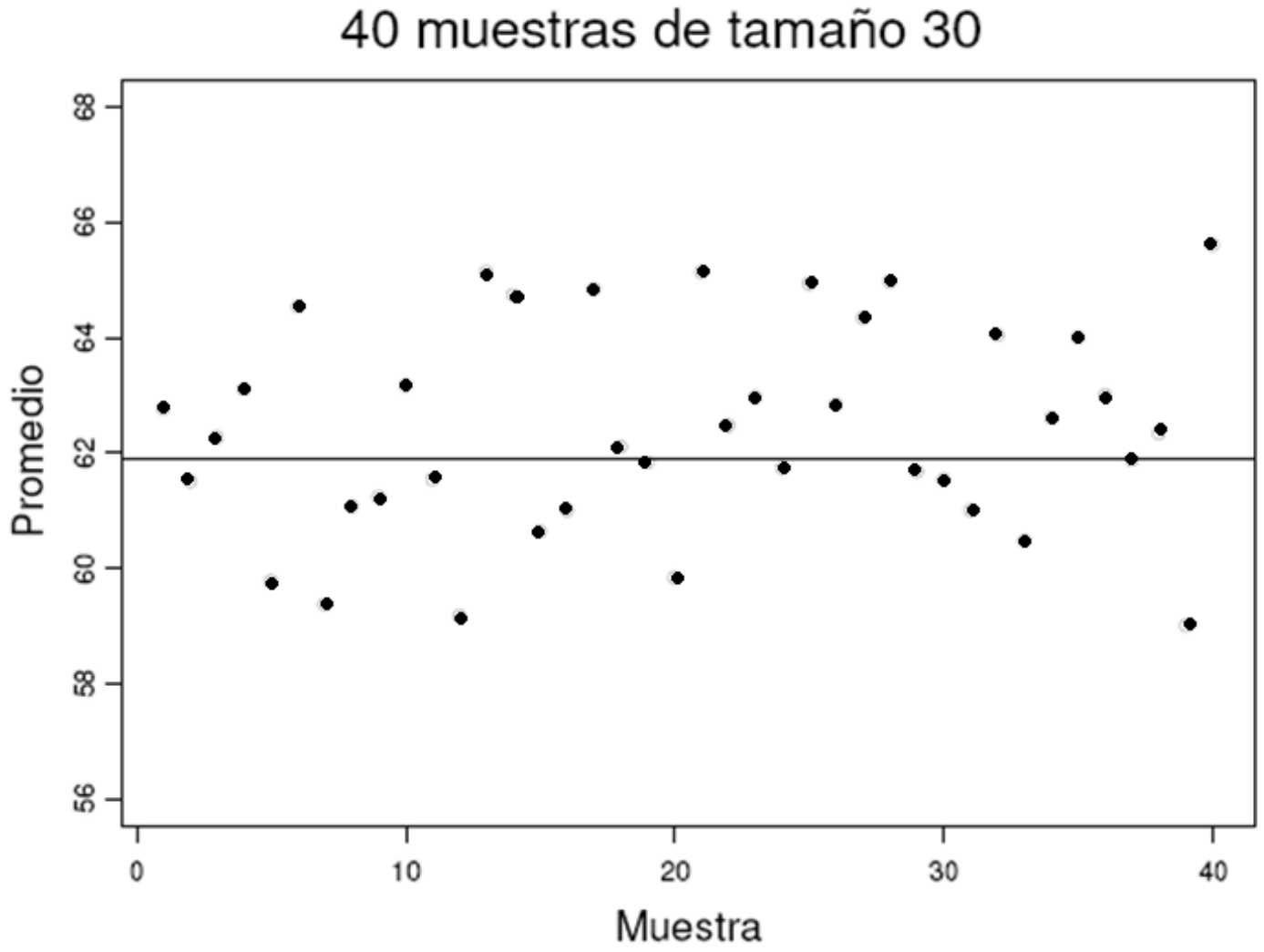
- La media de estos 350 datos es 61.9, lo que corresponde a la media poblacional
- Si calculamos el promedio de la muestra de tamaño 10, obtenemos 63.7
- Al repetir 40 veces el experimento se obtienen los siguientes resultados:

63.7	62.4	56.5	66.9	61.7	55.7	67.4	58.9	62.7	59.1
61.6	70.1	58.8	59.6	57.3	59.3	59.0	60.6	61.6	65.8
65.2	57.9	53.6	65.0	59.5	57.1	66.3	57.2	66.2	57.7
68.0	65.0	65.3	65.5	57.8	63.1	60.1	66.0	59.8	60.5

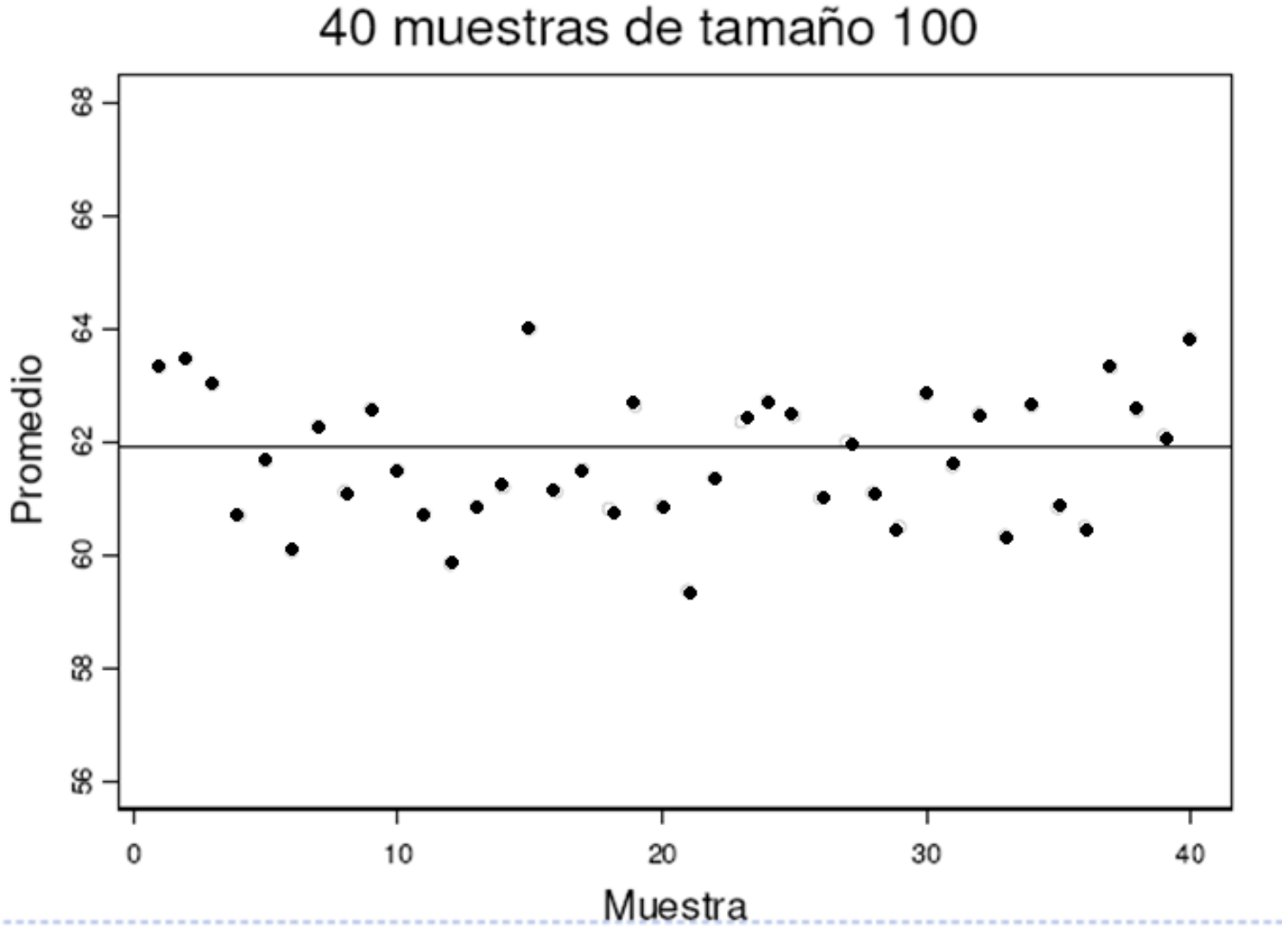
Conceptos de Variabilidad cont.



Conceptos de Variabilidad cont.



Conceptos de Variabilidad cont.



Distribución de Frecuencias

Variable X_i	Frecuencia Absoluta n_i	Frecuencia Absoluta Acumulada N_i	Frecuencia Relativa h_i	Frecuencia Relativa Acumulada H_i
X_1	n_1	$N_1 = n_1$	$h_1 = \frac{n_1}{N}$	$H_1 = h_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_k	n_k	$N_k = \sum_{i=1}^{k} n_i$	$h_k = \frac{n_k}{N}$	$H_k = \sum_{i=1}^{k} h_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_K	n_K	$N_K = N$	$h_K = \frac{n_K}{N}$	$H_K = 1$

Distribución de Frecuencias cont.

Intervalo $[L_i - L_{i+1}[$	Marca de Clase C_i	Frecuencia Absoluta n_i	Frecuencia Absoluta Acumulada N_i	Frecuencia Relativa h_i	Frecuencia Relativa Acumulada H_i
$[L_1 - L_2[$	C_1	n_1	$N_1 = n_1$	$h_1 = \frac{n_1}{N}$	$H_1 = h_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_k - L_{k+1}[$	C_k	n_k	$N_k = \sum_{i=1}^{k} n_i$	$h_k = \frac{n_k}{N}$	$H_k = \sum_{i=1}^{k} h_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_K - L_{K+1}[$	C_K	n_K	$N_K = N$	$h_K = \frac{n_K}{N}$	$H_K = 1$

Medidas de Resumen de Centralización

- Media:
  - Conocida como promedio, se calcula como:

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

  - En datos agrupados en una tabla de distribución de frecuencias la media se puede estimar como:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} c_i n_i = \sum_{i=1}^{n} c_i h_i$$

Medidas de Resumen de Centralización cont.

- **Mediana**:
  - En un conjunto de datos ordenados de menor a mayor, la mediana corresponde al dato central. Aquel que deja un 50% de la información bajo él y el otro 50% es mayor o igual.

Si $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ es la muestra ordenada:

$$Me = \begin{cases} \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2}, & \text{si n es par} \\ x_{(\frac{n+1}{2})}, & \text{si n es impar} \end{cases}$$

Medidas de Resumen de Centralización cont.

- **Mediana**:
  - **En datos agrupados la mediana puede ser estimada como**:

$$Me = L_{me} + \frac{A}{h_{me}}(0.5 - H_{me-1})$$

Donde:

L_{me} = Límite inferior de la clase de la mediana

A = Amplitud de la Clase

h_{me} = Frecuencia Relativa de la clase de la mediana

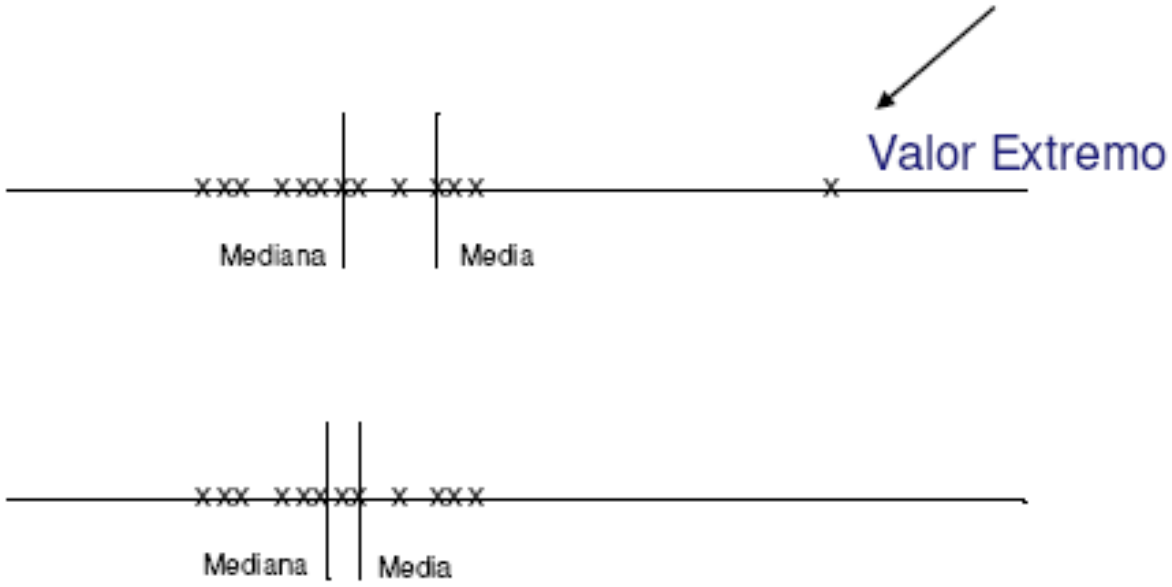
H_{me} = Frecuencia Relativa acumulada de la clase que precede a la mediana.

Medidas de Resumen de Centralización cont.

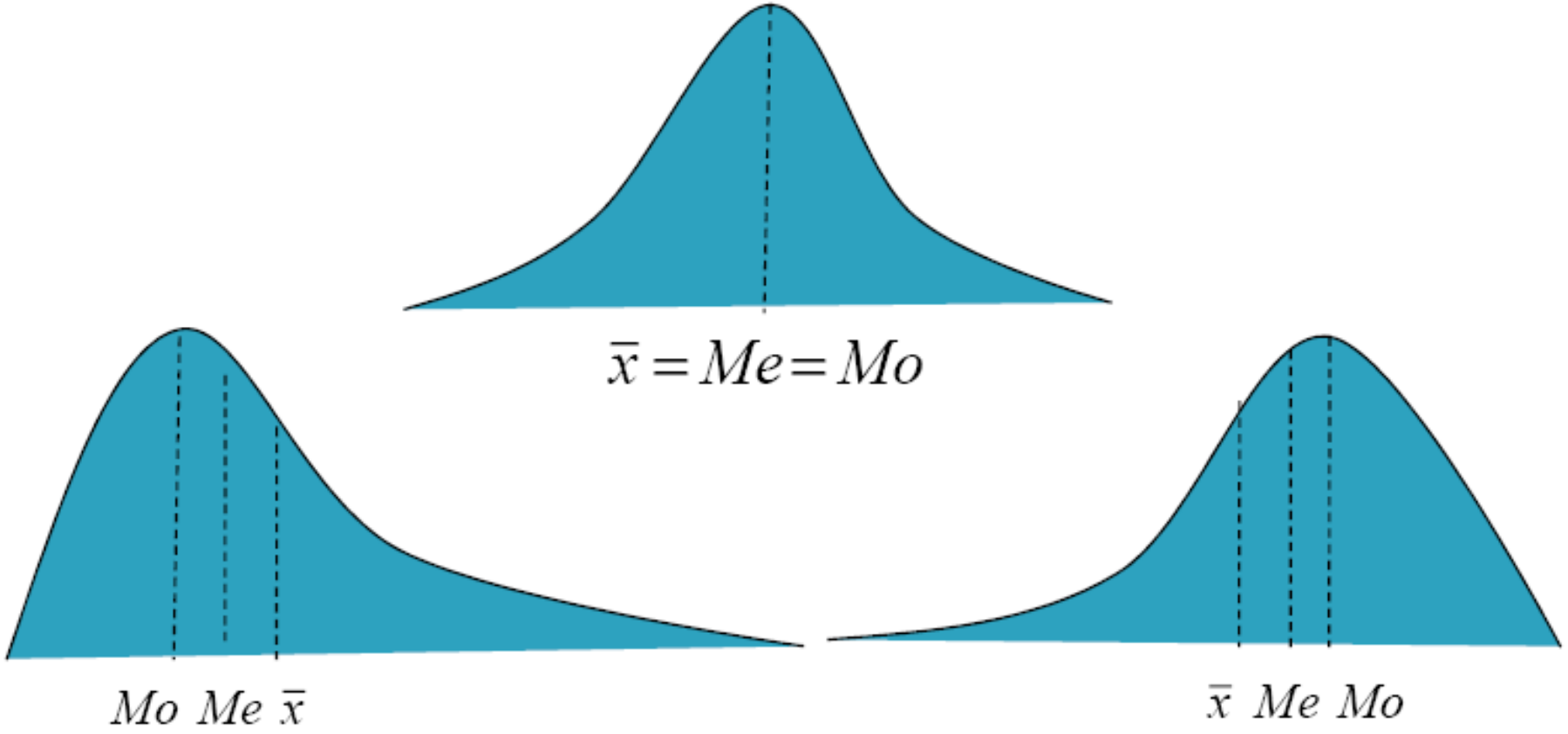
- **Moda**:
  - Es aquel valor observado que tienen mayor frecuencia.
  - En datos agrupados se puede considerar como moda a la marca de clase de la categoría con mayor frecuencia.
  - Cuando dos valores ocurren con la misma frecuencia y ésta es la más alta, ambos valores son modas, por lo que el conjunto de datos es Bimodal.
  - Cuando ningún valor se repite se dice que no hay moda.

Medidas de Resumen de Centralización cont.

- La media es sensible a la presencia de datos extremos.
- La mediana es muy útil cuando la distribución de la variable es poco simétrica.



Medidas de Resumen de Centralización cont.



Medidas de Resumen de Dispersión

- **Varianza:**

  Cuantifica la dispersión de los datos con respecto a la media. Se obtiene como la media de las desviaciones cuadráticas de cada dato con respecto a la media.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

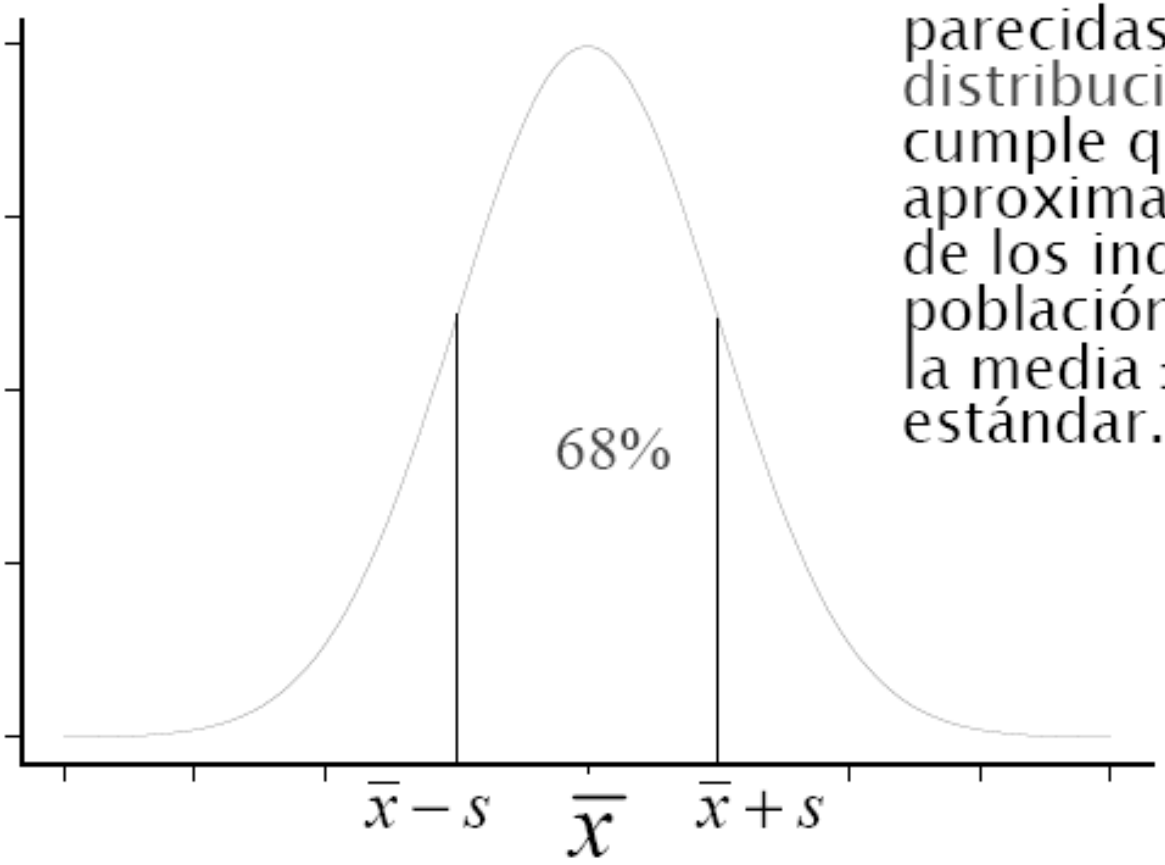
Medidas de Resumen de Dispersión cont.

- **Desviación Estándar**

  Es la raíz cuadrada de la varianza. Es la más usada de las medidas de dispersión.

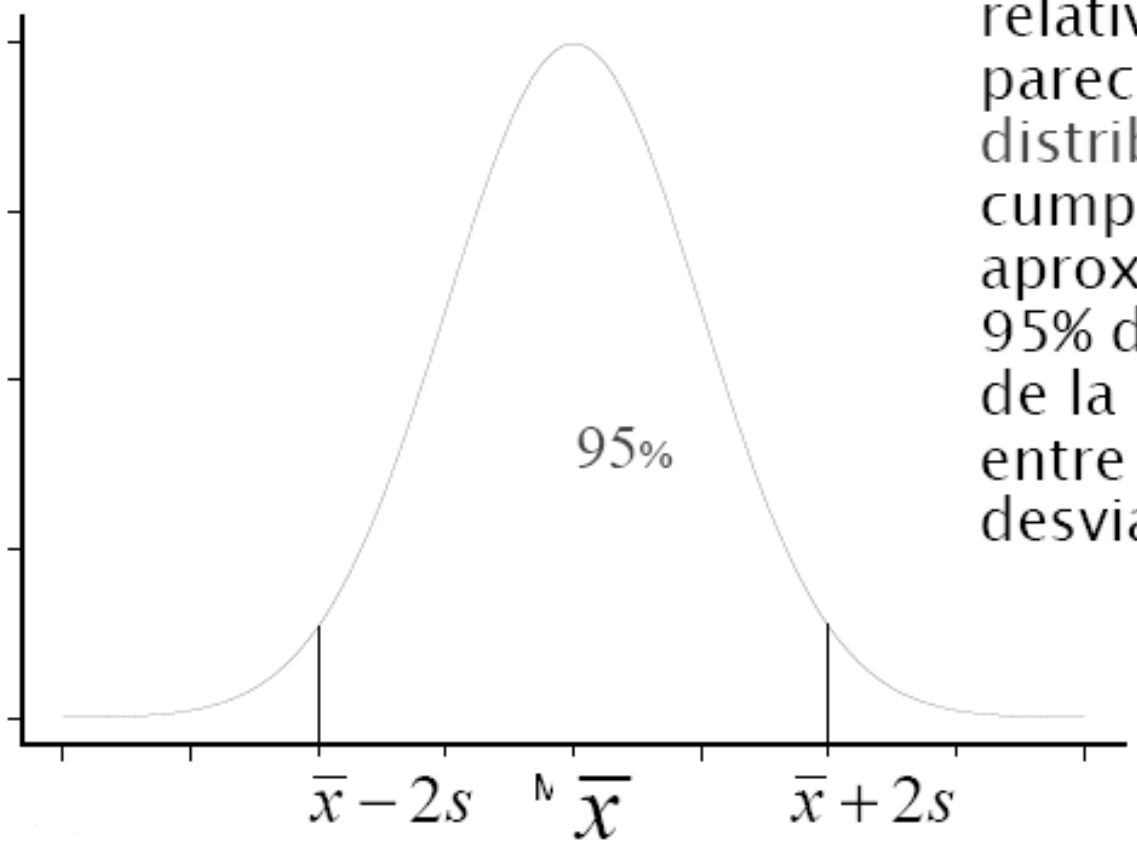
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Medidas de Resumen de Dispersión cont.



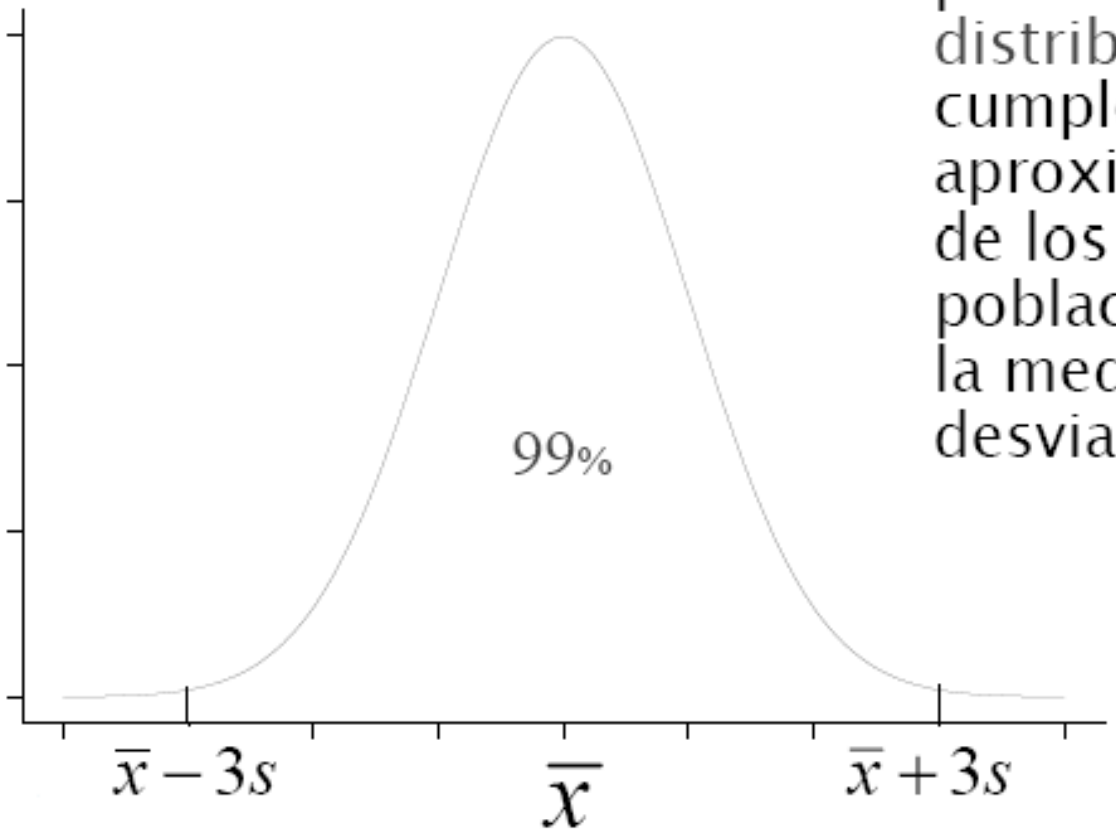
- En distribuciones relativamente simétricas parecidas a la distribución normal, se cumple que aproximadamente el 68% de los individuos de la población se sitúa entre la media ± una desviación estándar.

Medidas de Resumen de Dispersión cont.



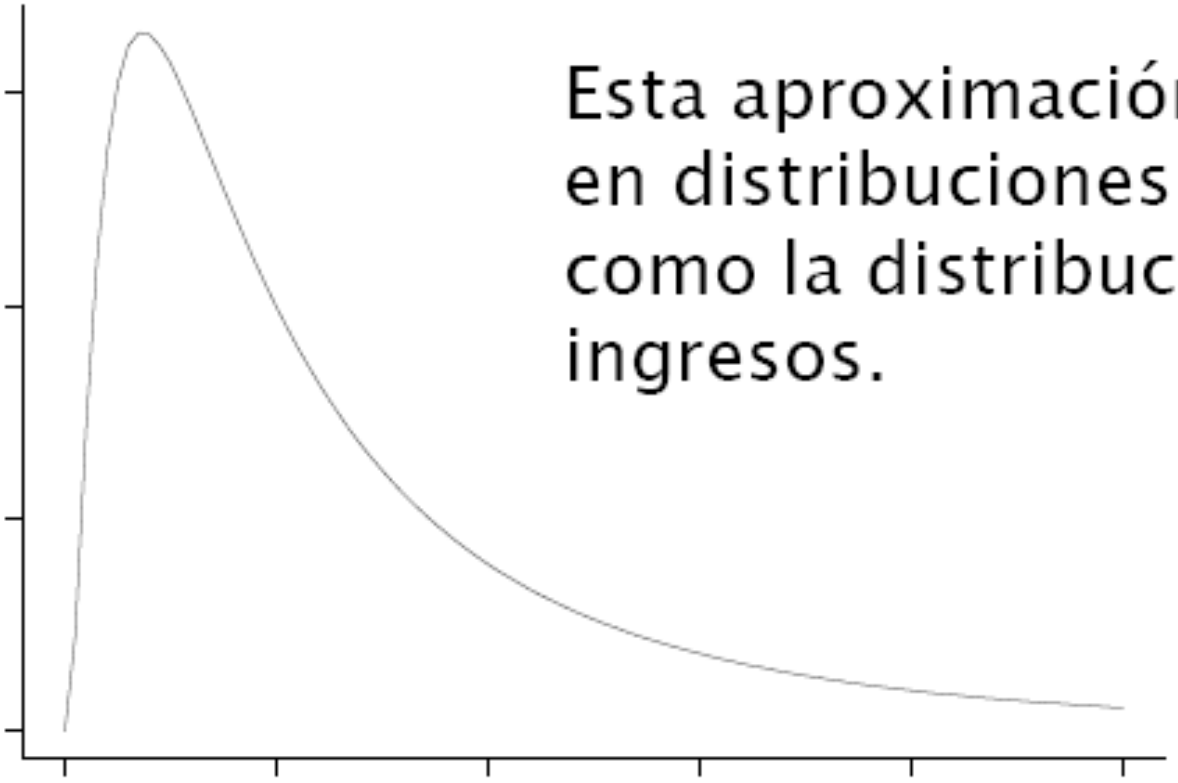
- En distribuciones relativamente simétricas parecidas a la distribución normal, se cumple que aproximadamente el 95% de los individuos de la población se sitúa entre la media ± 2 desviación estándar.

Medidas de Resumen de Dispersión cont.



- En distribuciones relativamente simétricas parecidas a la distribución normal, se cumple que aproximadamente el 99% de los individuos de la población se sitúa entre la media ± 3 desviaciones estándar.

Medidas de Resumen de Dispersión cont.



Esta aproximación no funciona en distribuciones asimétrica como la distribución de los ingresos.

Medidas de Resumen de Dispersión cont.

- **Coeficiente de Variación:**

  Describe la desviación estándar relativa a la media, sirve para comparar la variación en diferentes poblaciones. Se calcula de la siguiente forma:

$$CV = \frac{s}{\bar{x}}$$

- # Coeficiente de variación
- Es la razón entre la desviación típica y la media.

$$CV = \frac{S}{\bar{x}}$$

  - Mide la desviación típica en forma de "qué tamaño tiene con respecto a la media"
  - También se la denomina variabilidad relativa.
  - Es frecuente mostrarla en porcentajes
    - Si la media es 80 y la desviación típica 20 entonces CV=20/80=0,25=25% (variabilidad relativa)

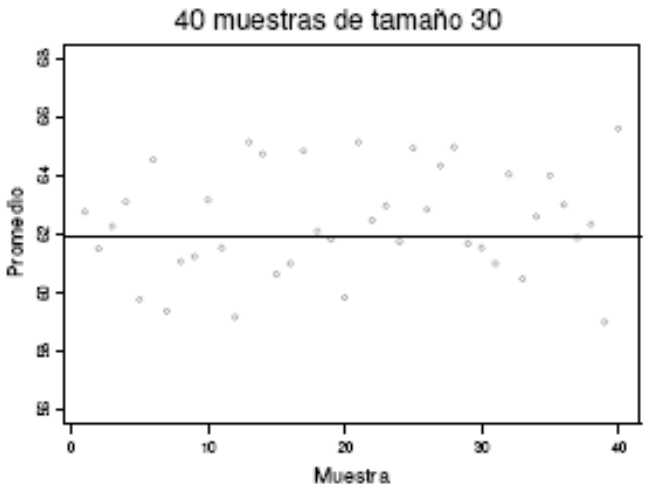
- Es una cantidad **adimensional**. Interesante para comparar la variabilidad de diferentes variables.
  - Si el peso tiene CV=30% y la altura tiene CV=10%, los individuos presentan más dispersión en peso que en altura.

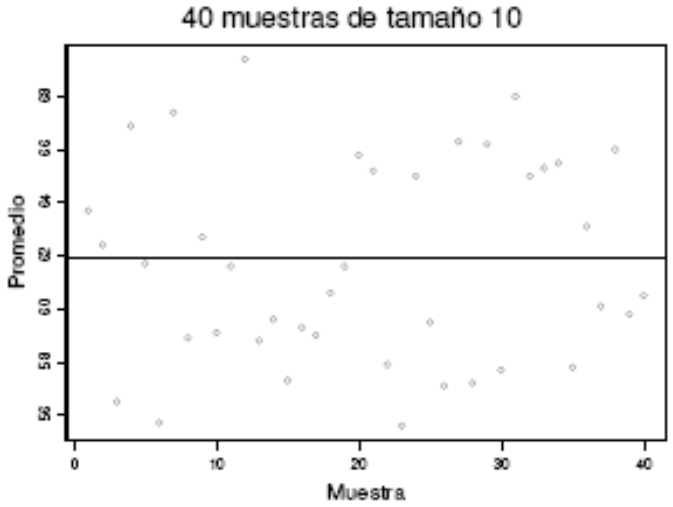
- No debe usarse cuando la variable presenta valores negativos o donde el valor 0 sea una cantidad fijada arbitrariamente
  - Por ejemplo 0ºC ≠ 0ºF

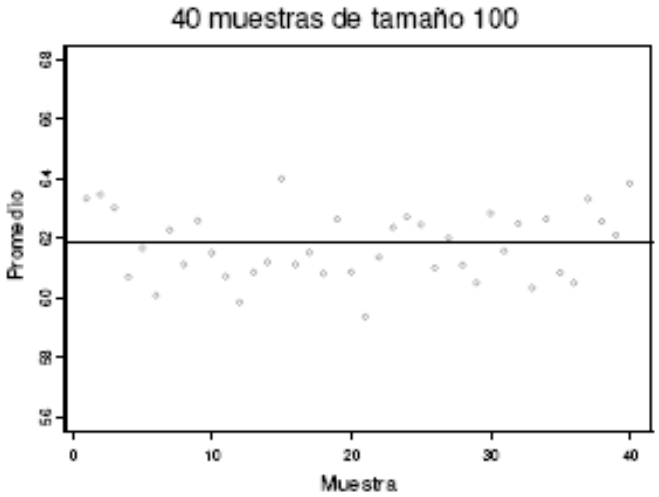
Medidas de Resumen de Dispersión cont.

- El error estándar mide la variabilidad esperada del promedio muestral como estimación de la media poblacional.

$$SEM = \frac{s}{\sqrt{n}}$$ ← *Depende de n*

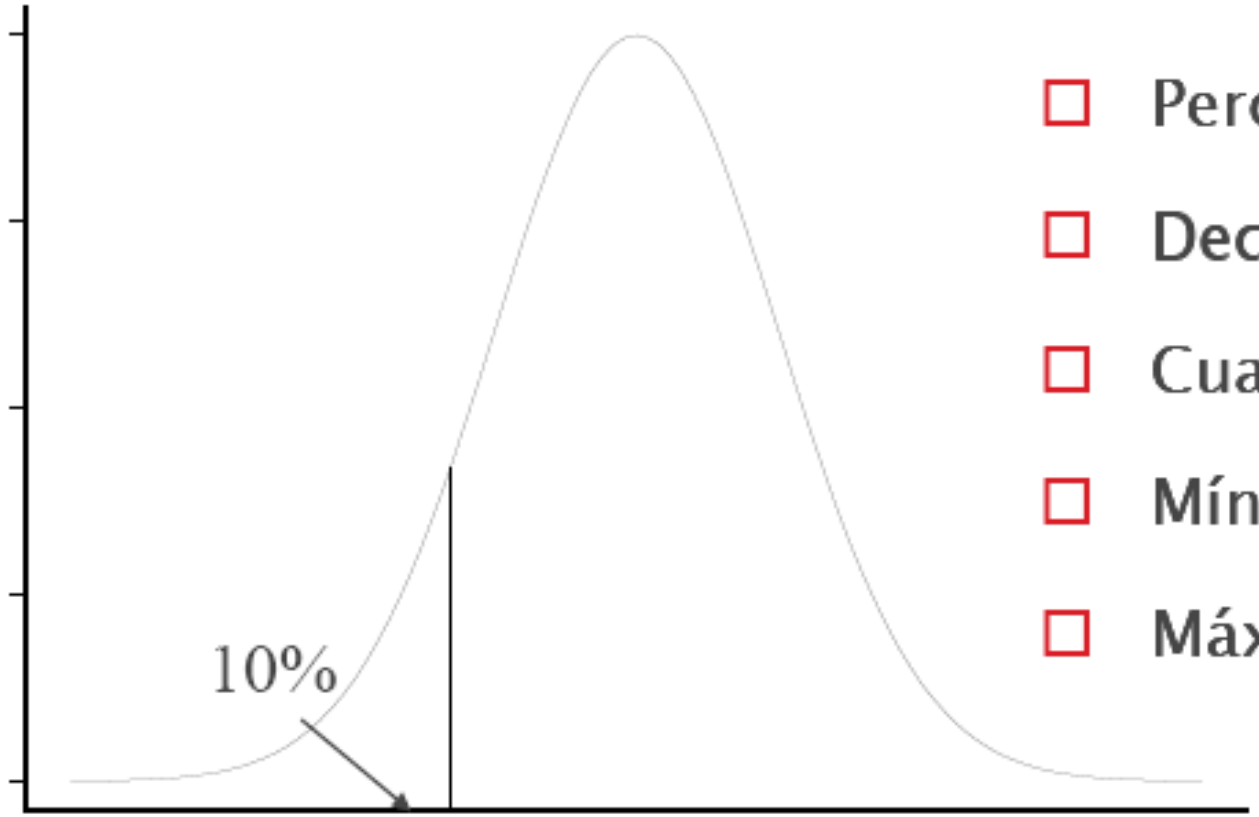






Medidas basadas en el Orden (Posición)

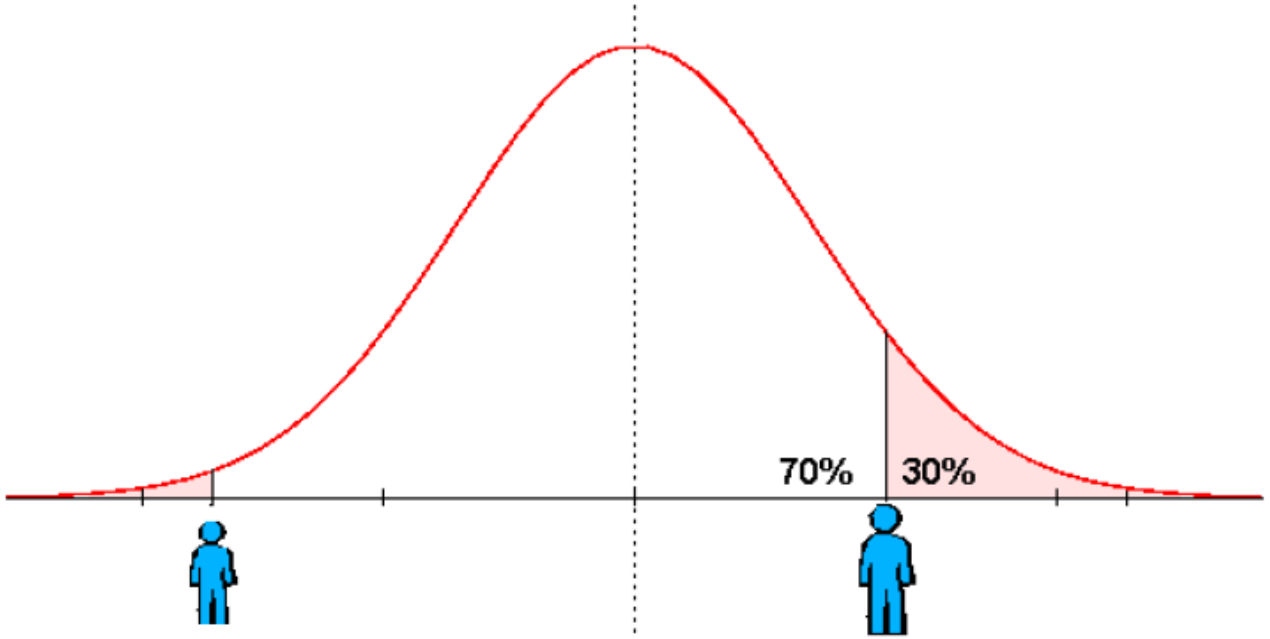
- Mediana
- Percentiles
- Deciles
- Cuartiles
- Mínimo
- Máximo



Percentil 10%

Estadísticos de Posición

- Se define el **cuantil** de orden α como un valor de la variable por debajo del cual se encuentra una frecuencia acumulada α.
- Casos particulares son los percentiles, cuartiles, deciles, quintiles,...



Estadísticos de Posición cont.

- **Percentil** de orden k = cuantil de orden k/100
  - La **mediana** es el percentil 50.
  - El percentil de orden 15 deja por debajo al 15% de las observaciones. Por encima queda el 85%.

- **Cuartiles**: Dividen a la muestra en 4 grupos con frecuencias similares.
  - Primer cuartil = Percentil 25 = Cuantil 0.25.
  - Segundo cuartil = Percentil 50 = Cuantil 0.5 = **mediana.**
  - Tercer cuartil = Percentil 75 = cuantil 0.75.

Estadísticos de Posición cont.

Ejemplos: El 5% de los recién nacidos tiene un peso demasiado bajo. ¿Qué peso se considera "demasiado bajo"?

- Percentil 5 o cuantil 0.05.

¿Qué peso es superado sólo por el 25% de los individuos?

- Percentil 75.

El colesterol se distribuye simétricamente en la población. Se considera patológico los valores extremos. El 90% de los individuos son normales. ¿Entre qué valores se encuentran los individuos normales?

- Entre el percentil 5 y el 95.

¿Entre qué valores se encuentran la mitad de los individuos "más normales" de una población?

- Entre el cuartil 1º y 3º.

Estadísticos de Posición cont.

Son valores de la variable que dividen a la muestra en partes de igual porcentaje.

Los **percentiles** separan la muestra en grupos de 1% cada uno (son 99).

- **Cuartiles**: agrupan 25% c/u (son 3).
- **Quintiles**: agrupan 20% c/u (son 4).
- **Deciles**: agrupan 10% c/u (son 9).

Estadísticos de Posición cont.

Se calculan de la siguiente forma:

Ordenar de menor a mayor los n datos.

Obtener D = n * k /100

a) Si **D es entero**, entonces el percentil k corresponde al valor medio de las observaciones ubicadas en las posiciones **D** y **D+1**.

b) Si **D no es un entero**, el percentil k corresponde a la observación ubicada en la posición entera siguiente, es decir, **[D+1]**

Estadísticos de Posición cont.

Ejemplo

Determinar los percentiles 25 y 60 de los siguientes datos: 3, 5, 5, 8, 12, 15, 21, 23, 25, 26, 29, 35

P25 **D= 12 x 25 /100 = 3**

resulta un entero, por tanto el P25 corresponde al promedio de las observaciones en las posiciones 3º y 4º, es decir, P25= (5+8)/2 = 6.5

P60 **D = 12 x 60 / 100 = 7.2**

Dado que no es un entero, nos "movemos" al entero siguiente.

Es decir, P60 = 23 (observación en la 8ª posición)

Estadísticos de Posición cont.

- **En datos agrupados el K-esimo percentil puede ser estimado como:**

$$P_k = L_{p_k} + \frac{A}{h_{p_k}}\left(\frac{k}{100} - H_{p_k - 1}\right)$$

Donde:

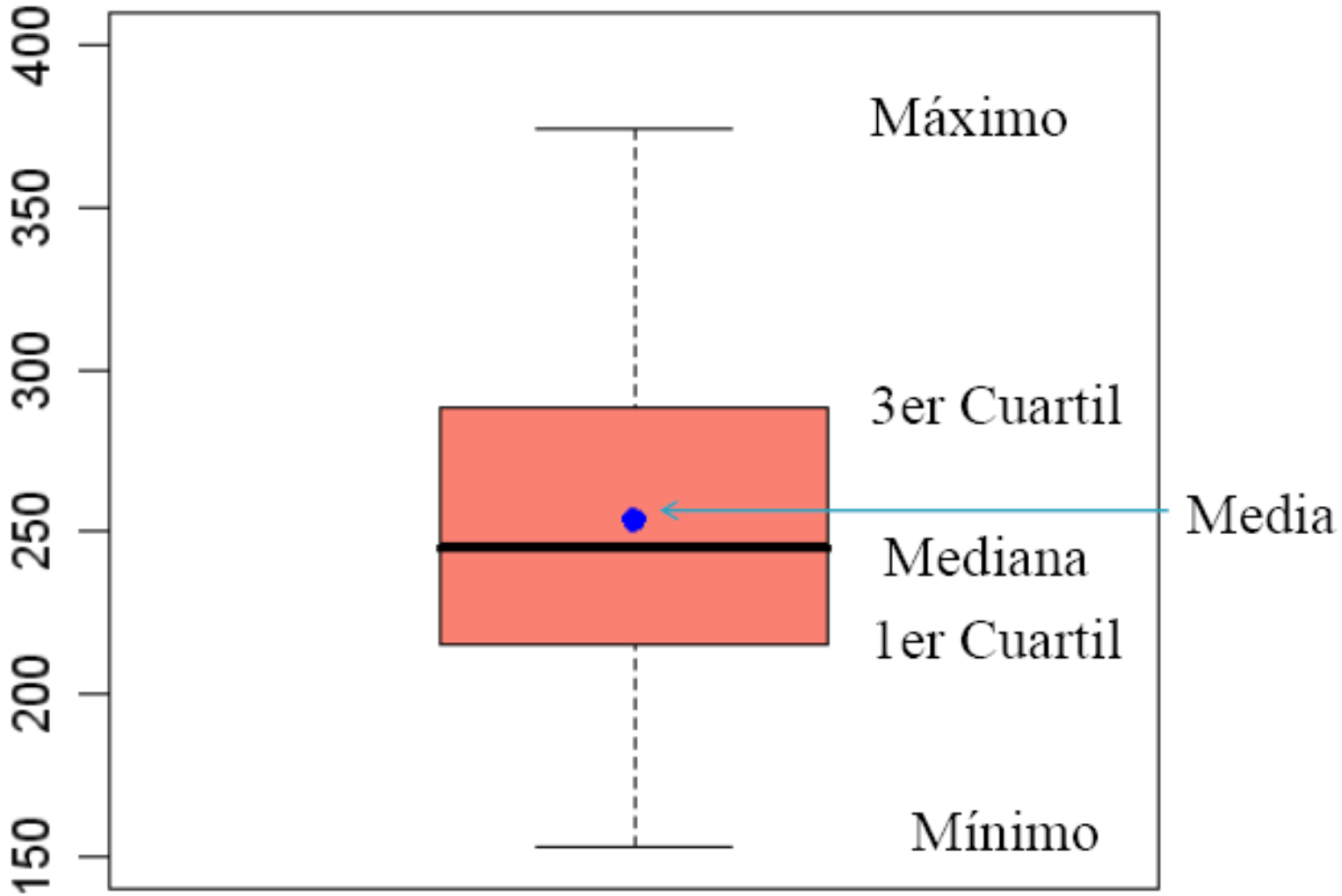
L_{p_k} = Límite inferior de la clase del percentil k-ésimo.

A = Amplitud de la Clase

h_{p_k} = Frecuencia Relativa de la clase del percentil k-ésimo

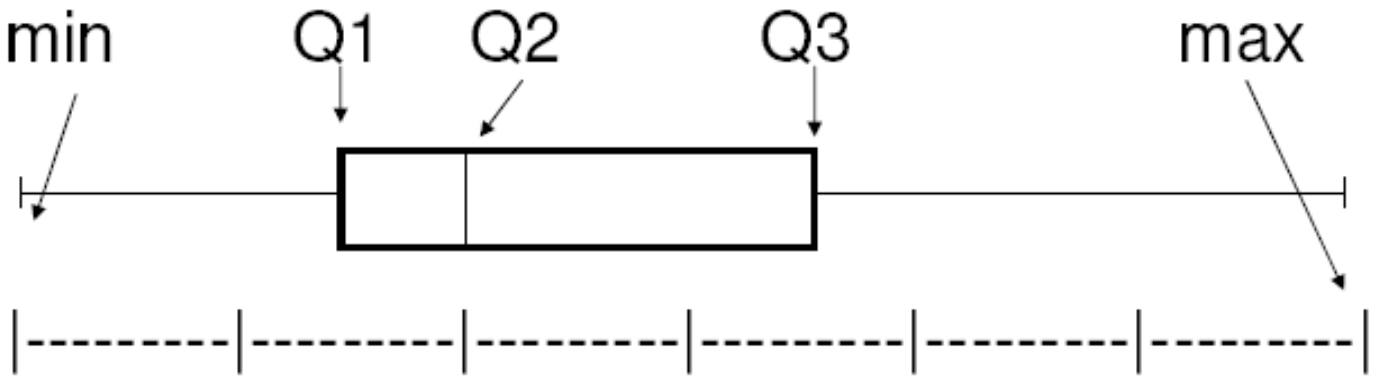
H_{p_k} = Frecuencia Relativa acumulada de la clase que precede al percentil k-ésimo.

Box-plot (Caja con bigotes)



Box-plot cont.

Un gráfico asociado a los cuartiles es el **box-plot**: en un eje se ubican los siguientes 5 números extraídos de una muestra: mínimo, cuartil 1, cuartil 2, cuartil 3 y máximo.



Una regla para determinar si un dato es **anómalo** (outlier) es:

- Si un dato es < Q1 – 1.5(Q3-Q1)
- Si un dato es > Q3 + 1.5(Q3-Q1)

Niveles de Hb en 61 adultos normales

105	110	112	112	118	119	120	120	120
125	126	127	128	130	132	133	134	135
138	138	138	138	141	142	144	145	146
148	148	148	149	150	150	150	151	151
153	153	154	154	154	154	155	156	156
158	160	160	160	163	164	164	165	166
168	168	170	172	172	176	179		

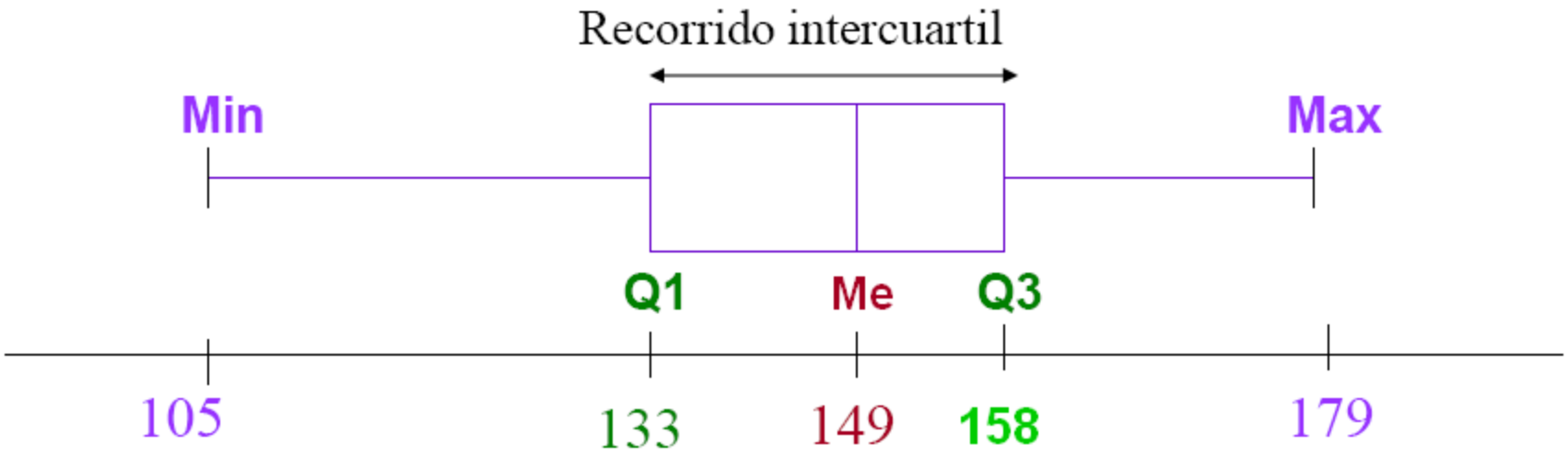
133

149

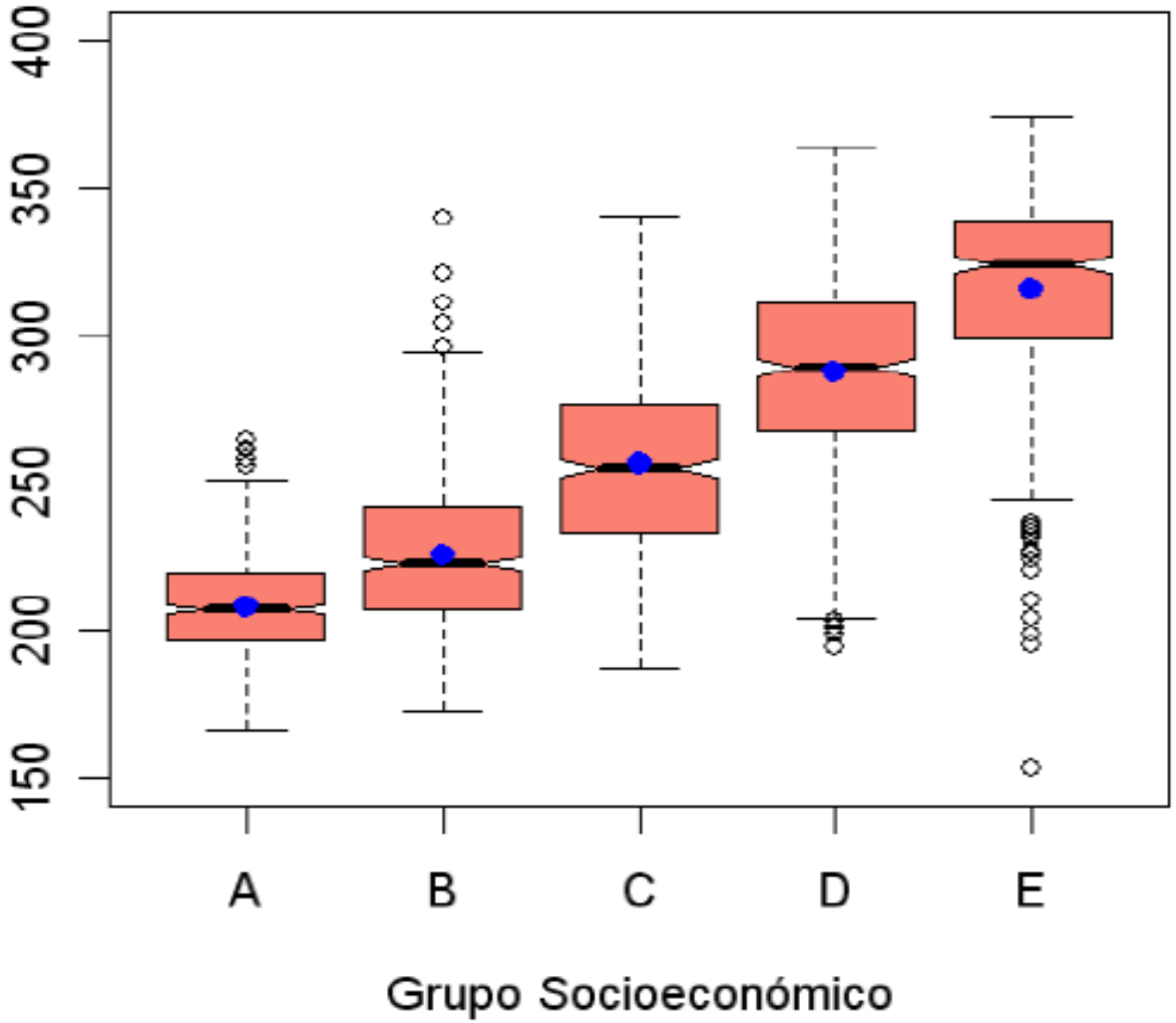
158

Un resumen de esta serie en 5 valores

Min = 105 ; Max =179; Q1 = 133 ; Q3 = 158 ; Q2 = Me = 149



Box-plot comparación de grupos



Estadísticos de Forma: Asimetría y Curtosis

Momentos de una distribución

- Los momentos de una distribución son medidas obtenidas a partir de todos sus datos y de sus frecuencias absolutas. Estas medidas caracterizan de tal forma a las distribuciones que si los momentos de dos distribuciones son iguales, diremos que las distribuciones son iguales. Podemos decir que dos distribuciones son más semejantes cuanto mayor sea el número de sus momentos que coinciden.
- Se define el **momento de orden h respecto al origen de una variable** estadística como:

$$a_h = x_1^h \frac{n_1}{N} + x_2^h \frac{n_2}{N} + \ldots + x_r^h \frac{n_r}{N} = \sum_{i=1}^{r} x_i^h \frac{n_i}{N}$$

- Es inmediato observar que, para h=1, a_1 es la **media** de la distribución.

Estadísticos de Forma cont.

- Se define el **momento central de orden h o momento respecto a la media** aritmética de orden h como:

$$m_h = (x_1 - \bar{x})^h \frac{n_1}{N} + (x_2 - \bar{x})^h \frac{n_2}{N} + \ldots + (x_r - \bar{x})^h \frac{n_r}{N} = \sum_{i=1}^{r} (x_i - \bar{x})^h \frac{n_i}{N}$$

- Es inmediato observar que $m_1 = 0$ y que $m_2 = S^2$
- **Relaciones entre los momentos**:

1. $$m_2 = a_2 - \bar{x}^2$$

2. Los momentos respecto a la media se ven afectados por los cambios de escala, pero no por los cambios de origen. El resto, por ambos.

Estadísticos de Forma cont.

Forma de una distribución

Cuando dos distribuciones coinciden en sus medidas de posición y dispersión, no tenemos datos analíticos para ver si son distintas. Una forma de compararlas es mediante su forma. Bastará con comparar la forma de sus histogramas o diagramas de barras para ver si se distribuyen o no de igual manera.

Para efectuar este estudio de la forma en una sola variable, hemos de tener como referencia una distribución modelo. Como convenio, se toma para la comparación la distribución **normal de media 0 y varianza 1**. En particular, es conveniente estudiar si la variable en cuestión está más o menos apuntada que la Normal. Y si es más o menos simétrica que ésta, para lo que se definen los conceptos de **Asimetría y Curtosis**, y sus correspondientes formas de medida.

La asimetría y su medida

- El objetivo de la medida de la **asimetría** es, sin necesidad de dibujar la distribución de frecuencias, estudiar la deformación horizontal de los valores de la variable respecto al valor central de la media. Las medidas de forma pretenden estudiar la concentración de la variable hacia uno de sus extremos.
- Una distribución es **simétrica** cuando a la derecha y a la izquierda de la media existe el mismo número de valores, equidistantes dos a dos de la media, y además con la misma frecuencia.

La asimetría y su medida cont.

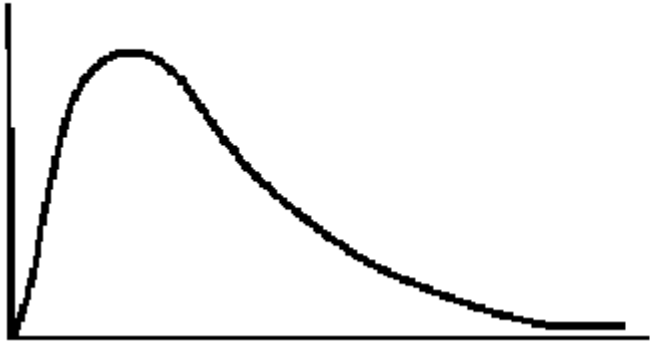
Una distribución es **Simétrica** si $\bar{x} = Me = Mo$

En caso contrario, decimos que la distribución es **Asimétrica**, y entonces puede ser de dos tipos:

Asimétrica a la izquierda. Es el caso en que $Mo \geq Me \geq \bar{x}$

Curva Asimétrica a la izquierda

Asimétrica a la derecha. Es el caso en que $Mo \leq Me \leq \bar{x}$



Curva Asimétrica a la derecha

La asimetría y su medida cont.

Coeficiente de asimetría de Fisher

- En una distribución **simétrica** los valores se sitúan en torno a la media aritmética de forma simétrica. El coeficiente de asimetría de Fisher se basa en la relación entre las distancias a la media y la desviación típica.

En una distribución simétrica $\bar{x} = Me = Mo$ y $m_3 = 0$. Por eso define como:

$$g_1 = \frac{\frac{\sum_{i=1}^{r}(x_i - \bar{x})^3 n_i}{N}}{S^3} = \frac{m_3}{S^3}$$

- Si $g_1 > 0$, la distribución es asimétrica positiva o a la derecha.
- Si $g_1 = 0$, la distribución es simétrica.
- Si $g_1 < 0$, la distribución es asimétrica negativa o a la izquierda.

La asimetría y su medida cont.

Coeficiente de asimetría de Pearson

- Se basa en el hecho de que en una distribución simétrica, la media coincide con la moda. A partir de este dato se define el **coeficiente de asimetría de Pearson** como:

$$A_P = \frac{\bar{x} - Mo}{S}$$

- Si $A_P > 0$, la distribución es asimétrica positiva o a la derecha.
- Si $A_P = 0$, la distribución es simétrica.
- Si $A_P < 0$, la distribución es asimétrica negativa o a la izquierda.

Este coeficiente no es muy bueno para medir asimetrías leves.

La curtosis y su medida

- El concepto de **curtosis o apuntamiento** de una distribución surge al comparar la forma de dicha distribución con la forma de la distribución Normal. De esta forma, clasificaremos las distribuciones según sean más o menos apuntadas que la distribución Normal.
- **Coeficiente de Curtosis de Fischer**

  El **coeficiente de curtosis o apuntamiento de Fischer** pretende comparar la curva de una distribución con la curva de la variable Normal, en función de la cantidad de valores extremos e la distribución. Basándose en el dato de que en una distribución normal se verifica que:

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

La curtosis y su medida cont.

Se define el coeficiente de curtosis de Fisher como:

$$K = g_2 = \frac{\sum_{i=1}^{r} (x_i - \bar{x})^4 n_i}{\frac{N}{S^4}} - 3 = \frac{m_4}{S^4} - 3$$

• Si $g_2 = 0$, la distribución es **Mesocúrtica:** Al igual que en la asimetría es bastante difícil encontrar un coeficiente de curtosis de cero, por lo que se suelen aceptar los valores cercanos (0.5 aprox.).

• Si $g_2 > 0$, la distribución es **Leptocúrtica**

• Si $g_2 < 0$, la distribución es **Platicúrtica**