Estadísitica

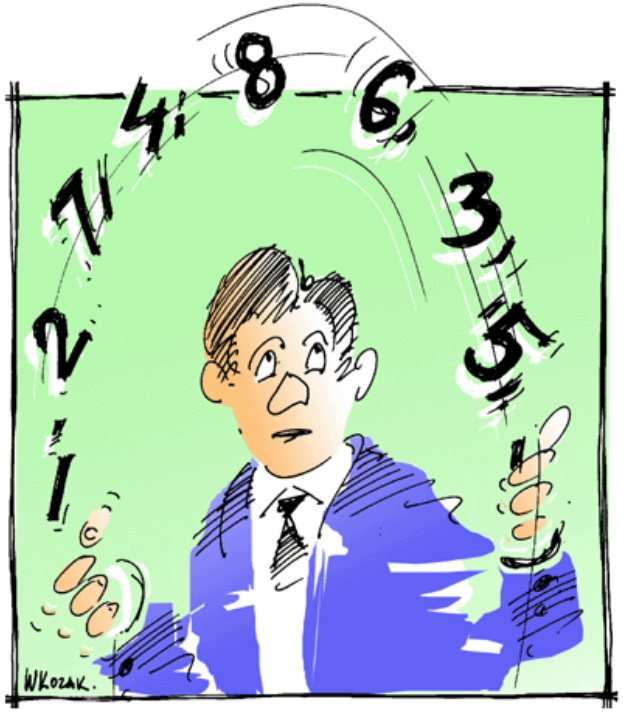


Imagen popular de la estadística:

"Existen medias mentiras, mentiras y estadísticas".

Dos significados:

(1) Colección de datos numéricos (una estadística).

(2) Ciencia: obtener regularidades de fenómenos de masas (la estadística).

"Más del 75% de los americanos blancos son propietarios de su casa y menos del 50% de los hispanos y afroamericanos no son propietarios de su casa. Aquí hay un abismo, el abismo de la propiedad de la casa".
George W. Bush, Cleveland, 1 de julio de 2002
www.bushisms.com

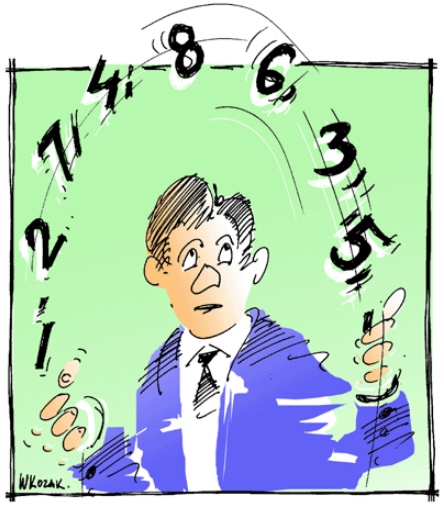
La estadística surgió como una necesidad del estado: el censo y su descripción política, geográfica y económica.

En el siglo XVII y XVIII nace la probabilidad aplicada a los juegos de azar que ejerce una fuerte influencia sobre la estadística.

En el XIX empieza a aplicarse a cuestiones sociales.

Y actualmente se aplica a la historia, psicología, pedagogía, ingeniería, biología, economía, periodismo, política, medicina...

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**.

Pasos en un estudio estadístico

- **Plantear hipótesis sobre una *población:***
  - **Los fumadores tienen *"más bajas"* 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 bajas.
    - Tiempo de duración de cada baja.
    - ¿Sexo? ¿Sector laboral? ¿Otros factores?

Pasos en un estudio estadístico (y 2)

- **Recoger los datos (*muestreo*):**
  - ¿Estratificado? ¿Sistemáticamente?

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

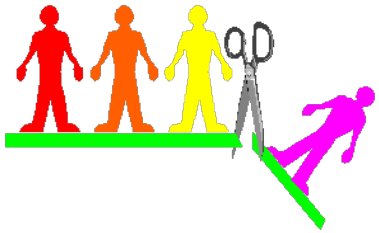
- **Realizar una inferencia sobre la población:**
  - Los fumadores están de baja al menos 10 días/año más *de media* que los no fumadores.

- **Cuantificar la confianza en la inferencia:**
  - *Nivel de confianza del 95%*
  - *Significación del contraste: $p = 2\%$*

Población y muestra

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

- **Muestra** (‘*sample*’) 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).



Variables

- 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**.

- En los individuos de la *población* española, de uno a otro ***es variable***:
  - El grupo sanguíneo
    - {A, B, AB, O} ← Var. **Cualitativa**
  - Su nivel de felicidad "declarado"
    - {Deprimido, Ni fu ni fa, Muy Feliz} ← Var. **Ordinal**
  - El número de hijos
    - {0,1,2,3,...} ← Var. **Numérica discreta**
  - La altura
    - {1,62 ; 1,74; ...} ← Var. **Numérica continua**

- Es buena idea codificar las variables como números para poder procesarlas con facilidad en un ordenador.
- Es conveniente asignar "etiquetas" a los valores de las variables para recordar qué significan los códigos numéricos.
  - Sexo (Cualit: Códigos arbitrarios)
    - 1 = Hombre
    - 2 = Mujer
  - Raza (Cualit: 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...
- Estas situaciones deberán ser tenidas en cuentas en el análisis. Datos perdidos ('missing data')

- Los posibles valores de una variable suelen denominarse modalidades. Las modalidades pueden agruparse en clases (intervalos)
  - Edades:
    - Menos de 20 años, de 20 a 50 años, más de 50 años
  - Hijos:
    - Menos de 3 hijos, De 3 a 5, 6 o más hijos
- Las modalidades/clases deben formar un sistema exhaustivo y excluyente
  - Exhaustivo: No podemos olvidar ningún posible valor de la variable
    - **Mal: ¿Cuál es su color del pelo: (Rubio, Moreno)?**
    - Bien: ¿Cuál es su grupo sanguíneo?
  - Excluyente: Nadie puede presentar dos valores simultáneos de la variable
    - Estudio sobre el ocio
      - Mal: De los siguientes, qué le gusta: (deporte, cine)
      - Bien: Le gusta el deporte: (Sí, No)
      - Bien: Le gusta el cine: (Sí, No)

Ejemplo:

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.

Tablas de frecuencia

- 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).

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 $A = 120 - 60 =$ 60

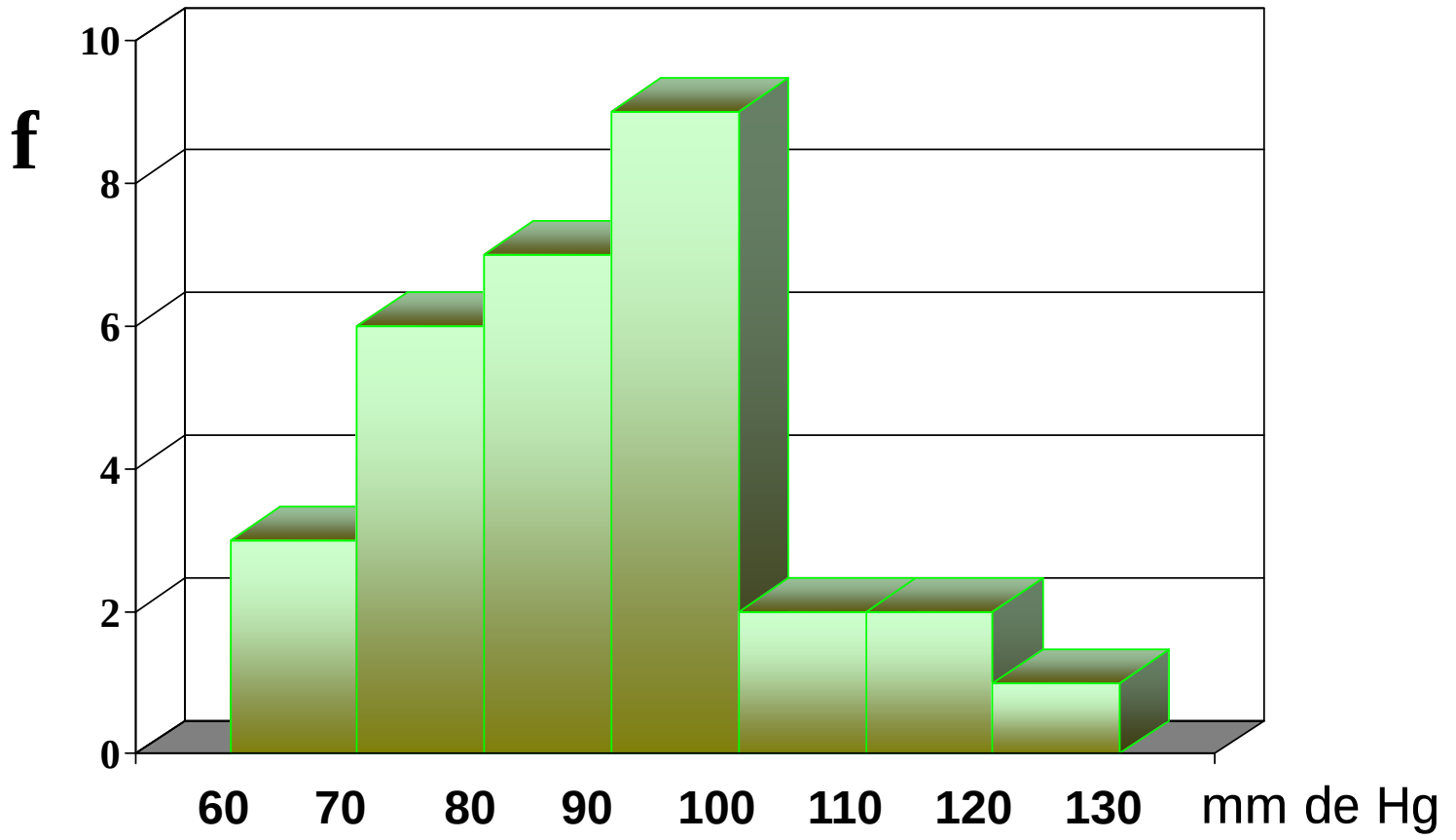
Número de clases: $K = \sqrt{30} = 5.48$ Aprox. 6 clases

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.

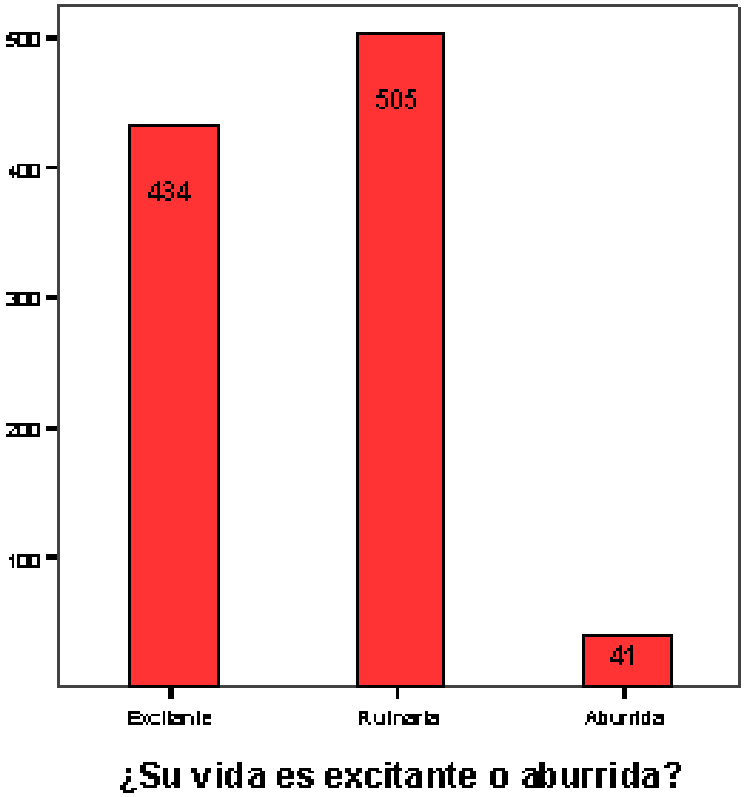
Variable	Frecuencia	Frecuencia normalizada	Frecuencia absoluta	Frecuencia absoluta norm.
x	f	fr	F	Fr
60 - 70	3	0.1	3	0.1
70 - 80	6	0.2	9	0.3
80 - 90	7	0.23	16	0.53
90 - 100	9	0.3	25	0.83
100 - 110	2	0.07	27	0.90
110 - 120	2	0.07	29	0.97
120 - 130	1	0.03	30	1.00
total	30	1.0		

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

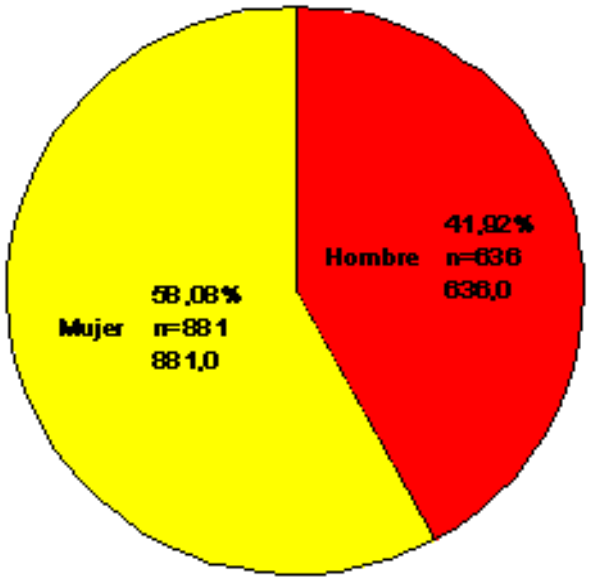


Gráficos para variables cualitativas

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



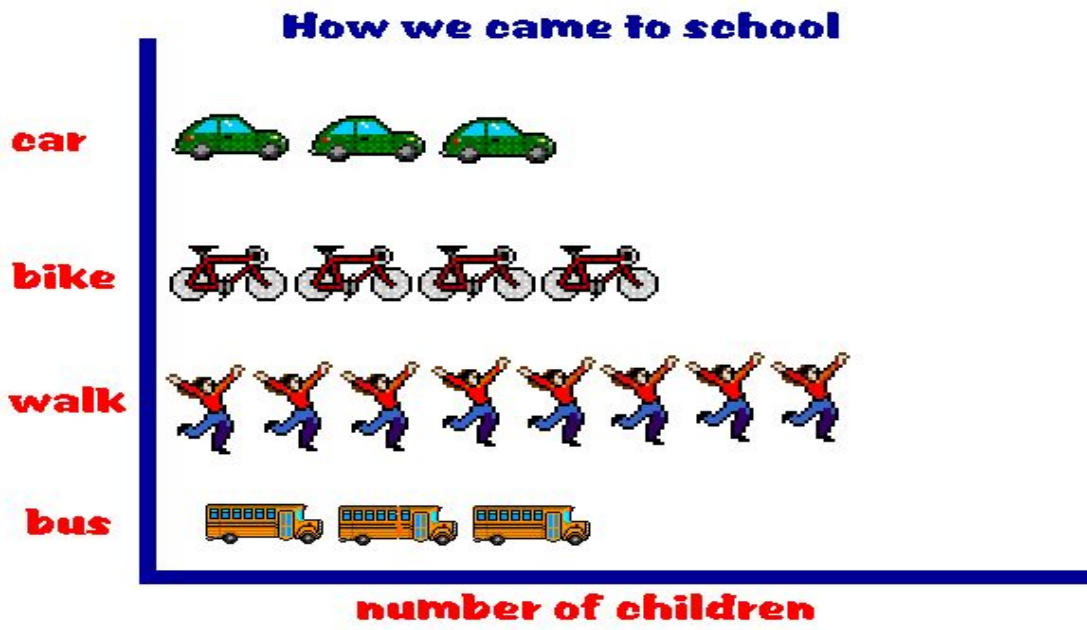
- Diagramas de sectores (tartas, polares)
  - El área de cada sector es proporcional a su frecuencia (abs. o rel.)



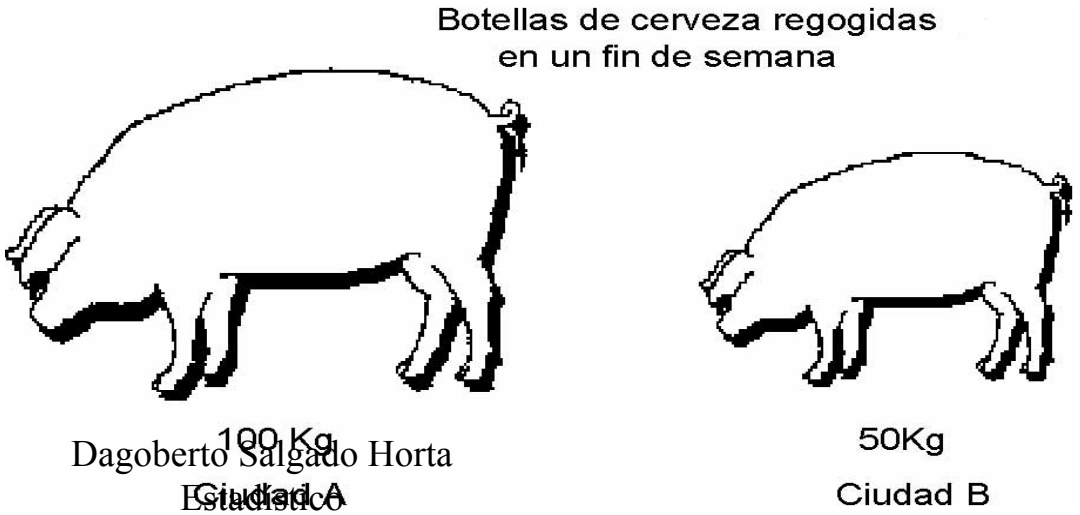
Dagoberto Salgado Horta
Estadístico

Gráficos para variables cualitativas (y 2)

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



De los dos pictogramas, ¿cuál dirías que es incorrecto?



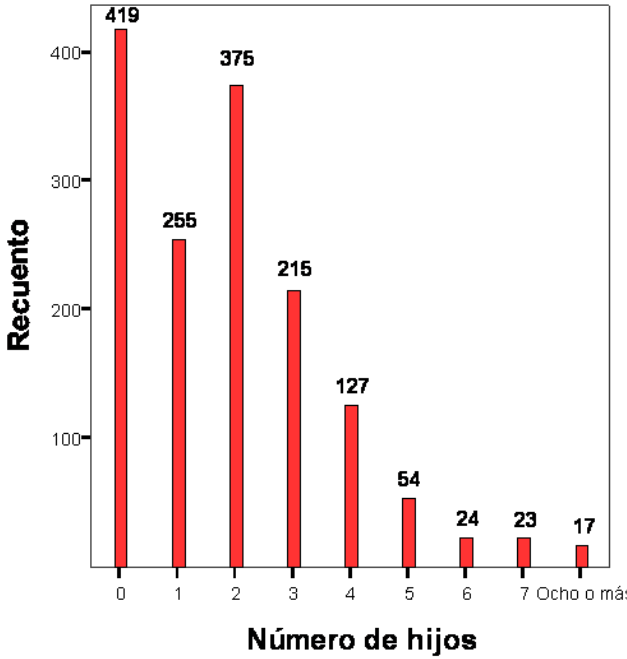
Gráficos diferenciales para variables numéricas

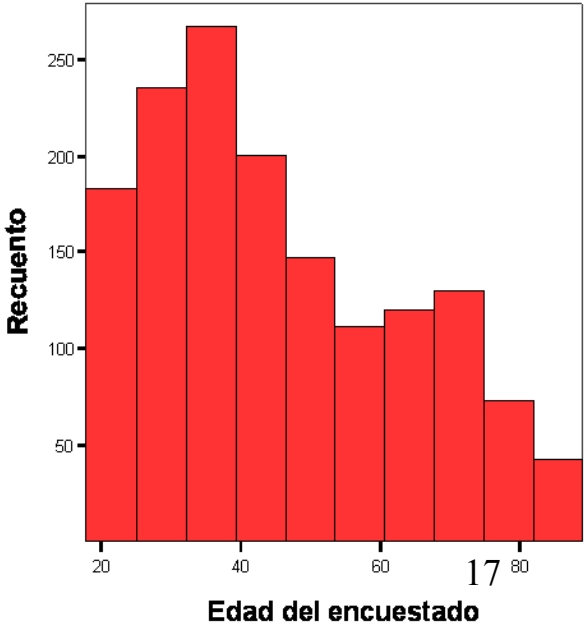
Son diferentes en función de que las variables sean discretas o continuas.

Valen con frec. absolutas o relativas.

- **Diagramas barras para v. discretas**
  - Se deja un hueco entre barras para indicar los valores que no son posibles

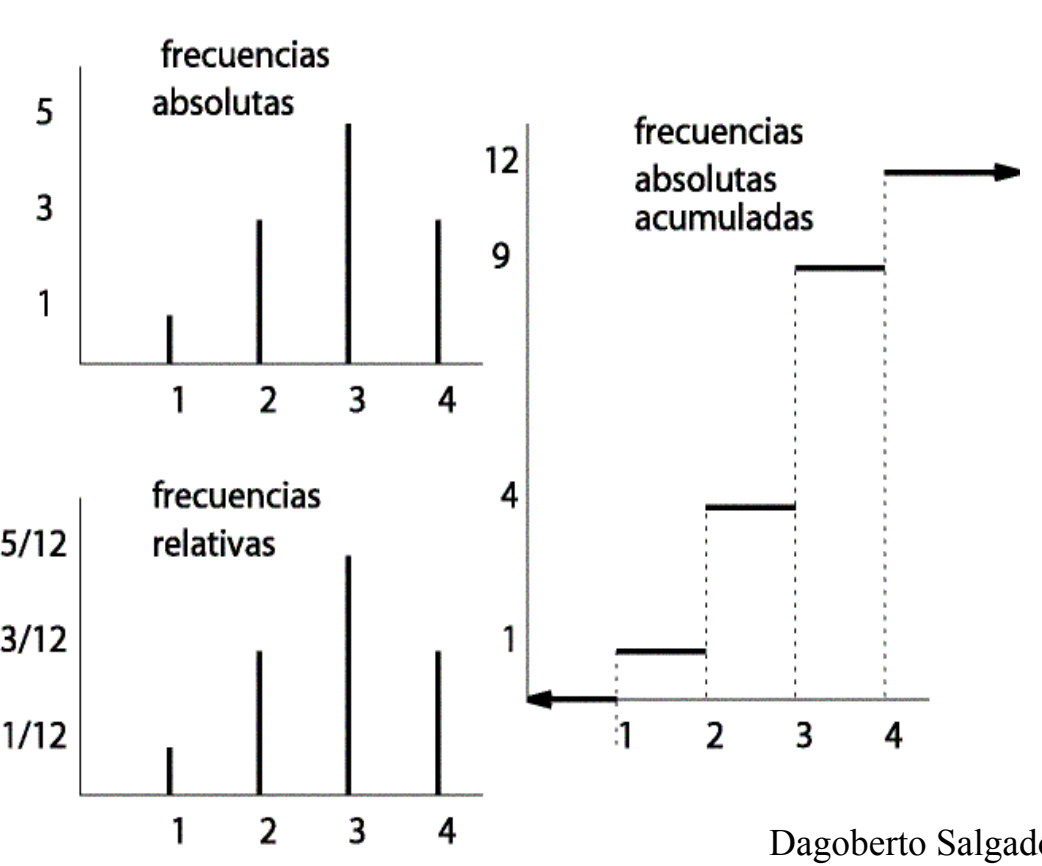
- **Histogramas para v. continuas**
  - El área que hay bajo el histograma entre dos puntos cualesquiera indica la cantidad (porcentaje o frecuencia) de individuos en el intervalo.

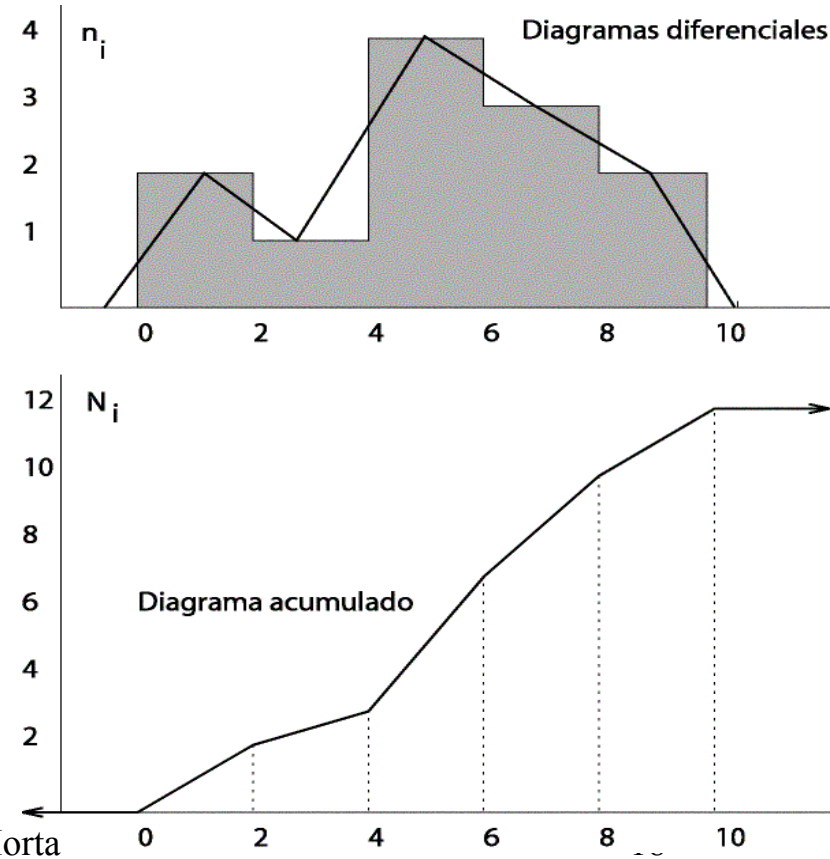




Diagramas integrales

- 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.



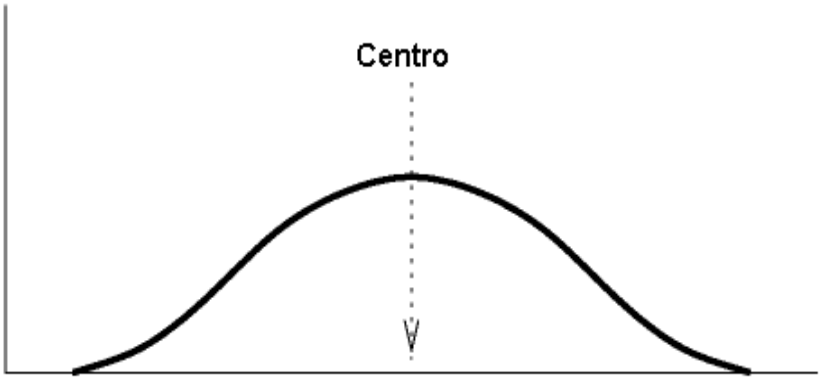


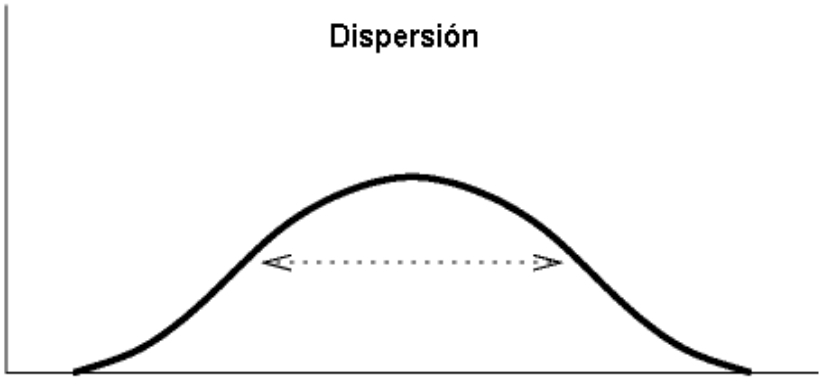
Parámetros y estadísticos

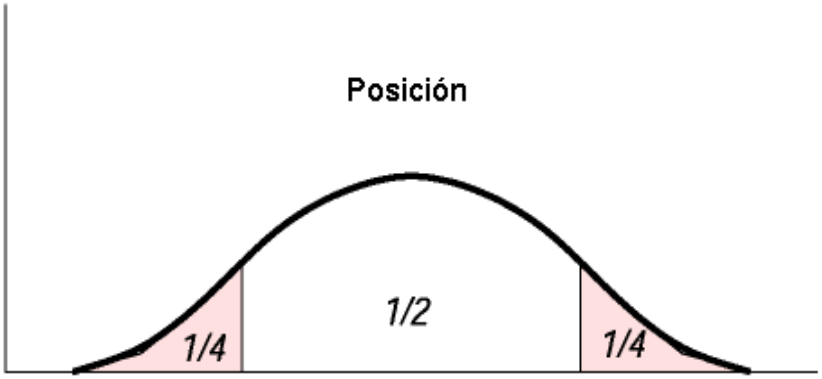
- Parámetro: Es una cantidad numérica calculada sobre una población.
  - La altura media de los individuos de un país.
  - La idea es resumir toda la información que hay en la población en unos pocos números (parámetros).

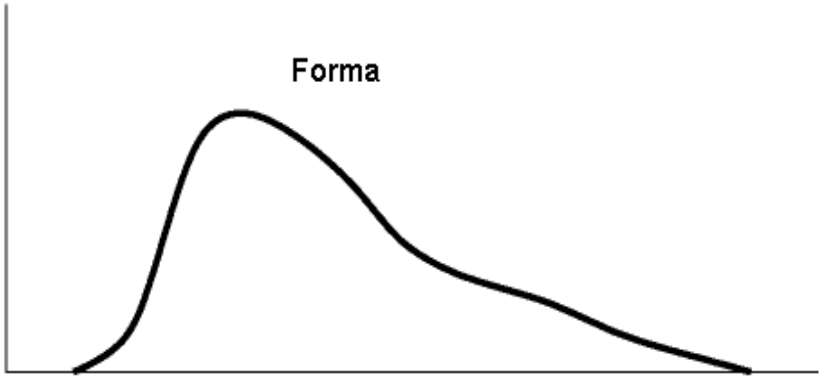
- Estadístico: Ídem (cambiar población por muestra).
  - La altura media de los que estamos en este aula.
    - Somos una muestra (¿representativa?) de la población.
  - Si un estadístico se usa para aproximar un parámetro también se le suele llamar estimador.

Estadísticos de forma intuitiva







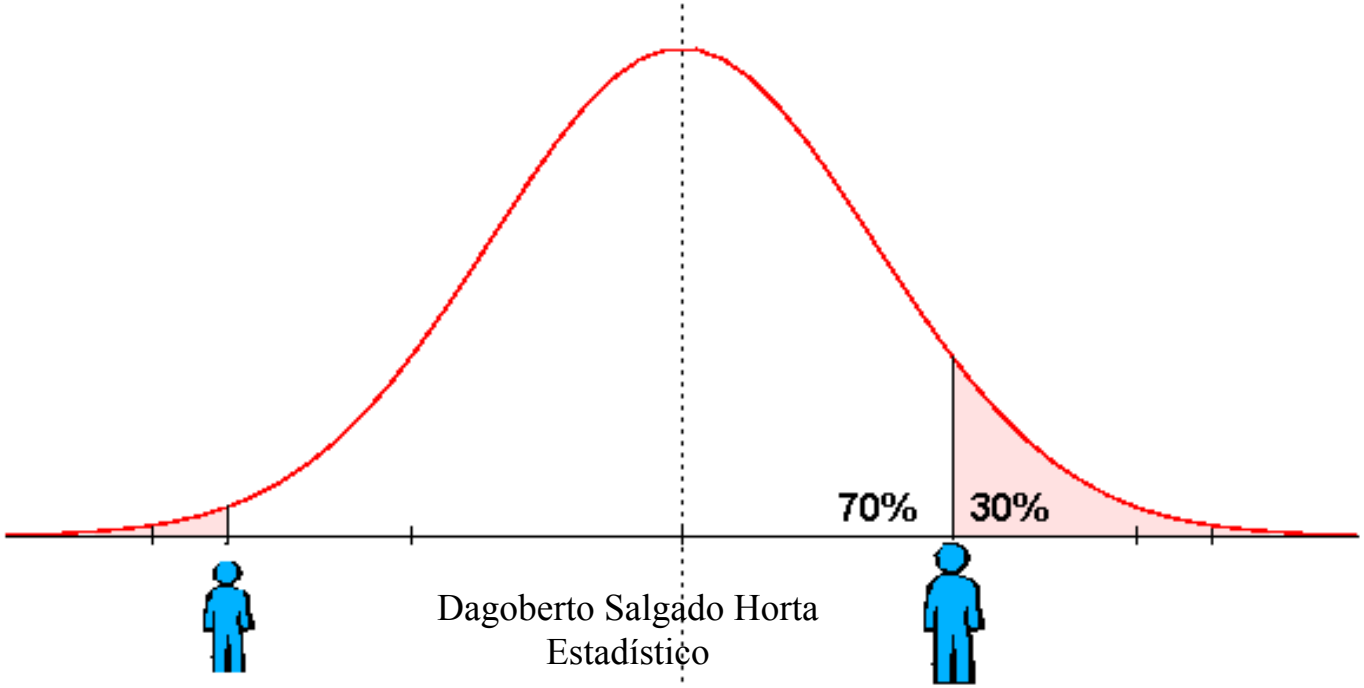


Estadísticos

- Posición
  - 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 típica, coeficiente de variación, rango, varianza
- Forma
  - Asimetría
  - Apuntamiento o curtosis

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,...



- **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.

- ***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 1º y 3º cuartil (Q_1 y Q_3).

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	159 160	160	160	163	164	164	165	166
168	168	170	172	172	176	179		

133.5

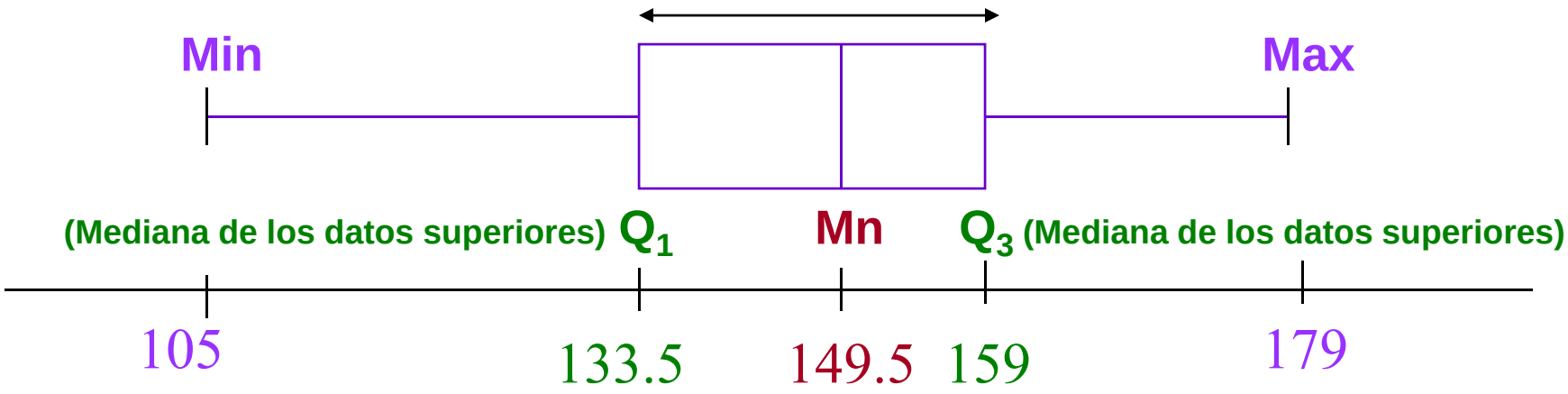
149.5

Un resumen de esta serie en 5 valores

Min = 105 ; Max =179; Q_1 = 133.5 ; Q_3 = 159 ; Q_2 = Mn = 149.5

$IQR = Q_3 - Q_1$

Recorrido intercuartílico



("Box-and-Whisker" plot)

Centralización

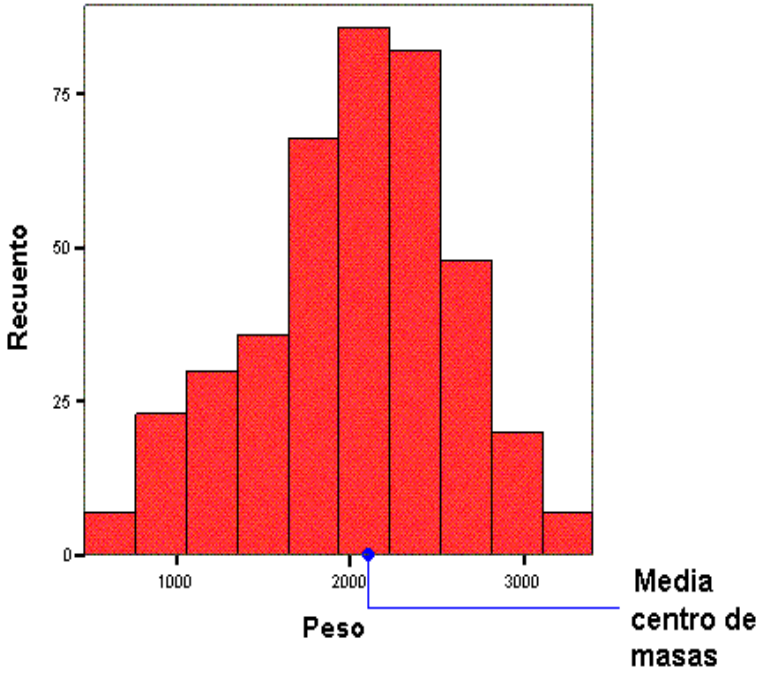
- Añaden unos cuantos casos particulares a las medidas de posición. Son medidas que buscan posiciones (valores) con respecto a los que los datos muestran tendencia a agruparse.

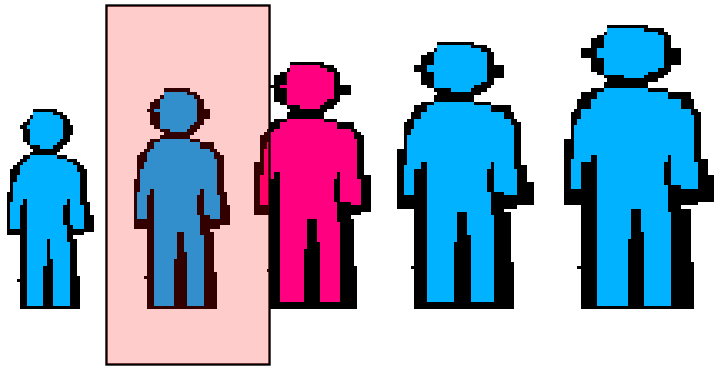
- Media (‘mean’) Es la media aritmética (promedio) de los valores de una variable. Suma de los valores dividido por el tamaño muestral.
  - Media de {2, 2, 3, 7} es (2+2+3+7)/4 = 3,5
  - Conveniente cuando los datos se concentran simétricamente con respecto a ese valor. Muy sensible a valores extremos.
  - Centro de gravedad de los datos.

Centralización

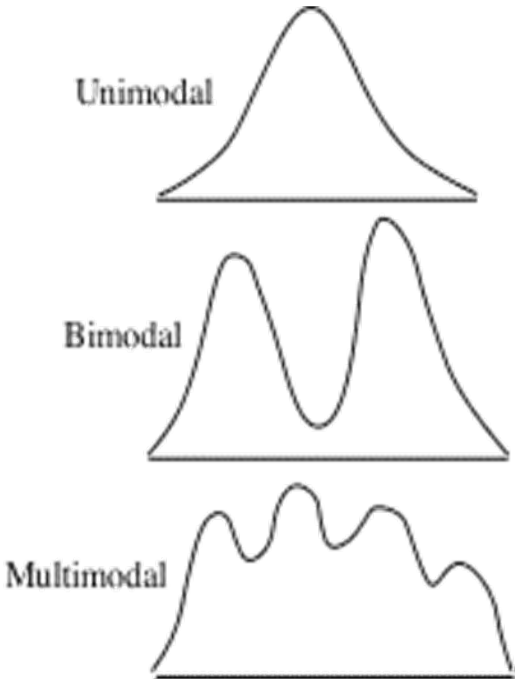
- Mediana (‘median’) Es un valor que divide a las observaciones en dos grupos con el mismo número de individuos (percentil 50). Si el número de datos es par, se elige la media de los dos datos centrales.
  - Mediana de 1, 2, 4, **5**, 6, 6, 8 es 5
  - Mediana de 1, 2, 4, **5**, **6**, 6, 8, 9 es (5+6)/2 = 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!

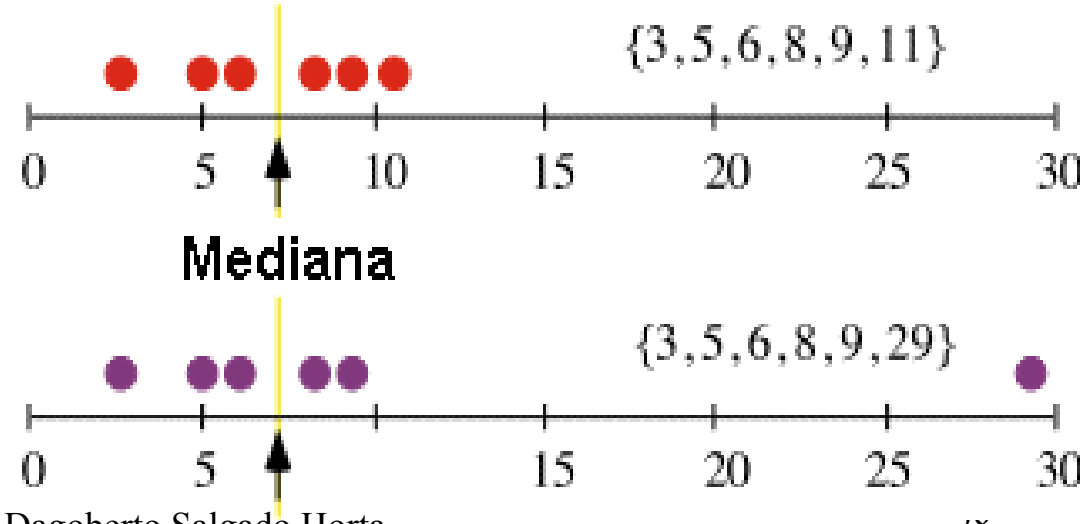
- Moda (‘mode’) Es el/los valor/es donde la distribución de frecuencia alcanza un máximo.





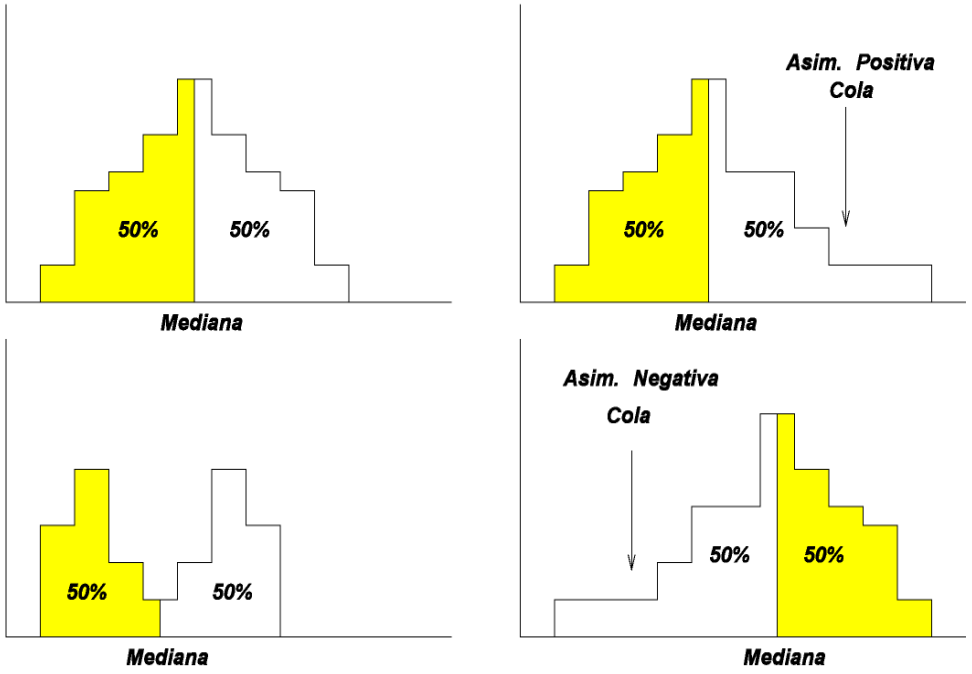
Altura mediana

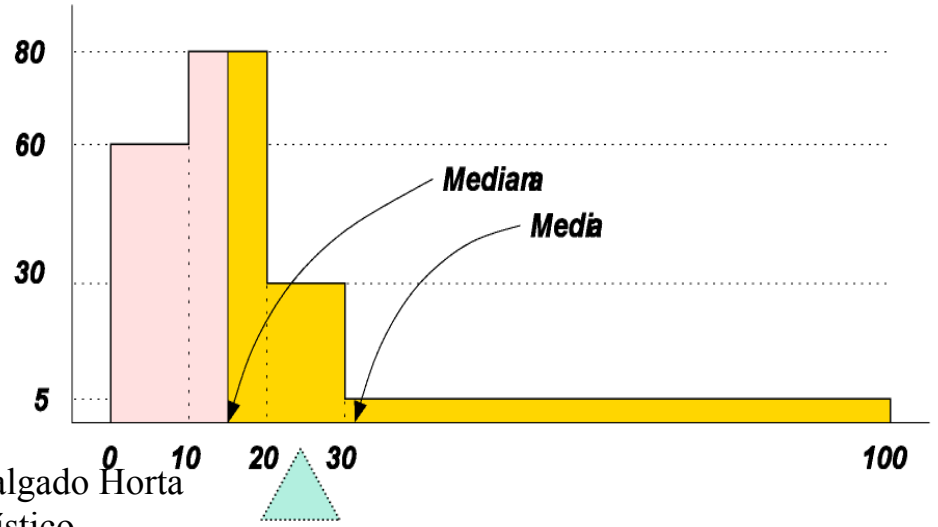




Asimetría o sesgo

- 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.

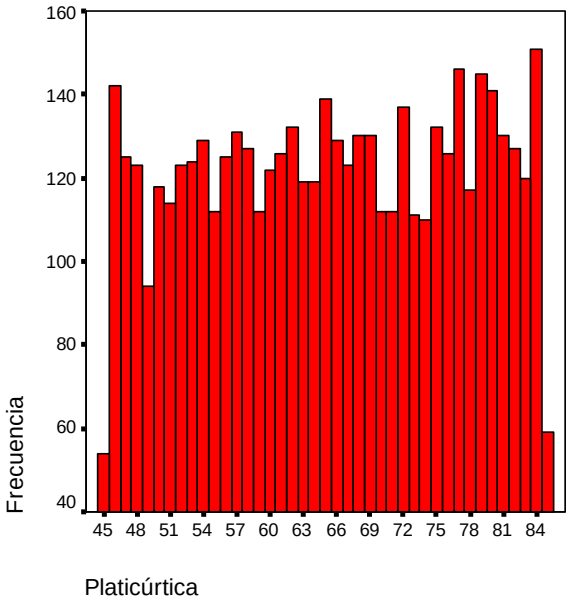




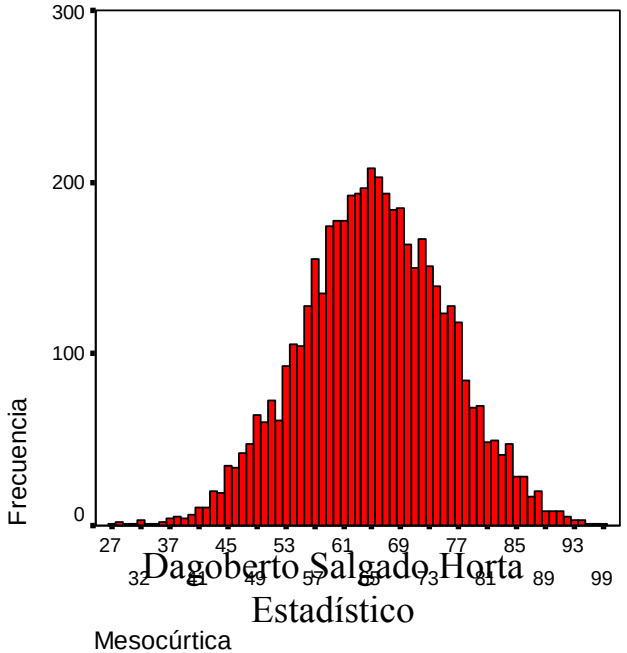
Apuntamiento o curtosis (kurtosis)

- 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

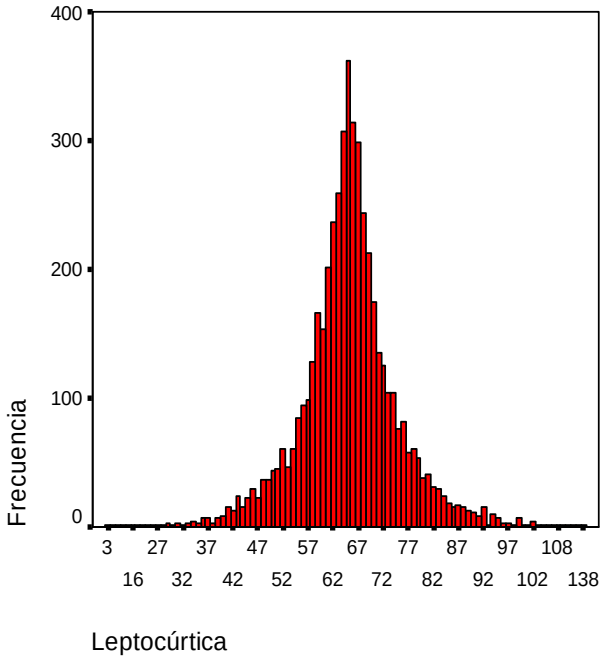
Los gráficos poseen la misma media y desviación típica, pero diferente grado de apuntamiento o curtosis.



Platicúrtica



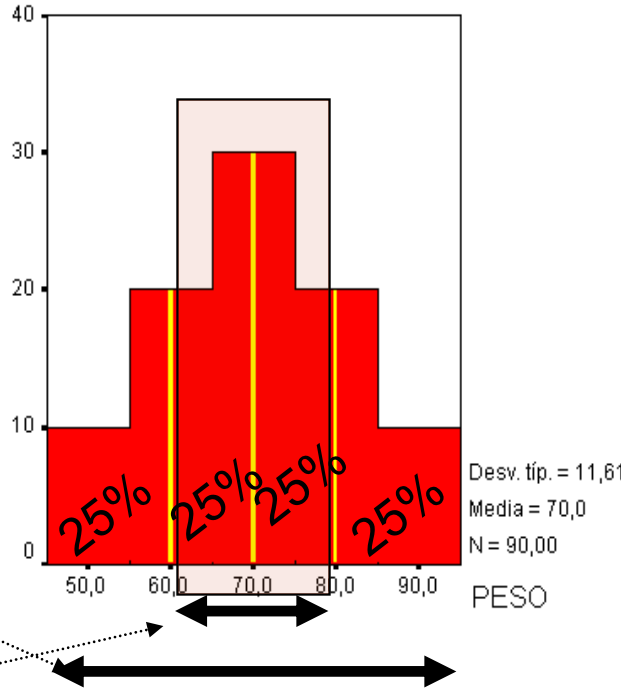
Mesocúrtica



Leptocúrtica

Medidas de dispersión

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

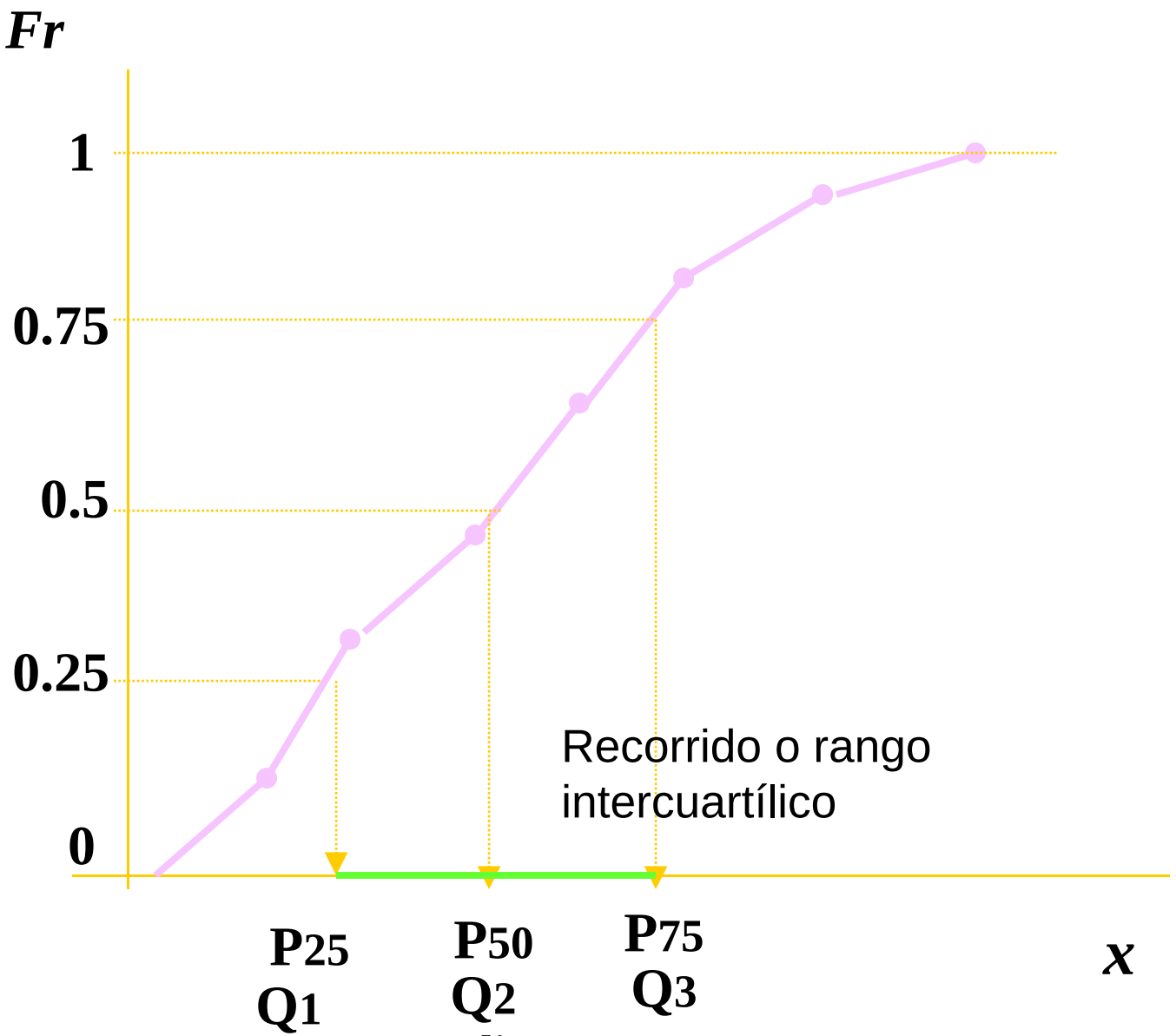


•**Amplitud o Rango** (‘range’): La diferencia entre las observaciones extremas.

- 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

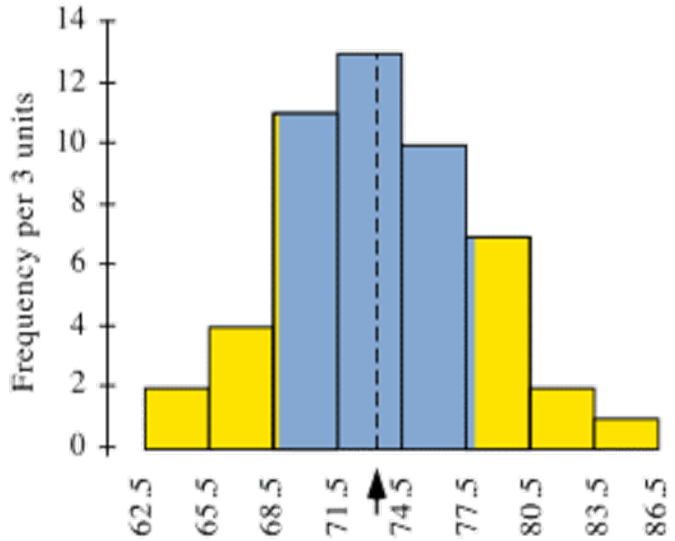
- Varianza S^2 ('Variance'): Mide el promedio de las desviaciones (al cuadrado) de las observaciones con respecto a la media.

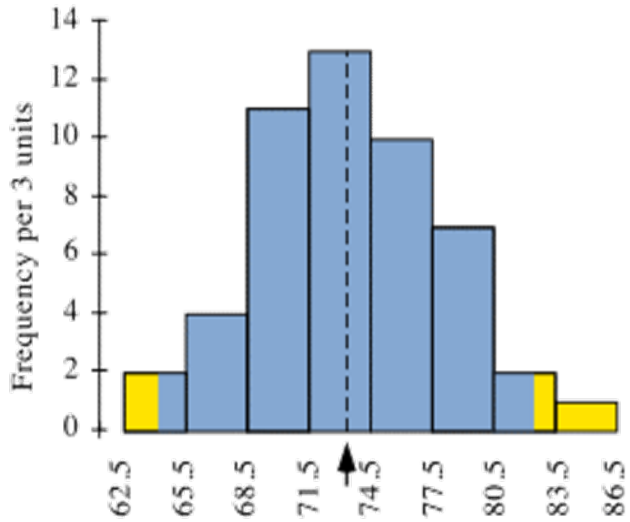
$$S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

  - Es sensible a valores extremos (alejados de la media).
  - Sus unidades son el cuadrado de las de la variable.

- Desviación típica ('standard deviation')
  Es la raíz cuadrada de la varianza. Tiene las misma dimensionalidad (unidades) que la variable.

$$S = \sqrt{S^2}$$





- Centrados en la media y a una desviación típica de distancia tenemos más de la mitad de las observaciones (izq.)

- A dos desviaciones típicas las tenemos a casi todas (dcha.)

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
- Los ingenieros electrónicos hablan de la razón 'señal/ruido' (su inverso).

Desigualdad de Chebyshev (1821-1894)

Si un conjunto de datos posee una varianza pequeña no existirán "muchos valores" alejados de la media. Precisemos: sea el intervalo alrededor de la media:

$$\bar{x} - k\sigma < x_i < \bar{x} + k\sigma$$

$$S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \cdot f_i$$

$$S^2 = \underbrace{\frac{1}{n}\sum_{\substack{i\,dentro \\ del\ entorno}} (x_i - \bar{x})^2 \cdot f_i}_{>0} + \underbrace{\frac{1}{n}\sum_{\substack{i\ fuera \\ del\ entorno}} (x_i - \bar{x})^2 \cdot f_i}_{>0}$$

Demostración:

$$S^2 \geq \frac{1}{n} \sum_{\substack{i \text{ fuera} \\ \text{del entorno}}} (x_i - \bar{x})^2 \cdot f_i \geq \frac{1}{n} \sum_{\substack{i \text{ fuera} \\ \text{del entorno}}} k^2 S^2 \cdot f_i =$$

$$= k^2 S^2 \frac{1}{n} \sum_{\substack{i \text{ fuera} \\ \text{del entorno}}} f_i$$

$$\boxed{\frac{1}{n} \sum_{\substack{i \text{ fuera} \\ \text{del entorno}}} f_i \leq \frac{1}{k^2}}$$

La frecuencia relativa de los datos que caen fuera del intervalo de centro media y radio k veces la varianza es igual o menor que $1/k^2$

•Han vuelto a pedirle una millonada al decano de la facultad de fisicas para hacer un experimento.
- ¡Otra vez ! Pero bueno, ¿por qué no podéis ser como los matemáticos, que se apañan solo con papel, lápiz y una papelera ? ¿O como los filósofos, que sólo necesitan papel y lápiz ?

•En cierta ocasión le preguntaron a un vendedor que como podia vender tan baratos sus sandwiches de conejo, a lo que respondió :
-"bueno, tengo que admitir que hay un poco de carne de caballo. Pero la mezcla es solo 50:50 ; uso el mismo numero de conejos que de caballos".
[Darrel Huff, "Como mentir con la estadística".]

Gráficos de tallos y hojas del estadístico John Tukey

La enfermera Florence Nightingale recopiló datos estadísticos sobre mortalidad en los hospitales militares británicos... guerra de Crimea. Consecuencia disminución de la tasa de mortalidad.

Fue John Tukeyquien inventó en bigote.

Se extiende a 1.5 IQR de los cuartiles. Así vemos los datos atípicos.

En un gráfico de caja es muy útil para representar diferencias entre grupos.

Filtrado: tenemos tendencia fuerte a olvidar los fracasos y concentrarnos en los éxitos y aciertos. Tragaperras, fracasos bursátiles y financieros, curanderos

El valor medio de unas medidas normalmente es igual para un pequeño conjunto que para uno grande, pero los valores extremos varían muchísimo. Pensemos en el caudal de un río. El caudal medio de un año coincide con el de 25 años. Un desbordamiento se recuerda fuertemente...

Como siempre nos quedamos con los extremos no es extraño que en deportes, ciencia o arte denigremos las figuras de hoy en comparación con las del pasado.

Otra consecuencia: las noticias internacionales son peores que las nacionales, peores que las regionales, peores que las locales que son peores que las del entorno inmediato.

La desviación típica es menor a todas las desviaciones cuadráticas respecto a cualquier promedio m (mirar librito de bachillerato)

$$\frac{d}{dy}\sum_i (x_i - y)^2 = -2\sum_i (x_i - y) = 0$$

$$-2\sum_i x_i + 2\sum_i y = 0$$

$$\sum_i x_i = \sum_i y = ny$$

$$y = \frac{1}{n}\sum_i x_i = \bar{x}$$