



Madrid Internet
of Things Institute
Make-build-learn

27 de febrero de 2020

Estadística para Data Science

Sesión 3: Estadística Descriptiva (I)

Jesús Hernando Corrochano



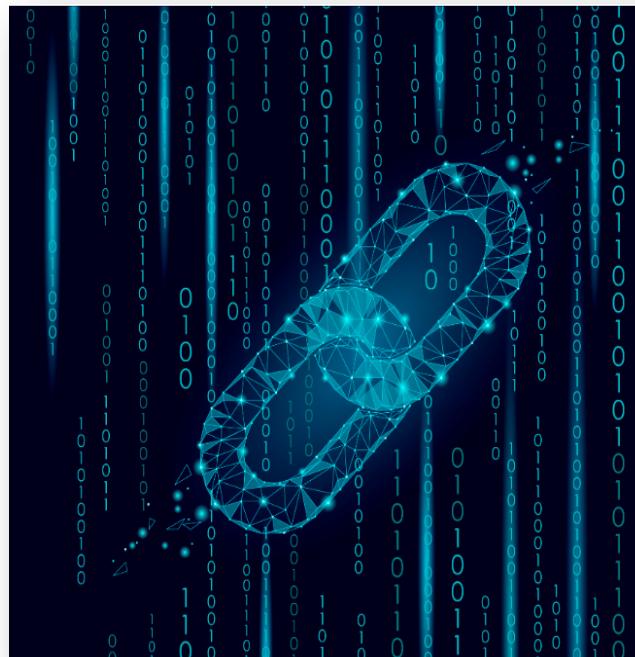
Estadística para Data Science

● Programa



	Sesión 1 13/2	Sesión 2 20/2	Sesión 3 27/2	Sesión 4 5/3	Sesión 5 12/3	Sesión 6 19/3	Sesión 7 26/3	Sesión 8 2/3
Introducción a la estadística	■							
Introducción a la combinatoria y la probabilidad		■						
Estadística descriptiva				■				
Regresión y correlación					■			
Estadística inferencial						■		
Probabilidad Total. Teorema de Bayes. Test A/B							■	

- 1. Estadística descriptiva**
- 2. Frecuencias**
- 3. Medidas de Centralización**
- 4. Medidas de Posición**
- 5. Medidas de Dispersion**
- 6. Distribución Normal**
- 7. Distribución Binomial**
- 8. Aproximación: Teorema de Moivre-Laplace**
- 9. Asimetría o sesgo y curtosis**



Estadística Descriptiva

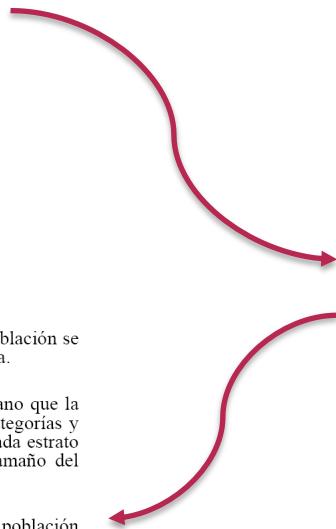
Pasos en un estudio estadístico

- **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?

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 sí.



- **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% ↗?

La **estadística descriptiva** trata de describir y analizar algunos caracteres de los individuos de un grupo dado sin extraer conclusiones para un grupo mayor. Se suele hacer con ayuda de tablas y gráficos y con algunos parámetros estadísticos.

La estadística descriptiva es la técnica matemática que obtiene, organiza, presenta y describe un conjunto de datos con el propósito de facilitar el uso, generalmente con el apoyo de tablas, medidas numéricas o gráficas.

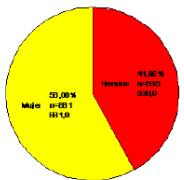
https://es.wikipedia.org/wiki/Estad%C3%ADstica_descriptiva

La estadística descriptiva es la rama de la estadística que recolecta, analiza y caracteriza un conjunto de datos (peso de la población, beneficios diarios de una empresa, temperatura mensual,...) con el objetivo de describir las características y comportamientos de este conjunto mediante medidas de resumen, tablas o gráficos.

<https://www.universoformulas.com/estadistica/descriptiva/>

● Estadística Descriptiva

- **Estadística 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.
- **Estadística 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.



Descriptiva

Probabilidad

Inferencia

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

• **deducir las leyes** que rigen esos fenómenos

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

● Estadística Resumen

- **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).



- **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**.

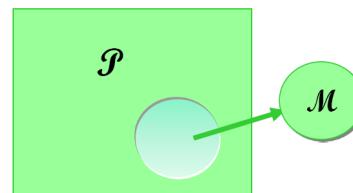
- **Variable:** una **variable** es una característica observable *que varía entre los diferentes individuos* de una población. La información que disponemos de cada individuo es resumida en **variables**.

- **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, ...}

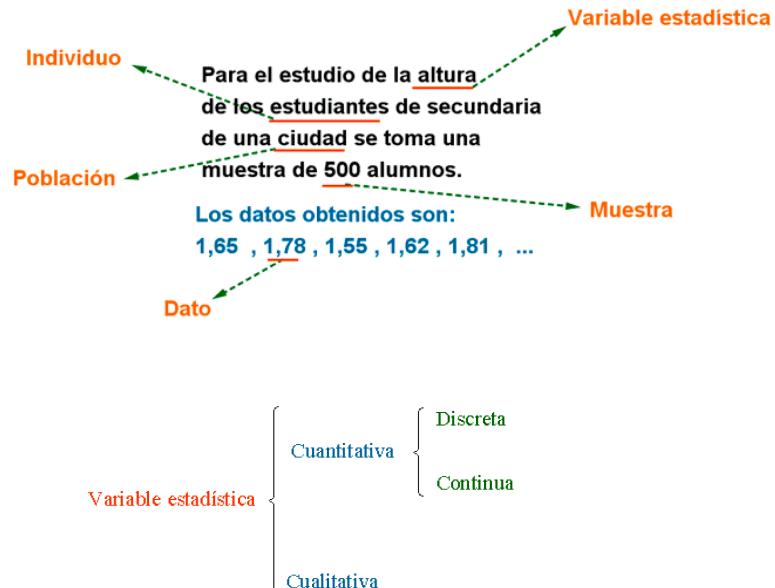
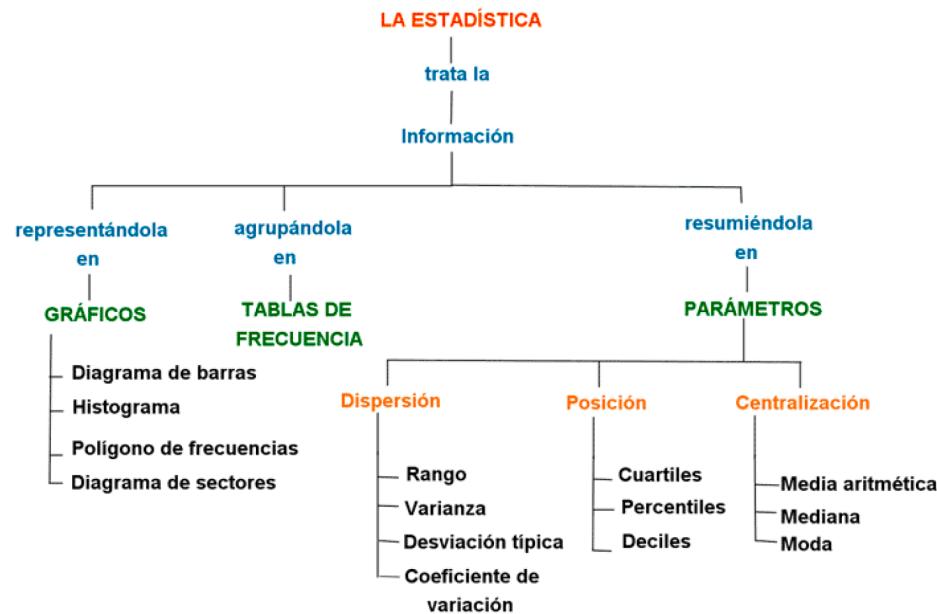
- **Muestra Aleatoria:** es una muestra bien representativa de la población. Se considera que cada elemento de la población ha tenido la misma oportunidad de formar parte de la muestra. Las conclusiones basadas en una muestra aleatoria son confiables.



\mathcal{P} : población

\mathcal{M} : muestra

● Estadística Resumen



Frecuencias



Absoluta (f_i): es el número de veces que aparece un valor x_i de una variable estadística.

La suma de todas las frecuencias absolutas es necesariamente el tamaño de la muestra o la población de estudio.

$$f_1 + f_2 + \dots + f_{n-1} + f_n = \sum_{i=1}^n f_i = N$$

- 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).

Relativa (h_i): cociente de la frecuencia absoluta correspondiente, f_i , entre el tamaño N de la población.

La suma de todas las frecuencias relativas acumuladas es la unidad.

$$h_1 + h_2 + \dots + h_{n-1} + h_n = \sum_{i=1}^n h_i = 1$$

Frecuencias (II)

Absoluta Acumulada (F_i): suma de todas las frecuencias absolutas de los valores menores o iguales que él.

$$F_i = f_1 + f_2 + \dots + f_i$$

Relativa Acumulad (H_i): suma de todas las frecuencias relativas de los valores menores o iguales que él.

$$H_i = H_1 + H_2 + \dots + H_i$$

En caso de variables continuas, los datos los agrupamos en intervalos que llamamos intervalos de clase I_i . Los intervalos de clase serán semiabiertos $[a, b)$ excepto el último que será cerrado $[a, b]$

El punto medio de cada intervalo de clase es lo que llamamos **marca** de clase x_i

Frecuencias (III)

..... cuando el número de datos es grande se organizan los datos en una tabla de frecuencias, agrupándolos previamente en intervalos y representándolos en un gráfico estadístico. Si N es el número de datos:

$$\text{Número de intervalos} = \sqrt{N}$$

$$\text{Amplitud del intervalo} = \frac{\text{Máx.} - \text{Mín.}}{\sqrt{N}}$$

<https://www.lifeder.com/regla-sturges/>

Intervalo (i)	Edades	Frecuencia Absoluta (fi)	Frecuencia Acumulada (Fi)	Frecuencia Relativa (hi)	Frecuencia Relativa Acumulada (Hi)
1	1 - 10	7	7	7 / 42 = 0,17	7 / 42 = 0,17
2	11 - 20	6	7+6= 13	6 / 42 = 0,14	13 / 42 = 0,31
3	21 - 30	8	13+8= 21	8 / 42 = 0,19	21 / 42 = 0,5
4	31 - 40	6	21+6= 27	6 / 42 = 0,14	27 / 42 = 0,64
5	41 - 50	5	27+5= 32	5 / 42 = 0,12	32 / 42 = 0,76
6	51 - 60	4	32+4= 36	4 / 42 = 0,1	36 / 42 = 0,86
7	61 - 70	4	36+4= 40	4 / 42 = 0,1	40 / 42 = 0,95
8	71 - 80	2	40+2= 42	2 / 42 = 0,05	42 / 42 = 1
	Amplitud 9	N: 42		hi= fi / N	Hi= Fi / N

Intervalo	Centro de clase	Amplitud	F	f	FAA	fra
I ₁	c ₁	a ₁				
I ₂	c ₂	a ₂				
.						
I _k	c _k	a _k			n	1
Total			n	1		
			c _j = (L _{ij} + L _{Sj}) / 2			
				a _j = (L _{Sj} - L _{ij})		

Peso	Marca	Recuento	f relativa (%)	F acumulada	F. a. relativa (%)
[36,40)	38	2	2,22	2	2,22
[40,44)	42	3	3,33	5	5,56
[44,48)	46	8	8,89	13	14,44
[48,52)	50	10	11,11	23	25,56
[52,56)	54	13	14,44	36	40,00
[56,60)	58	20	22,22	56	62,22
[60,64)	62	12	13,33	68	75,56
[64,68)	66	9	10,00	77	85,56
[68,72)	70	7	7,78	84	93,33
[72,76)	74	4	4,44	88	97,78
[76,80)	78	2	2,22	90	100,00
		90	100,00		

Frecuencias (III)

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.

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 $A = 120 - 60$

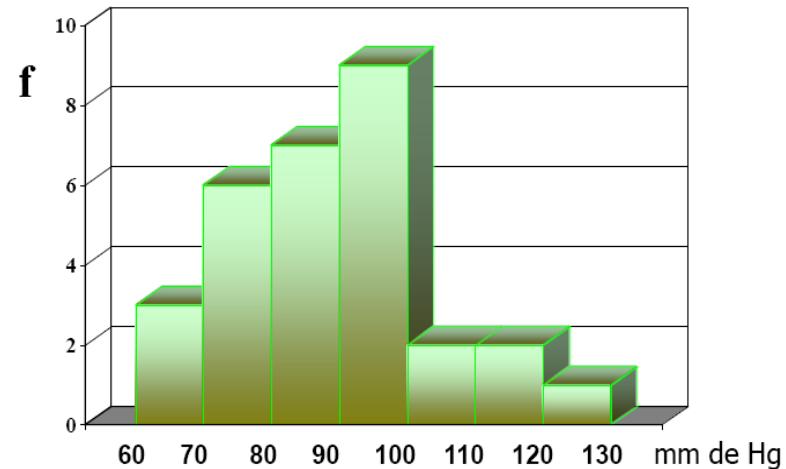
Número de clases: $K = 30^{1/2} = 5.48$. Aprox. 6 clases

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.

Frecuencias (III)

Variable	Frecuencia Absoluta	Frecuencia Absoluta Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
X_i	n_i	N_i	h_i	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	



Frecuencias (III)

Variable X_i	Frecuencia Absoluta n_i	Frecuencia Absoluta Acumulada $N_i = n_1 + \dots + n_i$	Frecuencia Relativa $h_i = \frac{n_i}{N}$	Frecuencia Relativa Acumulada $H_i = h_1 + \dots + h_i$
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	$N_K = N$	$h_K = \frac{n_K}{N}$	$H_K = 1$

Intervalo $[L_i - L_{i+1}[$	Marca de Clase C_i	Frecuencia Absoluta n_i	Frecuencia Absoluta Acumulada $N_i = n_1 + \dots + n_i$	Frecuencia Relativa $h_i = \frac{n_i}{N}$	Frecuencia Relativa Acumulada $H_i = h_1 + \dots + h_i$
$[L_1 - L_2[$	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 Centralización

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

● Medidas de Centralización

Media o promedio: suma de todos los valores de la muestra o población divididos por el número de casos.

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

Media Ponderada. media modificada, donde cada uno de los valores tienen un peso ó ponderación específica, de tal manera que algunos valores pesan más que otros.

$$\bar{X}_p = \frac{\sum_{i=1}^n p_i * x_i}{\sum_{i=1}^n p_i}$$

Media Geométrica

$$\bar{x}_g = \sqrt[n]{x_1 x_2 x_3 \dots x_n}$$

Media Cuadrática

$$\bar{x}_c = \sqrt[n]{\frac{x_1^2 + x_2^2 + x_3^2 + \dots + x_n^2}{n}}$$

{ Ámbito y uso de las mismas?

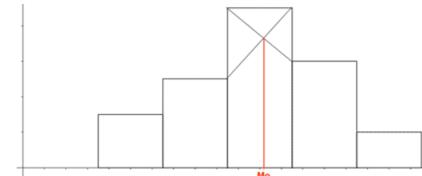
● Medidas de Centralización

La moda (Mo) es el valor x_i que tiene la mayor frecuencia absoluta.

En caso de una variable continua tomando la clase modal:

$$Mo = L_i + \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} \cdot c$$

L_i : límite inferior de la clase modal
c: amplitud del intervalo de la variable estadística
 f_{Mo} , f_{Mo+1} , f_{Mo-1} son, respectivamente , las frecuencias absolutas de la clase modal, la clase anterior y la posterior



La mediana (Me) es la primera variable estadística cuya frecuencia absoluta acumulada (F_i) excede a la mitad del número de datos.

$$\begin{cases} \text{impar} & : \text{la mediana está en el lugar } \frac{N+1}{2} \\ \text{par} & : \text{la mediana está en el lugar } \frac{N}{2} \end{cases}$$

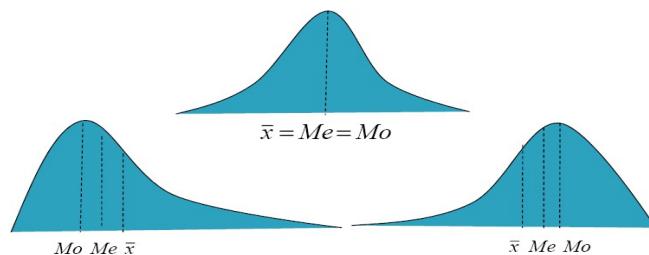
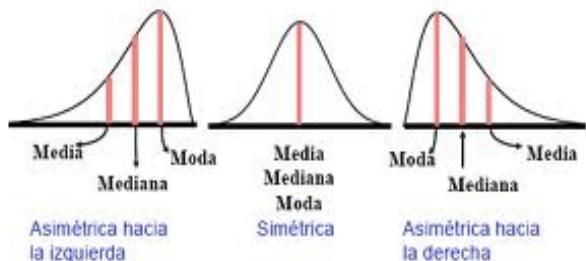
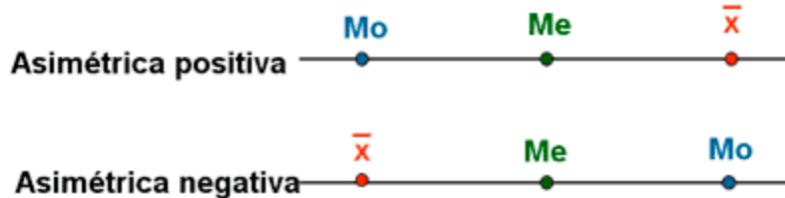
Si el número de datos es impar la mediana coincide con el dato que ocupa el lugar central. Y si es par, es la media aritmética de los dos datos que ocupan los lugares centrales

$$Me = L_i + h \cdot \frac{\frac{N}{2} - F_{i-1}}{f_i}$$

L_i : extremo inferior de la clase medianal
h : amplitud de la clase medianal
N : número total de datos
 F_{i-1} : frecuencia absoluta acumulada del intervalo anterior a la clase medianal
 f_i : frecuencia absoluta de la clase medianal

En caso de que los datos estén agrupados intervalos, se define la clase mediana o intervalo mediano como el intervalo que contiene la mediana.

RELACIONES entre la media, la mediana y la moda

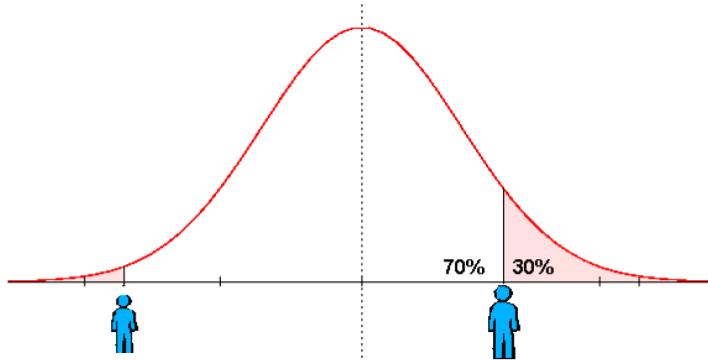


Medidas de Posición



● Medidas de posición: cuartiles, deciles y percentiles

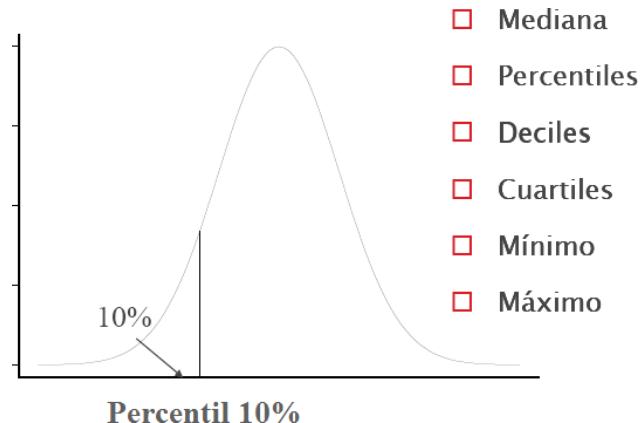
- Se define el **cuantil** de orden α como un valor de la variable por debajo del cual se encuentra una frecuencia acumulada α .
- Casos particulares son los percentiles, cuartiles, deciles, quintiles,...



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).



● Medidas de posición: cuartiles, deciles y percentiles

Las medidas de posición son valores de la variable que informan del lugar que ocupa un dato dentro del conjunto ordenado de valores.

Los **cuartiles** Q1 , Q2 y Q3 son tres valores de la variable estadística que divide en cuatro partes el número de datos. Es decir, que cada tramo será el 25% de los datos recogidos en el estudio.

$$Q_1 = L_i + \frac{\frac{N}{4} - F_{Q_1-1}}{f_{Q_1}} \cdot c$$

$$Q_3 = L_i + \frac{\frac{3 \cdot N}{4} - F_{Q_3-1}}{f_{Q_3}} \cdot c$$

La mediana coincide con el cuartil dos ($Me = Q_2$)

Cuartil inferior: Q1 es un valor de la variable que deja por debajo de él al 25% de la población y por encima al 75%.

Cuartil superior: Q3 es un valor de la variable que deja por debajo de él al 75% de la población y por encima al 25%.

El **rango intercuartílico (r)** es la diferencia entre el tercero y el primer cuartil. $r = Q_3 - Q_1$

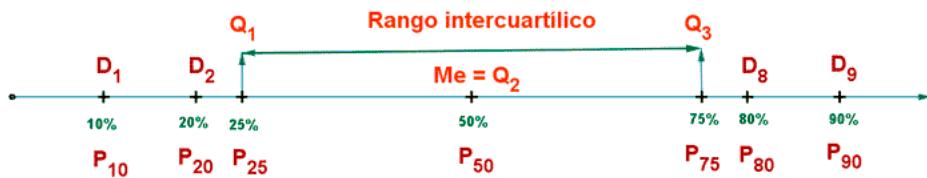
● Medidas de posición: cuartiles, deciles y percentiles

Los **Deciles**, $D_1, D_2, D_3, \dots, D_9$ son nueve valores de la variable estadística que divide en diez partes el número de datos. Es decir, que cada tramo será el 10% de los datos recogidos en el estudio.

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{D_{k-1}}}{f_{D_k}} \cdot c \quad k = 1, 2, \dots, 9$$

Los **Percentiles o Centiles (P_x)**, son 99 valores de la variable estadística que dividen en 100 partes el número de datos

$$P_x = L_i + \frac{\frac{x \cdot N}{100} - F_{P_{x-1}}}{f_{P_x}} \cdot c \quad x = 1, 2, \dots, 99$$



$$\begin{cases} Q_2 \equiv P_{50} \equiv Me \\ Q_1 \equiv P_{25} \\ Q_3 \equiv P_{75} \end{cases}$$

$\left\{ \begin{array}{l} \text{El segundo cuartil } Q_2 \text{ coincide con la mediana y con } P_{50} \\ \text{El cuartil } Q_1 \text{ coincide con } P_{25} \\ \text{El cuartil } Q_3 \text{ coincide con } P_{75} \end{array} \right.$

Medidas de Dispersión

● Varianza y Desviación Típica

Varianza de una variable estadística es la media aritmética de los cuadrados de las desviaciones de todos los datos o marcas de clase respecto de la media.

$$\sigma^2 = \frac{\sum_{i=1}^n f_i \cdot x_i^2}{N} - \bar{x}^2$$

$$\sigma^2 = \frac{\sum_{i=1}^n f_i \cdot (x_i - \bar{x})^2}{N}$$

Desviación típica: raíz cuadrada positiva de la varianza.

$$\sigma = \sqrt{\frac{\sum_{i=1}^n f_i \cdot (x_i - \bar{x})^2}{N}}$$

Coeficiente de variación o coeficiente de variación de Pearson

Es el cociente entre la desviación típica y la media aritmética de un conjunto de valores.

El CV se interpreta como el número de veces que la desviación típica contiene a la media.

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

● Varianza y Desviación Típica

▶ 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.

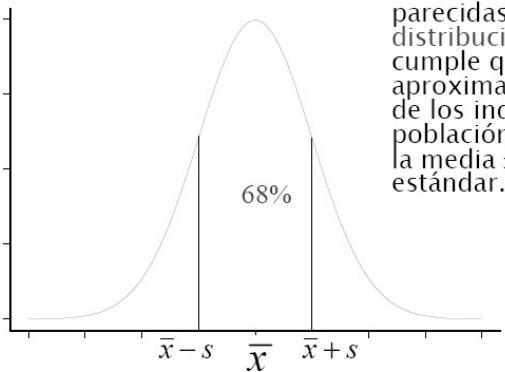
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

▶ Desviación Estándar

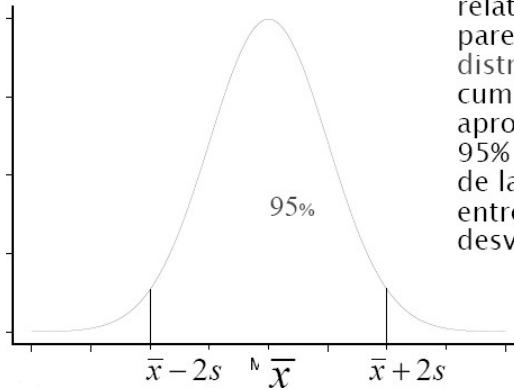
Es la raíz cuadrada de la varianza. Es la más usada de las medidas de dispersión.

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

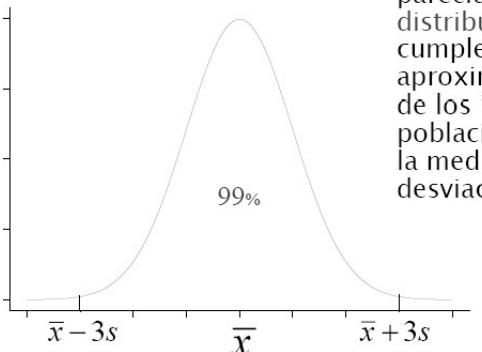
● Varianza y Desviación Típica



- 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 \pm una desviación estándar.



- 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.



- 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.

▶ 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:

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

● Varianza y Desviación Típica

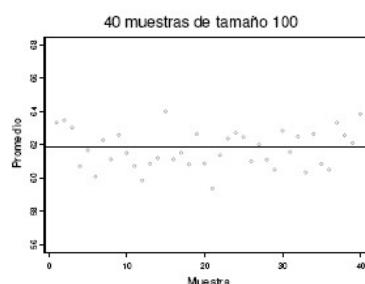
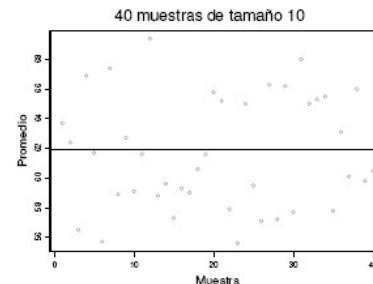
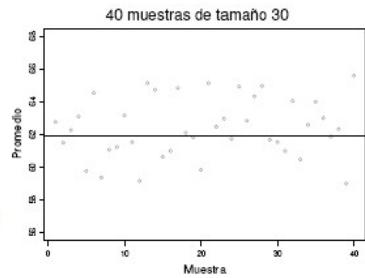
- **Coeficiente de variación**
- Es la razón entre la desviación típica y la media.
 - Mide la desviación típica en forma de “qué tamaño tiene con respecto a la media”
 - También se la denomina **variabilidad relativa**.
 - Es frecuente mostrarla en porcentajes
 - Si la media es 80 y la desviación típica 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 $0^\circ\text{C} \neq 0^\circ\text{F}$

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

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

$$SEM = \frac{s}{\sqrt{n}}$$

Depende de n



Distribución Normal



Distribución Normal (distribución gaussiana)

Una variable X sigue una distribución $N(\mu, \sigma)$ si se verifica que:

- 1) La variable toma valores en toda la recta real
- 2) Su función de densidad es (μ es la media y σ la desviación típica):

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

donde $-\infty < x < +\infty$

1. **Dominio o campo de existencia:** toda la recta real : $(-\infty, +\infty)$

2. **Simetrias:** la función es simétrica respecto a la media $x = \mu$

3. **Corte con los ejes:**

a) No tiene puntos de corte con el eje X.

b) Con el eje Y :

$$x = 0 \Rightarrow f(0) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{\mu^2}{2\sigma^2}}$$

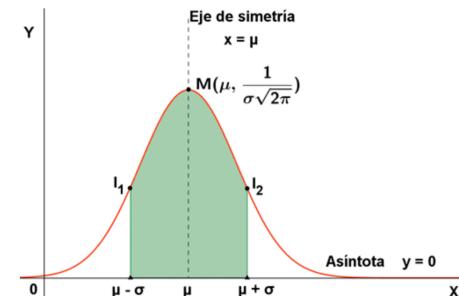
4. **Asíntotas:** $\lim_{x \rightarrow \infty} f(x) = 0$ por tanto el eje X es una asíntota

5. **Crecimiento y decrecimiento:** la función crece hasta $x = \mu$ y decrece hasta $x = \mu$

6. **Máximos y mínimos :** la función $f(x)$ presenta un máximo en $x = \mu$

7. **Puntos de inflexión:** la función presenta dos puntos de inflexión I_1 e I_2 en $x = \mu - \sigma$ y en $x = \mu + \sigma$.

Gráfica conocida como campana de Gauss



Al ser $f(X)$ una función de densidad, el área comprendida entre la gráfica y el eje X es 1.

Distribución Normal (distribución gaussiana)

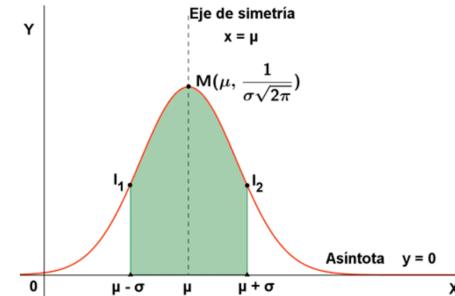
En la mayoría de las ocasiones un conjunto de datos tienen un comportamiento normal cuando:

En el intervalo $(x - \mu, x + \mu)$ el área encerrada es 0,6826, es decir, el 68,25% del total.

En el intervalo $(x - 2\mu, x + 2\mu)$ el área encerrada es 0,9544, es decir, el 95,44% del total.

En el intervalo $(x - 3\mu, x + 3\mu)$ el área encerrada es 0,9973, es decir, el 99,73% del total.

Un dato es atípico cuando está fuera de estos intervalos



Tipificación de una Normal

Si tenemos una distribución normal $N(\mu, \sigma)$, llamamos tipificar la variable al proceso de convertirla en una Normal Estándar $N(0,1)$

$$z = \frac{x - \bar{x}}{\sigma}$$

$$\text{Si } X \rightarrow N(\mu, \sigma) \text{ entonces } Z = \frac{X - \mu}{\sigma} \rightarrow N(0, 1)$$

Distribución Normal Estándar

Una distribución normal estándar o tipificada es la que tiene $\mu = 0$ y $\sigma = 1$. La variable se representa con la letra z. Y la función de densidad correspondiente es:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

$$F(x) = P(X \leq x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

1. **Dominio o campo de existencia:** toda la recta real : $(-\infty, +\infty)$

2. **Simetrías:** la función es simétrica respecto al eje de ordenadas ya que la función es impar, es decir que $f(x) = f(-x)$

3. **Corte con los ejes:**

a) No tiene puntos de corte con el eje X.

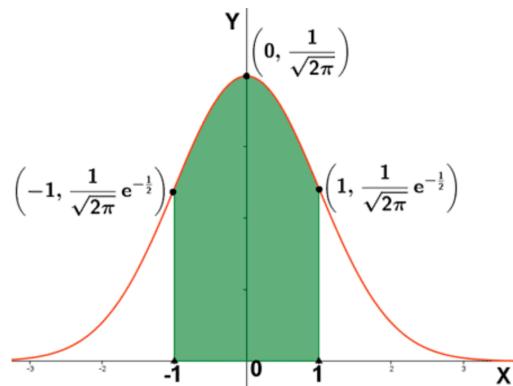
b) Con el eje Y : $x = 0 \Rightarrow f(0) = \frac{1}{\sqrt{2\pi}} \Rightarrow \left(0, \frac{1}{\sqrt{2\pi}}\right)$

4. **Asíntotas:** $\lim_{x \rightarrow \infty} f(x) = 0$ por tanto el eje X es una asíntota

5. **Crecimiento y decrecimiento:** la función crece en el intervalo $\left(-\infty, \frac{1}{\sqrt{2\pi}}\right)$ y decrece en $\left(\frac{1}{\sqrt{2\pi}}, +\infty\right)$

6. **Máximos y mínimos:** la función $f(x)$ presenta un máximo en $\left(0, \frac{1}{\sqrt{2\pi}}\right)$

7. **Puntos de inflexión:** la función presenta dos puntos de inflexión I_1 e I_2 en $\left(1, \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}}\right)$ y en



Distribución Binomial

Distribución Binomial

En estadística , la distribución binomial es una distribución de probabilidad discreta que indica el número de éxitos al realizar una secuencia de n ensayos independientes entre sí, con una probabilidad fija (p) de ocurrencia del éxito entre esos ensayos.
Sólo determina éxito o fracaso.

Ejemplos:

- Producción una pieza, puede salir bien o defectuosa.
- Al nacer puede ser varón o hembra.
- Un equipo de baloncesto puede ganar o perder.
- En un test psicotécnico hay preguntas de verdadero o falso, es decir sólo hay dos alternativas.
- Un tratamiento médico, como por ejemplo la vacuna de la gripe A, puede ser efectivo o inefectivo.
- El objetivo de ventas al año de coches en un concesionario se puede o no lograr

Distribución Binomial

Una distribución binomial de n pruebas o ensayos es una distribución discreta que se representa por $B(n, p)$ y tiene las siguientes características:

- a) El resultado de cada prueba sólo tiene dos opciones, que por ser contrarios son incompatibles. . El suceso A que se llama éxito y el suceso contrario que se llama fracaso.
- b) Cada resultado de cada prueba es independiente de los resultados obtenidos anteriormente.
- c) La probabilidad de éxito se representa $P(A)=p$ y la del fracaso por $P(\text{ }) = q$, siendo:

$$q = 1 - p$$

Llamamos X a la variable aleatoria binomial que describe el número de éxitos, se tiene que:

$$P(\text{obtener } r \text{ éxitos}) = P(X = r) = \binom{n}{r} \cdot p^r \cdot q^{n-r}$$

La expresión anterior se denomina función de probabilidad de la distribución binomial $B(n, p)$

Aproximación



Teorema de De Moivre-Laplace

Establece que la distribución binomial del número de éxitos en n pruebas independientes de Bernoulli con probabilidad de éxito p en cada intento es, aproximadamente, una distribución normal de media np y desviación típica la raíz cuadrada de npq .

https://es.wikipedia.org/wiki/Teorema_de_De_Moivre-Laplace

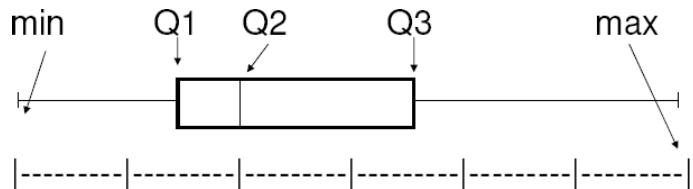
Una distribución $B(n, p)$ se puede aproximar a una normal :

$$N(np, \sqrt{npq}) \xrightarrow{\text{Tipificación}} Z = \frac{X - np}{\sqrt{npq}} \quad \text{es } N(0, 1)$$

$$np \geq 5 \quad n(1-p) \geq 5$$

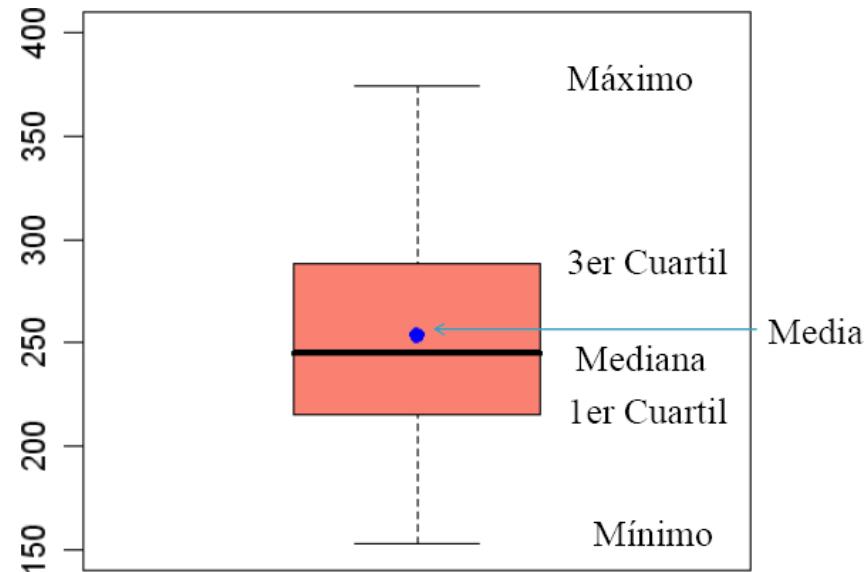
● Box-plot (Caja con bigotes)

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 $> Q3 + 1.5(Q3-Q1)$

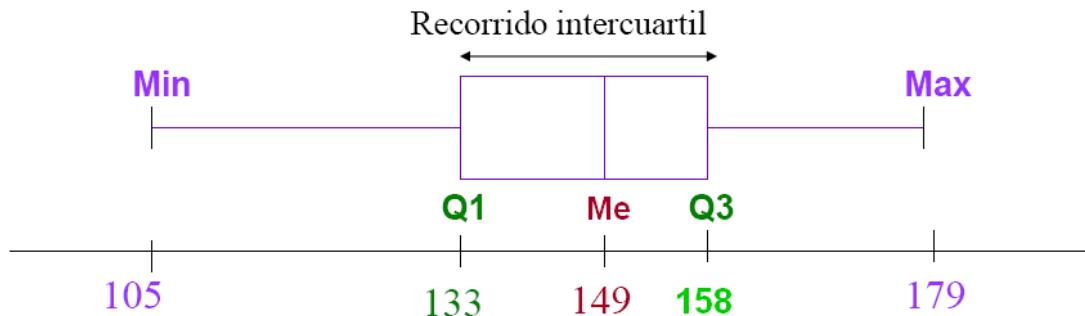


Niveles de Hb en 61 adultos normales

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

Un resumen de esta serie en 5 valores

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



Asimetría o sesgo y curtosis

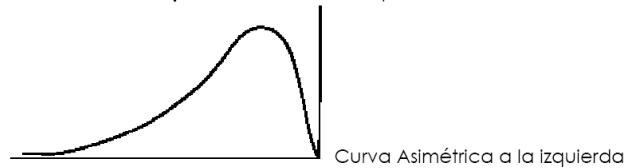
Asimetría

- 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.

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

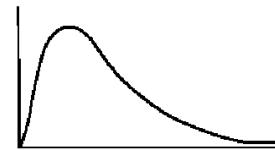
En caso contrario, decimos que la distribución es **Asimétrica**, y entonces puede ser de dos tipos:

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

Asimetría

La **Asimetría** es una medida necesaria para conocer cuánto se parece nuestra distribución a la distribución teórica de una “curva normal”, y constituye un indicador del lado de la curva donde se agrupan las frecuencias.

Índice de simetría 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.

Media – Moda = 3 (Media-Mediana)

Índice de simetría 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

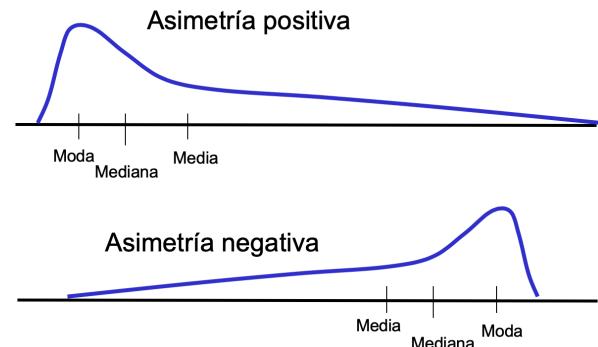
En una distribución simétrica $\bar{x} = Me = Mo$ y $m_3 = 0$. Por eso define como:

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

- 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.

Asimetría

- Si el valor del sesgo es cero (asimetría = 0), la curva de distribución es simétrica, en este caso coinciden los valores de la media, la mediana y la moda.
- Cuando es positiva, el promedio es mayor que la mediana, quiere decir que hay valores agrupados hacia la izquierda de la curva, por debajo del valor de la media.
- Cuando es negativa, la media es menor a la mediana, significa que los valores tienden a agruparse hacia la derecha de la curva, por encima de la media.



- 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*" a 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 de la distribución. Basándose en el dato de que en una distribución normal se verifica que:

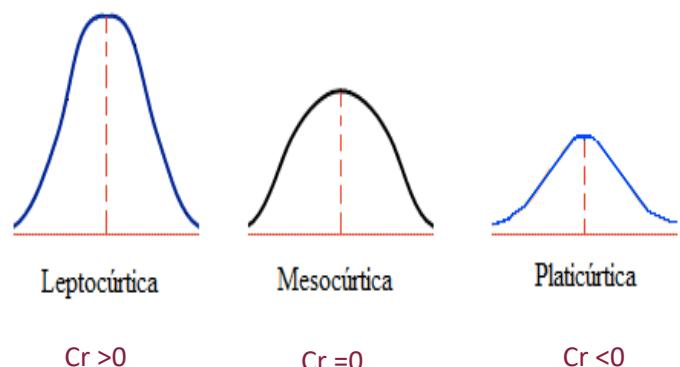
$$\frac{m_4}{s^4} = 3$$

Curtosis

Se define el coeficiente de curtosis de Fisher como:

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

- 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**



Curtosis

Según el coeficiente de curtosis las distribuciones pueden ser:

- **Leptocúrticas ($C_r > 0$):** Cuando un conjunto de datos tiene una mayor concentración alrededor de la media que la distribución normal (mas puntiaguda).
- **Mesocurtica($C_r = 0$):** Cuando las distribución de datos es media alrededor de la media (como la curva normal).
- **Platicurtica($C_r < 0$):** Cuando la distribución de datos alrededor de la media, es menor a la existente en una distribución normal (forma achatada, aplastada).

Para una distribución normal (mesocúrtica) sabemos que

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^4 / n}{s_x^4} = 3$$

Y esta va a ser la referencia para el índice de curtosis que vamos a emplear:

$$C_r = \frac{\sum_{i=1}^n (X_i - \bar{X})^4 / n}{s_x^4} - 3$$



Calle Rufino González 25
28037 Madrid
+34810527241
www.mioti.es

