Tema 1: Estadística Descriptiva

Descripción de una variable

Departamento Matemática Aplicada

Universidad de Málaga

Curso 2020-2021

Tema 1: Estadística Descriptiva

La **estadística** es la ciencia de los datos; implica la colección, clasificación, síntesis, organización, análisis e interpretación de los datos.
Suele aplicarse a dos tipos de problemas:

- Resumir, describir y explorar datos referidos a un colectivo.
- Utilizar datos de muestras para deducir conclusiones sobre un colectivo más amplio del cual se escogieron las muestras.

La rama de la estadística que se dedica a la organización, síntesis y descripción de conjuntos de datos es la **estadística descriptiva**.

La rama de la estadística que se ocupa de utilizar datos de muestras, para inferir algo acerca de la población de la que provienen, se denomina **estadística inferencial**.

Errores básicos mostrando datos estadísticos

- La muestra de un porcentaje suelto nunca puede servir para inferir una relación entre dos variables
- Mostrar Rankings absolutos, no relativos, para intentar clasificar.

Muestra de un porcentaje suelto

Muestra de un porcentaje suelto

POBLACIÓN | Estudio de la Universidad de Almería

4.000 menores son detenidos cada año en Andalucía, el 20% de toda España

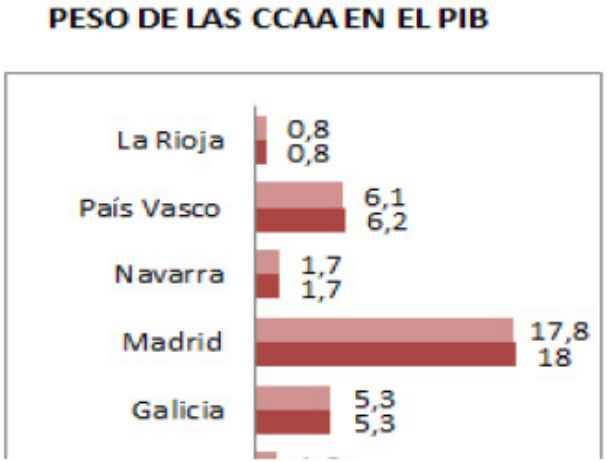
Rankings absolutos, no relativos

¿Desigualdad? Las siete provincias más ricas de España generan la mitad del PIB

En España, la actividad económica está más que concentrada. Según datos del INE, casi la mitad del PIB se genera en sólo siete provincias: Madrid, Barcelona, Valencia, Sevilla, Alicante, Vizcaya y Málaga. Sólo la capital del Estado concentra casi uno de cada cinco euros que se produce en España, más de lo que aporta cualquier comunidad autónoma, excepto Cataluña.

Inés Calderón / www.invertia.com

Viernes, 21 de Febrero de 2014 - **8:28 h.**



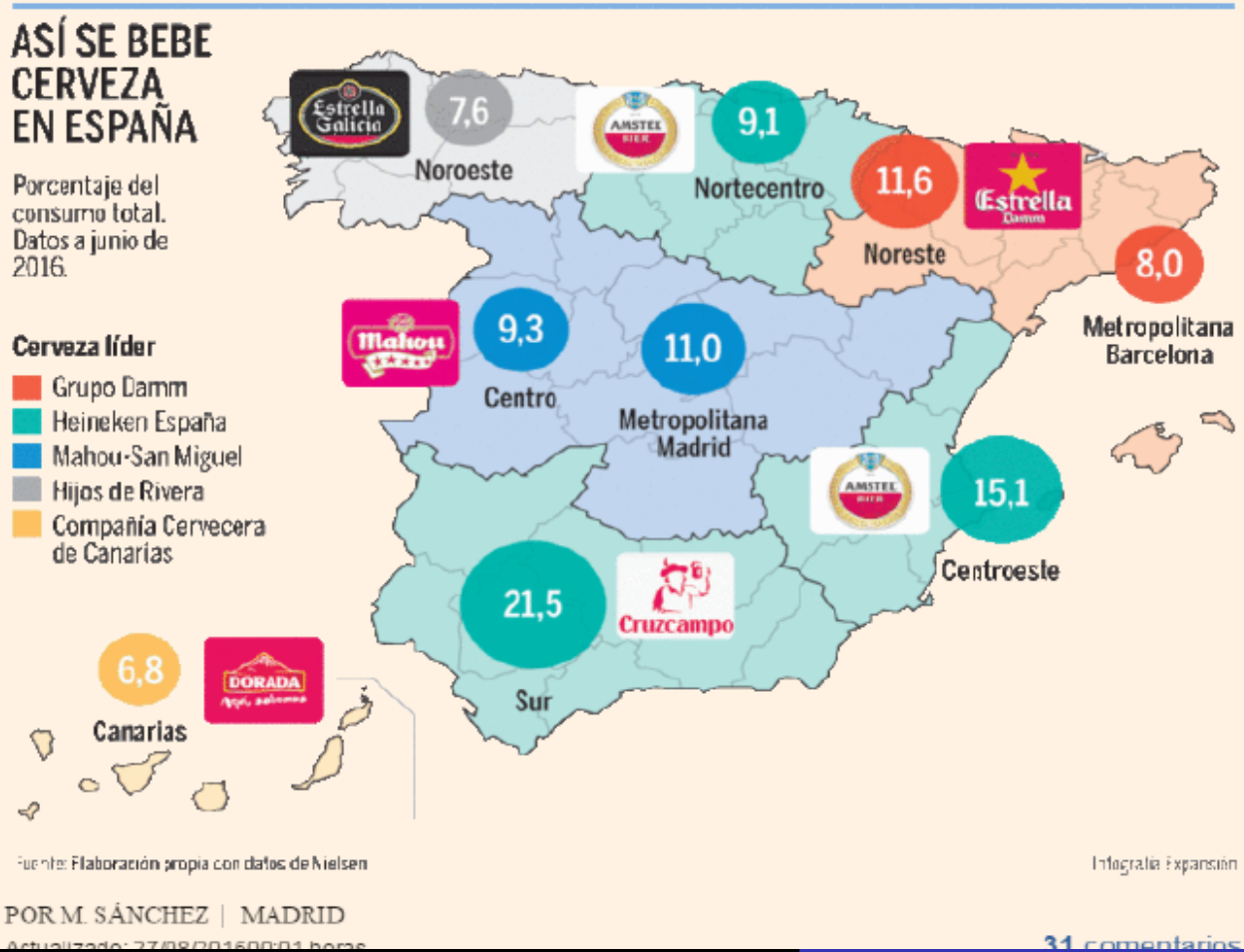
Madrid concentra el 18% de la actividad económica española, según datos del INE, que fija en 188.444 millones de euros el PIB madrileño de 2011. A pesar de la crisis, la provincia ha ganado peso en la economía española, ya que en 2008 aportaba el 17,8% de la actividad económica nacional.

Después de Madrid, Barcelona es la segunda provincia con más peso al generar el 13,6% del PIB. Valencia (5,1%), Sevilla (3,3%), Alicante (3,2%), Vizcaya (3,1%) y Málaga (2,7%) son las siguientes. Entre las siete primeras provincias de España se genera el 49% del PIB, mientras que el 51% restante se reparte ente las otras 43 provincias y las dos ciudades autónomas.

Rankings absolutos, no relativos

DISTRIBUCIÓN Y CONSUMO

El sur y el este de España, los más cerveceros



POR M. SÁNCHEZ | MADRID

31 comentarios

Conceptos previos

Población: Conjunto de elementos que son objeto de estudio.
Individuo: Cada uno de los elementos de la población descrito mediante una serie de características a las que se refiere el estudio estadístico.
Muestra: Una muestra es un subconjunto no vacío de individuos de la población. El número de elementos que componen la muestra se denomina tamaño muestral.
Caracteres o Variables: Las cualidades de los individuos de la población que son objeto de estudio. Pueden ser cualitativos (nominales u ordinales) o cuantitativos (discretos o continuos).
Modalidades: Las diferentes situaciones posibles de una variable cualitativa. Un individuo debe pertenecer a una y sólo a una modalidad.

Tipos de variables: Ejemplos

Cualitativa nominal:

País={Francia, España,...}

Color={Rojo, Verde, Amarillo, . . . }

Cualitativa ordinal: (Común en escalas jerárquicas)

{Todo, Mucho, Regular, Poco, Nada}

{Muy alto, Alto, Regular, Bajo, Muy Bajo}

Cuantitativa discreta:

Número de hijos={0, 1, . . . };

Número de símbolos en un mensaje={2, 3, . . . }

Cuantitativa continua:

Altura en cm.

Peso en Kg.

Ruido en decibelios (dB).

Frecuencias

Consideremos un conjunto de datos de tamaño N y sea X un carácter con modalidades $x_1, x_2, \ldots x_k$ (ordenadas si son cuantitativas):

Definición

La **frecuencia absoluta** *(n_i) de la modalidad x_i es el número de individuos observados que presentan esa modalidad.*

Definición

La **frecuencia relativa** *(f_i) de la modalidad x_i es el cociente entre la frecuencia absoluta y el número total de individuos*

$$f_i = \frac{n_i}{N}$$

Frecuencias-2

Ejemplo

Los precios (en euros) de los menús servidos durante un día en un restaurante determinado son: $6, 8, 6, 8, 6, 8, 12, 6, 8, 8, 6, 8, 8, 8, 12, 12, 8, 8, 12, 6, 8, 6, 6, 8, 12, 6, 6, 6, 6, 6.$

Frecuencias-2

Ejemplo

Los precios (en euros) de los menús servidos durante un día en un restaurante determinado son: $6, 8, 6, 8, 6, 8, 12, 6, 8, 8, 6, 8, 8, 8, 12, 12, 8, 8, 12, 6, 8, 6, 6, 8, 12, 6, 6, 6, 6, 6.$

Expresado en forma de tabla de frecuencias absolutas:

x_i	n_i	f_i
6	13	$f_1 = n_1 / N = 0{,}43$
8	12	$f_2 = n_2 / N = 0{,}40$
12	5	$f_3 = n_3 / N = 0{,}17$
Total	30	1,00

Figura: Tabla de frecuencias

Frecuencias Acumuladas

Definición

La **frecuencia absoluta acumulada** *(N_i) de una modalidad x_i de la variable X es la suma de las frecuencias de los valores que son inferiores o iguales a él.*

$$N_i = \sum_{j=1}^{i} n_j$$

Definición

La **frecuencia relativa acumulada** *(F_i) de una modalidad x_i de X es el cociente entre la frecuencia absoluta acumulada y el número total de individuos:*

$$F_i = \frac{N_i}{N}$$

Distribuciones de frecuencia. Tablas estadísticas

La **distribución de frecuencias** de un carácter, sea cualitativo (atributo) o sea cuantitativo (variable estadística), está constituida por las distintas modalidades del carácter junto a las correspondientes frecuencias.
Generalmente, las distribuciones se presentan en forma de tabla estadística o de frecuencias. Esta forma de representación permite tener organizada y resumida la información contenida en el conjunto de datos y presentada de forma más comprensible y significativa.

Tablas de frecuencias

En el ejemplo anterior, la distribución de frecuencias es discreta y su tabla estadística es:

x_i	n_i	f_i	N_i	F_i
6	13	0,43	13	0,43
8	12	0,40	25	0,83
12	5	0,17	30	1,00
Total	30	1,00		

En las distribuciones de variables cualitativas nominales no tendrán sentido las frecuencias acumuladas.

Tablas de frecuencias acumuladas

Cuando el número de observaciones y el número de modalidades es muy grande es comun mostrar los datos agrupados en intervalos (*clases*) y se determina el número de individuos que pertenecen a cada intervalo. Usualmente los intervalos serán de la forma $I_i = (L_{i-1}, L_i]$.

Por convenio: Los intervalos extremos de la forma $(-\infty, L_1]$, o (L_{k-1}, ∞) se consideran de igual amplitud que sus adyacentes.

Los intervalos tienen que formar una partición. Su uso supone una pérdida de información y es importante elegir un número adecuado de intervalos que no suponga una pérdida significativa.

Los puntos medios de las clases son llamados **marcas de clases**.

Ejemplo

Los precios pagados por las consumiciones realizadas en una cafetería a lo largo de un determinado día vienen dadas en la tabla

Precio ($L_{i-1} - L_i$)	Número de consumiciones (n_i)
0 – 3	40
3 – 6	30
6 – 9	10
9 – 12	5
12 – 15	5

Amplitud de un intervalo: $a_i = L_i - L_{i-1}$

Marca de clase: $x_i = \frac{L_{i-1}+L_i}{2}$

$L_{i-1} - L_i$	x_i	n_i
0 – 3	1,5	40
3 – 6	4,5	30
6 – 9	7,5	10
9 – 12	10,5	5
12 – 15	13,5	5

Representaciones gráficas

Muestran la distribución de frecuencias y deben ser capaces de transmitir información de la muestra permitiendo observar algunas características de los datos. Tratan de facilitar una síntesis visual y conviene cuidar la presentación (colores, formas, . . .). El tipo de carácter establece una clasificación de las representaciones gráficas.

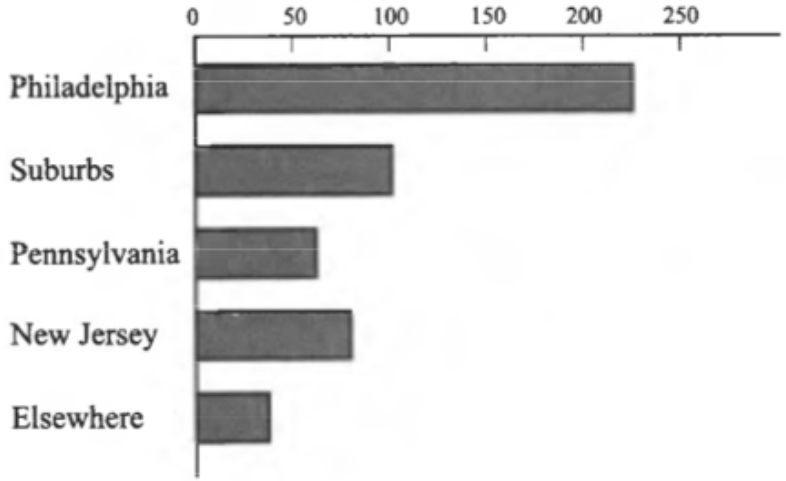
- **Caracteres cualitativos.** No hay orden numérico.
  - **Diagrama de rectángulos o barras:** Cada modalidad se representa mediante una barra cuya altura es la frecuencia absoluta o relativa.
  - **Diagrama de sectores:** Círculo con sectores de área proporcional a la frecuencia de la modalidad correspondiente.
  - **Pictograma y cartograma:** Representación icónica con dibujos simbólicos o mapas.

Ejemplo

Los estudiantes de una universidad se dividen en 5 grupos según su procedencia de acuerdo a la siguiente distribución de frecuencias:

	Philadelphia	Suburbs	PA	NJ	Elsewhere	Sum
Number of students:	225	100	60	75	40	500

Dicha información podemos representarla mediante un diagrama de barras o un diagrama de sectores:



- **Caracteres cuantitativos.** Se realizan sobre los ejes de coordenadas
  - **Diagrama de barras o puntos:** Caso discreto. Con barras verticales o puntos en los extremos; la longitud de la barra queda determinada por la frecuencia y el valor de la variable determina el lugar del eje horizontal donde se apoya.
  - **Histograma:** Datos agrupados en intervalos. En cada clase dibujamos un rectángulo sobre el eje X con base el intervalo y área proporcional a la frecuencia a representar.
  - **Polígono de frecuencias:** Se obtiene uniendo los extremos de las barras en el diagrama de barras o los puntos medios superiores de los rectángulos en el histograma. Al inicio y final se consideran 2 nuevos intervalos de igual amplitud que el primero y último.

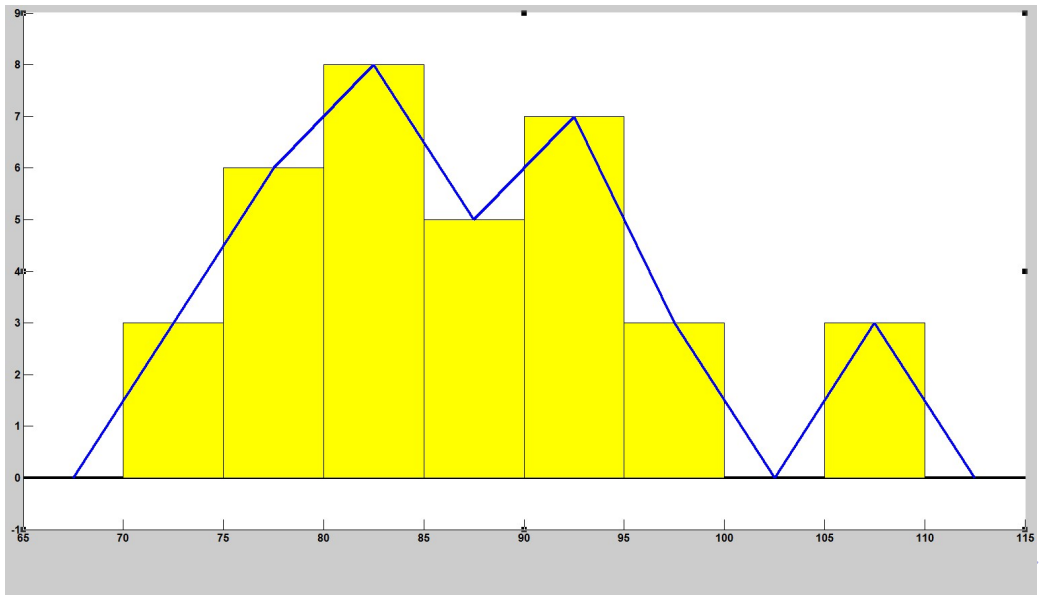
Ejemplo

Durante un periodo de 35 días las temperaturas (en grados Fahrenheit) a las 6 de la mañana han sido:

72	78	86	93	106	107	98	82	81	77	87	82
91	95	92	83	76	78	73	81	86	92	93	84
107	99	94	86	81	77	73	76	80	88	91	

Dicha información podemos representarla mediante:

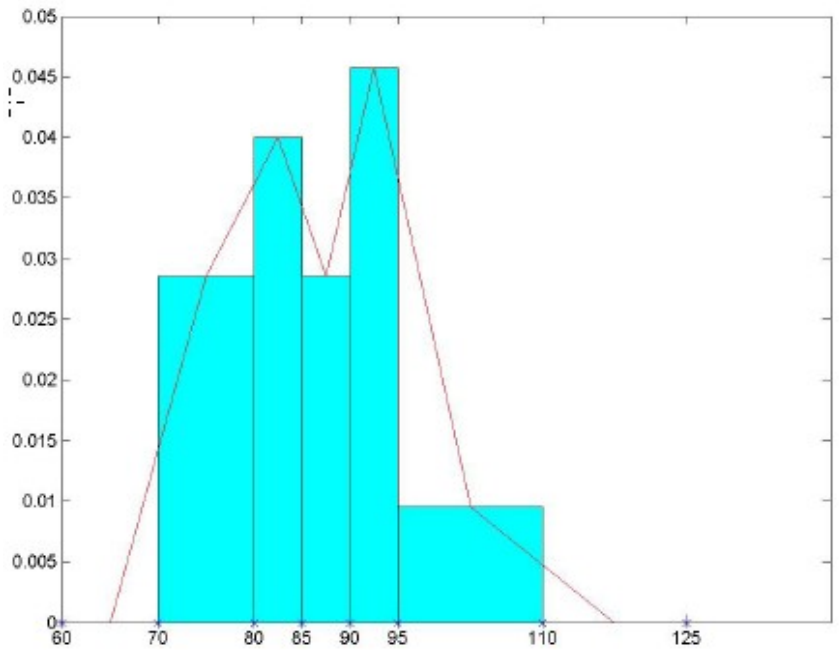
Class boundaries, °F	Class value, °F	Frequency	Cumulative frequency
70–75	72.5	3	3
75–80	77.5	6	9
80–85	82.5	8	17
85–90	87.5	5	22
90–95	92.5	7	29
95–100	97.5	3	32
100–105	102.5	0	32
105–110	107.5	3	35
	Sum	35	



Ejemplo-cont.

Pero también:

Límites de clase	Marcas de clase	Frecuencias n_i	Frecuencias $f_i = \frac{n_i}{35}$	Ampl. a_i	Altura $h_i = \frac{f_i}{a_i}$
70-80	75.0	9	0.2571	10	0.0257
80-85	82.5	8	0.2286	5	0.0457
85-90	87.5	5	0.1429	5	0.0286
90-95	92.5	7	0.2000	5	0.0400
95-110	102.5	6	0.1714	15	0.0114
	Suma	35	1		



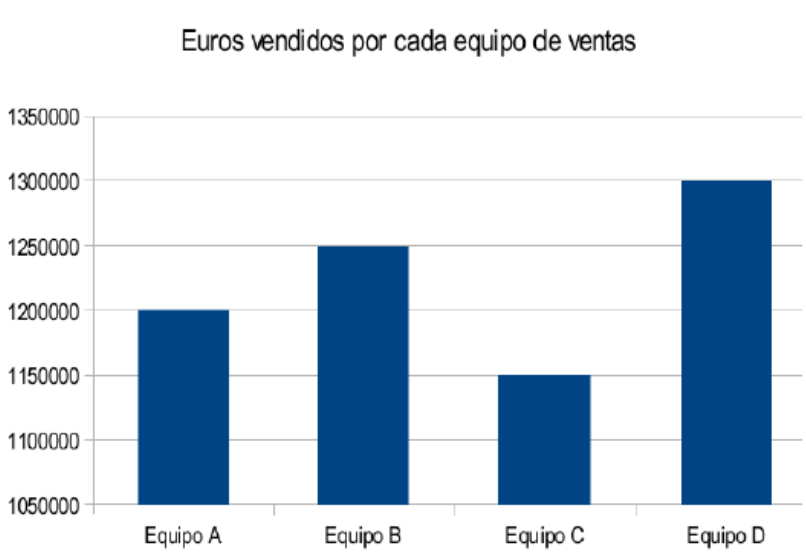
Nota sobre los histogramas

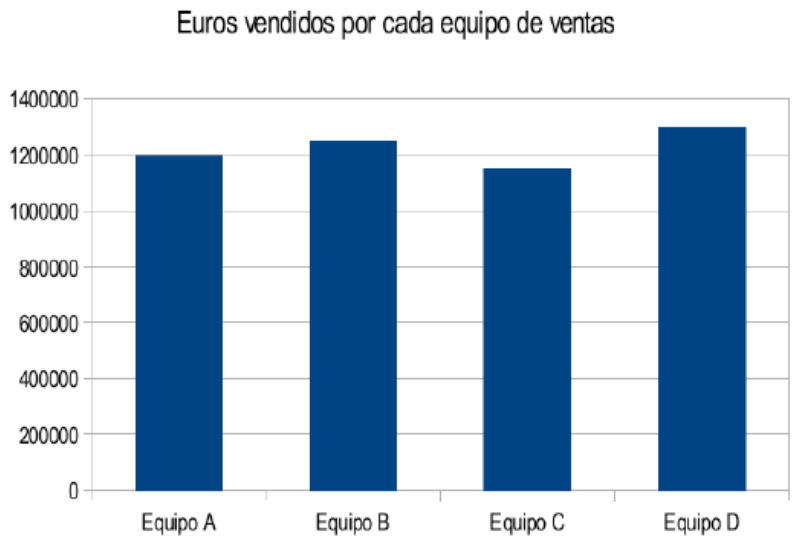
Las alturas de los histogramas se calculan para que las superficies de los rectángulos sean proporcionales a las frecuencias de las clases.
$S_i = Kn_i \Rightarrow h_i a_i = Kn_i$, es decir, $h_i = K\frac{n_i}{a_i}$ y por comodidad se usa K=1.

Pero si ponemos $K = \frac{K'}{N}$, tenemos: $h_i = \frac{K'}{N}\frac{n_i}{a_i} = K'\frac{f_i}{a_i}$ y estaríamos usando las frecuencias relativas.

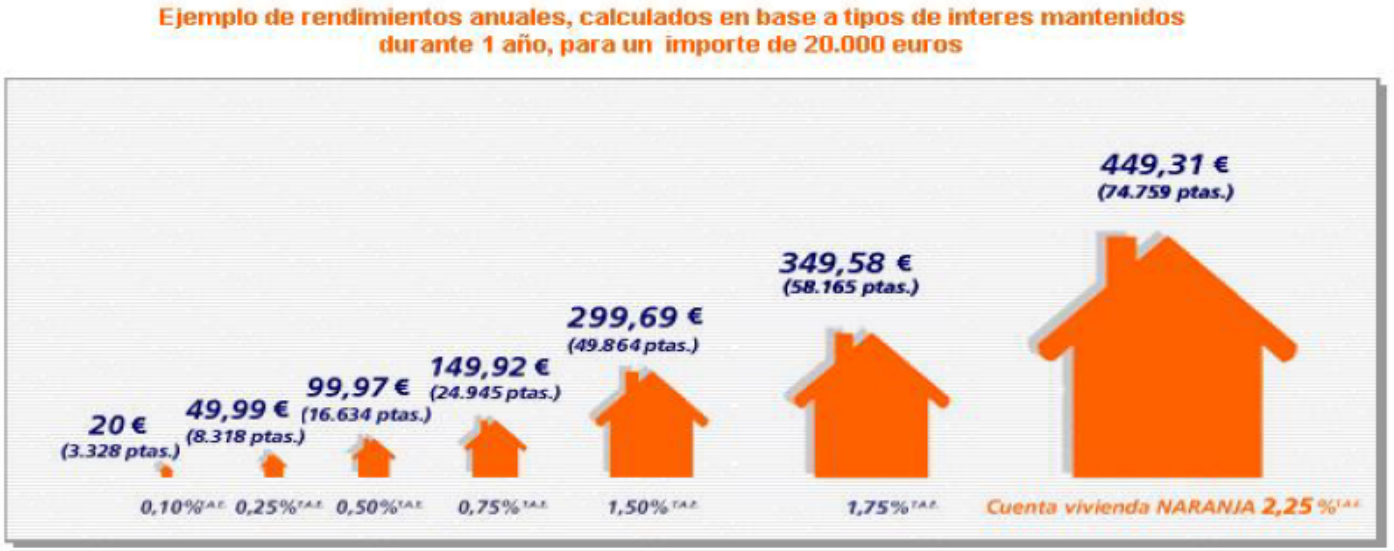
Por tanto, **es indiferente usar la proporcionalidad con las frecuencias absolutas como con las relativas.**

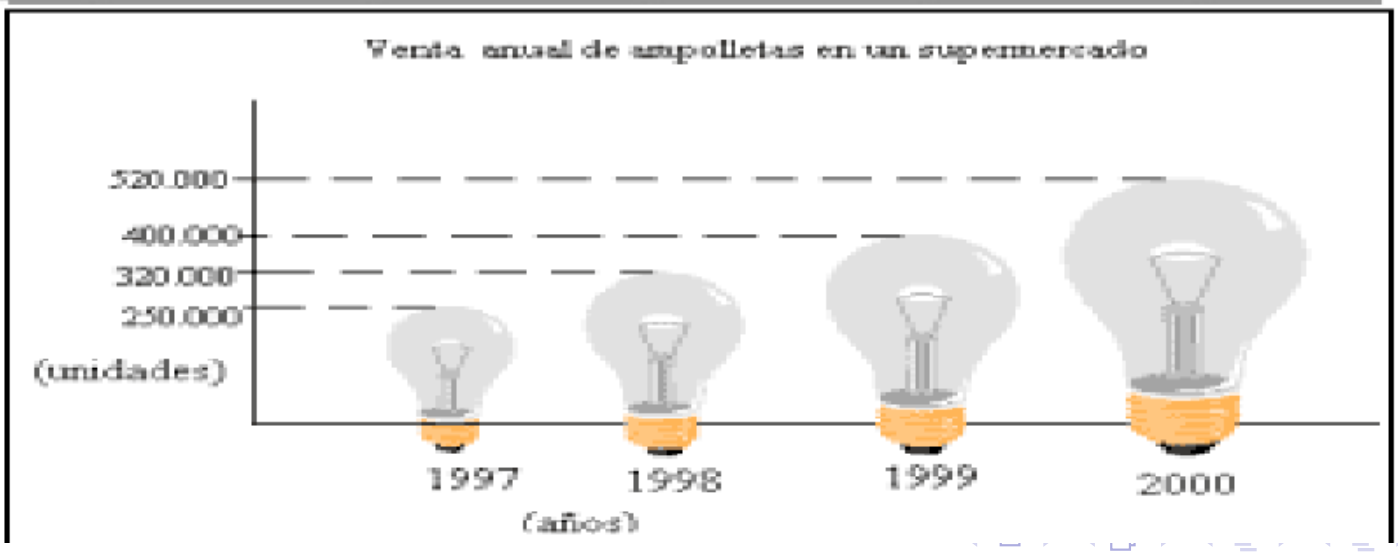
Problemas con gráficos cortados





Problemas con Ideogramas





Medidas de tendencia central: Medias, mediana y moda

Ayudan a encontrar el centro de la distribución o la posición relativa de una observación dentro de los datos.
Consideremos la variable X con valores $x_1, x_2, \ldots x_k$ y frecuencias $n_1, n_2, \ldots n_k$; siendo N el número total de datos.

Definición

La **media aritmética simple** *es la suma de todos los valores divididos por el número total de datos*

$$\mu = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \sum_{i=1}^{k} x_i f_i$$

Notación: La media aritmética de la población se denota por μ.
La media aritmética de la muestra se denota por $\bar{x}$.

Ejemplo 1

Ejemplo

Si consideramos los valores numéricos $7, 11, 11, 8, 12, 7, 6, 6$
La media aritmética simple viene dada por

$$\mu = \frac{7 + 11 + 11 + 8 + 12 + 7 + 6 + 6}{8} = 8.5$$

Cuando los valores están agrupados por intervalos, consideraremos las marcas de clase como los valores de la variable y la frecuencia absoluta al número de datos contenidos en el intervalo.

Observación

- *La media puede no corresponderse con ningún valor de la variable.*
- *La media es muy sensible a valores extremos (inusuales) de la variable*

Ejemplo: Dados los datos 5.3; 4.7; 5.2; 4.9; 49 la media aritmética es $\mu = 13.82$ que no representa nada.
Por otra parte, el último dato 49 puede ser erróneo y ser, en realidad, 4.9. La media sería entonces $\mu = 5$.
Depende demasiado de algún dato erróneo. Se dice que es poco robusta.

Outliers y Estadística robusta

- A menudo tenemos datos extraños que se separan bastante del resto. Estos datos son denominados **outliers**. Los outliers pueden aparecer por diversas razones:
  - Porque son datos erróneamente recolectados.
  - Porque corresponden a situaciones excepcionales como fraudes.
  - Porque corresponden a hechos excepcionales como las ventas de un black friday.
- Hay técnicas específicas para detectar los outliers que no estudiaremos aquí.
- La aparición de un outlier puede hacer que el valor de un estadístico cambie muchísimo su valor.
- Cuando un estadístico es poco sensible a modificar su valor ante la presencia de un outlier decimos que se trata de un **estadístico robusto**.

La Ciencia de datos y la Estadística robusta

- Hoy en día manejamos bases de datos enormes y es muy normal encontrarnos con outliers.
- A menudo es muy importante estudiar estos outliers y separarlos. Pero en cualquier caso es muy importante realizar calculos que no estén muy influenciados por los outliers.
- Existe toda una rama de la estadística llamada **Estadística robusta** dedicada al estudio de este tipo de cálculos. La estadística robusta es de gran importancia en la era de la Ciencia de Datos.
- Casi todos los cálculos que vamos a estudiar en este curso pueden ser sustituidos por cálculos robustos, pero no nos vamos a ocupar de ello ya que primero debe conocerse la estadística ordinaria y es un tema de profundización posterior.

Transformación afín

Si a partir de los valores de una variable X, construimos otra $Y = aX + b$, es decir: $y_i = a \cdot x_i + b$, entonces:

$$\mu_Y = a \cdot \mu_X + b$$

Ejemplo: Hallar la media de la variable X en la tabla:

x_i	n_i
13.725	7
13.975	14
14.225	18
14.725	6

Podemos facilitar los cálculos haciendo el cambio:
$X = 0.25Y + 13.975 \Leftrightarrow Y = \frac{X-13.975}{0.25} \Rightarrow \mu_X = 0.25\mu_Y + 13.975$

Solución:

x_i	n_i	$y_i = \frac{x_i - 13.975}{0.25}$	$n_i y_i$
13.725	7	-1	-7
13.975	14	0	0
14.225	18	1	18
14.725	6	3	18
	45		29

$$\mu_Y = \frac{29}{45} = 0.64444444 \Rightarrow$$

$$\mu_X = 0.25\frac{29}{45} + 13.975 \approx 14.1361111$$

Media ponderada

Definición

La **media ponderada** *de los datos* x_i *por los pesos* w_i *se define como:*

$$MP = \frac{\sum_i x_i w_i}{\sum_i w_i}$$

Cuando la variable es continua y los valores están agrupados por intervalos, consideraremos las marcas de clase como los valores de la variable y la frecuencia absoluta al número de datos contenidos en el intervalo.

Ejemplo

Ejemplo

Las calificaciones de un alumno son 2.6; 3.7; 5.1, 4.9 y 6.4. Las 3 primeras corresponden a controles con ponderación 1, la cuarta es la nota de prácticas con ponderación 2 y la última es el examen final con ponderación 3. ¿Cuál es la nota media?

Las ponderaciones son $\bar{w} = \{1, 1, 1, 2, 3\}$

$$MP = \frac{2.6 \cdot 1 + 3.7 \cdot 1 + 5.1 \cdot 1 + 4.9 \cdot 2 + 6.4 \cdot 3}{1 + 1 + 1 + 2 + 3} = 5.05$$

Media cuadrática o Valor cuadrático medio (RMS)

Definición

La **media cuadrática** *de los datos x_i se obtiene mediante la expresión:*

$$RMS = \sqrt{\frac{\sum_i x_i^2}{N}}, \text{ o bien para datos agrupados: } RMS = \sqrt{\frac{\sum_i n_i x_i^2}{N}}$$

Ejemplo

Al contabilizar durante una semana el número de llamadas recibidas en un servicio técnico debido a algún tipo de avería, se obtuvieron los valores: 2, 3, 1, 0, 4, 3. Hallar la media cuadrática.

$$RMS = \sqrt{\frac{2^2 + 3^2 + 1^2 + 0^2 + 4^2 + 3^2}{6}} = \sqrt{\frac{39}{6}} = \sqrt{6.5} \approx 2.54951$$

Moda

Definición

La moda (Mo) de un conjunto de datos es el valor de la variable que presenta mayor frecuencia. Puede no ser única o puede que no exista si todos los valores tienen la misma frecuencia

Si la variable es continua, debemos tener en cuenta la amplitud de los intervalos y llamaremos **intervalo modal** al que tenga mayor $h_i = \frac{n_i}{a_i}$ (mayor altura en el histograma).

Ejemplo

Si consideramos el conjunto de datos $A = \{7, 11, 11, 8, 12, 9, 6, 6\}$, *tenemos dos modas que corresponden a los valores* 6 *y* 11
En $B = \{7, 11, 11, 8, 12, 12, 12, 7, 6, 6\}$ *la moda es el* 12
En $C = \{7, 11, 8, 12, 9, 6\}$ *no hay moda.*

Mediana

La mediana es una medida central mucho más robusta que la media, aunque debe tenerse en cuenta que su cálculo es más lento y no fácilmente paralelizable.

Definición

La **mediana** *(Me) es aquel valor que divide a la población en dos partes de igual tamaño. Si N es impar la mediana coincidirá con un término de la población, si N es par, se toman los dos valores centrales y se calcula su media.*

Ejemplo

Consideremos las listas de números ordenados $List_A = \{11, 11, 16, 17, 25\}$ *y* $List_B = \{1, 4, 8, 8, 10, 16, 16, 19\}$*; la mediana de la primera lista es* 16 *y la de la segunda lista es* $\frac{8+10}{2} = 9$.

Cuantiles:

Constituyen una generalización del concepto de mediana.

Definición

Cuantiles *Dado un valor* $c \in (0,1)$ *se define el* **cuantil c** *como el valor* $X(c)$ *que divide a la variable dejando una proporción* c *menor y una proporción* $1-c$ *mayor que él.*

Evidentemente la mediana coincide con el cuantil $c = 0.5$.

Cuartiles, deciles y percentiles

Definición

Cuartiles. *Son tres valores con las siguientes características:*
$Q_1 = X(0.25)$: *Valor que deja por debajo* $1/4$ *de la población.*
$Q_2 = X(0.5) = Me$: *Deja por debajo la mitad de la población.*
$Q_3 = X(0.75)$: *Deja por debajo* $3/4$ *de la población.*

Definición

Deciles *Hay* 9 *deciles que dividen a la población en* 10 *partes iguales.* $D_k = X(\frac{k}{10})$.

Definición

Percentiles *Hay* 99 *percentiles que dividen a la población en* 100 *partes iguales. Se denotan por* $P_k = X(\frac{k}{100})$ *que será el valor que divide a la población dejando por debajo el* $k\,\%$ *de los valores y por encima el* $(100 - k)\,\%$.

Cálculo del cuantil c

- **Caso discreto:** Realizamos la descomposición de cN en su parte entera (E) y decimal (D): $\mathrm{cN} = \mathrm{E} + \mathrm{D}$
  - Si $D \neq 0$, $X(c)$ es el valor que ocupa el lugar $(E+1)$
  - Si $D = 0$, $X(c) = \frac{\text{valor de lugar (E)+valor de lugar (E+1)}}{2}$
- **Caso continuo:** Cálculo cN. En la columna de las frecuencias acumuladas N_i busco la primera que rebasa ese valor $N_{i-1} \leq cN < N_i$, a continuación aplico:

$$\mathrm{X(c)} = \mathrm{L_{i-1}} + \frac{\mathrm{cN} - \mathrm{N_{i-1}}}{\mathrm{n_i}}\mathrm{a_i}$$

donde:

- L_{i-1} : Límite inferior del intervalo.
- N_{i-1} : Frecuencia absoluta acumulada correspondiente al intervalo anterior.
- a_i : Amplitud del intervalo.
- n_i : Frecuencia absoluta del intervalo.

Ejemplo

Ejemplo

Calcular los cuartiles y los percentiles: P_{37} *y* P_{68} *para los siguientes valores numéricos:* $2, 5, 3, 4, 7, 0, 11, 2, 3, 8$

En primer lugar, ordenamos los $N = 10$ valores en orden creciente:
$0, 2, 2, 3, 3, 4, 5, 7, 8, 11$
Q_1 : Calculamos N/4=2.5. Como no es entero se trata del que ocupa el tercer lugar, luego $Q_1 = 2$.
Mediana = Q_2 : Calculo 10/2=5 que es entero, luego es la media entre el 5º y 6º valor: $Me = \frac{3+4}{2} = 3.5$
Q_3 : Nc=10*3/4=7.5, no es entero luego tomo el 8º valor: $Q_3 = 7$.
P_{37} : Calculo 10(37/100)=3.7, no es entero, tomo el 4º
$\Rightarrow P_{37} = 3$.
P_{68} : Calculo 10(68/100)=6.8, no es entero, tomo el 7º
$\Rightarrow P_{68} = 5$.

Ejercicio

Consideremos los siguientes 50 *valores ordenados:*

10	20	35	44	55	64	75	81	87	99
11	22	36	48	56	68	76	82	89	101
13	23	38	49	57	69	76	83	90	102
15	23	41	50	60	70	78	83	94	105
18	30	44	50	63	73	80	85	96	107

Calcular P_5, P_{95}, Q_1, Me *y* Q_3.

$\mathrm{P}_5 : v = 50 * \frac{5}{100} = 2.5 \Rightarrow$ busco el $3^{\underline{o}}$, por lo que $P_5 = 13$
$\mathrm{P}_{95} : v = 50 * \frac{95}{100} = 47.5 \Rightarrow$ busco el $48^{\underline{o}}$, por lo que $P_{95} = 102$
$\mathrm{Q}_1 : v = 50 * \frac{1}{4} = 12.5 \Rightarrow$ busco el $13^{\underline{o}}$, por lo que $Q_1 = 38$
$\mathrm{Me} : v = 50 * \frac{1}{2} = 25 \Rightarrow$ saco la media entre el $25^{\underline{o}}$ y $26^{\underline{o}}$, por lo que $Me = \frac{63+64}{2} = 63.5$
$\mathrm{Q}_3 : v = 50 * \frac{3}{4} = 37.5 \Rightarrow$ busco el $38^{\underline{o}}$, por lo que $Q_3 = 83$

Ejercicio

Ejercicio

La duración en horas de un tipo de lámpara incandescente viene reflejado en la tabla adjunta. Calcular:

a) Dibujar el histograma.

b) Porcentaje que duran menos de 950 h.

c) Q_1, Q_3, $D_1 = P_{10}$ *y* $D_9 = P_{90}$.

d) Media, moda y mediana.

Int	$(200-600]$	$(600-800]$	$(800-1000]$	$(1000-1200]$	$(1200-1400]$	$(1400-1700]$
n_i	4	31	136	165	67	14

Medidas de desviación y dispersión

Ayudan a determinar la variación de los datos. Sirven para determinar lo agrupada o dispersa que está una población y si la medida de tendencia central calculada es representativa.

Rango: Recorrido o intervalo (R) es la diferencia entre el mayor y el menor valor observado de la variable.

Otros rangos usados son:

Rango intercuartílico: $R_Q = Q_3 - Q_1$

Rango intercentílico: $R_P = P_{99} - P_1$

El rango es muy sensible a un error en los datos, no así los rangos intercuartílico e intercentílico.

Desviación media

La desviación d_i de un valor x_i de la variable respecto a un parámetro p es la diferencia $d_i = |x_i - p|$ entre esos valores. Usualmente p es una medida de tendencia central.
La **desviación media respecto a un promedio** p es la media del valor absoluto de las desviaciones a una determinada medida de tendencia central p.

$$DM(p) = \frac{\sum_{i=1}^{k} |x_i - p| \cdot n_i}{N} = \sum_{i=1}^{k} |x_i - p| \cdot f_i$$

Si el parámetro p es la media aritmética simple lo llamamos **desviación media:**

$$DM = \frac{\sum_{i=1}^{k} |x_i - \mu| \cdot n_i}{N} = \sum_{i=1}^{k} |x_i - \mu| \cdot f_i$$

Tiene el inconveniente de usar valores absolutos (no derivable).

Error cuadrático medio

Definición

Llamamos **error cuadrático medio** *a la media de las desviaciones al cuadrado:* $$ECM(p) = \frac{\sum_i n_i (x_i - p)^2}{N}$$

Ejemplo: Dados los valores $\{5, 2, 3, 3, 3, 5, 7\}$ hallar la desviación media y error cuadrático medio respecto a la media y la mediana.
Respecto a la media: $\mu = \frac{5+2+3+3+3+5+7}{7} = 4$, las desviaciones absolutas son: $|\vec{d_i}| = \{1, 2, 1, 1, 1, 1, 3\}$, luego
$DM = \frac{5(1)+1(2)+1(3)}{7} = \frac{10}{7}$ y $ECM = \frac{5(1)^2+1(2)^2+1(3)^2}{7} = \frac{18}{7}$.
Respecto a la mediana: $Me = 3$, $|\vec{d_i}| = \{2, 1, 0, 0, 0, 2, 4\}$, luego
$DM = \frac{3(0)+1(1)+2(2)+1(4)}{7} = \frac{9}{7}$, $ECM = \frac{3(0)^2+1(1)^2+2(2)^2+1(4)^2}{7} = \frac{25}{7}$
NOTA: La mediana es el valor que hace mínimo la desviación media, mientras la media hace mínimo el error cuadrático medio.

Ejemplo

Ejemplo

Los valores previstos (x_i), reales (x_i^*) y frecuencia absoluta (n_i) vienen dados en la tabla. Hallar la media cuadrática, la desviación media y la desviación cuadrática media del error cometido en la estimación.

x_i	0	0	0	1	1	1	3	3	3
x_i^*	0	1	3	0	1	3	0	1	3
n_i	6	3	2	3	5	1	1	3	7

x_i	x_i^*	n_i	$d_i = x_i - x_i^*$	$\lvert d_i \rvert$	d_i^2	$n_i \lvert d_i \rvert$	$n_i d_i^2$
0	0	6	0	0	0	0	0
0	1	3	-1	1	1	3	3
0	3	2	-3	3	9	6	18
1	0	3	1	1	1	3	3
1	1	5	0	0	0	0	0
1	3	1	-2	2	4	2	4
3	0	1	3	3	9	3	9
3	1	3	2	2	4	6	12
3	3	7	0	0	0	0	0
		31				23	49

$\Rightarrow$

Media cuadrática$=\sqrt{\frac{49}{31}} =$
$= 1.25723711$

Desviación media$=$
Error medio $=\frac{23}{31} =$
$= 0.74193548$

Desviación cuadrática media$=$
$=$Error cuadrático medio$=$
$= \frac{49}{31} = 1.58064516$

La varianza y la desviación típica

La **varianza** de un conjunto de datos de tamaño N viene dada por:

$$V = \sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \mu)^2 \cdot n_i}{N} = \sum_{i=1}^{k}(x_i - \mu)^2 \cdot f_i$$

Es decir, es la media de los cuadrados de las desviaciones respecto a la media.
Otra forma equivalente para calcular la varianza es:

$$V = \sigma^2 = \sum_{i=1}^{k} x_i^2 \cdot f_i - \mu^2 = \frac{\sum_{i}^{k} n_i x_i^2}{N} - \mu^2$$

La **desviación típica o estándar** es la raíz cuadrada de la varianza.

$$\sigma = +\sqrt{V} = \sqrt{\sum_{i=1}^{k}(x_i - \mu)^2 \cdot f_i}$$

Notación: La varianza de la población se denota por σ^2. La desviación típica de la población se denota por σ. La varianza muestral se denota por s_n^2 siendo n el tamaño de la muestra. La desviación típica muestral se denota por s_n.

Ejemplo

Consideremos las listas de valores numéricos:

$$Lista_A = \{12, 10, 9, 9, 10\} \text{ y } Lista_B : \{5, 10, 16, 15, 4\}.$$

Calcular la desviación típica e interpretar los resultados.

Observamos que en ambos casos la media es 10
Para la lista A:

$$V_A = \sigma_A^2 = \frac{12^2 + 2 \cdot 10^2 + 2 \cdot 9^2}{5} - 10^2 = 1.2 \quad \sigma_A = \sqrt{1.2}$$

Para B:

$$V_B = \sigma_B^2 = \frac{5^2 + 10^2 + 16^2 + 15^2 + 4^2}{5} - 10^2 = 24.4 \quad \sigma_B = \sqrt{24.4}$$

Luego los datos están más dispersos en la lista B que en la A.

Media y varianza muestral

Normalmente no podemos medir toda la población y tenemos que conformarnos con una muestra, sin embargo queremos inferir el valor para la población completa.

- El mejor estimador para la media poblacional μ es la media muestral $\bar{x} = \frac{\sum_i x_i n_i}{n}$.
- El mejor estimador de la varianza de una población no es la varianza de la muestra, es la cuasivarianza de la muestra:

La **cuasi-varianza muestral (s^2)**, para una muestra de tamaño n vale: $s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i}{n-1}$

No confundir "cuasi-varianza muestral" (s^2), con "varianza de la muestra" (s_n^2):

$$s^2 = \frac{n}{n-1} \cdot s_n^2 \quad \text{donde} \quad s_n^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i}{n}$$

Medidas de comparación

Se usan para comparar información obtenidas de distintas muestras o distintas poblaciones.

Variable tipificada:

Haciendo uso de la media y de la desviación típica de la variable X, podemos considerar una nueva variable dada por:

$$Z = \frac{X - \mu}{\sigma} \quad \text{con valores} \quad z_i = \frac{x_i - \mu}{\sigma} \quad i = 1, 2, ..., k$$

La variable tipificada es adimensional y, por tanto, independiente de las unidades usadas. Mide la desviación de la variable respecto de su media en términos de la desviación típica.

Ejemplo

El alumno A ha obtenido una puntuación de 8.5 *en un examen cuya puntuación media ha sido* 7.9 *y desviación típica* 0.8*. El alumno B ha obtenido como puntuación* 7.4 *en otro examen cuya puntuación media ha sido* 7.0 *y desviación típica* 0.5*. Compara las puntuaciones de ambos alumnos.*

Para proceder a la comparación tipificamos las variables:

$$z_A = \frac{8.5 - 7.9}{0.8} = 0.75 \qquad z_B = \frac{7.4 - 7.0}{0.5} = 0.8$$

Observamos que la nota del alumno B es mejor que la del A. La nota de A se encuentra a 0.75 desviaciones típicas por encima de la nota media, siendo inferior a la nota de B que supera a la nota media en 0.8 desviaciones.

Coeficiente de variación de Pearson

Un problema de la desviación típica como medida de dispersión es que depende de las unidades de la variable y de la muestra. Por tanto no resulta útil para comparar dispersiones entre dos muestras distintas o expresadas con unidades distintas.
Por ello se define el **coeficiente de variación de Pearson** , como el cociente entre la desviación típica y el valor absoluto de la media:

$$CV = \frac{\sigma}{|\mu|}$$

Normalmente se expresa en tanto por ciento, para ello basta multiplicar el cociente por 100.
Tiene el problema de no estar definido cuando $\mu = 0$.

Ejemplo

Un fabricante de tubos de televisión produce dos tipos de tubos, A y B, que tienen vidas medias respectivas μ_A=1495 horas y μ_B=1875 horas, y desviación típica σ_A=280 horas y σ_B=310. Comparar las dispersiones de las dos poblaciones.

Los coeficientes de variación para cada tipo de tubos

$$CV_A = \frac{280}{1495} \cdot 100 \approx 18'73\,\% \qquad CV_B = \frac{310}{1875} \cdot 100 \approx 16'53\,\%$$

indican que, en términos relativos, la dispersión es mayor en la población A, a pesar de que las desviaciones típicas sugieran lo contrario.

Momentos respecto a un punto

Definición

Se define el **momento de orden** *r* **respecto al punto** *c como:*

$$m_r(c) = \sum_{i=1}^{k}(x_i - c)^r f_i = \frac{\sum_i^k n_i(x_i - c)^r}{N}$$

Ejemplo: Hallar $m_1(Me)$ y $m_2(Me)$ de los datos:
$\{0, 0, 2, 2, 2, 2, 3, 3, 4, 4, 7, 7\}$.
Hay $N = 12$ datos el $6^{\underline{o}}$ vale 2 y el $7^{\underline{o}}$ vale 3 $\Rightarrow Me = \frac{2+3}{2} = 2.5$

$$m_1(Me) = \frac{2(0-2.5)+4(2-2.5)+2(3-2.5)+2(4-2.5)+2(7-2.5)}{12} = \frac{6}{12}$$

$$m_2(Me) = \frac{2(0-2.5)^2+4(2-2.5)^2+2(3-2.5)^2+2(4-2.5)^2+2(7-2.5)^2}{12} = \frac{59}{12}$$

Momentos ordinarios

Definición

Se define el **momento ordinario de orden** *r como la media aritmética de las potencias de orden r de los datos de la variable:*

$$m_r = \sum_{i=1}^{k} x_i^r f_i = \frac{\sum_i^k n_i x_i^r}{N}$$

Se verifica que:

- El momento ordinario de orden 0 vale 1, $m_0 = 1$.
- El momento ordinario de orden 1 es la media aritmética:

$$\mathrm{m}_1 = \mu$$

- El momento ordinario de orden 2 , $m_2 = \sigma^2 + \mu^2$.

Uso informático de los momentos ordinarios

- La mayoría de estadísticos y cálculos básicos que hacemos en una distribución pueden ser calculados a partir de los momentos ordinarios.
- Los momentos ordinarios multiplicados por N pueden ser calculados de manera paralela y/o acumulativa.
- Estos hechos son de una gran importancia, porque nos permiten de una manera más o menos inmediata poder paralelizar cálculos estadísticos y poder añadir datos de manera incremental sin tener que recorrer los datos anteriores de nuevo.
- En el análisis de Big Data estas técnicas son fundamentales.

Momentos centrales

Definición

Se define el **momento central de orden** *r como la media aritmética de las potencias de orden r de las desviaciones de los datos respecto de la media:*

$$\mu_r = \sum_{i=1}^{k} (x_i - \mu)^r f_i = \frac{\sum_{i=1}^{k} n_i (x_i - \mu)^r}{N}$$

Propiedades:

- Los momentos centrales $\mu_0 = 1$ y $\mu_1 = 0$.
- El momento central de orden 2 es la varianza:

$$\mu_2 = \mathrm{V} = \sigma^2 = \mathrm{m}_2 - \mu^2$$

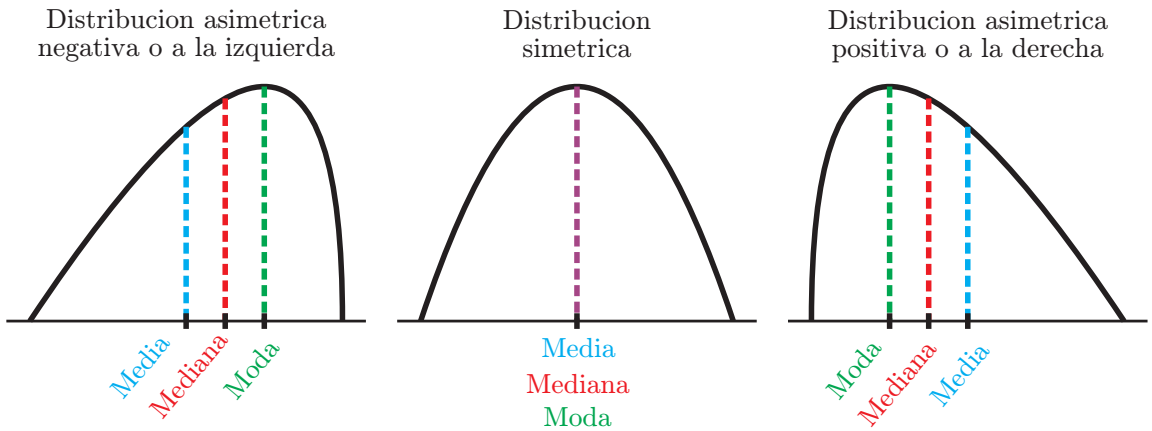
- $\mu_3 = \mathrm{m}_3 - 3\mathrm{m}_2\mu + 2\mu^3$
- $\mu_4 = \mathrm{m}_4 - 4\mathrm{m}_3\mu + 6\mathrm{m}_2\mu^2 - 3\mu^4$

Medidas de forma: simetría y apuntamiento

Otras medidas que nos permiten clasificar la forma de una distribución son las medidas de asimetría (o sesgo) y las medidas de apuntamiento (o curtosis).

Medidas de asimetría Una distribución de frecuencias es simétrica cuando los valores de la variable que equidistan de un valor central tienen las mismas frecuencias.

Las distribuciones simétricas verifican: $\mu = Me$, y usualmente $\mu = Me = Mo$.



Sesgo negativo | Simétrica | Sesgo positivo

Coeficientes de asimetría

Hay dos coeficientes que nos permiten estudiar el grado de asimetría de una distribución.
Definimos el **coeficiente de asimetría de Pearson** como:

$$A_P = \frac{\mu - Mo}{\sigma}$$

Siendo su interpretación:

- $A_P > 0$ Asimetría a la derecha o positiva
- $A_P = 0$ Simetría
- $A_P < 0$ Asimetría a la izquierda o negativa

Coeficiente de asimetría de Fisher

Definimos el **coeficiente de asimetría de Fisher** como:

$$g_1 = \frac{\mu_3}{\sigma^3}$$

Siendo su interpretación:

- $g_1 > 0$ Asimétrica (o sesgada) a la derecha o positiva
- $g_1 = 0$ Simétrica o insesgada.
- $g_1 < 0$ Asimétrica (o sesgada) a la izquierda o negativa

Lo que se hace es comparar con la distribución normal que es simétrica y tiene $g_1 = 0$

Coeficiente de apuntamiento

El aplastamiento, apuntamiento o curtosis de una distribución es el grado de achatamiento o afilamiento en comparación con la distribución normal con igual media y varianza.
El **coeficiente de aplastamiento de Fisher** es:

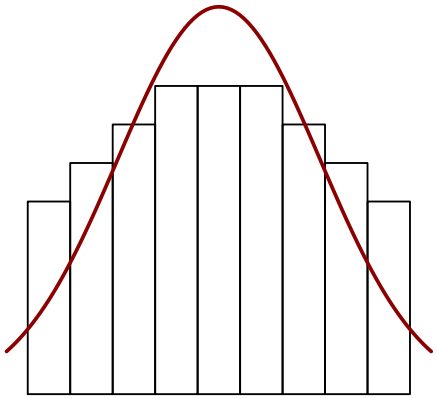
$$g_2 = \frac{\mu_4}{\sigma^4} - 3$$

Siendo su interpretación:

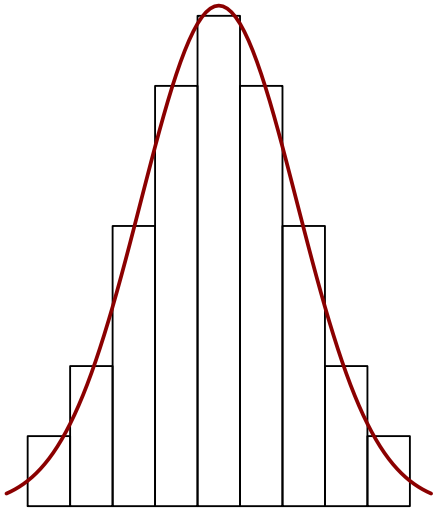
- $g_2 < 0$ Menos apuntamiento que la normal (Platicúrtica).
- $g_2 = 0$ Igual apuntamiento que la normal (Mesocúrtica).
- $g_2 > 0$ Más apuntamiento que la normal (Leptocúrtica).

Lo que se hace es comparar con la distribución normal que tiene $g_2 = 0$

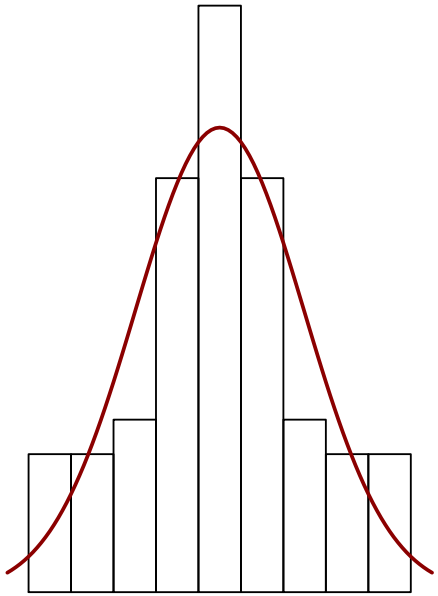
Significado de la curtosis



Platicúrtica



Mesocúrtica



Leptocúrtica

La curva superpuesta al histograma es una normal con igual media y varianza.

Transparencias basadas en la siguiente bibliografía:

- Lipschutz, S; Schiller, J.J. Schaum's **Outline of Theory and Problems of Introduction to Probability and Statistics**. McGraw-Hill, 1998.
- **Apuntes de Estadística** elaborados por el profesor Sixto Sánchez Merino del Dpto. de Matemática Aplicada de la Universidad de Málaga
- V. Quesada, A. Isidoro, L. A. López. **Curso y ejercicios de Estadística**. Ed. Alhambra Universidad.