Tema 1: Estadística Descriptiva

Descripción de una variable

Departamento Matemática Aplicada

Universidad de Málaga

Curso 2015-2016

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**.

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 (N) y si coincide con el tamaño de la población decimos que se está realizando un **censo**.
Caracteres: Los caracteres son 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 del carácter se denominan modalidades. Tales situaciones deben estar bien definidas de tal manera que cada individuo "pertenezca" a una y sólo una modalidad.

Tipos de caracteres: 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 una población y una muestra 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*

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

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.*

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

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:*

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

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.
Las distribuciones de frecuencias son básicamente de dos tipos: **discretas y continuas**.

Distribuciones discretas

Consideramos que la distribución de frecuencias es **discreta** si el carácter es cualitativo o si el carácter es cuantitativo, pero el número de modalidades es "pequeño" en relación con el número de observaciones.
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		

Las distribuciones de caracteres cualitativos se analizan como discretas, aunque para las nominales no tendrán sentido las frecuencias acumuladas.

Distribuciones continuas.

Consideramos que la distribución de frecuencias es **continua** cuando el número de observaciones y el número de modalidades es muy grande.

Los datos se agrupan 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**.

La tabla estadística se obtiene a partir de los intervalos, sus marcas de clase y las frecuencias correspondientes.

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.

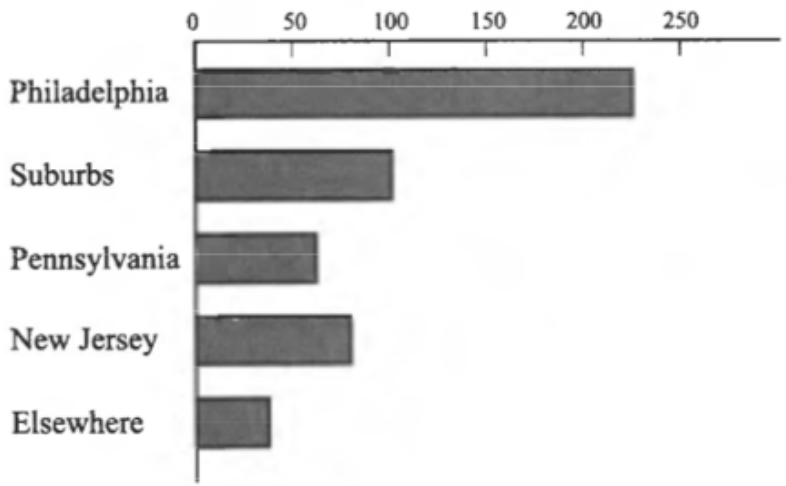
- **Caracteres cualitativos.** No hay orden numérico.
  - **Diagrama de rectángulos o barras:** Cada modalidad se representa mediante una barra cuya altura es la frecuencia absoluta o relativa.
  - **Diagrama de sectores:** Círculo con sectores de área proporcional a la frecuencia de la modalidad correspondiente.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

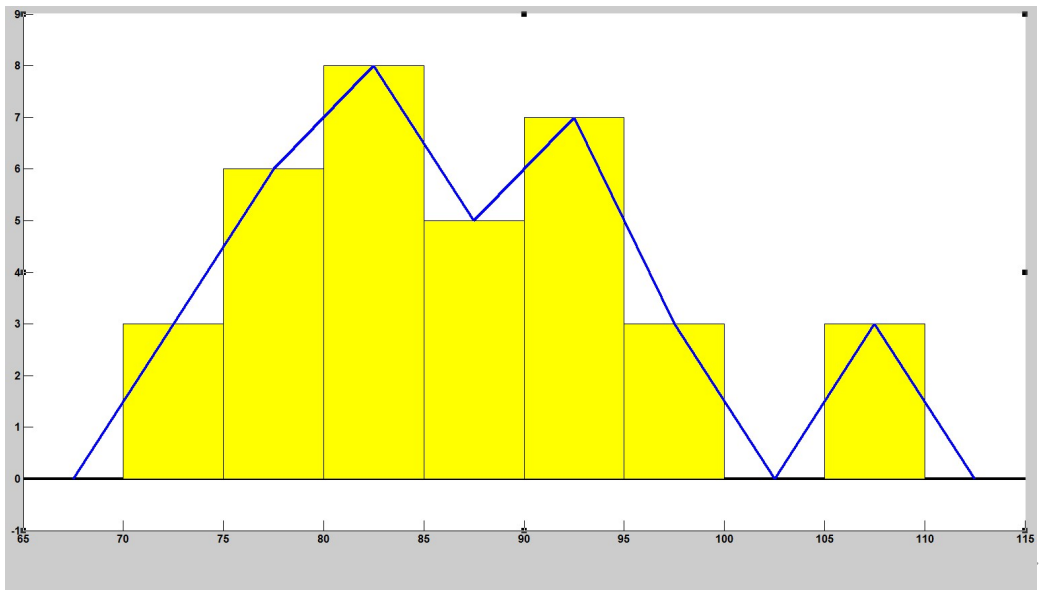
Ejemplo

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:

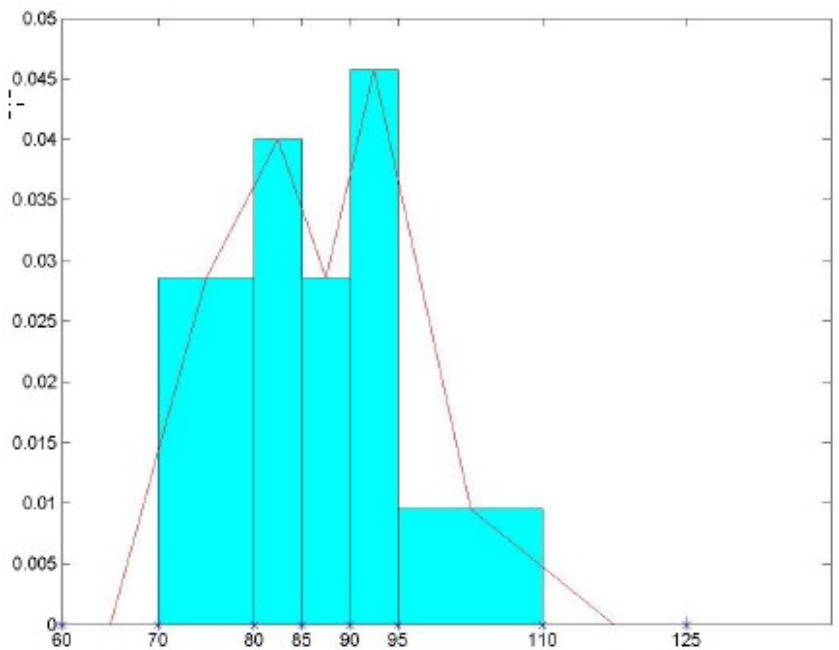
Class boundaries, °F	Class value, °F	Frequency	Cumulative frequency
70–75	72.5	3	3
75–80	77.5	6	9
80–85	82.5	8	17
85–90	87.5	5	22
90–95	92.5	7	29
95–100	97.5	3	32
100–105	102.5	0	32
105–110	107.5	3	35
	Sum	35	



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		



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*

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

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

Ejemplo 1

Ejemplo

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

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

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

Calcular la media aritmética en el ejemplo de las temperaturas.

Solución

Intervalo $(L_{i-1} - L_i]$	*Frecuencia* absoluta (n_i)	Marca de clase (x_i)	$n_i x_i$
$70 - 75$	3	72.5	217.5
$75 - 80$	6	77.5	465.0
$80 - 85$	8	82.5	660.0
$85 - 90$	5	87.5	437.5
$90 - 95$	7	92.5	647.5
$95 - 100$	3	97.5	292.5
$100 - 105$	0	102.5	0.0
$105 - 110$	3	107.5	322.5
	35		**3042.5**

La media aritmética simple vale:

$$\bar{x} = \frac{3042.5}{35} = \mathbf{86.9285714}$$

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 $\bar{x} = 13.82$ que no representa nada.
Por otra parte, el último dato 49 puede ser erróneo y ser, en realidad, 4.9. La media sería entonces $\bar{x} = 5$.
Depende demasiado de algún dato erróneo. Se dice que es poco robusta.

Transformación afín

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

$$\bar{y} = a\bar{x} + b$$

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:
$X = 0.25Y + 13.975 \Leftrightarrow Y = \frac{X-13.975}{0.25} \Rightarrow \bar{x} = 0.25\bar{y} + 13.975$

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

$$\bar{y} = \frac{29}{45} = 0.64444444 \Rightarrow$$

$$\bar{\mathbf{x}} = \mathbf{0.25}\frac{\mathbf{29}}{\mathbf{45}} + \mathbf{13.975} \approx \mathbf{14.1361111}$$

Media ponderada

Definición

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

$$\bar{x}_w = \frac{\sum_i x_i w_i}{\sum_i w_i}$$

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 $\bar{w} = \{1, 1, 1, 2, 3\}$

$$\bar{x}_w = \frac{2.6 \cdot 1 + 3.7 \cdot 1 + 5.1 \cdot 1 + 4.9 \cdot 2 + 6.4 \cdot 3}{1 + 1 + 1 + 2 + 3} = \mathbf{5.05}$$

Media armónica

Definición

La **media armónica (H)** *de los datos x_i se define mediante:*

$$H = \frac{N}{\sum_i \frac{1}{x_i}}$$

Si los datos están agrupados puede usarse:

$$H = \frac{N}{\sum_i \frac{n_i}{x_i}}$$

donde $N = \sum_i n_i$

Ejemplo 1

Ejemplo

Tenemos 10 condensadores, 5 de $1\mu F$, 3 de $2\mu F$ y los otros 2 de $5\mu F$ conectados en serie. Queremos usar un único tipo de condensador (los 10 iguales). ¿Cuál debe ser su capacidad?

Evidentemente debe tener una capacidad intermedia. Pero teniendo en cuenta que la capacidad equivalente C a varios condensadores conectados en serie cumple: $\frac{1}{C} = \sum_i \frac{1}{C_i}$ tenemos:

$$\frac{1}{C} = \frac{5}{1} + \frac{3}{2} + \frac{2}{5} = \frac{10}{C_H} \Rightarrow \mathbf{C_H} = \frac{10}{\frac{5}{1} + \frac{3}{2} + \frac{2}{5}} = \mathbf{1.44927536}$$

que es la media armónica.
También vale para resistencias en paralelo.

Ejemplo 2

Ejemplo

Se quiere comparar la duración del establecimiento de conexión entre dos protocolos. El protocolo A produce los valores (en ms.): $\{10.8; 5.3; \infty; 12.4; 8.5; 7.7; 7.9; 4.4\}$*, mientras que el protocolo B produce:* $\{6.6; \infty; 4.6; 2.3; 1.3; 3.3; 4.9; \infty; 2.8; 2.2; 3.5; 1.7; 2.1; 3.9; 3.4; 3.8; 4.0\}$

Podemos usar la media armónica, pero no la aritmética:

$$H_A = \frac{8}{\frac{1}{10.8} + \frac{1}{5.3} + \frac{1}{\infty} + \frac{1}{12.4} + \frac{1}{8.5} + \frac{1}{7.7} + \frac{1}{7.9} + \frac{1}{4.4}} = \mathbf{8.3048788}$$

$$H_B = \frac{17}{\frac{1}{6.6} + \frac{1}{\infty} + \frac{1}{4.6} + \frac{1}{2.3} + \ldots + \frac{1}{3.8} + \frac{1}{4}} = \mathbf{3.20419429}$$

Así pues, parece mejor el B, a pesar de que no establece la llamada en el 11.76 % de los casos estudiados.

Media cuadrática o Valor cuadrático medio (RMS)

Definición

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

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

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.

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

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

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.*

Cálculo de la moda:

- Si la variable es cualitativa o discreta será la clase con mayor frecuencia (absoluta o relativa).
- 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). Posteriormente la moda será:

$$\mathbf{Mo} = \mathbf{L_{i-1}} + \frac{\mathbf{\Delta_1}}{\mathbf{\Delta_1} + \mathbf{\Delta_2}} \mathbf{a_i}$$

donde:

  - L_{i-1} = Extremo inferior del intervalo modal.
  - a_i = Amplitud del intervalo modal.
  - $\Delta_1 = h_i - h_{i-1}$
  - $\Delta_2 = h_i - h_{i+1}$

NOTA: Si no existe la clase $i-1$ ó $i+1$ consideraremos como 0 el valor de h_{i-1} ó h_{i+1} respectivamente.

Mediana

Es una medida central más robusta frente a los datos que la media.

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): $\mathbf{cN = E + D}$
  - 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:

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

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.

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

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 $\mathbf{d_i = |x_i - p|}$ 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.

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

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

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

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: $\bar{x} = \frac{5+2+3+3+3+5+7}{7} = 4$, 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.

La varianza y la desviación típica

La **varianza poblacional o varianza** de un conjunto de datos viene dada por:

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

Es decir, es la media de los cuadrados de las desviaciones respecto a la media.

Otra forma equivalente para calcular la varianza es:

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

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

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

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 $\bar{x} = \frac{\sum_i x_i}{N}$.
- El mejor estimador de la varianza de una población no es la varianza de la muestra, es la cuasivarianza de la muestra:

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

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

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

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:

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

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:

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

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

Ejemplo

Un fabricante de tubos de televisión produce dos tipos de tubos, A y B, que tienen vidas medias respectivas $\bar{x}_A$=1495 horas y $\bar{x}_B$=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.

Entropía

Definición

Llamamos **entropía** *de una variable (cualitativa o discreta) a:*

$$H = -\sum_i f_i log(f_i)$$

La base de logaritmo puede ser cualquiera, pero se suele hacer coincidir con el número de clases, si no se conoce se toma 2.
Mide la dispersión de una variable, mientras que la mayoría de las anteriores medidas miden la desviación respecto a un parámetro.
La entropía es mínima ($H = 0$) si todas las observaciones son iguales.
La entropía es máxima si todas las f_i son iguales ($f_i = \frac{1}{K}$).

$$H = -\sum_i \frac{1}{K} log_K \left(\frac{1}{K}\right) = \log_K(K) = 1$$

Ejemplo

Ejemplo

Dada la tabla de valores:

	Rojo	*Verde*	*Azul*	*Amarillo*
n_i	14	23	12	11

¿Cuál es su entropía?

$N = 14 + 23 + 12 + 11 = 60 \Rightarrow f_i = \{\frac{14}{60}, \frac{23}{60}, \frac{12}{60}, \frac{11}{60}\}$. Entonces:
$H = -\left[\frac{14}{60}\log_4\left(\frac{14}{60}\right) + \frac{23}{60}\log_4\left(\frac{23}{60}\right) + \frac{12}{60}\log_4\left(\frac{12}{60}\right) + \frac{11}{60}\log_4\left(\frac{11}{60}\right)\right]$
$\Rightarrow \mathbf{H \approx 0.96662703}$
NOTA: Podemos usar: $\log_K(x) = \frac{\log_b(x)}{\log_b(K)}$, pero MATLAB contiene las funciones **log**, **log2** y **log10** que calculan el de base e, 2 y 10. El de base 4 se obtiene mediante: $\log_{\mathbf{4}}(\mathbf{x}) = \log(\mathbf{x})/\log(\mathbf{4})$.

Momentos ordinarios respecto a un punto

Definición

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

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

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 6º vale 2 y el 7º 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:

$$\mathbf{m_1 = \bar{x}}$$

- El momento ordinario de orden 2 , $m_2 = \sigma^2 + \bar{x}^2$.

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:*

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

Propiedades:

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

$$\mu_2 = \mathbf{V} = \sigma^2 = \mathbf{m_2} - \bar{\mathbf{x}}^2$$

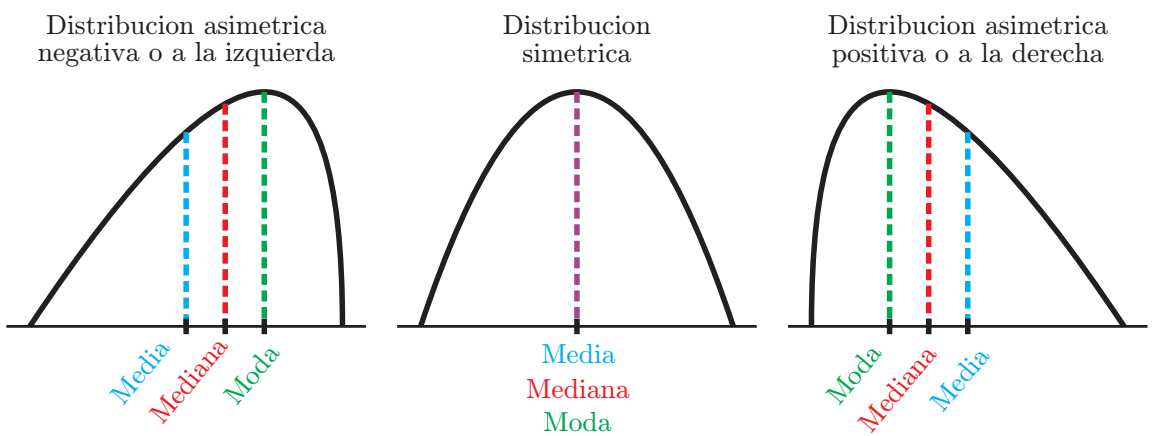
- $\mu_3 = \mathbf{m_3} - \mathbf{3m_2\bar{x}} + \mathbf{2\bar{x}^3}$
- $\mu_4 = \mathbf{m_4} - \mathbf{4m_3\bar{x}} + \mathbf{6m_2\bar{x}^2} - \mathbf{3\bar{x}^4}$

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: $\bar{x} = Me$, y usualmente $\bar{x} = Me = Mo$.



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:

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

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:

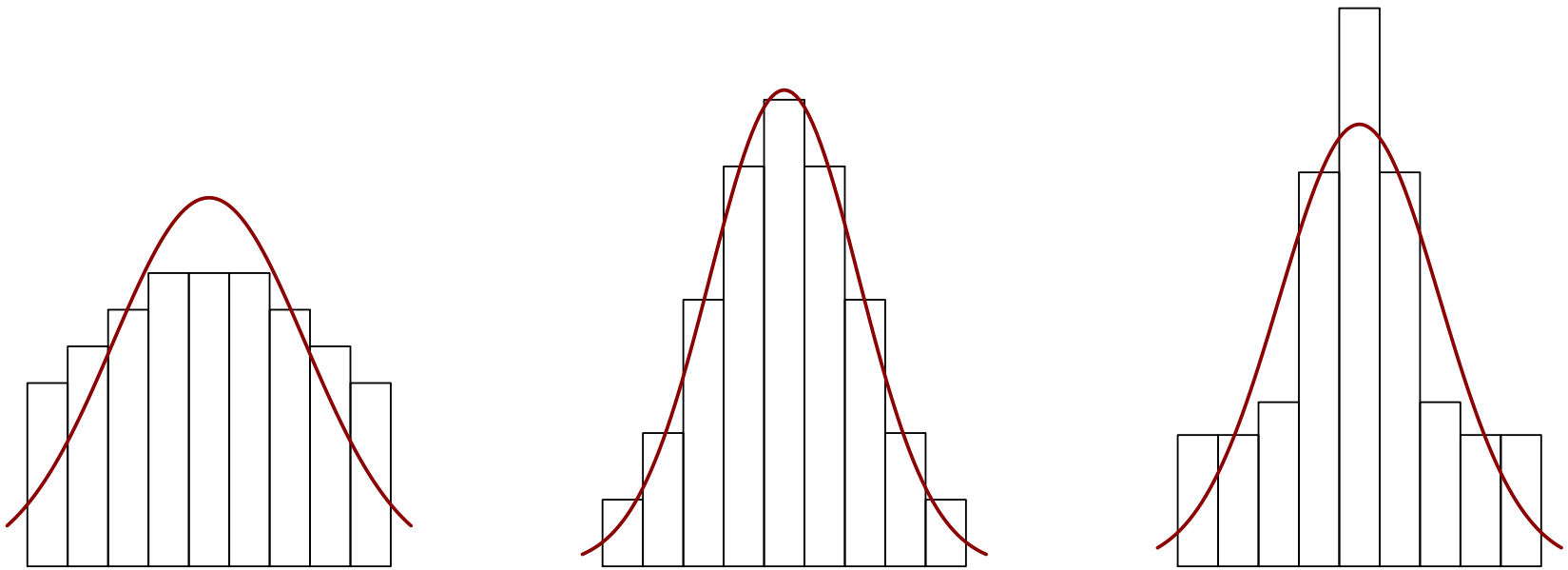
$$g_2 = \frac{\mu_4}{\sigma^4} - 3$$

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.

Ejemplo 1

Ejemplo

Los valores previstos (x_i), *reales* (x_i^*) *y frecuencia absoluta* (n_i) *vienen dados en la tabla. Hallar la media cuadrática (MC) y la desviación media (DM) 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$=\sqrt{\frac{49}{31}} =$
$= \mathbf{1.25723711}$

Error medio o DM$=\frac{23}{31} =$
$= \mathbf{0.74193548}$

Error cuadrático$=\frac{49}{31} =$
$= \mathbf{1.58064516}$

Ejemplo 2

Ejemplo

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.

e) Varianza y desviación típica.

f) Sesgo y curtosis.

g) Medias armónica y cuadrática.

h) Desviación media.

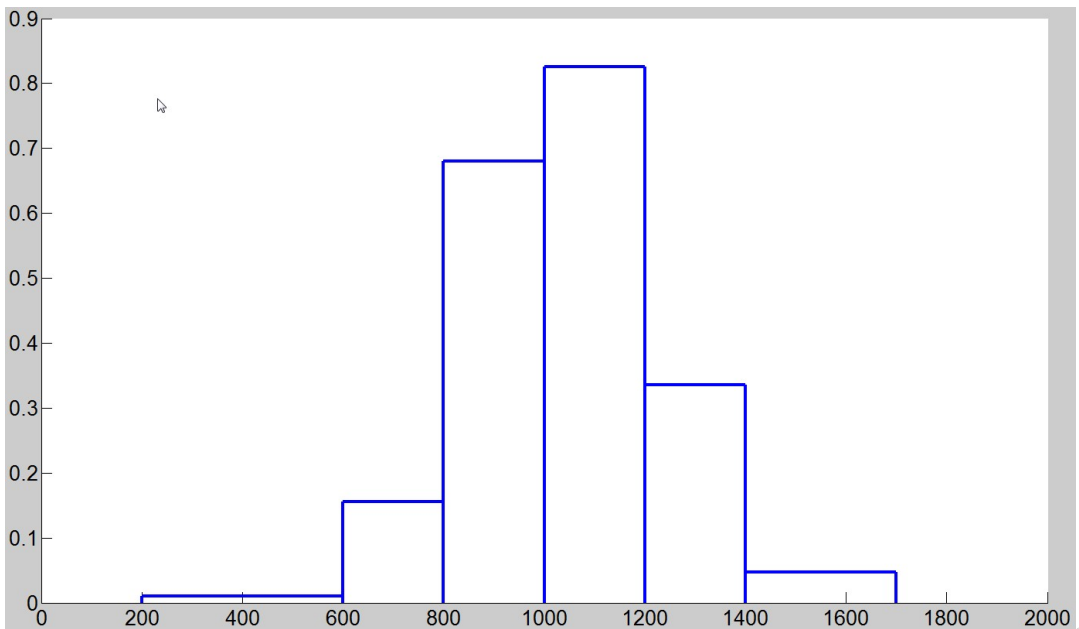
i) Error cuadrático medio (ECM) respecto media.

j) ECM respecto a la mediana.

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

Solución a) Histograma frec. absolutas

Int.	$\mathbf{n_i}$	a_i	$h_i = \frac{n_i}{a_i}$
$\mathbf{200 - 600}$	**4**	400	0.01
$\mathbf{600 - 800}$	**31**	200	0.155
$\mathbf{800 - 1000}$	**136**	200	0.68
$\mathbf{1000 - 1200}$	**165**	200	0.825
$\mathbf{1200 - 1400}$	**67**	200	0.335
$\mathbf{1400 - 1700}$	**14**	300	0.0467
	417		



b) Porcentaje que duran menos de 950.

Necesito las frecuencias absolutas acumuladas.

Int.	$\mathbf{n_i}$	a_i	h_i	N_i
$\mathbf{200-600}$	**4**	400	0.01	4
$\mathbf{600-800}$	**31**	200	0.155	35
$\mathbf{800-1000}$	**136**	200	0.68	171
$\mathbf{1000-1200}$	**165**	200	0.825	336
$\mathbf{1200-1400}$	**67**	200	0.335	403
$\mathbf{1400-1700}$	**14**	300	0.0467	417
	417			

Serán los 35 individuos que son menores o iguales que 800, más la parte proporcional de los que se encuentran en el intervalo (800-1000], es decir:

$$\mathbf{P} = 35 + 136 \cdot \frac{950-800}{1000-800} = 137 \Rightarrow \frac{137}{417} = \mathbf{0.329} \Rightarrow \mathbf{32.9\,\%}$$

c) Q_1

Int.	$\mathbf{n_i}$	a_i	h_i	N_i
$\mathbf{200-600}$	**4**	400	0.01	4
$\mathbf{600-800}$	**31**	200	0.155	35
$\mathbf{800-1000}$	**136**	200	0.68	171
$\mathbf{1000-1200}$	**165**	200	0.825	336
$\mathbf{1200-1400}$	**67**	200	0.335	403
$\mathbf{1400-1700}$	**14**	300	0.0467	417
	$N = 417$			

$\mathbf{Q_1}$: $c = \frac{1}{4} \Rightarrow Nc = \frac{417}{4} = 104.3$. Busco en la columna N_i el primero que rebasa el valor 104.3 $\Rightarrow$ el intervalo que contiene a Q_1 es el (800-1000] $\Rightarrow$

$$\mathbf{Q_1 = X\left(\frac{1}{4}\right) = L_{i-1} + \frac{N \cdot c - N_{i-1}}{n_i} a_i} = 800 + \frac{104.3 - 35}{136} 200 = \mathbf{901.9}$$

c) Q_3

Int.	$\mathbf{n_i}$	a_i	h_i	N_i
$\mathbf{200-600}$	**4**	400	0.01	4
$\mathbf{600-800}$	**31**	200	0.155	35
$\mathbf{800-1000}$	**136**	200	0.68	171
$\mathbf{1000-1200}$	**165**	200	0.825	336
$\mathbf{1200-1400}$	**67**	200	0.335	403
$\mathbf{1400-1700}$	**14**	300	0.0467	417
	$N = 417$			

Aplico la fórmula: $$\mathbf{X(c) = L_{i-1} + \frac{N \cdot c - N_{i-1}}{n_i} a_i}$$

$\mathbf{Q_3}$: $c = \frac{3}{4} \Rightarrow Nc = \frac{417}{4} = 312.9$. El primero en rebasar ese valor es el intervalo (1000-1200]: $\Rightarrow$

$$\mathbf{Q_3} = 1000 + \frac{312.9 - 171}{165} 200 = \mathbf{1172}$$

P_{10}, P_{90} y Mediana

Int.	$\mathbf{n_i}$	a_i	h_i	N_i
$\mathbf{200-600}$	**4**	400	0.01	4
$\mathbf{600-800}$	**31**	200	0.155	35
$\mathbf{800-1000}$	**136**	200	0.68	171
$\mathbf{1000-1200}$	**165**	200	0.825	336
$\mathbf{1200-1400}$	**67**	200	0.335	403
$\mathbf{1400-1700}$	**14**	300	0.0467	417
	$N = 417$			

$\mathbf{P_{10}}$: $c = \frac{1}{10} \Rightarrow Nc = 41.7$. El primero en rebasarlo es el (800-1000].
$\Rightarrow \mathbf{D_1} = \mathbf{P_{10}} = 800 + \frac{41.7-35}{136}200 = \mathbf{809.9}$

$\mathbf{P_{90}}$: $c = \frac{9}{10} \Rightarrow Nc = 375.3$. El primero en rebasarlo es el (1200-1400].
$\Rightarrow \mathbf{D_9} = \mathbf{P_{90}} = 1200 + \frac{375.3-336}{67}200 = \mathbf{1317}$

Mediana=$\mathbf{Q_2}$: $c = 0.5 \Rightarrow Nc = 208.5$. El primero en rebasarlo es el (1000-1200]. $\Rightarrow \mathbf{Me} = 1000 + \frac{208.5-171}{165}200 = \mathbf{1045.45}$

d) Moda

Int.	$\mathbf{n_i}$	a_i	h_i	N_i
$\mathbf{200-600}$	**4**	400	0.01	4
$\mathbf{600-800}$	**31**	200	0.155	35
$\mathbf{800-1000}$	**136**	200	0.68	171
$\mathbf{1000-1200}$	**165**	200	0.825	336
$\mathbf{1200-1400}$	**67**	200	0.335	403
$\mathbf{1400-1700}$	**14**	300	0.0467	417
	$N = 417$			

El intervalo modal es el (1000-1200] pues tiene el h_i mayor.

$$\mathbf{Mo} = \mathbf{L_{i-1}} + \frac{\mathbf{\Delta_1}}{\mathbf{\Delta_1} + \mathbf{\Delta_2}}\mathbf{a_i} = 1000 + \frac{0.145}{0.145 + 0.49}200 \approx \mathbf{1045.67}$$

donde $\Delta_1 = h_i - h_{i-1} = 0.825 - 0.68 = 0.145$ y
$\Delta_2 = h_i - h_{i+1} = 0.825 - 0.335 = 0.49$.

d) Media y e) Varianza y desviación típica.

Calculamos las marcas de clase (x_i) y las columnas $n_i x_i$ y $n_i x_i^2$.

Int.	$\mathbf{n_i}$	a_i	h_i	N_i	x_i	$n_i x_i$	$n_i x_i^2$
$\mathbf{200-600}$	**4**	400	0.01	4	400	1600	$64(10)^4$
$\mathbf{600-800}$	**31**	200	0.155	35	700	21700	$1519(10)^4$
$\mathbf{800-1000}$	**136**	200	0.68	171	900	122400	$11016(10)^4$
$\mathbf{1000-1200}$	**165**	200	0.825	336	1100	181500	$19965(10)^4$
$\mathbf{1200-1400}$	**67**	200	0.335	403	1300	87100	$11323(10)^4$
$\mathbf{1400-1700}$	**14**	300	0.0467	417	1550	21700	$33635(10)^3$
	$N = 417$					436000	$472505(10)^3$

Media: $\bar{\mathbf{x}} = \frac{436000}{417} = \mathbf{1045.56}$

Varianza: Calculo primero $\mathbf{m_2} = \frac{472505000}{417} = 1133105.516 \Rightarrow$

$\mathbf{V} = \mathbf{m_2} - \bar{\mathbf{x}}^2 = 1133105.516 - (1045.6)^2 = \mathbf{39902.38}$

Desviación típica: $\sigma = \sqrt{\mathbf{V}} = \sqrt{39902.38} = \mathbf{199.756}$

f) Sesgo y curtosis

Calculamos las columnas $n_i x_i^3$ y $n_i x_i^4$. Borro las no usadas.

Int.	$\mathbf{n_i}$	x_i	$n_i x_i$	$n_i x_i^2$	$n_i x_i^3$	$n_i x_i^4$
$\mathbf{200-600}$	**4**	400	1600	$64(10)^4$	$25600(10)^4$	$1024(10)^8$
$\mathbf{600-800}$	**31**	700	21700	$1519(10)^4$	$1063300(10)^4$	$74431(10)^8$
$\mathbf{800-1000}$	**136**	900	122400	$11016(10)^4$	$9914400(10)^4$	$892296(10)^8$
$\mathbf{1000-1200}$	**165**	1100	181500	$19965(10)^4$	$21961500(10)^4$	$2415765(10)^8$
$\mathbf{1200-1400}$	**67**	1300	87100	$11323(10)^4$	$14719900(10)^4$	$1913587(10)^8$
$\mathbf{1400-1700}$	**14**	1550	21700	$33635(10)^3$	$5213425(10)^4$	$808080.875(10)^8$
	417		436000	$472505(10)^3$	$52898125(10)^4$	$6105183.875(10)^8$

Momentos ordinarios: $\mathbf{m_1} = \bar{\mathbf{x}} = \frac{436000}{417} = \mathbf{1045.6}$

$\mathbf{m_2} = \frac{472505000}{417} \approx 1133105.52$, $\mathbf{m_3} = \frac{528981250000}{417} \approx 1268540167.87$

$\mathbf{m_4} = \frac{610518387500000}{417} \approx 1464072871702.64$

$\mu_3 = \mathbf{m_3} - \mathbf{3m_2\bar{x}} + \mathbf{2\bar{x}^3} =$

$= 1268540167.87 - 3(1133105.52)(1045.6) + 2(1045.6)^3 = 365394.78 \Rightarrow$

$$\mathbf{g_1} = \frac{\mu_3}{\sigma^3} = \frac{365394.78}{199.756^3} \approx 0.04584 \Rightarrow \textbf{Débilmente sesgada a derecha}$$

Sesgo y curtosis-2

$$\mu_4 = \mathbf{m_4} - \mathbf{4m_3\bar{x}} + \mathbf{6m_2\bar{x}^2} - \mathbf{3\bar{x}^4} =$$
$$= 14640728717702.64 - 4(1268540167.87)(1045.6) +$$
$$+6(1133105.52)(1045.6)^2 - 3(1045.6)^4 = 5723159551.29 \Rightarrow$$

$$\mathbf{g_2} = \frac{\mu_4}{\sigma^4} - \mathbf{3} = \frac{5723159551.29}{199.756^4} - 3 \approx 0.594498 \Rightarrow \textbf{Leptocúrtica}$$

NOTA: Cuando se pida el sesgo se está pidiendo el coeficiente de Fisher, excepto que se pida expresamente el de Pearson.

g) Media armónica, cuadrática y h) desviación media

Calculamos $\frac{n_i}{x_i}$ para la media armónica y $|di| = |x_i - \bar{x}|$ para la desviación media.

Int.	$\mathbf{n_i}$	x_i	$n_i x_i$	$n_i x_i^2$	$\frac{n_i}{x_i}$	$\lvert d_i \rvert$	$n_i \lvert d_i \rvert$
200 – 600	**4**	400	1600	$64(10)^4$	0.010000	645.563	2582.25
600 – 800	**31**	700	21700	$1519(10)^4$	0.044286	345.564	10712.470
800 – 1000	**136**	900	122400	$11016(10)^4$	0.151111	145.564	19796.643
1000 – 1200	**165**	1100	181500	$19965(10)^4$	0.150000	54.436	8982.014
1200 – 1400	**67**	1300	87100	$11323(10)^4$	0.051538	254.436	17047.242
1400 – 1700	**14**	1550	21700	$33635(10)^3$	0.009032	504.436	7062.110
	417		436000	$472505(10)^3$	0.415968		66182.734

Media armónica: $\mathbf{H = \frac{N}{\sum_i \frac{n_i}{x_i}} = \frac{417}{0.415968} \approx 1002.482}$

Media cuadrática:

$\mathbf{MC = \sqrt{m_2} = \sqrt{\frac{472505000}{417}} = \sqrt{1133105.516} \approx 1064.474}$

Desviación media: $\mathbf{DM = \frac{\sum_i n_i |d_i|}{N} = \frac{66182.734}{417} = 158.712}$

i) $ECM(\bar{x})$ y j) $ECM(Me)$ (Me=1045.45).

Int.	$\mathbf{n_i}$	x_i	$d_i^* = \lvert x_i - 1045.45 \rvert$	$n_i d_i^*$	$n_i (d_i^*)^2$
$\mathbf{200 - 600}$	**4**	400	645.45	2581.80	1666422.8
$\mathbf{600 - 800}$	**31**	700	345.45	10789.50	3699406.7
$\mathbf{800 - 1000}$	**136**	900	145.45	19781.20	2877175.6
$\mathbf{1000 - 1200}$	**165**	1100	54.55	9000.75	490990.9
$\mathbf{1200 - 1400}$	**67**	1300	254.55	17054.85	4341312.1
$\mathbf{1400 - 1700}$	**14**	1550	504.55	7063.70	3563989.8
	417			66191.25	16639297.9

Error cuadrático medio respecto a la media: Es la varianza que ya se ha calculado. $ECM(\bar{x}) = V = 39902.38$

Error cuadrático medio respecto a la mediana (ECM(Me)):

$$ECM(1053) = \frac{\sum_i n_i |x_i - 1045.45|^2}{N} = \frac{16639297.9}{417} \approx \mathbf{39902.393}$$

Desviación media respecto a la mediana:

$$\mathbf{DM(Me)} = \frac{\sum_{\mathbf{i}} \mathbf{n_i} |\mathbf{x_i} - \mathbf{1045.45}|}{\mathbf{N}} = \frac{66191.25}{417} = \mathbf{158.732}$$

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.