TEMA 1. ESTADÍSTICA DESCRIPTIVA

1. Concepto y origen de la estadística.

La estadística es la parte de las matemáticas que se ocupa de los procedimientos que permiten el tratamiento sistemático de diversos tipos de datos con el fin de darles una interpretación a partir de la cual tomar una decisión.

En sus orígenes históricos, la Estadística estuvo ligada a cuestiones de Estado (censos, recuentos, etc.) y de ahí su nombre. Hoy en día la estadística es una de las ramas matemáticas más usadas en todo tipo de ciencias (medicina, economía, biología, etc.).

La estadística ha llegado a los medios de comunicación, donde se nos presentan numerosos estudios estadísticos relativos a diversos temas, políticos, audiencias, deportivos…

En todo el tema trabajaremos con las siguientes tres estadísticas, que como veremos usan los tres tipos de variables estadísticas:

Ejemplo 1. Variable cuantitativa discreta: la siguiente lista representa el número de mensajes recibidos en los teléfonos móviles de 40 personas en un día:

3, 2, 1, 2, 0, 2, 1, 3, 2, 1, 1, 0, 2, 2, 1, 1, 3, 2, 1, 3, 2, 3, 1, 1, 0, 2, 2, 1, 2, 2, 0, 2,2,1, 2, 3, 2, 0, 1, 2.

Ejemplo 2. Variable cualitativa continua: Pesos de 20 asistentes a una reunión: 72, 63, 88, 91, 65,77, 81, 60, 84, 70, 75, 73, 78, 88, 64, 69, 86, 77, 90.

Ejemplo 3. Variable cualitativa: colores de los coches del claustro de profesores (20 profesores): rojo, rojo, blanco, negro, azul, gris, gris, negro, verde, amarillo, blanco, rojo, gris, amarillo, azul, azul, verde, amarillo, blanco, gris.

2. Conceptos básicos.

Para entender mejor los conceptos básicos que aparecen en cualquier estudio estadístico pongamos un ejemplo, el estudio de la altura media en España:

- **Población:** es el conjunto formado por todos los elementos que existen para el estudio de un determinado fenómeno y a los cuales nos referimos en el estudio. En nuestro ejemplo es la población de España
- **Individuo u objeto estadístico:** es cada uno de los elementos de la población. Cada uno de los españoles

- **Muestra:** es el subconjunto de individuos que tomamos de la población para realizar el estudio. Como elegir esta muestra será un tema de estudio más adelante. Puede ocurrir (en poblaciones pequeñas generalmente) que la muestra coincida con la población. En nuestro ejemplo es el conjunto de españoles a los cuales medimos para hacer el estudio.
- **Tamaño de la muestra:** es el número de individuos que forman la muestra elegida. Se denota generalmente como N.
- **Variable estadística:** cada una de las cualidades o propiedades referidas a la población y que son objeto de estudio. En nuestro ejemplo será la altura. Las variables estadísticas pueden ser de dos tipos:
    - ***Variables cualitativas*** o atributos: no se pueden medir numéricamente (por ejemplo: nacionalidad, color de la piel, sexo).
    - ***Variables cuantitativas***: tienen valor numérico (edad, precio de un producto, ingresos anuales). Por su parte, las variables cuantitativas se pueden clasificar en discretas y continuas:
        - *Discretas*: sólo pueden tomar un número finito y tratable de valores numéricos (por ejemplo: número de hijos de una familia, número de habitaciones en la casa)
        - *Continuas*: pueden tomar cualquier valor real dentro de un intervalo. (por ejemplo, la velocidad de un vehículo, altura de una persona)

3. Tablas estadísticas: recuento.

Como vemos en los tres ejemplos del tema los datos tal como están presentados no nos dan gran información, es por esto que la forma usual y útil de presentar los datos es en forma de **tabla *estadística***, una vez realizado el ***recuento***.

El recuento en Estadística se realiza de la forma siguiente:

1. En una columna (fila) se ponen los distintos valores que toma la variable, x_i (agrupados en intervalos si son continuos).
2. En la siguiente columna (fila) se pone la ***frecuencia absoluta,*** f_i**,** de cada valor de la variable***:*** número de veces que aparece dicho valor.

3. Generalmente se añaden otros parámetros estadísticos en las sucesivas columnas (filas) como la frecuencia relativa, frecuencias acumuladas y tanto por cien.

La frecuencia relativa (h_i): es el cociente entre la frecuencia absoluta y el número total de elementos de la encuesta, N. Se puede entender como el tanto por uno

$$\boldsymbol{h_i = \frac{f_i}{N} \qquad (0 \leq h_i \leq 1) \;\rightarrow\; \sum_{i=1}^{k} h_i = 1}$$

Tanto por cien (p_i): como su nombre indica nos indica el porcentaje relativo a 100 de la característica respecto del total:

$$\boldsymbol{p_i = \frac{f_i}{N} \cdot 100 = h_i \cdot 100 \qquad (0 \leq p_i \leq 100) \;\rightarrow\; \sum_{i=1}^{k} p_i = 100}$$

La frecuencia absoluta acumulada (F_i): es la suma de todas las frecuencias absolutas hasta la i-esima (incluida), es decir

$$\boldsymbol{F_i = \sum_{t=1}^{i} f_t = f_1 + f_2 + \cdots + f_i}$$

La frecuencia relativa acumulada (H_i): es la suma de todas las frecuencias relativas hasta la i-esima (incluida), es decir

$$\boldsymbol{H_i = \sum_{t=1}^{i} h_t = h_1 + h_2 + \cdots + h_i = \frac{F_i}{N}}$$

El porcentaje acumulado (P_i): es la suma de todos los porcentajes hasta el i-esimo (incluido), es decir

$$\boldsymbol{P_i = \sum_{t=1}^{i} p_t = p_1 + p_2 + \cdots + p_i}$$

Para calcular las frecuencias acumuladas utilizar la relación entre dos frecuencias acumuladas sucesivas: $F_{i+1}=F_i+f_{i+1}$, $H_{i+1}=H_i+h_{i+1}$, $P_{i+1}=P_i+p_{i+1}$

Veamos en los ejemplos anteriores como quedaría la tabla de frecuencias:

Ejemplo 1. Variable cuantitativa discreta: la siguiente lista representa el número de mensajes recibidos en los teléfonos móviles de 40 personas en un día:

3, 2, 1, 2, 0, 2, 1, 3, 2, 1, 1, 0, 2, 2, 1, 1, 3, 2, 1, 3, 2, 3, 1, 1, 0, 2, 2, 1, 2, 2, 0, 2,2,1, 2, 3, 2, 0, 1, 2.

x_i=nºsms	f_i	h_i	p_i	F_i	H_i	P_i
0	5	0,125	12,5%	5	0,125	12,5%
1	12	0,3	30%	17	0,425	42,5%
2	17	0,425	42.5%	34	0,85	85%
3	6	0,15	15%	40	1	100%
Total	40	1	100%			

Ejemplo 3. Variable cualitativa: colores de los coches del claustro de profesores (20 profesores): rojo, rojo, blanco, negro, azul, gris, gris, negro, verde, amarillo, blanco, rojo, gris, amarillo, azul, azul, verde, amarillo, blanco, gris.

En las *variables cualitativas no tiene sentido hablar de las frecuencias acumuladas*, ya que las características no son números y por tanto no se pueden ordenar

x_i=color	f_i	h_i	p_i
Rojo	3	0,15	15%
Blanco	3	0,15	15%
Negro	2	0,1	10%
Gris	4	0,2	20%
Verde	2	0,1	10%
Amarillo	3	0,15	15%
Azul	3	0,15	15%
Total	20	1	100%

Ejemplo 2. Variable cualitativa continua: Pesos de 20 asistentes a una reunión: 72, 63, 88, 91, 65,77, 81, 60, 84, 70, 75, 73, 78, 88, 64, 69, 86, 77, 90, 80.

Hemos dejado esta para el final, pues hay que elaborar los intervalos. Para hacerlos debemos conocer el rango, que es la diferencia máxima entre dos valores, y el número de intervalos en los que deseamos clasificar la variable.

Rango=R=x_{max}-x_{min}=91-60=31. Y vamos a agruparlos en 4 intervalos. Si queremos hacerlo exacto el numero rango de cada intervalos será 31/4=7,75, aunque es más lógico ampliar el rango con el fin de que este número sea exacto. En nuestro caso ampliaremos el rango a 32, con lo que cada intervalo tendrá un recorrido de 32/4=8. Al ampliar dicho rango en 1 tendremos que comenzar 1 unidad antes o acabar 1 después. Hagamos lo segundo (puede hacerse una u otra indistintamente)

Intervalo I_i	**Marca de clase (x_i)**	**f_i**	**h_i**	**p_i**	**F_i**	**H_i**	**P_i**
[60,68)	64	4	0,2	20%	4	0,2	20%
[68,76)	72	5	0,25	25%	9	0,45	45%
[76,84)	80	5	0,25	25%	14	0,7	70%
[84,92]	88	6	0,3	30%	20	1	100%
Total		20	1	100%			

Las ***marcas de clase*** son los puntos medios de los intervalos.

Nota: las amplitudes de las clases no tienen por qué ser iguales, esto lo tendremos muy en cuenta cuando representamos la gráfica del histograma.

4. Representación de graficas.

4.1. Variables cualitativas

Las representaciones de las variables cualitativas son:

- Diagrama de barras
- Diagrama de sectores
- Pictogramas
- Cartogramas (variables relativas a zonas)
- Pirámides de población (estudio de edad de una población)

Diagrama de barras: consiste en dibujar un rectángulo por cada una de las modalidades de la variable, de forma que las bases sean todas iguales y apoyadas en el eje OX, donde se indican los valores de la variable y la altura de cada rectángulo (barra) es proporcional a la frecuencia (relativa, absoluta o porcentaje es la misma proporción).

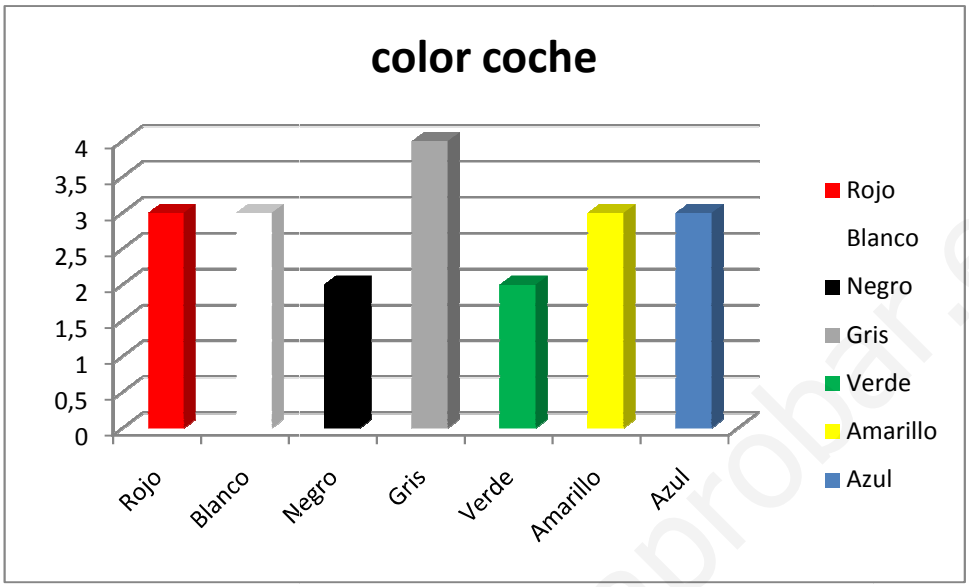
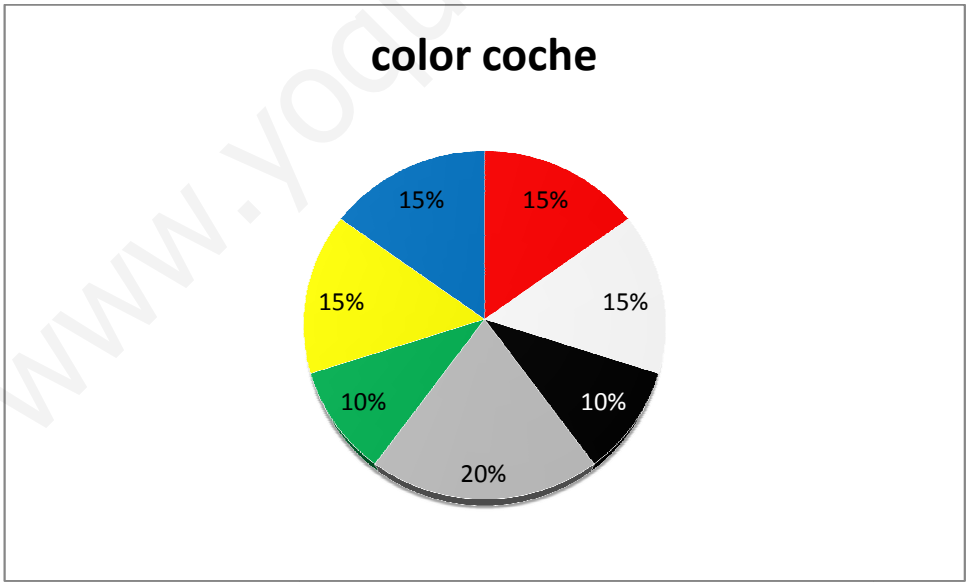
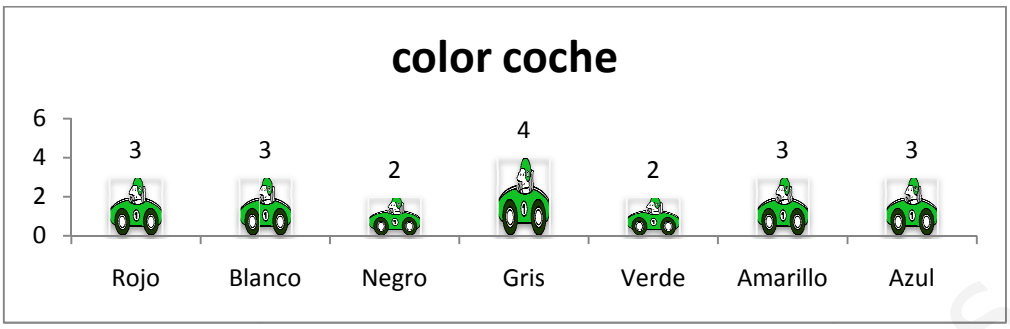


Diagrama de sectores: Consiste en dividir un circulo en sectores circulares, con ángulo proporcional a la frecuencia ($\alpha_i = h_i \cdot 360º$).

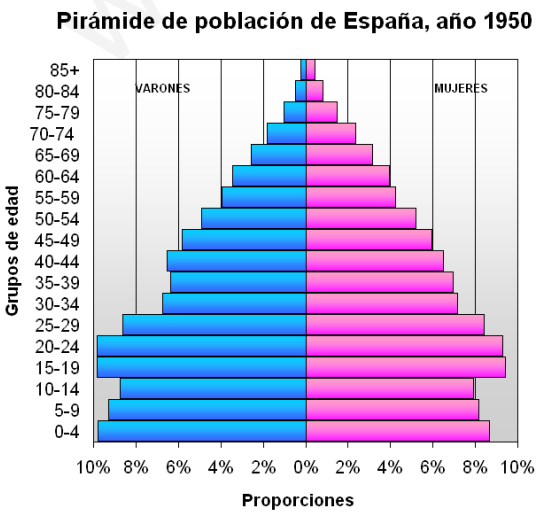


Pictograma: consiste en realizar dibujos alusivos a la distribución que se desea presentar. Son gráficos poco precisos pero fáciles de interpretar a simple vista.

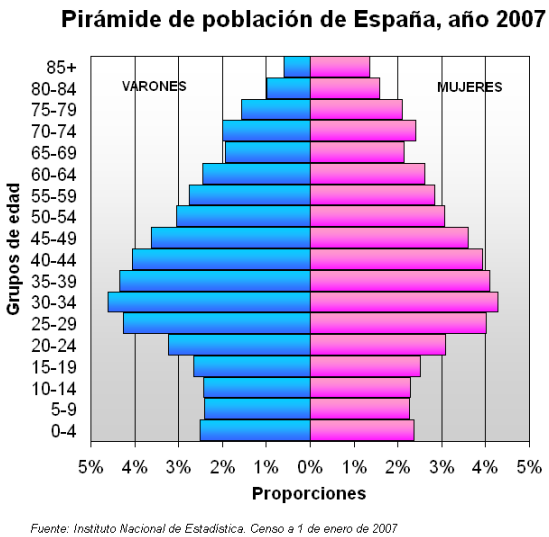


Cartogramas: consiste en representar en un mapa cualquier tipo de datos relacionados con un área geográfica. Ejemplo:

Pirámides de población: se utilizan para estudiar conjuntamente el carácter cuantitativo edad y el cualitativo sexo. Según la forma de la pirámide se puede deducir si se trata de una población joven, madura o vieja. Veamos dos ejemplos



Fuente: Instituto Nacional de Estadística. Censo de 1950



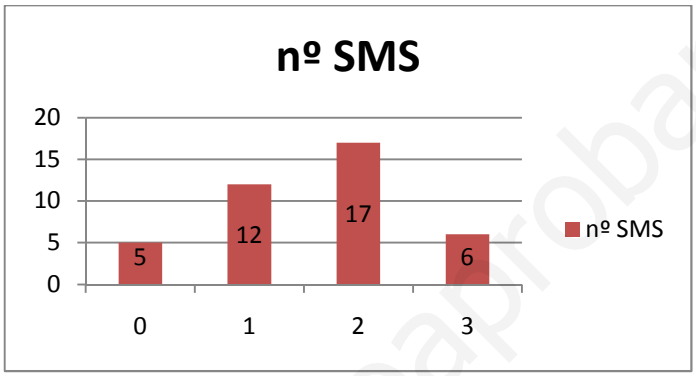
Fuente: Instituto Nacional de Estadística. Censo a 1 de enero de 2007

4.2. Variables cuantitativas discretas

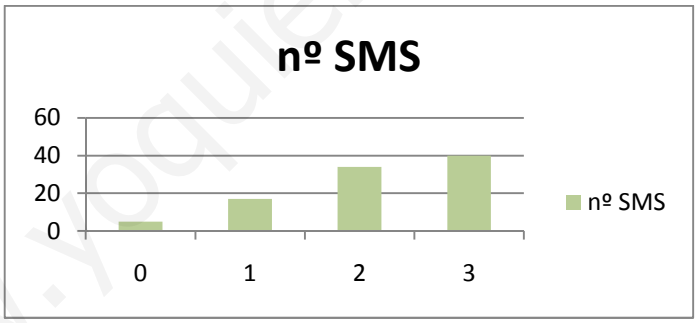
Los gráficos más utilizados para representar distribuciones de variable cuantitativas discretas son:

- Diagrama de barras o columnas
- Diagrama de frecuencia o polígono de frecuencia

Diagrama de barras: se representan por barras o columnas independientes y de igual anchura situadas encima del eje de la variable. La altura de las barras (o longitud de las columnas) es proporcional a la frecuencia. Veamos en nuestro ejemplo



A veces los datos presentados son las frecuencias acumuladas



Nota: En muchas ocasiones se superponen dos diagramas de barras con el fin de comparar dos variables cuantitativas discretas. Veamos el siguiente ejemplo:

ABANDONO DE NIÑOS

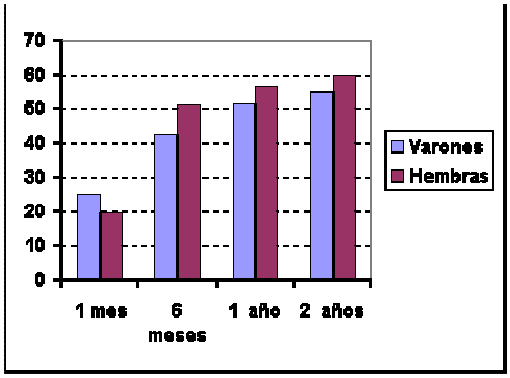
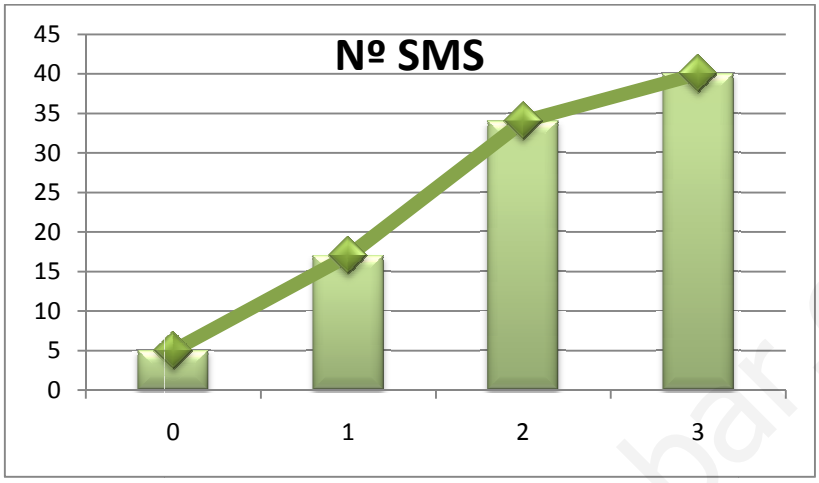


Diagrama de frecuencia o polígono de frecuencia: Se forman uniendo los extremos de las barras o columnas mediante una línea quebrada. Son muy utilizados en las frecuencias acumuladas en el estudio de determinados fenómenos:

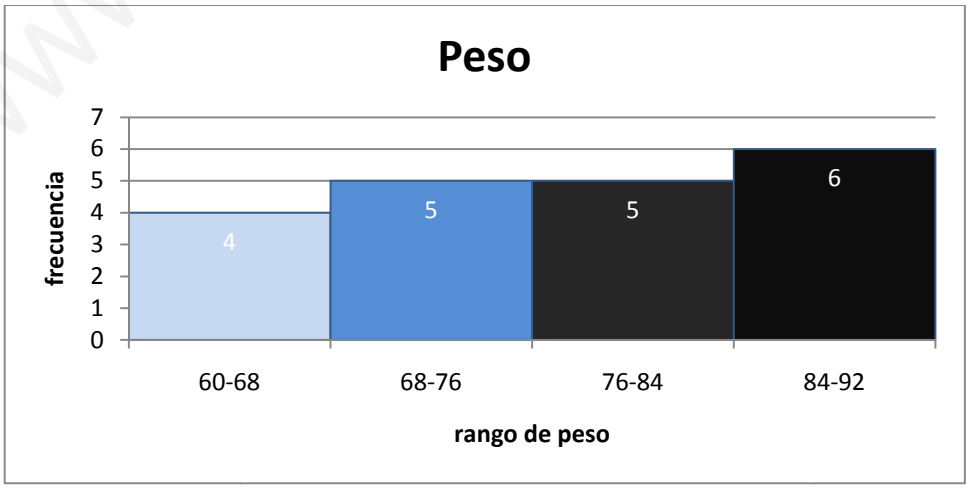


4.3. Variables cuantitativas continuas.

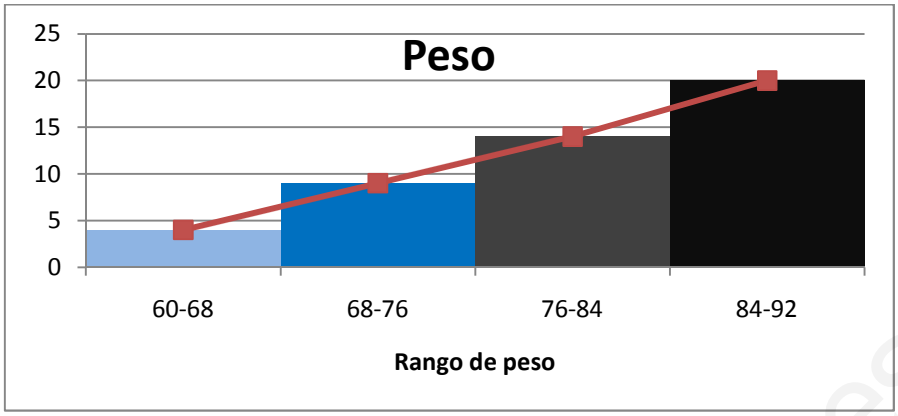
Los gráficos más utilizados para representar distribuciones de variable cuantitativas continua son:

- Histograma
- Diagrama de frecuencia o polígono de frecuencia

Histograma: son análogos a los diagrama de barras pero se utilizan para representar variables continuas. La diferencia es que en los histogramas las bases de los rectángulos son los distintos intervalos. La altura de los rectángulo son proporcionales a las frecuencias siempre y cuando sean intervalos de misma amplitud, en caso contrario las alturas serán tales que las áreas de los rectángulos sean proporcionales a las frecuencias.



Polígono de frecuencia: igual que en las variables cuantitativas discretas



5. Parámetros estadísticos.

5.1. Parámetros de centralización.

Estos parámetros nos indican en torno a que puntos se encuentran los valores de la variable cuantitativa en estudio. Es la forma de representar un conjunto de datos mediante un solo valor, tratando de resumir o sintetizar la distribución de frecuencias. Los parámetros más importantes son:

- Media (aritmética y geométrica)
- Moda
- Mediana

1. Media: es el valor medio ponderado de la serie de datos. Se pueden calcular diversos tipos de media, siendo las más utilizadas:

Media aritmética: se calcula multiplicando cada valor por el número de veces que se repite. La suma de todos estos productos se divide por el total de datos de la muestra. La media aritmética es el parámetro de centralización más importante y más usada. La media aritmética de un conjunto de datos x_i se representa por $\bar{x}$. Su cálculo se realiza de la siguiente forma:

a) Datos sin frecuencia:

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

b) Si tenemos k datos distintos con sus frecuencias (tabla de frecuencias):

$$\bar{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \ldots + x_k \cdot f_k}{N} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{N}$$

c) Con datos ponderados: es cuando queremos dar más "peso" a algunos datos que otro. Si llamamos l_i al peso en tanto por cien ($\sum l_i = 100$) la ***media ponderada*** es:

$$\bar{x} = \frac{x_1 \cdot l_1 + x_2 \cdot l_2 + ... + x_N \cdot l_N}{100} = \frac{\sum_{i=1}^{N} x_i \cdot l_i}{100}$$

Ejemplo: nota media ponderada de 3 exámenes, el primero pondera 30% el segundo 30% y el tercero 40% → $\bar{x} = \frac{x_1 \cdot 30 + x_2 \cdot 30 + x_3 \cdot 40}{100}$, siendo x_1, x_2, x_3 las notas de los tres exámenes.

Veamos la media en los dos ejemplos cuantitativos que desarrollamos en el tema:

Ejemplo 1: $\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{N} = \frac{0 \cdot 5 + 1 \cdot 12 + 2 \cdot 17 + 3 \cdot 6}{40} = 1.6 \ sms$

Ejemplo 2: en las variables continuas se suele aproximar utilizando las marcas de clase en vez de los verdaderos valores, a fin de simplificar los cálculos.

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{N} = \frac{64 \cdot 4 + 72 \cdot 5 + 80 \cdot 5 + 88 \cdot 6}{20} = 77.2 \ kg$$

Para el cálculo de la media muchas veces se realiza una tabla con las siguientes tres columnas: los valores x_i, las frecuencias absolutas f_i, el producto $x_i \cdot f_i$. En las celdas inferiores se hace la suma de todos los productos $x_i \cdot f_i$, siendo la media por tanto esta suma entre N:

x_i=nºsms	f_i	$x_i \cdot f_i$
0	5	0
1	12	12
2	17	34
3	6	18
Total	40	64

$$\bar{x} = \frac{64}{40} = 1.6$$

x_i=peso	f_i	$x_i \cdot f_i$
64	4	256
72	5	360
80	5	400
88	6	528
	20	1544

$$\bar{x} = \frac{15444}{20} = 77.2$$

***Media geométrica*:** se eleva cada valor al número de veces que se ha repetido. Se multiplican todo estos resultados y al producto final se le calcula la raíz "N" (siendo "N" el total de datos de la muestra).

$$\overline{x_g} = \sqrt[N]{\prod_{i=1}^{k} (x_i)^{f_i}} = \sqrt[N]{x_1^{f_1} \cdot x_2^{f_2} \cdot \ldots \cdot x_k^{f_k}}$$

La media geométrica se suele utilizar en series de datos como tipos de interés anuales, inflación, etc., donde el valor de cada año tiene un efecto multiplicativo sobre el de los años anteriores. En todo caso, la media aritmética es la medida de posición central más utilizada.

Las medias (tanto en el caso de la media aritmética como geométrica) presenta el problema de que su valor se puede ver muy influido por valores extremos, que se aparten en exceso del resto de la serie. Estos valores anómalos podrían condicionar en gran medida el valor de la media, perdiendo ésta representatividad

2. ***Moda (M_0):*** es el valor que más se repite en la muestra.

Calculo en las variables cuantitativas discretas (también cualitativas): para calcularlo basta con buscar el valor de la variable que presenta más frecuencia. Puede ocurrir que la moda no sea única, es decir, la distribución tenga 2, 3,… modas, recibiendo el nombre de bimodales, trimodales, etc.

En nuestro *ejemplo 1* la moda es 2 sms, pues es el de mayor frecuencia absoluta (17)

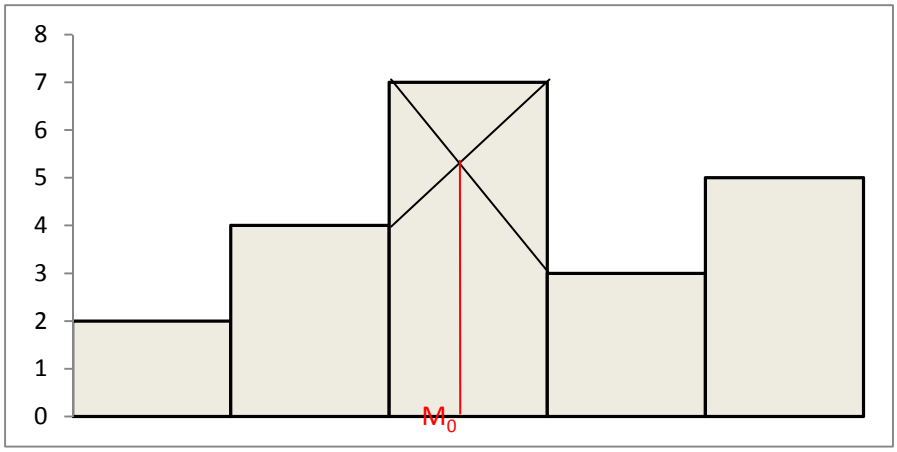
Calculo en la variable continua: se puede hacer de forma aproximada con las marcas de clase, aunque si se quiere ser más preciso se puede obtener mediante la expresión:

$$M_0 = L_i + \frac{f_{M_0} - f_{M_{0-1}}}{\left(f_{M_0} - f_{M_{0-1}}\right) + \left(f_{M_0} - f_{M_{0+1}}\right)} \cdot c$$

siendo:

- L_i el límite inferior de la clase modal
- c la amplitud del intervalo modal
- f_{M_0}, $f_{M_{0-1}}$, $f_{M_{0+1}}$ las frecuencias absolutas de la clase modal, la anterior y la siguiente.

Este valor M_0 es la intersección de las rectas que unen los extremos de la clase modal con los extremos más próximos de las clases anterior y siguiente:



En nuestro ejemplo 2, el valor aproximado de la moda es el intervalo [84,92], pues es el que tiene mayor frecuencia absoluta (6). Si queremos calcular M_0 de forma exacta usemos la formula (L_i=84, e=8, f_i=6, f_{i-1}=5, f_{i+1}=0)→ M_0=85,14.

3. Mediana (M_e): ordenados los N elementos en orden creciente es el que ocupa la posición intermedia, siendo el 50% de los datos menores o iguales que M_e y el restante 50% mayores o iguales que M_e.

Calculo para variable cuantitativa discreta: es el primer valor que supera el 50% en porcentaje acumulado (o N/2 en frecuencia absoluta acumulada). Puede ocurrir cuando N es par que un dato tenga frecuencia acumulada de 50%, en este caso la mediana se considera la media entre el dato con dicha frecuencia acumulada y el siguiente dato. En nuestro *ejemplo 1* la mediana es 2 mensajes.

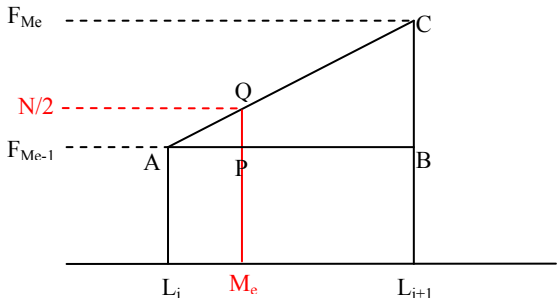
Calculo para variable cuantitativa continua: de forma aproximada se hace igual que para la variable discreta usándolas marcas de clase. Si se quiere ser más exacto se debe buscar el valor de la recta frecuencia acumulada que valga N/2. La formula es la siguiente:

$$M_e = L_i + \frac{\frac{N}{2} - F_{Me-1}}{f_{Me}} \cdot c$$

siendo:

- L_i el límite inferior de la clase mediana
- c la amplitud del intervalo mediana
- f_{Me} la frecuencia absolutas de la clase modal.
- N el número total de datos.
- F_{Me-1} la frecuencia absoluta acumulad hasta llegar a la mediana sin incluirla.

La formula se puede obtener gráficamente por semejanza de triángulos ABC y APQ:



En nuestro *ejemplo 2,* la mediana aproximada es M_e=80, y si la calculamos de forma exacta: $M_e = 76 + \frac{10-9}{5}\cdot 8 = 77.6$.

5.2. Parámetros de posición

Sirven para determinar en qué posición de la distribución se encuentra un individuo, supuestos ordenados de forma creciente. Los parámetros de posición más importantes son:

- Cuartiles.
- Percentiles.

1. ***Cuartiles:*** son 3 valores (Q_1, Q_2, Q_3) que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cuatro tramos iguales, en los que cada uno de ellos concentra el 25% de los resultados.
   - Q_1 (el primer valor que supere su frecuencia acumulada el 25%). En variable continua: $Q_1 = L_i + \frac{\frac{N}{4} - F_{Q_1-1}}{f_{Q_1}}\cdot c$
   - Q_3 (el primer valor que supere su frecuencia acumulada el 75%). En variable continua: $Q_3 = L_i + \frac{\frac{3\cdot N}{4} - F_{Q_3-1}}{f_{Q_3}}\cdot c$

   Nota: Q_2=M_e

2. ***Percentiles:*** son 99 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cien tramos iguales, en los que cada uno de ellos concentra el 1% de los resultados. Se representan por P_1,P_2,…,P_{99} siendo el valor de la variable que primero supere el porcentaje acumulado el 1%, 2%...,99%.

5.3. Parámetros de dispersión.

Estudia la distribución de los valores de la serie, analizando si estos se encuentran más o menos concentrados, o más o menos dispersos.

Existen diversas medidas de dispersión, entre las más utilizadas podemos destacar las siguientes:

- Rango o recorrido
- Desviación media
- Varianza
- Desviación típica
- Coeficiente de variación.

1. ***Rango o recorrido:*** mide la amplitud de los valores de la muestra y se calcula por diferencia entre el valor más elevado y el más bajo. Se representa por R

   $$R=x_{mx}-x_{min}$$

   En nuestros ejemplos: ejemplo 1→ R=3-0=3, ejemplo 2 → R=91-60=31

2. ***Desviación media:*** es la media de los valores absolutos de las desviaciones de los datos o marcas de clase respecto la media aritmética. Se representa por DM

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

   En nuestros ejemplos:

   - Ejemplo 1: $DM = \frac{|0-1.6|\cdot 5+|1-1.6|\cdot 12+|2-1.6|\cdot 17+|3-1.6|\cdot 6}{40} = 0.76$
   - Ejemplo 2: $DM = \frac{|64-77.2|\cdot 4+|72-77.2|\cdot 5+|80-77.2|\cdot 5+|88-77.2|\cdot 6}{20} = 25{,}38$

3. ***Varianza:*** es la media aritmética de los cuadrados de las desviaciones de todos los datos o marcas de clase respecto a la media. Se representa por σ^2 o Var(x):

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

La varianza no tiene las mismas unidades que x (si x es metros σ^2 será metros cuadrados). Es por eso que se utiliza más la desviación típica.

En nuestros ejemplos:

- Ejemplo 1: $\sigma^2 = 0{,}79$
- Ejemplo 2: $\sigma^2 = 78{,}56$

Para calcularla se suele añadir la columna $x_i^2 \cdot f_i$ a la derecha de la variable y de sus frecuencias absolutas. La suma de esta columna nos permite calcular dividiendo entre N el valor de $\overline{(x^2)}$. Veamos con el ejemplo de los mensajes y de los pesos:

x_i=nºsms	f_i	$x_i^2 \cdot f_i$
0	5	0
1	12	12
2	17	68
3	6	54
Total	40	134

x_i=peso	f_i	$x_i^2 \cdot f_i$
64	4	16.384
72	5	25.920
80	5	32.000
88	6	46.464
	20	120.768

Ejemplo 1: $\overline{(x^2)} = \frac{134}{40} = 3.35 \rightarrow \sigma^2 = 3.35 - 1.6^2 = 0.79$

Ejemplo 2: $\overline{(x^2)} = \frac{120738}{20} = 6038.4 \rightarrow \sigma^2 = 6038.4 - 77.2^2 = 78.56$

4. ***Desviación típica:*** es la raíz cuadrada de la varianza. Tiene mismas dimensiones que la variable estadística en estudio. Se denota por σ

$$\sigma = \sqrt{Var(x)} = \sqrt{\sigma^2}$$

Ejemplo 1: $\sigma = \sqrt{0.79} = 0.89$

Ejemplo 2: $\sigma = \sqrt{78.56} = 8.86$

En la medida en que los parámetros de dispersión tomen valores más o menos grandes esto nos indicara el grado de dispersión o alejamiento de los datos respecto de la media. En el caso trivial que todos los datos centrados en un mismo valor todos estos parámetros valdrían cero. Para distribuciones normales (que veremos más adelante) se cumple:

- El 68,27% datos en el intervalo $[\bar{x}-\sigma, \bar{x}+\sigma]$
- El 95,45% datos en el intervalo $[\bar{x}-2\sigma, \bar{x}+2\sigma]$
- El 99,73% datos en el intervalo $[\bar{x}-3\sigma, \bar{x}+3\sigma]$

5. ***Coeficiente de variación:*** las medidas de dispersión estudiadas hasta ahora se expresan en la misma medida que la variable estadística, designando medidas de dispersión absolutas respecto de la media. Esto presenta los siguientes problemas:
    - No podemos comparar distribuciones de distinta naturaleza (peso y altura) o incluso de la misma naturaleza expresadas en distintas unidades.
    - No es relativa al valor de la media: la variación de 100€ respecto de 1.000€ es mucho más significativa que la de los mismos 100€ respecto a 1.000.000€.

Estos problemas se resuelven con el coeficiente de variación, que es el cociente entre la desviación típica y la media, siendo por tanto adimensional.

$$CV = \frac{\sigma}{|\bar{x}|} \quad \text{o en tanto por cien} \quad CV(\%) = \frac{\sigma}{|\bar{x}|}\cdot 100\%$$

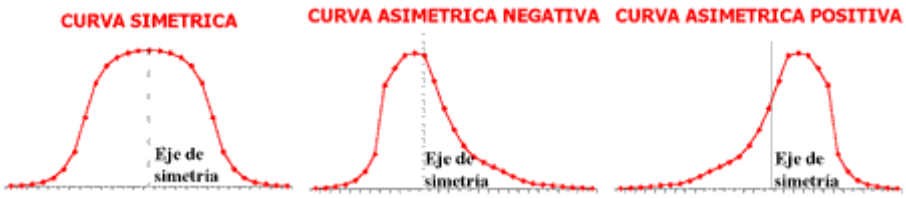
Cuanto más se aproxime el CV a cero más representativa será la media en la distribución.

En nuestros ejemplos:

- Ejemplo 1: $CV = \frac{0.89}{1.6} = 0.56$
- Ejemplo 2: $CV = \frac{8.86}{77.2} = 0.11$

5.4. Coeficientes de forma. Medida de asimetría y curtosis

El concepto de asimetría se refiere a si la curva que forman los valores de la serie presenta la misma forma a izquierda y derecha de un valor central (media aritmética)



Para medir el nivel de ***asimetría*** se utiliza el llamado ***Coeficiente de Asimetría de Fisher***, que viene definido:

$$g_1 = \frac{\overline{(x-\bar{x})^3}}{\sigma^3} = \frac{\frac{1}{N}\sum_{i=1}^{k}(x_i-\bar{x})^3 \cdot f_i}{\sigma^3}$$

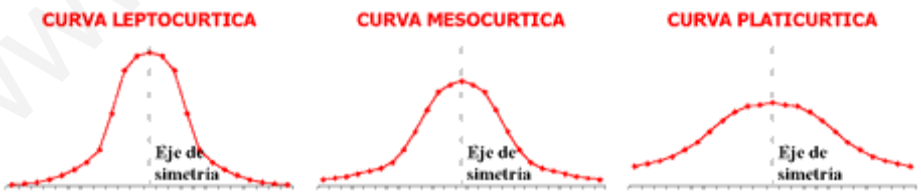
Los resultados pueden ser los siguientes:

- $g_1 = 0$ (distribución simétrica; existe la misma concentración de valores a la derecha y a la izquierda de la media)
- $g_1 > 0$ (distribución asimétrica positiva; existe mayor concentración de valores a la derecha de la media que a su izquierda)
- $g_1 < 0$ (distribución asimétrica negativa; existe mayor concentración de valores a la izquierda de la media que a su derecha)

El ***Coeficiente de Curtosis analiza el grado de concentración*** que presentan los valores alrededor de la zona central de la distribución.

Se definen 3 tipos de distribuciones según su grado de curtosis:

- Distribución *mesocúrtica*: presenta un grado de concentración medio alrededor de los valores centrales de la variable (el mismo que presenta una distribución normal).
- Distribución *leptocúrtica*: presenta un elevado grado de concentración alrededor de los valores centrales de la variable.
- Distribución platicúrtica: presenta un reducido grado de concentración alrededor de los valores centrales de la variable.



El Coeficiente de Curtosis viene definido por la siguiente fórmula:

$$g_2 = \frac{\overline{(x-\bar{x})^4}}{\sigma^4} - 3 = \frac{\frac{1}{N}\sum_{i=1}^{k}(x_i-\bar{x})^4 \cdot f_i}{\sigma^4} - 3$$

Los resultados pueden ser los siguientes:

- $g_2 = 0$ (distribución mesocúrtica).
- $g_2 > 0$ (distribución leptocúrtica).
- $g_2 < 0$ (distribución platicúrtica).

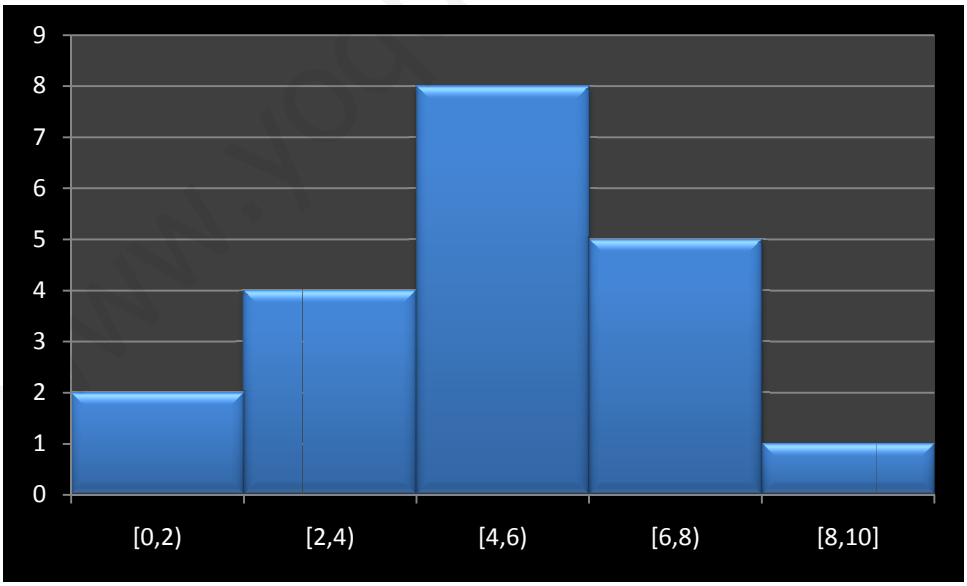
Las medidas de asimetría, sobre todo el coeficiente de asimetría de Fisher, junto con las medidas de apuntamiento o curtosis se utilizan para contrastar si se puede aceptar que una distribución estadística sigue la distribución normal. Esto es necesario para realizar numerosos contrastes estadísticos en la teoría de inferencia estadística.

Ejercicios finales

Ejercicio 1. Completar los datos que faltan en la siguiente tabla estadística. Calcular todos los parámetros estadísticos explicados en el tema e interpretar la distribución estadística.

x_i	f_i	h_i	F_i	H_i	$f_i \cdot x_i$	$f_i \cdot x_i^2$
1	4	0,08				
2	4					
3			16			
4	7					
5	5					
6			38			
7	7					
8						
	N=					

Ejercicio 2. Las puntuaciones obtenidas por una clase en un examen de estadística quedan reflejadas en el siguiente histograma de frecuencias absolutas. Calcular la media, la moda, la varianza y el coeficiente de variación. Interpretar con los datos la distribución.



Ejercicio 3. Las notas de dso grupos de 10 alumnos en la segunda evaluación de matemáticas se recogen en la siguiente tabla:

Grupo A	1	1	3	5	5	6	8	8	9	9
Grupo B	2	2	4	4	4	5	5	6	6	8

Contestar razonadamente las siguientes preguntas:

a) ¿Cuál de los dos grupos obtuvo mejores resultados?

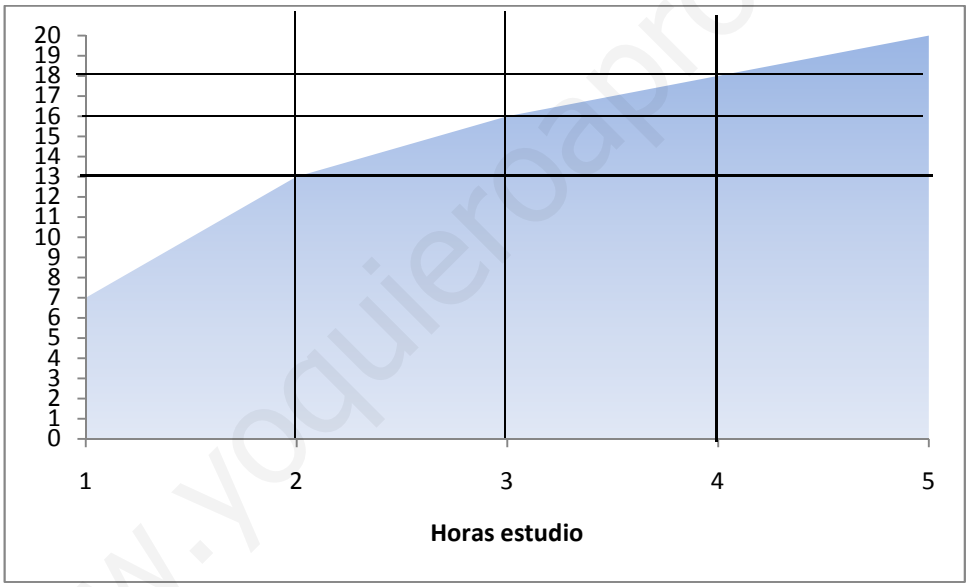
b) ¿Cuál es el grupo más homogéneo?

Ejercicio 4: La siguiente grafica representa la frecuencia acumulada de horas de estudio en una clase de 20 alumnos.

a) Construir la tabla de frecuencias absolutas y relativas

b) Calcular los Cuartiles y P_{90} y P_{60}

c) Calcular los coeficientes de forma e interpretarlos



Ejercicio 5. Calcular los coeficientes de forma de los ejemplos 1 y 2 y explicar los resultados comparándolos con sus graficas (diagrama de barras e histograma respectivamente).