Tema 9: Estadística Unidimensional.

1.0.- Introducción

Estadística es la ciencia que se ocupa de ocupa de la recogida de datos, su organización y análisis, así como de las predicciones que, a partir de estos datos, pueden hacerse. Es decir, permite el tratamiento sistemático de datos, la búsqueda de las conclusiones de los mismos y la toma de decisiones tras su análisis.

Existen dos clases de Estadística, según el problema que se estudie y el método utilizado:

- Estadística Descriptiva:

Se ocupa de tomar los datos de un conjunto, organizarlos en tablas o en representaciones gráficas y del cálculo de unos números que nos informen de manera global del conjunto estudiado. (Nos centraremos en ésta)

- Estadística Inferencial:

Trata sobre la elaboración de conclusiones para la población, partiendo de los resultados de una muestra y del grado de fiabilidad de estas conclusiones.

2.0.- Conceptos básicos

El vocabulario específico que aparece en cualquier estudio estadístico es:

- ***Población:*** Es el conjunto formado por todos los elementos que existen para el estudio de un determinado fenómeno.
- ***Individuo u objeto:*** Es cada uno de los elementos de la población.
- ***Muestra:*** Es el subconjunto extraído de la población ya sea por necesidad o por obligación, cuyo estudio sirve ara inferir características de toda la población.
- ***Tamaño de la muestra (N):*** Es el número de individuos que componen la muestra (en algunos casos coincide con el de la población)

2.1.- Variables o caracteres estadísticos

Es cada una de las cualidades o propiedades que permiten clasificar a los individuos de una población (objeto de estudio estadístico). Los valores de la variable se suelen representar por $x_1, x_2,\ldots\ldots,x_n$.

Las clases de variables estadísticas que aparecen en cualquier estudio estadístico son:

- ***Variables cualitativas:*** Aquellas que no se pueden medir y se describen con palabras. Por ejemplo: *Raza de un perro, Estado civil de una persona, color favorito….*
- ***Variables cuantitativas:*** Aquellas que se pueden medir y expresar con números. A su vez, las variables cuantitativas pueden ser:
  - ✓ ***Discretas:*** Aquellas que pueden tomar solamente un número finito de valores numéricos aislados. Por ejemplo: *número de hermanos, Número de asignaturas aprobadas, Frutos de un árbol, etc.*
  - ✓ ***Continuas:*** Aquellas que pueden tomar cualquier valor en un intervalo dado. Por ejemplo: Estatura de los alumnos, el peso de cada una de los frutos de un árbol, número de teléfono….

Cuando se realiza un estudio estadístico, todos los conceptos de los que hemos hablado se suelen reflejar en la que se llama ***ficha técnica***.

3.0.- Tablas Estadísticas

Una vez recogidos los datos, mediante encuestas por ejemplo, se suelen ordenar. La forma usual de hacerlo es realizando un recuento y posteriormente formar una tabla con distintas columnas.

3.1.- Primera columna, x_i:

Está formada por los distintos valores que puede tomar la variable estadística (ordenados de menor a mayor si se trata de una variable cuantitativa). Si la variable a estudio es cuantitativa continua, o discreta pero con muchos valores distintos, los datos deben agruparse en clases o intervalos. En estos casos, esta primera columna se divide en dos, una para los intervalos y otra para la marca de clase, que es el valor medio de cada intervalo y se calcula como la semisuma de los extremos del intervalo $\left(\frac{a'+b'}{2}\right)$

Para ***construir los intervalos*** hay que tener en cuenta:

- Se localizan los valores ***extremos a y b*** y se halla su diferencia (recorrido) $r = b - a$
- Se ***decide el número de intervalos*** que se quiere formar, teniendo en cuenta la cantidad de datos que se poseen. Es conveniente que el número de intervalos esté ***entre 6 y 15.***
- Se busca un ***número entero un poco mayor que el recorrido*** y que sea múltiplo del número de intervalos, r'.
- ***Se forman los intervalos***, de modo que el extremo inferior del primero sea algo menor que a y el extremo superior del último sea algo mayor que b. Es deseable que los extremos de los intervalos no coincidan con ninguno de los datos, pero si esto ocurriera, se forman los intervalos teniendo presente que el límite inferior de una clase pertenece al intervalo, pero el límite superior no pertenece intervalo, se cuenta en el siguiente intervalo.

Ejemplo 1: ***En una maternidad se han tomado los pesos (en kilogramos) de 50 recién nacidos:***

2,8	3,2	3,8	2,5	2,7	3,7	1,9	2,6	3,5	2,3
3,0	2,6	1,8	3,3	2,9	2,1	3,4	2,8	3,1	3,9
2,9	3,5	3,0	3,1	2,2	3,4	2,5	1,9	3,0	2,9
2,4	3,4	2,0	2,6	3,1	2,3	3,5	2,9	3,0	2,7
2,9	2,8	2,7	3,1	3,0	3,1	2,8	2,6	2,9	3,3

INTERVALOS	MARCA DE CLASE (x_i)	f_i
1,65 - 2,05	1,85	4
2,05 - 2,45	2,25	5
2,45 - 2,85	2,65	13
2,85 - 3,25	3,05	17
3,25 - 3,65	3,45	8
3,65 - 4,05	3,85	3
		50

a) ¿Cuál es la variable y de qué tipo es?

Variable: peso de los recién nacidos. Tipo: cuantitativa continua.

b) Construye una tabla con los datos agrupados en 6 intervalos de 1,65 a 4,05.

Localizamos los valores extremos: 1,8 y 3,9. Por tanto, el Recorrido será: R = 3,9 – 1,8 = 2,1

3.2.- Segunda columna, f_i:

Se sitúan en ella las frecuencias absolutas: f_i, que es el ***número total de veces que aparece el valor x_i*** de la variable estadística, o el número de valores de la variable que hay en un determinado intervalo.

La suma de todas las frecuencias absolutas es el ***tamaño de la muestra*** o ***población*** a estudio (***N***), es decir:

$$f_1 + f_2 + f_3 + \ldots\ldots\ldots + f_n = \sum_{i=1}^{n} f_i = N$$

3.3.- Tercera Columna, F_i:

En ella se colocan las ***frecuencias absolutas acumuladas***: *Fi* que son las sumas de todas las frecuencias absolutas correspondientes a los valores anteriores a x_i y la suya propia.

$$F_i = f_1 + f_2 + f_3 + ... + f_i$$

3.4.- Cuarta columna, h_i:

La colocamos cuando nos interesa saber cuál es la ***proporción del número de individuos*** con un valor determinado respecto al total. Para ello calculamos la frecuencia relativa o proporción: hi que es el cociente que resulta de dividir su frecuencia absoluta (f_i) entre el número total, N, de individuos.

$$h_i = \frac{f_i}{N}$$ y además siempre cumple que: $0 \le h_i \le 1$

La suma de todas las frecuencias relativas es la unidad.

$$h_1 + h_2 + h_3 + ... + h_n = \sum_{i=1}^{n} h_i = 1$$

3.5.- Quinta columna, H_i:

En ella se colocan las ***frecuencias relativas acumuladas***: H_i que son las sumas de todas las frecuencias relativas correspondientes a los valores anteriores a x_i y la suya propia.

$$H_i = h_1 + h_2 + h_3 + ... + h_i$$

3.6.- Sexta columna, p_i:

Se calculan los ***porcentajes***, que es el tanto por ciento que representa el valor xi respecto del total. Se calcula multiplicando la frecuencia relativa h_i por 100 (o mediante una regla de tres)

$$p_i = h_i \cdot 100$$ Siempre cumple que: $0 \le p_i \le 100$

La suma de todos los porcentajes es 100:

$$p_1 + p_2 + p_3 + ... + p_n = \sum_{i=1}^{n} p_i = 100$$

3.7.- Séptima columna, $D_{\bar{x}}$

En ella escribiremos la diferencia en valor absoluto entre la medida x_i y la media, que luego nos servirá para calcular la desviación media.

$$|x_i - \bar{x}|$$

Si se trata de una tabla de datos agrupados en intervalos, entonces en esta columna aparecería:

$$|x_i - \bar{x}| \cdot f_i$$

Además de estas 7 columnas vamos a utilizar otras dos, que nos vendrán muy bien para calcular algunos de los parámetros estadísticos.

Estas columnas van a ser, la columna séptima, en la que aparecerá el producto $x_i \cdot f_i$ y la columna octava en la que aparece el producto $x_i^2 \cdot f_i$.

Con todo esto, nuestra tabla de recogida y análisis de datos sería de la forma:

Intervalos	x_i	f_i	F_i	h_i	H_i	P_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$	$\lvert x_i - \bar{x} \rvert \cdot f_i$
					1				
Totales:		$N = \sum f_i$		1		100	$\sum x_i \cdot f_i$	$\sum x_i^2 \cdot f_i$	$\sum \lvert x_i - \bar{x} \rvert \cdot f_i$

Se puede colocar una columna formada por las frecuencias porcentuales acumuladas, pero no es muy habitual. A la hora de realizar un estudio estadístico, no es obligatorio que utilicemos todas las columnas, sino solo aquellas que sean necesarias para dicho estudio.

4.0.- Parámetros estadísticos

Tomada una muestra unidimensional $(x_1, x_2, \ldots, x_n)$ de tamaño n, interesa reducir la información encerrada en ella a sólo unos pocos parámetros, llamados ***parámetros estadísticos***.
Hay tres tipos parámetros estadísticos:

- ✓ De centralización.
- ✓ De posición
- ✓ De dispersión.

Para estudiar algunos de los parámetros estadísticos más importantes, nos vamos a ayudar del siguiente ejemplo:

Ejemplo 2: ***Las urgencias atendidas en las 7 primeras horas del día en un centro de salud son:***

1 5 3 6 4 5 3

4.1- Medidas de Centralización

Las medidas de centralización resumen la información de la muestra:

4.1.1.- La Media Aritmética:

La media aritmética o ***media***, $\bar{x}$, es el más conocido e intuitivo, siendo su objeto localizar alrededor de qué punto se sitúan todas las observaciones. Se calcula sumando todas las medidas y dividiendo dicho resultado entre el número de medidas. Su cálculo es bien sencillo:

$$\bar{x} = \frac{\sum f_i \cdot x_i}{\sum f_i} = \frac{\sum f_i \cdot x_i}{N}$$

En nuestro ejemplo, la media de urgencias atendidas será: $\bar{x} = \frac{\sum f_i \cdot x_i}{N} = \frac{27}{7} = 3{,}86$ **personas**

4.1.2.- La Mediana:

La Mediana, Me, como la media, es un parámetro de localización, siendo su objeto resumir en una sola cantidad los valores muestrales. Se define como el valor numérico que queda en el centro cuando se ordena toda la muestra. (Medida central).

En nuestro ejemplo, la mediana de las urgencias atendidas será:

Las ordenamos de menor a mayor: 1-3-3-**4**-5-5-6

Y Vemos que **la mediana Me= 4.**

Cálculo de la mediana:

1. Ordenamos los datos de menor a mayor.
2. Si la serie tiene un número impar de medidas la mediana es la puntuación central de la misma.

$$2, 3, 4, 4, 5, 5, 5, 6, 6 \qquad Me = 5$$

3. Si la serie tiene un número par de puntuaciones la mediana es la media entre las dos puntuaciones centrales.

$$7, 8, 9, 10, 11, 12 \qquad Me = 9.5$$

Cálculo de la mediana para datos agrupados

La mediana se encuentra en el intervalo donde la frecuencia acumulada llega hasta la mitad de la suma de las frecuencias absolutas.

Es decir, tenemos que buscar el intervalo en el que se encuentre $\frac{N}{2}$

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

Donde:

- ✓ L_i es el límite inferior de la clase donde se encuentra la mediana.
- ✓ $\frac{N}{2}$ es la semisuma de las frecuencias absolutas.
- ✓ F_{i-1} es la frecuencia acumulada anterior a la clase de la mediana.
- ✓ a_i es la amplitud de la clase (longitud del intervalo).

La mediana es independiente de las amplitudes de los intervalos.

Ejemplo 3: ***Calcular la mediana de una distribución estadística que viene dada por la siguiente tabla:***

	fi	Fi
[60, 63)	5	5
[63, 66)	18	23
[66, 69)	42	65
[69, 72)	27	92
[72, 75)	8	100

Clase de la mediana: [66, 69)
El Límite inferior de la clase es: 66
La semisuma de las frecuencias absolutas es: N=100/2 = 50
La frecuencia acumulada anterior es 23.
La amplitud de la clase es 3.

Con todos estos datos: $Me = L_i + \frac{\frac{N}{2} - F_{i-1}}{f_i} \cdot a_i = 66 + \frac{50 - 23}{42} \cdot 3 = 67,93$

4.1.3.- La Moda:

La Moda, Mo, es el dato que tiene mayor frecuencia absoluta, es decir el que más se repite. Si la variable es continua hablamos de intervalo modal. Podría ocurrir que hubiera varias modas, porque hubiera datos que se repiten lo mismo.

En nuestro ejemplo, las modas serían 3 y 5 porque ambas se repiten dos veces. **Mo=3 y 5**

Se puede hallar la moda para variables cualitativas y cuantitativas.

Hallar la moda de la distribución: 2, 3, 3, 4, 4, 4, 5, 5 **Mo = 4**

Si en un grupo hay dos o varias puntuaciones con la misma frecuencia y esa frecuencia es la máxima, la distribución es bimodal o multimodal, es decir, tiene varias modas.

1, 1, 1, 4, 4, 5, 5, 5, 7, 8, 9, 9, 9 **Mo= 1, 5, 9**

Cuando todas las puntuaciones de un grupo tienen la misma frecuencia, no hay moda.

2, 2, 3, 3, 6, 6, 9, 9 **No hay moda**

Si dos puntuaciones adyacentes tienen la frecuencia máxima, la moda es el promedio de las dos puntuaciones adyacentes.

0, 1, 3, 3, 5, 5, 7, 8 **Mo = 4**

Cálculo de la moda para datos agrupados:

- ✓ ***Si todos los intervalos tienen la misma amplitud.***

$$Mo = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \cdot a_i$$

Donde:

- ✓ L_i es el límite inferior de la clase modal.
- ✓ f_i es la frecuencia absoluta de la clase modal.
- ✓ f_{i-1} es la frecuencia absoluta inmediatamente inferior a la clase modal.
- ✓ f_{i+1} es la frecuencia absoluta inmediatamente posterior a la clase modal.
- ✓ a_i es la amplitud de la clase.

También se utiliza otra fórmula de la moda que da un valor aproximado de ésta: $Mo = L_i + \frac{f_{i+1}}{f_{i-1} + f_{i+1}} \cdot a_i$

Ejemplo 4: ***Calcular la moda de una distribución estadística que viene dada por la siguiente tabla:***

	fi
[60, 63)	5
[63, 66)	18
[66, 69)	42
[69, 72)	27
[72, 75)	8
	N=100

La moda está en la clase [66, 69)
El límite inferior de este intervalo es: 66
F_i=42 f_{i-1}=18 f_{i+1}=27
La amplitud de la clase es 3.
Así que si sustituimos en ambas expresiones:

$$Mo = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \cdot a_i = 66 + \frac{(42-18)}{(42-18) + (42-27)} \cdot 3 = 67{,}846$$

Y con la fórmula aproximada:

$$Mo = L_i + \frac{f_{i+1}}{f_{i-1} + f_{i+1}} \cdot a_i = 66 + \frac{27}{18+27} \cdot 3 = 67{,}8$$

Y como vemos, el error cometido es prácticamente despreciable.

✓ ***Si los intervalos tienen amplitudes distintas.***

En primer lugar, tenemos que hallar las alturas. $h_i = \frac{f_i}{a_i}$

La clase modal es la que tiene mayor altura.

Y para calcularla utilizaremos la expresión: $Mo = L_i + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} \cdot a_i$

La fórmula de la moda aproximada cuando existen distintas amplitudes es: $Mo = L_i + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot a_i$

Ejemplo 5: ***En la siguiente tabla se muestra las calificaciones (suspenso, aprobado, notable y sobresaliente) obtenidas por un grupo de 50 alumnos. Calcular la moda.***

	fi	hi
[0, 5)	15	3
[5, 7)	20	10
[7, 9)	12	6
[9, 10)	3	3
	N=50	

Calculamos la clase que tiene mayor altura mediante: $h_i = \frac{f_i}{a_i}$

$$h_1 = \frac{f_1}{a_1} = \frac{15}{5} = 3 \qquad h_2 = \frac{f_2}{a_2} = \frac{20}{2} = 10 \qquad h_3 = \frac{f_3}{a_3} = \frac{12}{2} = 6 \qquad h_4 = \frac{f_4}{a_4} = \frac{3}{1} = 3$$

Por tanto, el intervalo modal o la clase modal es [5, 7)

Y la moda será:

$$Mo = L_i + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} \cdot a_i = 5 + \frac{10-3}{(10-3)+(10-6)} \cdot 2 = 6{,}27$$

Si utilizamos la fórmula aproximada:

$$Mo = L_i + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot a_i = 5 + \frac{6}{3+6} \cdot 2 = 6{,}33$$

4.2.- Medidas de Posición

Las medidas de posición son valores de la variable que informan del lugar que ocupa un dato dentro del conjunto ordenado de valores. Para calcular estas medidas la variable debe ser cuantitativa.

4.2.1.- Los Cuartiles:

Los Cuartiles Q_1, Q_2 y Q_3, son medidas que dividen el conjunto de datos ordenados en cuatro partes iguales, es decir, en cada tramo está el 25% de los datos recogidos en el estudio.

- ✓ Q_1, Q_2 y Q_3 determinan los valores correspondientes al 25%, al 50% y al 75% de los datos.
- ✓ Q_2 coincide con la mediana.

Cálculo de los cuartiles:

1. Ordenamos los datos de menor a mayor.
2. Buscamos el lugar que ocupa cada cuartil mediante la expresión: $\frac{k \cdot N}{4}$ *con* $k = 1,2,3$

Cálculo de los cuartiles para Número impar de datos:

$X_i = 2, 5, 3, 6, 7, 4, 9$

2, 3, 4, 5, 6, 7, 9
↓ ↓ ↓
Q_1 Q_2 Q_3

Cálculo de los cuartiles para Número par de datos:

2, 5, 3, 4, 6, 7, 1, 9

1, 2, 3, 4, 5, 6, 7, 9
2.5 4.5 6.5
↓ ↓ ↓
Q_1 Q_2 Q_3

Cálculo de los cuartiles para datos agrupados:

En primer lugar, buscamos la clase donde se encuentra $\frac{k\cdot N}{4}$ *con* $k=1,2,3$, en la tabla de las frecuencias acumuladas.

Para calcular los 3 cuartiles nos ayudaremos de la fórmula: $Q_k = L_i + \frac{\frac{k\cdot N}{4} - F_{i-1}}{f_i}\cdot a_i \qquad k=1,2,3$

Donde:

- ✓ Li es el límite inferior de la clase modal.
- ✓ F_i es la frecuencia absoluta acumulada de la clase modal.
- ✓ F_{i-1} es la frecuencia absoluta acumulada inmediatamente inferior a la clase modal.
- ✓ f_i es la frecuencia absoluta de la clase modal.
- ✓ ai es la amplitud de la clase.

***Ejemplo:* Calcular los cuartiles de la distribución de la siguiente tabla:**

	fi	Fi
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
	N=65	

Cálculo del primer cuartil: $\frac{65}{4} = 16,25$ por tanto está en el intervalo [60,70) y utilizando la fórmula:

$$Q_k = L_i + \frac{\frac{k\cdot N}{4} - F_{i-1}}{f_i}\cdot a_i \quad \rightarrow \quad Q_1 = 60 + \frac{\frac{1\cdot 65}{4} - 8}{10}\cdot 10 = 68,25$$

Cálculo del segundo cuartil: $\frac{65\cdot 2}{4} = 32,5$ por tanto está en el intervalo [70,80) y utilizando la fórmula:

$$Q_k = L_i + \frac{\frac{k\cdot N}{4} - F_{i-1}}{f_i}\cdot a_i \quad \rightarrow \quad Q_2 = 70 + \frac{\frac{2\cdot 65}{4} - 18}{16}\cdot 10 = 79,06$$

Cálculo del tercer cuartil: $\frac{65\cdot 3}{4} = 48,75$ por tanto está en el intervalo [90,100) y utilizando la fórmula:

$$Q_k = L_i + \frac{\frac{k\cdot N}{4} - F_{i-1}}{f_i}\cdot a_i \quad \rightarrow \quad Q_3 = 90 + \frac{\frac{3\cdot 65}{4} - 48}{10}\cdot 10 = 90,75$$

4.2.2.- Los Deciles:

Los deciles son los nueve valores que dividen la serie de datos en diez partes iguales y dan los valores correspondientes al 10%, al 20%... y al 90% de los datos.

- ✓ D_{50} coincide con la mediana.

El cálculo de los Deciles no tiene mucho sentido en un estudio estadístico discreto.

Cálculo de los Deciles:

En primer lugar, buscamos la clase donde se encuentra, $\frac{k\cdot N}{10}$ *con* $k=1,2,3,.....,9$ en la tabla de las frecuencias acumuladas.

Y después nos ayudaremos de la fórmula:

$$D_k = L_i + \frac{\frac{k\cdot N}{10} - F_{i-1}}{fi}\cdot a_i \quad con \quad k=1,2,3,.....,9$$

Donde:

- ✓ Li es el límite inferior de la clase donde se encuentra el percentil.
- ✓ N es la suma de las frecuencias absolutas.
- ✓ F_{i-1} es la frecuencia acumulada anterior a la clase del percentil.
- ✓ a_i es la amplitud de la clase.

***Ejemplo:* Calcular el decil 60 de la siguiente distribución:**

	fi	Fi
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
	N=65	

Decil 60: Calculamos donde se encuentra: $\frac{6\cdot 65}{10} = 39$ por tanto está en el intervalo [80,90)

$$D_k = L_i + \frac{\frac{k\cdot N}{10} - F_{i-1}}{fi}\cdot a_i \quad \rightarrow \quad D_6 = 80 + \frac{\frac{6\cdot 65}{10} - 34}{14}\cdot 10 = 83{,}57$$

4.2.3.- Los Percentiles:

Los Percentiles o Centiles, P_k, son medidas que dividen el conjunto de todos los datos en 100 partes iguales.

- ✓ P_{25}, P_{50} y P_{75} determinan los valores correspondientes al 25%, al 50% y al 75% de los datos.
- ✓ P_{50} coincide con la mediana.

El cálculo de los percentiles no tiene mucho sentido en un estudio estadístico discreto.

Cálculo de los percentiles:

En primer lugar, buscamos la clase donde se encuentra, $\frac{k\cdot N}{100}$ *con* $k=1,2,3,.....,99$ en la tabla de las frecuencias acumuladas.

Y después nos ayudaremos de la fórmula: $P_k = L_i + \frac{\frac{k\cdot N}{100} - F_{i-1}}{fi}\cdot a_i$ *con* $k=1,2,3,.....,99$

Donde:

- ✓ Li es el límite inferior de la clase donde se encuentra el percentil.
- ✓ N es la suma de las frecuencias absolutas.
- ✓ F_{i-1} es la frecuencia acumulada anterior a la clase del percentil.
- ✓ a_i es la amplitud de la clase.

***Ejemplo:* Calcular el percentil 35, 60 y 95 de la distribución de la tabla:**

	fi	Fi
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
	N=65	

Percentil 35: Calculamos donde se encuentra: $\frac{35\cdot 65}{100} = 22,75$ por tanto está en el intervalo [70,80)

$$P_k = L_i + \frac{\frac{k\cdot N}{100} - F_{i-1}}{fi}\cdot a_i \quad \rightarrow \quad P_{35} = 70 + \frac{\frac{36\cdot 65}{100} - 18}{16}\cdot 10 = 72,97$$

Percentil 60: Calculamos donde se encuentra: $\frac{60\cdot 65}{100} = 39$ por tanto está en el intervalo [80,90)

$$P_k = L_i + \frac{\frac{k\cdot N}{100} - F_{i-1}}{fi}\cdot a_i \quad \rightarrow \quad P_{60} = 80 + \frac{\frac{60\cdot 65}{100} - 34}{14}\cdot 10 = 83,57$$

Percentil 95: Calculamos donde se encuentra: $\frac{95\cdot 65}{100} = 61,75$ por tanto está en el intervalo [100,110)

$$P_k = L_i + \frac{\frac{k\cdot N}{100} - F_{i-1}}{fi}\cdot a_i \quad \rightarrow \quad P_{60} = 100 + \frac{\frac{95\cdot 65}{100} - 58}{5}\cdot 10 = 107,5$$

4.3- Medidas de Dispersión

Las medidas de dispersión nos informan sobre cuanto se alejan del centro los valores de la distribución.

Las medidas de dispersión son:

- ✓ Rango
- ✓ Desviación Media
- ✓ Varianza
- ✓ Desviación típica
- ✓ Coeficiente de variación

4.3.1.- Rango o Recorrido:

El rango, *r*, (también conocido por recorrido o amplitud) es la diferencia entre el mayor y el menor de los datos de una distribución estadística, y que indica qué extensión de la recta de los números ocupan los datos de nuestra muestra.

4.3.2.- Desviación media:

La desviación respecto a la media es la diferencia en valor absoluto entre cada valor de la variable estadística y la media aritmética.

La desviación media se representa por signo $D_{\bar{x}}$ y la calcularemos mediante la expresión:

$$D_{\bar{x}} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + ... + |x_n - \bar{x}|}{N} = \frac{\sum |x_i - \bar{x}|}{N}$$

Donde:

- ✓ $\bar{x}$ es la media aritmética
- ✓ x_i hace referencia a la medida, tanto si es discreta como si es continua, solo que en la distribución continua representará la marca de clase del intervalo correspondiente.

Ejemplo: ***Calcular la desviación media de la distribución:***

	xi	***fi***	***xi · fi***	***|x - x|***	***|x - x| · fi***
[10, 15)	12.5	3	37.5	9.286	27.858
[15, 20)	17.5	5	87.5	4.286	21.43
[20, 25)	22.5	7	157.5	0.714	4.998
[25, 30)	27.5	4	110	5.714	22.856
[30, 35)	32.5	2	65	10.714	21.428
		21	457.5		98.57

Primero calculamos la media aritmética: $\bar{x} = \frac{\sum x_i \cdot f_i}{N} = \frac{457,5}{21} = 21,786$

Y después la desviación media: $D_{\bar{x}} = \frac{\sum |x_i - \bar{x}|}{N} = \frac{98,57}{21} = 4,69$

4.3.3.- Varianza:

La Varianza, *Var*, mide la dispersión alrededor de la media; cuanto más pequeña sea, más concentrados estarán los puntos alrededor de:

$$Var = \frac{\sum f_i \cdot x_i^2}{N} - \bar{x}^2$$

4.3.4.- Desviación típica:

Para evitar tener que trabajar con las unidades cuadradas de la varianza, se extrae su raíz cuadrada, con lo se obtiene la desviación típica, σ .

$$\sigma = \sqrt{Var} = \sqrt{\frac{\sum f_i \cdot x_i^2}{N} - \bar{x}^2}$$

4.3.5.- Coeficiente de Variación:

Las dispersiones de aquellas distribuciones que tienen medias aritméticas diferentes o cuyos datos vienen dados en unidades diferentes se pueden comparar mediante el coeficiente de variación, que se define como el cociente entre la desviación típica y la media.

$$C.V. = \frac{\sigma}{\overline{x}}$$

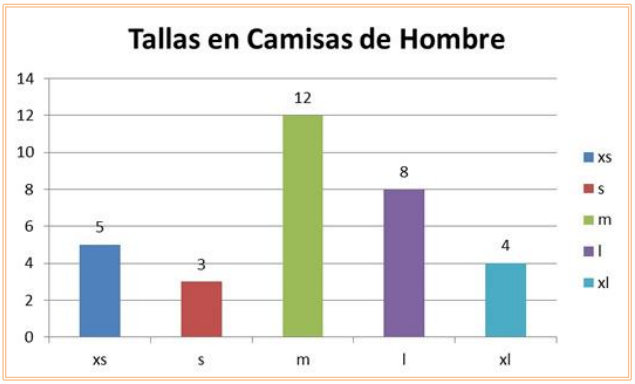
5.0.- Gráficos estadísticos

A la hora de presentar los resultados de un estudio estadístico, es muy útil y recomendable utilizar alguno u algunos gráficos estadísticos.

5.1.- Gráfico de barras:

El gráfico de barras, como su nombre lo indica, está constituido por barras rectangulares de igual ancho, conservando la misma distancia de separación entre sí. Se utiliza básicamente para mostrar y comparar frecuencias de variables cuantitativas o comportamientos en el tiempo, cuando el número de ítems es reducido.

Para elaborarlo deberíamos:



- ✓ Utilizar un sistema de coordenadas rectangulares y se llevan al eje de las "x" los valores que toma la variable en estudio y en el eje de las "y" se colocan las frecuencias de cada barra.
- ✓ Luego se construyen los rectángulos, tomando como base al eje de las abscisas, cuya altura será igual a cada una de las diferentes frecuencias que presentan las variables en estudio.
- ✓ La magnitud con que viene expresada la variable se observa en la longitud de las barras (rectángulos). Es importante destacar que solamente la longitud de las barras y no su anchura es lo que denota la diferencia de magnitud entre los valores de la variable.
- ✓ Todas las barras tienen que tener una anchura igual, separadas entre sí, preferiblemente por una longitud igual a la mitad del ancho de estas o distancias iguales entre barras.
- ✓ Las barras se pueden graficar tanto verticalmente como horizontalmente. Se pueden elaborar barras compuestas y barras agrupadas.

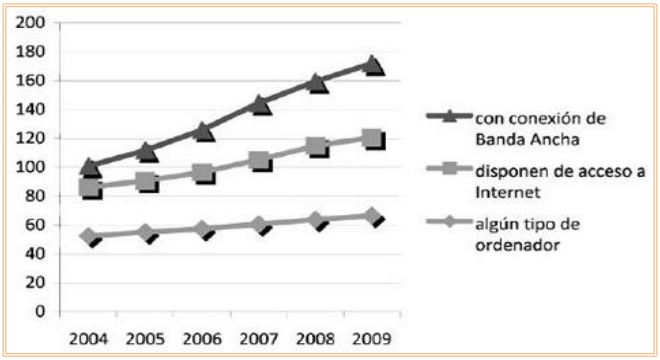
En este tipo de gráfico se pueden utilizar:

- Barras simples: Compara valores entre categorías de una variable
- Barras dobles: Compara valores entre categorías de dos variables
- Barras múltiples: Compara valores entre categorías de dos o más variables.
- Barras verticales: Las categorías de la variable deben ubicarse en el eje x.
- Barras horizontales: Las categorías de la variable deben ubicarse en el eje y.
- Barras Aplicadas: Compara entre categorías el aporte de cada valor en el total.

5.2.- Gráfico de líneas o Tendencias:

Usado básicamente para mostrar el comportamiento de una variable cuantitativa a través del tiempo. El gráfico de líneas consiste en segmentos rectilíneos unidos entre sí, los cuales resaltan las variaciones de la variable por unidad de tiempo.

Cuando se tienen varias variables a representar, con el fin de establecer comparaciones entre ellas (siempre que su unidad de medida sea la misma); se utiliza plasmarlos en un solo gráfico, el cual es el resultado de representar varias variables en un mismo plano. A este tipo de gráfico se le llama gráfico de líneas compuesto.



Criterios para elaborar un gráfico de líneas:

1. La utilización de la escala que se utilizará en el plano cartesiano puede variar tomando en cuenta el fenómeno que se va a graficar. No es necesario que las abscisas (ejes x) y las ordenadas (eje y) del plano cartesiano lleven la misma escala; sin embargo, cuando las magnitudes de las variables no se diferencian sustancialmente es recomendable utilizar escalas iguales para obtener un gráfico con mayor precisión.

2. Cuando una de las variables en estudio se inicia con valores muy altos es recomendable no comenzar el eje por el origen cartesiano sino por un valor próximo o por el mismo valor por donde comienza la variable.

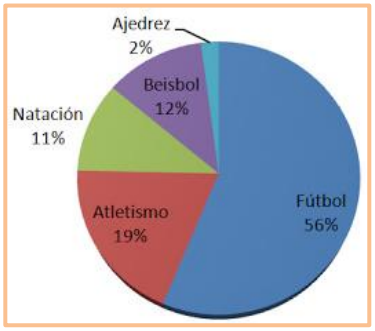
3. Es costumbre representar en el eje de las x del plano cartesiano la variable independiente del estudio que se realiza y en el eje de las y la variable dependiente. En aquellos casos que se dificulta distinguir el tipo de variable se recomienda colocar en la ordenada del plano cartesiano las frecuencias de las variables en estudio y sobre la abscisa la variable cronológica (años, semanas, días, horas, etc.)

5.3.- Gráfico de sectores circulares:

Usualmente llamado gráfico de torta, debido a su forma característica de una circunferencia dividida en sectores, por medio de radios que dan la sensación de un pastel cortado en porciones.

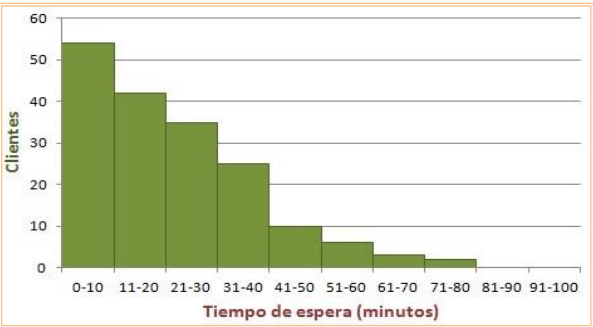
Se usa para representar variables cualitativas en porcentajes o cifras absolutas cuando el número de ítems no es superior a 5 y se quiere resaltar uno de ellos.

En el ejemplo de la derecha podemos observar el gráfico de sectores en el que aparecen los datos de los deportes preferidos de los españoles.



5.4.- Histograma de frecuencias:

El histograma es un diagrama en forma de columna, muy parecido a los gráficos de barras. Se define como un conjunto de rectángulos paralelos, en el que la base representa la clase de la distribución y su altura la magnitud que alcanza la frecuencia de la clase correspondiente. Son barras rectangulares levantadas sobre el eje de las abscisas del plano cartesiano utilizando escalas adecuadas para los valores que asume la variable en la distribución de frecuencia.

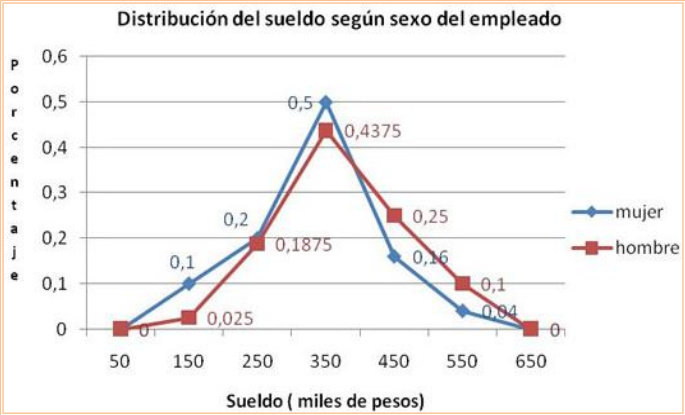


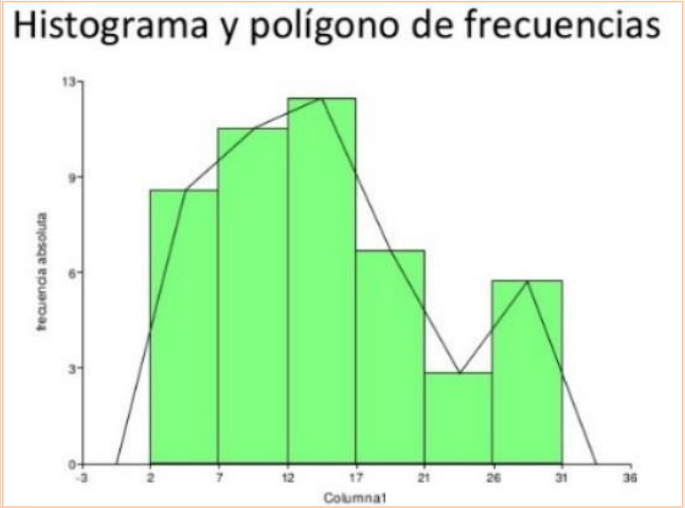
El ancho de la base de los rectángulos es proporcional a cada clase de la distribución, de tal manera que, cuando la distribución tiene clases de igual tamaño, el tamaño de todos los rectángulos tendrá bases iguales.
Los lados del rectángulo se levantan sobre los puntos del eje de las x que corresponden a los límites de cada clase y la longitud de los mismos será igual a la frecuencia que tenga esa clase, los lados por lo tanto corresponden a la frecuencia de cada clase de la distribución de frecuencia.

Cuando se elaboran gráficas estadísticas en el plano cartesiano es recomendable que en el eje de las ordenadas se representen las frecuencias y el eje de abscisas las variables independientes.

5.5.- Polígono de frecuencias:

Se utiliza básicamente para mostrar la distribución de frecuencias de variables cuantitativas, para construirlo se toma la marca de clase que coincide con el punto medio de cada rectángulo de un histograma.





Pasos para elaborar un polígono de frecuencias:

a) Se dibuja un plano cartesiano.
b) Se traza sobre el eje de las abscisas, a distancias iguales, los puntos medios de las diferentes clases de la distribución de frecuencias.
c) Se levantan perpendiculares por cada una de las marcas de clase, con una longitud igual a la frecuencia de cada una de las clases que integran la distribución de frecuencia. Al final de cada perpendicular se marca un punto.
d) Los puntos resultantes se unen por medio de una línea recta obteniéndose una línea poligonal.
e) Con la finalidad de cerrar la línea poligonal se agrega una clase imaginaria con frecuencia cero a cada extremo de la distribución de frecuencia, por tales motivos ambos extremos del polígono se cortan con el eje de las abscisas.

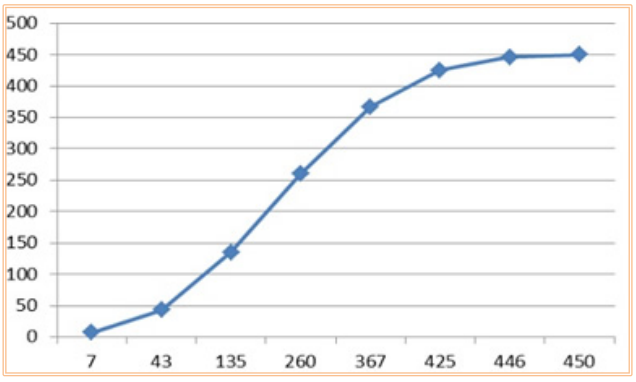
También se puede elaborar un polígono de frecuencia después de haber graficado un histograma; si se determina el punto medio de cada rectángulo de un histograma y esos puntos medios se unen por medio de segmentos de recta dan como resultado el polígono de frecuencia.
(Figura de la izquierda)

5.6.- Histograma de frecuencias acumuladas:

Se utiliza básicamente para mostrar la distribución de frecuencias acumulada de variables cuantitativas. Es una gráfica que se elabora con los valores de las frecuencias acumuladas (menor que y mayor que) y los límites de las clases de una distribución de frecuencia.

El polígono de frecuencias acumuladas es conocido comúnmente como ***ojiva***.

La ojiva es una representación gráfica que consiste en una línea, que puede ser ascendente o descendente y se utiliza para representar las distribuciones de frecuencias acumuladas menor que y mayor que, según los datos utilizados. En los estudios de análisis estadísticos la ojiva es de gran utilidad porque permite obtener con gran aproximación cierta información requerida, en un momento determinado.



6.0.- Ejercicios Resueltos:

1. Sea la siguiente distribución estadística:

x_i	f_i
[10, 15)	3
[15, 20)	5
[20, 25)	7
[25, 30)	4
[30, 35)	2

Hallar:

a) La moda, mediana y media.
b) El rango, desviación media, varianza, desviación típica y C.V.
c) Los cuartiles 1º y 3º.
d) Los percentiles 30, 70 y 97.

Lo primero es completar la tabla con las columnas necesarias:

x_i		fi	Fi	h_i	H_i	$xi \cdot fi$	$xi^2 \cdot fi$	$\lvert x_i - \bar{x}\rvert \cdot f_i$
Intervalo	*xi*							
[10, 15)	12.5	3	3	0,143	0,143	37.5	468.75	27.857
[15, 20)	17.5	5	8	0,238	0,381	87.5	1537.3	21.429
[20, 25)	22.5	7	15	0,333	0,714	157.5	3543.8	5
[25, 30)	27.5	4	19	0,190	0,904	110	3025	22.857
[30, 35)	32.5	2	21	0,095	1	65	2112.5	21.429
		N=21				$\sum x_i \cdot f_i = 457{,}5$	10681.25	$\sum \lvert x_i - \bar{x}\rvert \cdot f_i = 98{,}571$

Moda: La moda, que es el valor que más se repite, se encuentra en el intervalo [20, 25), así que:

$$Mo = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \cdot a_i = 20 + \frac{(7-5)}{(7-5)+(7-4)} \cdot 5 = 22$$

Mediana: La mediana, que es la medida central, N/2=10,5, se encuentra en el intervalo [20,25), por tanto:

$$Me = L_i + \frac{\frac{N}{2} - F_{i-1}}{f_i} \cdot a_i = 20 + \frac{10{,}5 - 8}{7} \cdot 5 = 21{,}786$$

Media: La media aritmética viene dada por: $\bar{x} = \frac{\sum f_i \cdot x_i}{\sum f_i} = \frac{\sum f_i \cdot x_i}{N} = \frac{457{,}5}{21} = 21{,}79$

Rango: El rango es la diferencia entre el mayor valor y el menor: $r = x_{max} - x_{min} = 35 - 10 = 25$

Desviación media: viene dada por: $D_{\bar{x}} = \frac{\sum \lvert x_i - \bar{x}\rvert}{N} = \frac{98{,}71}{21} = 4{,}694$

Varianza: La calculamos mediante la expresión $Var = \frac{\sum f_i \cdot x_i^2}{N} - \bar{x}^2 = \frac{10681{,}25}{21} - 21{,}79^2 = 33{,}827$

La **Desviación típica** es la raíz cuadrada de la varianza: $\sigma = \sqrt{Var} = \sqrt{\frac{\sum f_i \cdot x_i^2}{N} - \bar{x}^2} = \sqrt{33{,}827} = 5{,}816$

El **coeficiente de variación** es: $C.V. = \frac{\sigma}{\bar{x}} = \frac{5{,}816}{21{,}79} = 0{,}267$

Los cuartiles vienen dados por: $Q_k = L_i + \frac{\frac{k \cdot N}{4} - F_{i-1}}{f_i} \cdot a_i$

El primer cuartil, Q_1, está en 21/4=5,25 está en la clase [15, 20) y su valor es: $Q_1 = 15 + \frac{\frac{1 \cdot 21}{4} - 3}{5} \cdot 5 = 17{,}25$

El tercer cuartil, Q_3, está en $21\cdot 3/4 = 15{,}75$ está en la clase [25, 30) y su valor es: $Q_3 = 25 + \dfrac{\frac{3\cdot 21}{4} - 15}{4}\cdot 5 = 25{,}936$

Los Percentiles vienen dados por la expresión: $P_k = L_i + \dfrac{\frac{k\cdot N}{100} - F_{i-1}}{fi}\cdot a_i$

El percentil 30 $\dfrac{30\cdot 21}{100} = 6{,}3$ se encuentra en la clase [15, 20) y su valor es: $P_{30} = 15 + \dfrac{\frac{30\cdot 21}{100} - 3}{5}\cdot 5 = 18{,}3$

El percentil 70 $\dfrac{70\cdot 21}{100} = 14{,}7$ se encuentra en la clase [20, 25) y su valor es: $P_{70} = 20 + \dfrac{\frac{70\cdot 21}{100} - 8}{7}\cdot 5 = 24{,}786$

El percentil 97 $\dfrac{97\cdot 21}{100} = 20{,}37$ se encuentra en la clase [30, 35) y su valor es: $P_{97} = 30 + \dfrac{\frac{97\cdot 21}{100} - 19}{2}\cdot 5 = 33{,}425$

2.- Al laboratorio de la policía científica de Casablanca, han llegado 30 botellas de agua de distintas marcas, para analizar su contenido en sales minerales,

Se han obtenido los siguientes datos, expresados en mg.

46 25 27 30 48 40 76 75 49 59 33 52 21 32 45 27 44 37 62 56 29 54 45 66 69 34 53 45 75 56

1.- Clasifica la variable estadística de concentración de sales.

La concentración de sales, x_i, es una variable estadística cuantitativa continua.

2.- Agrupa los datos en una tabla de 7 intervalos.

El recorrido, r, es la diferencia entre los valores máximos y mínimos;

$$r = \max - \min = 76 - 21 = 55 \quad \rightarrow \quad r = 55$$

Elegimos un r', un poco más grande que el recorrido y que además sea múltiplo de 7. $r' = 56$

Calculamos $\Delta r = r' - r = 1 \quad \rightarrow \quad \dfrac{\Delta r}{2} = 0{,}5$ y calculamos la amplitud de cada intervalo: $a_i = \dfrac{r'}{7} = 8$

Con esto cada intervalo tiene una amplitud a=8 y empezamos en: $\min - a = 21 - 0{,}5 = 20{,}5$

Por tanto, el primer intervalo es (20,5-28,5).

3.- Completa la tabla con todas las columnas necesarias

x_i		f_i	F_i	h_i	H_i	$x_i\cdot f_i$	$x_i^2\cdot f_i$	$\lvert x_i - \bar{x}\rvert\cdot f_i$
Intervalos	x_i							
20,5 – 28,5	24,5	4	4	0,133	0,133	98	2401	91,72
28,5 – 36,5	32,5	5	9	0,166	0,3	162,5	5281,25	74,15
36,5 – 44,5	40,5	3	12	0,1	0,4	121,5	4920,75	20,79
44,5 – 52,5	48,5	7	19	0,233	0,633	339,5	16465,75	7,49
52,5 – 60,5	56,5	5	24	0,166	0,8	282,5	15961,25	45,35
60,5 – 68,5	64,5	2	26	0,066	0,866	129	8320,5	34,14
68,5 – 76,5	72,5	4	30	0,133	**1**	290	21025	100,28
Totales:			N=30			$\sum x_i\cdot f_i = 1423$	$\sum x_i^2\cdot f_i = 74375{,}5$	$\sum \lvert x_i - \bar{x}\rvert\cdot f = 373{,}92$

Calcula:

4.- La media y la moda

La media aritmética viene dada por: $Media: \bar{x} = \frac{\sum x_i \cdot f_i}{N} = \frac{1423}{30} = 47,43$

La moda, que es el valor que más se repite, el 7, se encuentra en el intervalo (44,5 – 52,5), así que:

$$Mo = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \cdot a_i = 44,5 + \frac{(7-3)}{(7-3)+(7-5)} \cdot 8 = 49,83$$

5.- La mediana y los cuartiles

La mediana, que es la medida central, N/2=15, se encuentra en el intervalo (44,5 – 52,5) por tanto:

$$Me = L_i + \frac{\frac{N}{2} - F_{i-1}}{f_i} \cdot a_i = 44,5 + \frac{15-12}{7} \cdot 8 = 47,93$$

Los cuartiles vienen dados por: $Q_k = L_i + \frac{\frac{k \cdot N}{4} - F_{i-1}}{f_i} \cdot a_i$

El primer cuartil, Q_1, está en 30/4=7,5 y se corresponde con la clase (28,5 – 36,5) y su valor es:

$$Q_1 = 28,5 + \frac{7,5-4}{5} \cdot 8 = 34,1$$

El tercer cuartil, Q_3, está en 30·3/4=22,5 y se corresponde con la clase (52,5 -60,5) y su valor es:

$$Q_3 = 52,5 + \frac{22,5-19}{5} \cdot 8 = 58,1$$

6.- Los percentiles P_{40}, P_{80} y P_{98}

Los Percentiles vienen dados por la expresión: $P_k = L_i + \frac{\frac{k \cdot N}{100} - F_{i-1}}{fi} \cdot a_i$

El percentil 40 $\frac{40 \cdot 30}{100} = 12$ se encuentra en la clase (36,5 – 44,5) y su valor es: $P_{40} = 36,5 + \frac{12-9}{3} \cdot 8 = 44,5$

El percentil 80 $\frac{80 \cdot 30}{100} = 24$ se encuentra en la clase (52,5 – 60,5) y su valor es: $P_{80} = 52,5 + \frac{24-19}{5} \cdot 8 = 60,5$

El percentil 98 $\frac{98 \cdot 30}{100} = 29,4$ se encuentra en la clase (68,5 – 76,5) y su valor es: $P_{98} = 68,5 + \frac{29,4-26}{4} \cdot 8 = 75,3$

7.- La desviación media

La desviación media viene dada por: $D_{\bar{x}} = \frac{\sum |x_i - \bar{x}|}{N} = \frac{373,92}{30} = 12,464$

8.- La varianza y la desviación típica

La Varianza la calculamos mediante la expresión $Var = \frac{\sum f_i \cdot x_i^2}{N} - \bar{x}^2 = \frac{74375,5}{30} - 47,43^2 = 229,58$

La Desviación típica es la raíz cuadrada de la varianza: $\sigma = \sqrt{Var} = \sqrt{\frac{\sum f_i \cdot x_i^2}{N} - \bar{x}^2} = \sqrt{229,58} = 15,15$

9.- El coeficiente de variación

El coeficiente de variación es: $C.V. = \frac{\sigma}{\bar{x}} = \frac{15,15}{47,43} = 0,319$

10.- Representa los datos mediante el gráfico que consideres más adecuado.

3.- Al lanzar 30 veces un dado, se obtienen los siguientes resultados:

2,5,4,3,1,6,4,5,4,2,4,6,1,3,6,3,1,2,4,1,5,4,6,4,1,2,3,4,1,4

a) Recuenta los datos y organízalos en una tabla.
b) Calcula la media, la mediana y la moda.
c) Calcula el recorrido.
d) Calcula la varianza y la desviación típica.
e) A la vista de la tabla, ¿Se puede sospechar que el dado está trucado?

Sol:

x_i	f_i	$f_i \cdot x_i$	$f_i \cdot x_i^2$
1	6	6	6
2	4	8	16
3	4	12	36
4	9	36	144
5	3	15	75
6	4	24	144
	N=30	$\sum_i f_i \cdot x_i = 101$	$\sum_i f_i \cdot x_i^2 = 421$

b) $\bar{x} = 3,36$; m_e=4; M_o=4; c) 5; d) 2,74 y 1,66; e) Trucado