Estadística
PREGUNTAS MÁS FRECUENTES

1. ¿A qué se denomina la frecuencia de un dato?

Se llama ***frecuencia absoluta*** (n_i) al número de veces que se repite un cierto dato (x_i) y ***frecuencia relativa*** (f_i) al cociente de la frecuencia absoluta entre el total de datos (N). Ordenando los datos de menor a mayor puede calcularse la ***frecuencia acumulada*** (F_i) de cada dato que se define como la suma de todas las frecuencias absolutas de los datos menores o iguales que él. Toda esta información se puede mostrar en forma de tabla.

Ejemplo:
Datos: 2, 2, 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 7, 8, 9, 9

x_i	n_i	f_i	F_i
2	3	0,1875	3
3	1	0,0625	4
4	3	0,1875	7
5	2	0,125	9
7	4	0,25	13
8	1	0,0625	14
9	2	0,125	16
	$N = 16$	1	

2. ¿Cómo se hallan la moda, la mediana y los cuartiles de un conjunto de datos?

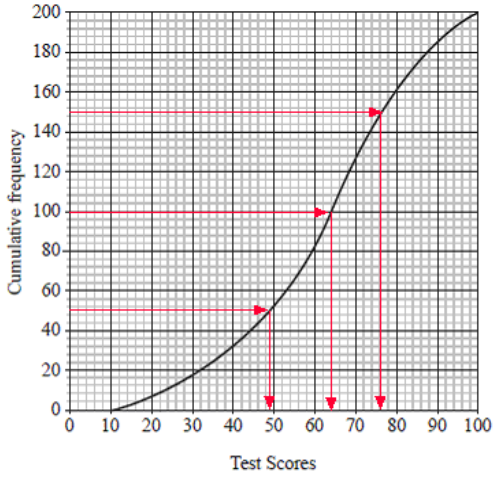
La moda es una medida de centralización y es el dato (o datos) con mayor frecuencia absoluta. Los cuartiles son medidas de dispersión que precisan que los datos aparezcan ordenados de menor a mayor. Son tres valores que reparten el total de los datos en cuatro partes. El primer cuartil (Q_1) es mayor que el 25% de los datos e inferior al 75%. El segundo cuartil (Q_2), también llamado mediana, es mayor que el 50% de los datos e inferior al otro 50%. El tercer cuartil (Q_3) es mayor que el 75% de los datos e inferior al 25%. Si el número de datos es pequeño se buscan los datos que cumplen dichas propiedades o la media aritmética de los dos datos que las cumplan. Si el número de datos es grande se buscan observando la tabla o la gráfica de las frecuencias.

Ejemplo:
Datos: 2, 2, 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 7, 8, 8, 9, 9, 9

x_i	n_i	F_i
2	3	3
3	1	4
4	3	7
5	2	9
7	4	13
8	2	15
9	3	18
	$N = 18$	

El total de datos es 18. La mediana (segundo cuartil) dejaría 9 datos a su izquierda y 9 datos a su derecha. Por lo tanto haremos la media aritmética del noveno y décimo dato. De los 9 datos más pequeños, el primer cuartil sería el quinto dato y de los 9 datos más grandes el tercer cuartil sería el decimocuarto. Observando la columna de frecuencias acumuladas obtenemos: $x_5 = 3$, $x_9 = 5$, $x_{10} = 7$, $x_{14} = 7$.
$Q_1 = 3$, $Q_2 = (5+7)/2 = 6$, $Q_3 = 7$.

Ejemplo: El siguiente gráfico muestra las frecuencias acumuladas de 200 datos con valores entre 10 y 100 correspondientes a las puntuaciones de un examen tipo test.



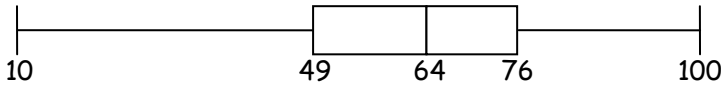
El total de datos es 200. Su cuarta parte es 50, su mitad 100 y sus tres cuartas partes 150. Llevando estos valores a la gráfica nos señala como valores de los tres cuartiles:

$Q_1 = 49$, $Q_2 = 64$, $Q_3 = 76$.

3. ¿Qué es un diagrama de *cajas y bigotes*?

Un diagrama de cajas y bigotes es un gráfico que visualiza la mediana y los cuartiles, de manera que nos permita hacer un análisis rápido de la manera en que se distribuyen los datos del estadístico. Sobre una recta se representan: el dato más pequeño, el primer cuartil, la mediana, el tercer cuartil y el mayor dato. Se le da una forma que recuerda a dos cajas con dos segmentos (bigotes) a izquierda y derecha. Cada uno de estos cuatro elementos incluirá un 25% del total de los datos por lo que la longitud de los mismos mostrará la dispersión o acumulación de datos en cada zona. Si alguno de estos elementos tiene más longitud, indicará que los datos están más dispersos en esa gama de valores mientras que una menor longitud indicará que están más concentrados.

El diagrama del ejemplo anterior es:



Comparando las cuatro longitudes concluiríamos que los datos se aprietan más entre 64 y 76 (segunda caja) y están mucho más dispersos entre 10 y 49 (primer bigote).

4. ¿Cómo se halla la media de un conjunto de datos?

La media aritmética es una medida de centralización que pretende representar al total de los datos mediante una única cantidad. Se define como la suma de todos los datos divididos por el número de ellos. Si, por repetirse, disponemos de sus frecuencias absolutas, se organiza la suma multiplicando cada dato distinto por su frecuencia: $\bar{x} = \dfrac{\sum n_i \cdot x_i}{\sum n_i}$. Se simboliza poniendo una raya sobre la misma letra de los datos.

Ejemplo:

Datos: 2, 2, 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 7, 8, 9, 9

x_i	n_i	$n_i \cdot x_i$
2	3	6
3	1	3
4	3	12
5	2	10
7	4	28
8	1	8
9	2	18
	N = 16	**85**

$$\bar{x} = \frac{\sum n_i \cdot x_i}{\sum n_i} = \frac{85}{16} = 5{,}3125$$

5. ¿Cómo se hallan la varianza y la desviación típica de un conjunto de datos?

La desviación típica es una medida de dispersión que pretende representar mediante una cantidad lo que separa, por término medio, a los datos de su media. Para calcular la desviación típica, simbolizada con la letra sigma, existen dos fórmulas alternativas equivalentes. Se suele usar más la segunda de ellas. La varianza es su cuadrado.

$$\sigma = \sqrt{\frac{\sum n_i \cdot (x_i - \bar{x})^2}{\sum n_i}} = \sqrt{\frac{\sum n_i \cdot x_i^{\,2}}{\sum n_i} - \bar{x}^2} \qquad \sigma^2 = \frac{\sum n_i \cdot (x_i - \bar{x})^2}{\sum n_i} = \frac{\sum n_i \cdot x_i^{\,2}}{\sum n_i} - \bar{x}^2$$

Ejemplo:

Datos: 2, 2, 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 7, 8, 9, 9

x_i	n_i	$n_i \cdot x_i$	$n_i \cdot x_i^2$
2	3	6	12
3	1	3	9
4	3	12	48
5	2	10	50
7	4	28	196
8	1	8	64
9	2	18	162
	N = 16	**85**	**541**

$$\sigma = \sqrt{\frac{\sum n_i \cdot x_i^{\,2}}{\sum n_i} - \bar{x}^2} = \sqrt{\frac{541}{16} - 5{,}3125^2} = 2{,}3643$$

$$\sigma^2 = \sqrt{\frac{\sum n_i \cdot x_i^{\,2}}{\sum n_i} - \bar{x}^2} = \sqrt{\frac{541}{16} - 5{,}3125^2} = 5{,}5898$$

6. ¿Qué información del conjunto de datos podríamos obtener sabiendo su media y su desviación típica?

Si el número de datos es suficientemente alto, el intervalo $(\bar{x}-\sigma, \bar{x}+\sigma)$ incluiría en su interior a un porcentaje *sustancial* del total de datos (aproximadamente el 73%)

Ejemplo: Datos: 2, 2, 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 7, 8, 9, 9

Aproximadamente, el 73% de los datos pertenece al intervalo:

$$(\bar{x}-\sigma, \bar{x}+\sigma) = (5'3125 - 2'3643, 5'3125 + 2'3643) = (2'9482, 7'6768)$$

El cociente $\frac{\sigma}{\bar{x}}$ se denomina coeficiente de variación. Carece de unidades medida y permite comparar las **dispersiones** de dos distribuciones distintas, siempre que sus medias sean positivas. Se calcula para cada una de las distribuciones y los valores que se obtienen se comparan entre sí. La mayor dispersión corresponderá a aquella distribución con mayor coeficiente de variación.

Ejemplo:

Datos: 5, 7, 9 $\rightarrow \bar{x} = 7$, $\sigma = 1{,}6330 \Rightarrow CV = 1{,}6330/7 = 0{,}2333$ Estos datos están más dispersos.

Datos: 13, 15, 23 $\rightarrow \bar{x} = 16$, $\sigma = 2{,}9439 \Rightarrow CV = 2{,}9439/16 = 0{,}1840$ Estos datos están menos dispersos.

7. ¿Cómo podemos aprovechar la información que nos aporta un muestreo para toda la población de la que ha sido elegido aleatoriamente?

De entrada hay que aclarar que el concepto de *población* no se refiere necesariamente a individuos, simplemente se refiere a un, normalmente numeroso, conjunto de datos que, en muchas ocasiones, es muy costoso o engorroso conocer en su totalidad. En la práctica lo que se hace es recabar aleatoriamente sólo algunos de los datos, es decir, elegir una *muestra*. La buena elección de la muestra es un elemento decisivo para que pueda representar con garantías a toda la población. El tamaño de la muestra y los distintos procedimientos que se pueden emplear en su elección aleatoria es un tema que necesita un amplio estudio englobado en la denominada '*Teoría de muestras*' que no procede detallar aquí.

Una vez elegida la muestra pasaríamos a obtener sus parámetros: media, varianza, mediana, cuartiles etc. Lo que nos interesa es cómo extrapolar los parámetros muestrales a toda la población. Los estimadores insesgados que emplearemos son los siguientes:

- Para estimar la media poblacional (μ) se utiliza la media muestral ($\bar{x}$)
- Para estimar la desviación típica poblacional (σ) se utiliza la cuasi-desviación típica media muestral (s_{n-1}), parámetro que normalmente nos facilita cualquier calculadora científica o gráfica.

Es decir, (llamando $N = \sum n_i$):

$$\mu \approx \bar{x} = \frac{\sum n_i \cdot x_i}{N} \qquad \sigma \approx s_{n-1} = \sqrt{\frac{\sum n_i \cdot (x_i - \bar{x})^2}{N-1}}$$

De hacer los cálculos a mano, hay que tener cuidado con la cuasi-desviación típica, ya que mientras que con la desviación típica hay dos fórmulas alternativas –más cómoda la segunda que la primera- que son equivalentes: $\sigma = \sqrt{\frac{\sum n_i \cdot (x_i - \bar{x})^2}{N}} = \sqrt{\frac{\sum n_i \cdot x_i^2}{N} - \bar{x}^2}$, **no** sucede lo mismo con la cuasi-desviación típica, es decir que: $s_{n-1} = \sqrt{\frac{\sum n_i \cdot (x_i - \bar{x})^2}{N-1}}$, pero $s_{n-1} \neq \sqrt{\frac{\sum n_i \cdot x_i^2}{(\sum n_i) - 1} - \bar{x}^2}$.

Si se desea utilizar el cálculo manual cómodo se puede hallar σ y después emplear la fórmula:

$$s_{n-1} = \sqrt{\frac{N}{N-1}} \cdot \sigma$$