Práctica 4 de Estadística

Análisis de una variable medible

Toda la información que se puede obtener sobre una variable está implícitamente contenida en su distribución de frecuencias. Sin embargo, es difícil retener toda esa información contenida en ella. Es conveniente por tanto intentar resumir esta información.

Los valores que resumen las propiedades de una distribución de frecuencias poblacional reciben el nombre de **parámetros poblacionales**. Estos valores son fijos y por lo general desconocidos. Nos interesa por tanto obtener un valor aproximado de los parámetros poblacionales a partir de muestras representativas de la población. A estos valores obtenidos a partir de una muestra concreta se les denominan **estimaciones del parámetro**, y la función a partir de la que se ha calculado se llama **estadístico**.

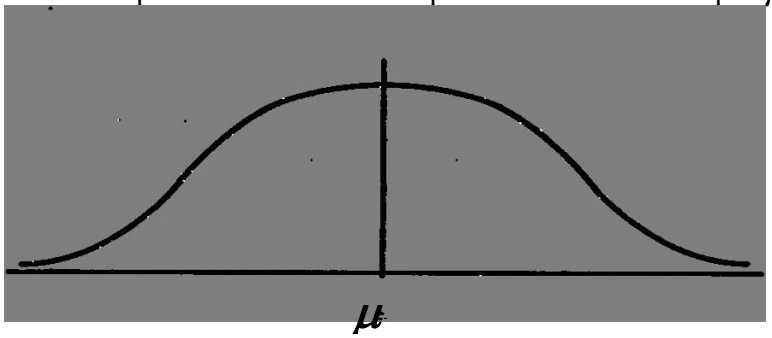
Distinguiremos entre:

a) Parámetros y estadísticos de **centralización**: dan un valor promedio que representa a la población o muestra, respectivamente.
b) Parámetros y estadísticos de **posición**: valores comprendidos en el rango muestral o poblacional, y que se usan para caracterizar la distribución.
c) Parámetros y estadísticos de **dispersión**: miden el grado de concentración de la distribución alrededor de un valor.

4.1 Parámetros y estadísticos de centralización

4.1.1 Media poblacional y media muestral

La **media poblacional** tiene una interpretación intuitiva simple. Si las frecuencias relativas se identifican con masas, entonces la media poblacional es el centro de masas de la distribución de frecuencias poblacional. La media poblacional se denota por μ.



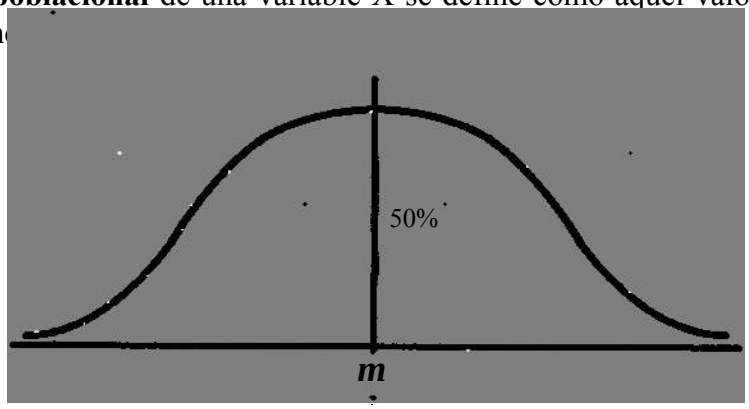
Este parámetro se estimará con la **media muestral** (average). Esta medida se denota por $\overline{X}$ y se interpreta intuitivamente como el centro de masas de la distribución de frecuencias muestral.

Si $x_1, x_2, \ldots, x_n$ es una muestra de n datos:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

4.1.2 Mediana poblacional y mediana muestral

La **mediana poblacional** de una variable X se define como aquel valor ***m*** que divide a la curva de frecuen

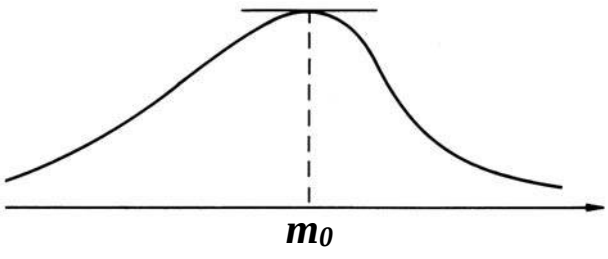


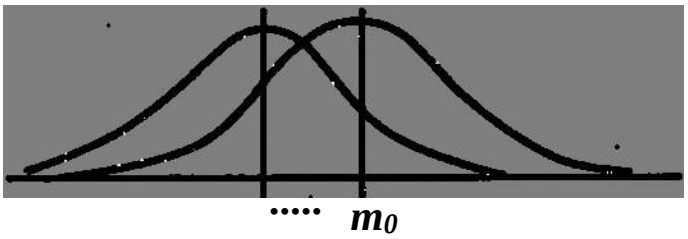
La **mediana muestral** denotada por ***M*** es el valor de la variable que ocupa el valor central (cuando los datos se presentan ordenados en forma creciente) cuando el tamaño de la muestra n es impar. Cuando el tamaño muestral es un número par, entonces la mediana muestral es la media aritmética de los dos valores centrales.

Nota: En distribuciones simétricas respecto al valor central, se cumple que la media y la mediana coinciden. En distribuciones asimétricas que presentan una cola larga debido a la existencia de valores atípicos, la mediana es preferible a la media; mientras que en distribuciones aproximadamente simétricas, la media es la medida de posición central más aconsejable.

4.1.3 Moda poblacional y muestral

La **moda poblacional**, denotada por m_0 , es un valor de la variable al que corresponde un máximo relativo de la curva de frecuencias. Es posible que algunas distribuciones presenten varios máximos relativos; en estos casos la moda absoluta es la mayor de las modas relativas.



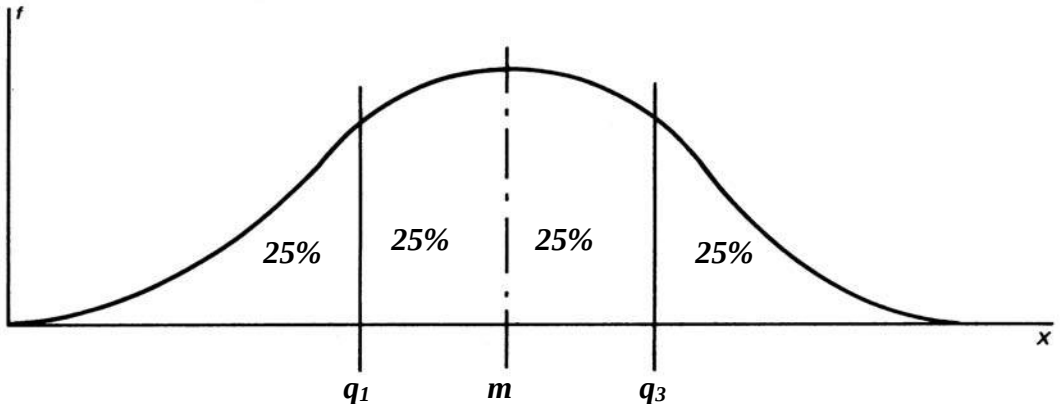


La **moda muestral**, denotada por M_0 , es el valor de la muestra que tiene mayor frecuencia.

4.2 Parámetros y estadísticos de posición

4.2.1 Cuartiles poblacionales y muestrales

Los **cuartiles poblacionales** dividen la distribución de frecuencias en cuartos. El segundo cuartil, q_2, coincide con la mediana.



Se define el **primer cuartil muestral Q_1** como el valor para el cual el 25% de las observaciones son menores o iguales que Q_1 y el 75% de las observaciones son mayores o iguales que Q_1.

Se define el **tercer cuartil muestral Q_3** como el valor para el cual el 75% de las observaciones son menores o iguales que Q_3 y el 25% de las observaciones son mayores o iguales que Q_3.

4.2.2 Percentiles poblacionales y muestrales

En general, para $0<p<1$ definimos un **percentil** (poblacional o muestral) de orden p y lo representamos por ***q(p)***, como aquel valor de la curva de frecuencias (poblacional o muestral) que deja a su izquierda un $p \cdot 100\%$ de la masa (de la población o de la muestra). Notar que:

- $q(0,5)$ = mediana.
- $q(0,25) = Q_1$ primer cuartil (Lower Quartile).
- *q(0,75) = Q3 tercer cuartil (Upper Quartile)*

4.3 Parámetros y estadísticos de dispersión

En general, al resumir los datos perdemos información y, en consecuencia, una tarea fundamental es medir la pérdida de información cometida al efectuar el resumen. Las medidas que cumplen dicho objetivo son las medidas de dispersión y, en definitiva, *miden la variabilidad de los datos*. Cuanto menor sea la variabilidad más representativo resultará el promedio considerado.

4.3.1 Varianza muestral y poblacional

- **Varianza muestral (Variance)**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

- **Desviación típica muestral (Standar Desviation):**

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

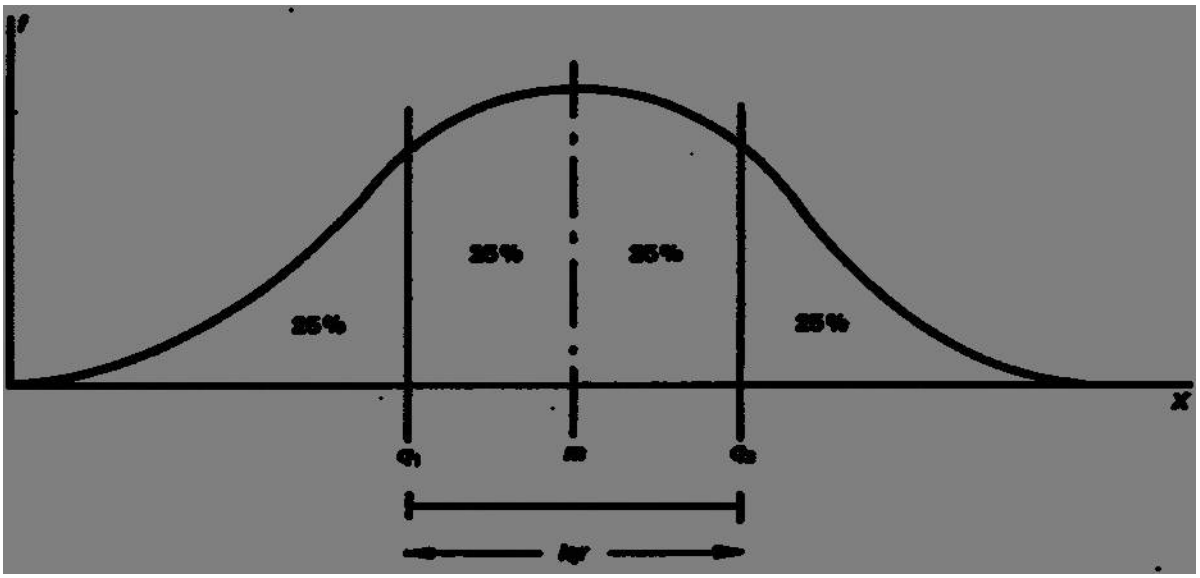
donde n es el tamaño de la muestra, x_i son los valores observados de la muestra y $\overline{X}$ es la media muestral.

Si la mayoría de los valores están próximos a la media muestral, la varianza muestral (desviación típica muestral) resultante será pequeña.

Los parámetros poblacionales respectivos se denotan por σ^2 y σ.

4.3.2 Recorrido intercuartílico

Es otra medida de dispersión, que corresponde con la distancia entre los cuartiles (poblacionales o muestrales según nos refiramos al parámetro o al estadístico).



El **recorrido intercuartílico muestral**, lo representamos como:

$$\boldsymbol{IQR} = Q_3 - Q_1$$

4.4 Ejemplos

Consideremos un edificio en el que viven 10 familias cuyos ingresos mensuales son, en €:

750	1.200	900	650	1.050	1.100	21.000	950	1.400	900

- **Media** $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$\overline{X} = \frac{1}{10}(750 + 1200 + 900 + 650 + 1050 + 1100 + 21000 + 950 + 1400 + 900) = 2990$$

Luego la media es 2.990 €. Esto es un ejemplo de cómo un valor atípico (21.000 €) distorsiona la realidad general del edificio; de hecho, si eliminamos el valor 21.000 resulta una media de 988 € que indica más la realidad.

- **Mediana** Ordenamos los valores de menor a mayor:

650	750	900	900	950	1.050	1.100	1.200	1.400	21.000

La mediana es el valor medio de los dos centrales $m = \frac{950 + 1050}{2} = 1000$.

Si eliminásramos el valor 21.000 obtendríamos m = 950, valor muy cercano al obtenido (1.000). Cuando hay valores *atípicos,* la mediana es un valor más significativo que la media ya que está mucho menos influenciada por los valores atípicos.

- **Primer cuartil** Q_1 deja el 25% de los datos por debajo de él. En este caso el 25% de 10 datos es 2,5, por tanto tomamos como primer cuartil el tercer dato (después de ordenarlos)

$$Q_1 = 900$$

- **Tercer cuartil** Q_3 Deja el 75% de los datos por debajo de él. En este caso el 75% de 10 datos es 7,5, por tanto tomamos como tercer cuartil el octavo dato (después de ordenarlos)

$$Q_3 = 1200$$

En este caso el recorrido intercuartílico es:

$$IQR = Q_3 - Q_1 = 1200 - 900 = 300$$

- **Varianza** $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

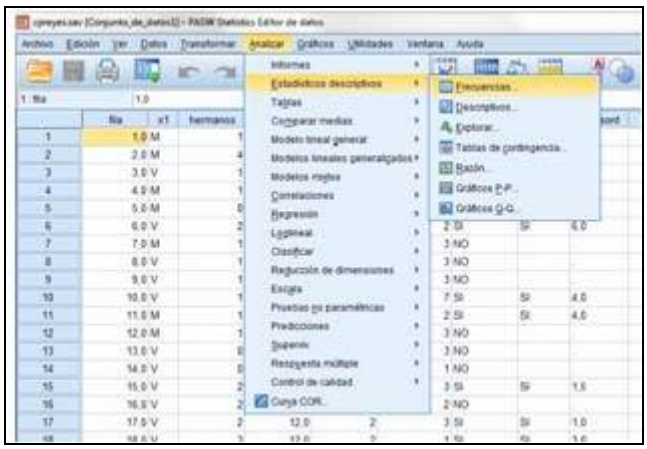
$$s^2 = \frac{1}{9}\left[\begin{array}{l}(650-2990)^2+(750-2990)^2+(900-2990)^2+(900-2990)^2+(950-2990)^2+\\+(1050-2990)^2+(1100-2990)^2+(1200-2990)^2+(1400-2990)^2+\\+(21000-2990)^2\end{array}\right] =$$

$$= \frac{1}{9}\ \square 360.819.000 = 40.091.000$$

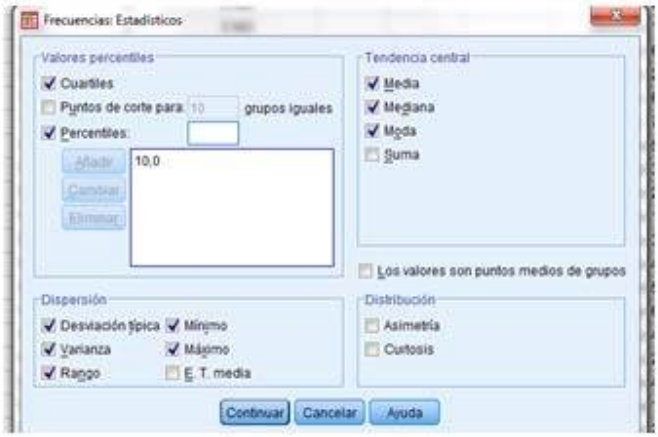
- **Desviación típica** $s = \sqrt{40.091.000} = 6.331,75$

4.5 Manejo de SPSS

En SPSS se pueden calcular los estadísticos para una variable determinada haciendo uso de la misma opción con la que se calculaban las tablas de frecuencias (‘Analizar\Estadísticos descriptivos\Frecuencias’). En dicho cuadro de diálogo, existe un botón ‘Estadísticos’, con el cual es posible especificar los valores que necesitemos obtener.



Estadísticos descriptivos



4.6 Práctica

1. En la siguiente tabla están representados los datos referidos al alquiler pagado mensualmente por 45 familias que habitan pisos de alquiler en una determinada ciudad:

Alquiler mensual €	Nº de familias
0 – 150	5
150 – 300	12
300 – 600	16
600 – 900	10
900-1200	2

a) ¿Qué medidas de centralización y dispersión son más adecuadas para resumir los datos? Razonar la respuesta.
b) Calcular el alquiler medio pagado por las familias analizadas. ¿En qué intervalo se sitúa la mediana? ¿Y la moda?
c) ¿Cuál es la proporción de familias que pagan un alquiler menor o igual a 600 €?

2. El número de mensajes basura (SPAM) recibidos diariamente por los empleados de una empresa lo puedes encontrar en el fichero PRACTICA4E17-datos-ejer2.sav de UACloud CV.

Calcular:

a) Agrupar los datos en intervalos de amplitud 25.
b) Número medio de mensajes SPAM por empleado y día. ¿En qué intervalo se encuentra?
c) Desviación típica, moda y mediana. ¿Qué medida de centralización es la más adecuada? Justifica la respuesta.
d) Tercer cuartil.

3. Se mide el tiempo que tienen que esperar los usuarios de un número de atención al cliente de cierta compañía. Después de varios días de recogida de datos, los resultados obtenidos (en segundos) se presentan en la tabla siguiente:

Tiempo de espera (segundos)	**0-125**	**125-250**	**250-375**	**375-500**
Nº de usuarios	66	21	10	3

a) Construir la tabla de frecuencias.
b) ¿De cuántos datos disponemos? Representar gráficamente la variable mediante un histograma y un polígono de frecuencias.
c) Calcular la media, mediana, desviación típica, moda y percentil 20.
d) ¿Por encima de qué tiempo de espera se encuentran el 75% de los usuarios?

4. El siguiente cuadro contiene algunos de los resultados del análisis descriptivo de la distribución de la variable X1 = Nº de DVD defectuosos en una caja de 50 unidades de la marca A, observada en una muestra de 100 cajas.

Estadísticos

X1

N	Válidos	100
	Perdidos	0
Media		4,00
Mediana		4,00
Moda		4
Desv. típ.		1,758
Varianza		3,091
Percentiles	25	3,00
	50	4,00
	75	5,00

Es cierto que:

a) El 50% de las cajas contiene como máximo 3 unidades defectuosas.
b) El 75% de las cajas contiene como máximo 5 unidades defectuosas.
c) El 75% de las cajas contiene más de 5 unidades defectuosas.
d) El 50% de las cajas contiene menos de 3 unidades defectuosas.