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**.

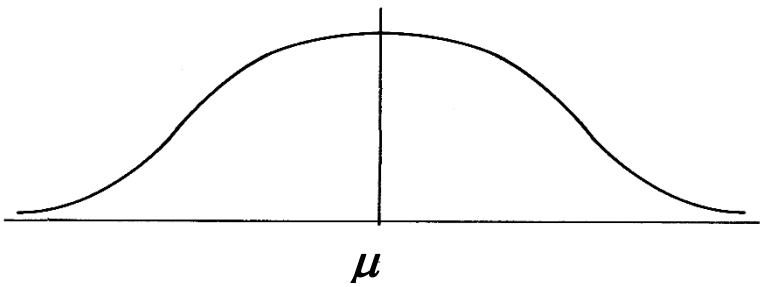
Distinguiremos entre:

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.



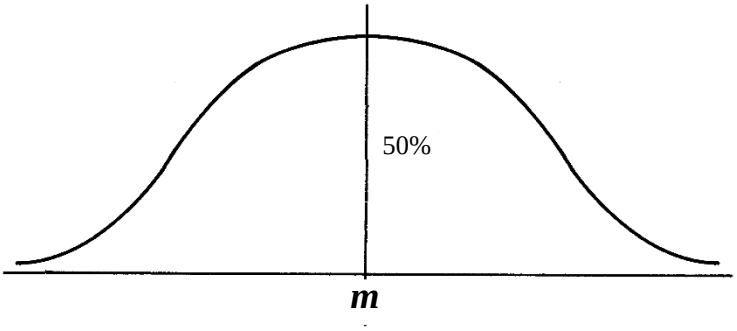
Este parámetro se estimará con la **media muestral** (average). Esta medida se denota por $\overline{X}$ y se interpreta intuitivamente como el centro de masas de la distribución de frecuencias muestral.

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 frecuencias en dos partes con idéntica masa.

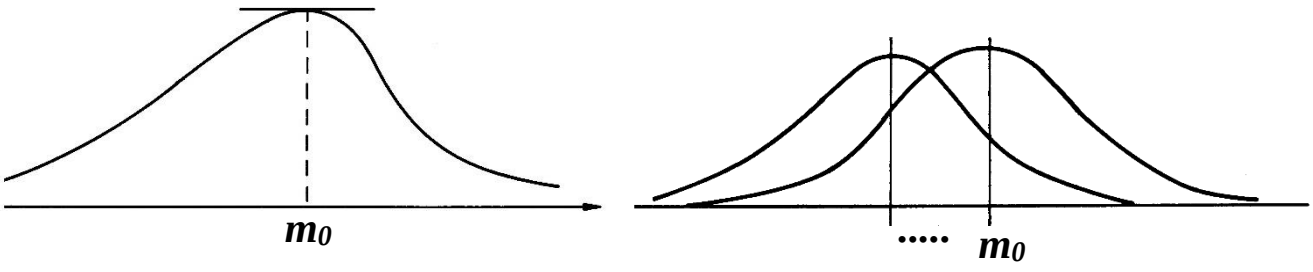


La **mediana muestral** denotada por ***M*** es el valor de la variable que ocupa el valor central (cuando los datos se presentan ordenados en forma creciente) cuando el tamaño de la muestra n es impar. Cuando el tamaño muestral es un número par, entonces la mediana muestral es la media aritmética de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

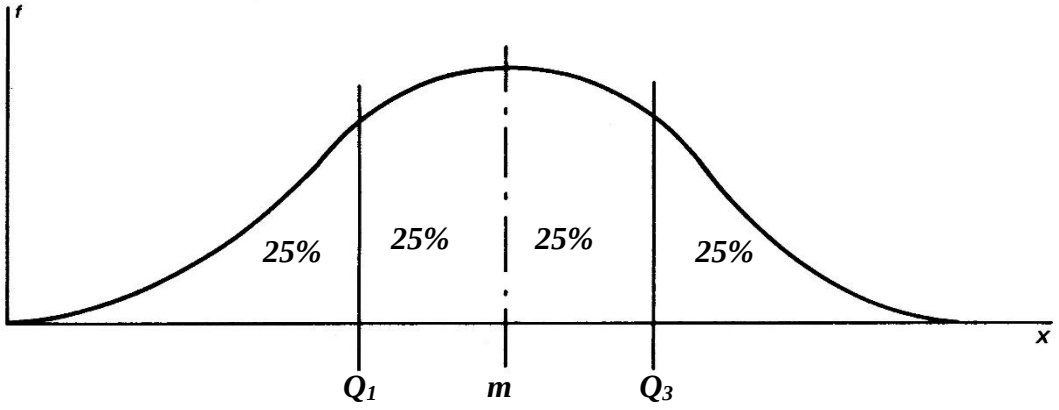


La **moda muestral**, denotada por M_0 , es el valor de la muestra que tiene mayor frecuencia.

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, $\boldsymbol{Q_2}$, coincide con la mediana.



Se define el **primer cuartil muestral** $\boldsymbol{Q_1}$ 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** $\boldsymbol{Q_3}$ 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 $0<i<100$ definimos un **percentil** (poblacional o muestral) **de orden** $\boldsymbol{i}$ y lo representamos por $\boldsymbol{P_i}$, como aquel valor de la curva de frecuencias (poblacional o muestral) que deja a su izquierda un i % de la masa (de la población o de la muestra). Notar que:

- $P_{25} = \boldsymbol{Q_1}$; *primer cuartil (Lower Quartile)*
- $P_{50} = \boldsymbol{Q_2}$; *segundo cuartil =* ***Mediana***
- $P_{75} = \boldsymbol{Q_3}$; *tercer cuartil (Upper Quartile)*

La forma de calcular los percentiles y, por consiguiente, los cuartiles no es única. Dependiendo de la fuente bibliográfica que se consulte, se puede encontrar una forma u otra. A continuación se explica la forma utilizada por SPSS para calcular los percentiles.

Cómo calcular un percentil

Para calcular el percentil de orden i de una muestra de tamaño N primero hemos de ordenar los datos de la muestra. A continuación aplicamos la siguiente fórmula para calcular su posición en la muestra:

$$P_i = \frac{i}{100} \cdot (N+1) = e,d$$

siendo *e* la parte entera del resultado y *d* la parte decimal.

Una vez tenemos la posición que el percentil ocupa en la muestra, calculamos su valor de la siguiente manera:

$$X_e + d \cdot (X_{e+1} - X_e)$$

siendo X_e el valor de la muestra en la posición *e*.

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)**

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

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

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

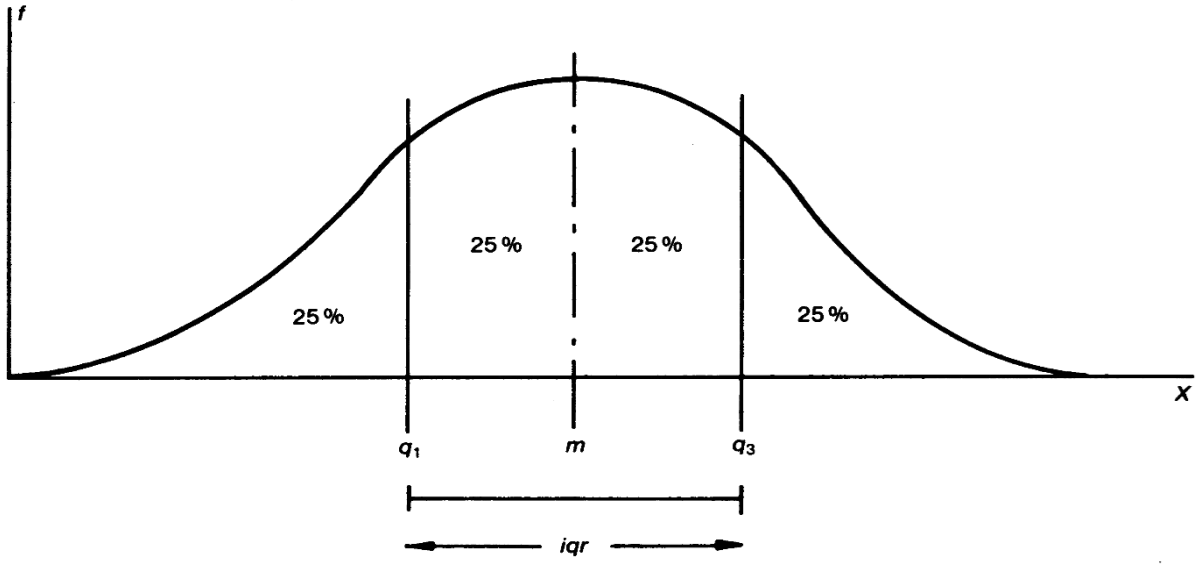
donde *n* es el tamaño de la muestra, x_i son los valores observados de la muestra y $\overline{X}$es la media muestral.

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 Ejemplo

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 coincide con el Q_2, es decir, el P_{50}. Por tanto, calculamos su valor como:

$$m \rightarrow P_{50} = \frac{50}{100} \cdot (10 + 1) = 5{,}5$$

Por tanto, $m = X_5 + 0{,}5 \cdot (X_6 - X_5) = 950 + 0{,}5 \cdot (1050 - 950) = 1000$

Como la muestra tiene un tamaño par, la mediana también se puede calcular como el valor medio de los dos valores centrales: $m = \frac{950+1050}{2} = 1000$. En el caso de que la muestra tuviera un número impar de valores, la mediana coincidiría con el valor central.

Si eliminásemos 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 Q_1, que es el P_{25}, se calcularía de la siguiente manera:

$$Q_1 \rightarrow P_{25} = \frac{25}{100} \cdot (10 + 1) = 2{,}75$$

Por tanto, $Q_1 = X_2 + 0{,}75 \cdot (X_3 - X_2) = 750 + 0{,}75 \cdot (900 - 750) = 862{,}5$

- **Tercer cuartil** Q_3 deja el 75% de los datos por debajo de él. En este caso Q_3, que es el P_{75}, se calcularía de la siguiente manera:

$$Q_3 \rightarrow P_{75} = \frac{75}{100} \cdot (10 + 1) = 8{,}25$$

Por tanto, $Q_3 = X_8 + 0{,}25 \cdot (X_9 - X_8) = 1200 + 0{,}25 \cdot (1400 - 1200) = 1250$

En este caso el **recorrido intercuartílico** es:

$$IQR = Q_3 - Q_1 = 1250 - 862{,}5 = 387{,}5$$

- **Varianza**

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

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

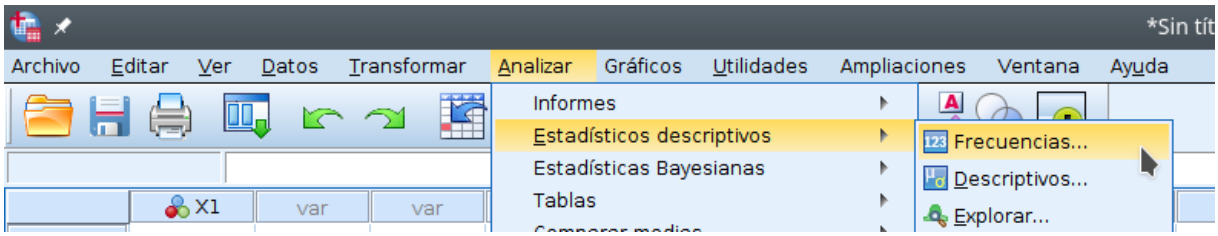
$$= \frac{1}{9}\ 360.819.000 = 40.091.000$$

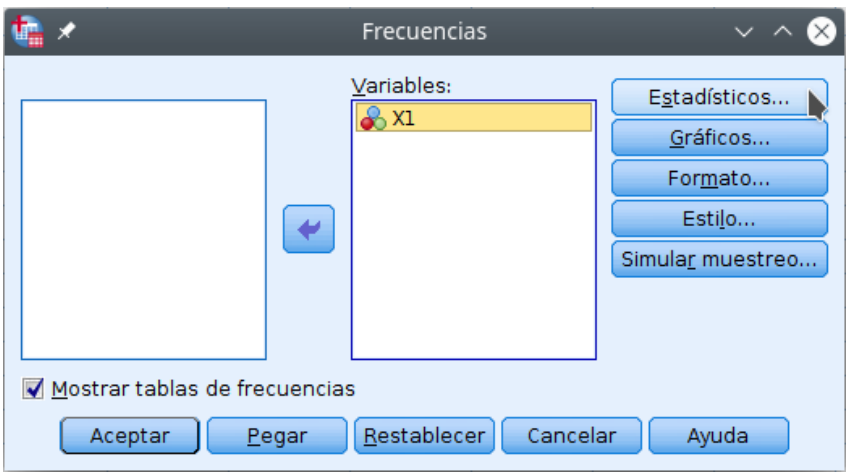
- **Desviación típica**

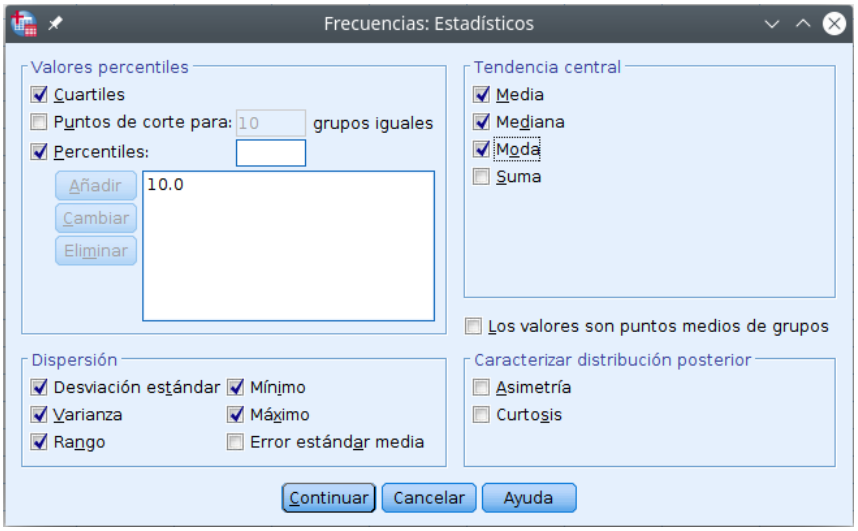
$$s = \sqrt{40.091.000} = 6.331,75$$

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.







4.6 Práctica

1. En la siguiente tabla se recogen 19 datos sobre el sexo, edad y nº de minutos de consumo de televisión (en las unidades correspondientes).

Sexo	Edad	Tv
1	19	30
2	17	60
2	20	120
1	22	30
1	21	90
1	25	0
2	19	100
2	15	180
1	22	120
2	24	30
2	29	20
1	18	60
2	18	90
2	14	90
2	19	240
1	30	180
1	50	260
1	25	90
2	17	50

Sexo: Sexo de la persona encuestada (1 = hombre, 2 = mujer).
Edad: Edad de la persona encuestada.
Tv: Nº de minutos de consumo de televisión (en el día de ayer).

a) Calcula la media, desviación típica y los cuartiles de la variable nº de minutos de consumo de televisión. Interpreta la información obtenida.
b) Calcula la media, desviación típica y varianza del tiempo que dedicaron a ver TV los chicos en el día de ayer. Compara dichos resultados con los obtenidos para las chicas.

2. Contesta a las siguientes preguntas, basándote en el cálculo de percentiles y atendiendo a los datos del ejercicio 1.

a) Se estima que el 20 por ciento de las personas consumen al menos ---------- minutos de TV.
b) Se estima que el 40 por ciento de las personas consumen menos de ---------- minutos de TV.
c) Se estima que, entre los hombres, el 80 por ciento consumen menos de ---------- minutos de TV.
d) Explica la resolución de los apartados anteriores

3. La distribución del importe de las facturas por reparación de carrocería (en €) de una muestra de 80 vehículos en un taller, viene dada por la siguiente tabla:

Importe (€)	Nº de vehículos
0 – 600	10
600 – 800	20
800 – 1200	40
1200 – 1800	10

a) Construir la tabla de frecuencias.
b) ¿Qué porcentaje de reparación de vehículos tiene un coste entre 800 y 1200 euros?
c) Calcular el importe medio de la reparación. Estudiar la representatividad de esta medida.
d) Calcular la mediana y estudiar su representatividad.
e) ¿Cuál es el importe más habitual?
f) ¿Cuál es el importe mínimo pagado por las 75 reparaciones más baratas?

4. Contesta a las siguientes preguntas marcando la opción correcta. Explica por qué la has elegido.

Los siguientes datos están ordenados de menor a mayor: 2.5, 2.6, 2.6, 2.7, 3, 3, 3, 3.3, 15 ¿Cuál de las siguientes afirmaciones es cierta?

a) La mediana es 3.
b) La moda es 2,6.
c) El rango o recorrido es 12.
d) No hay atípicos.

¿Cuál de las siguientes afirmaciones es cierta?

a) Si a todos los valores de una variable les sumamos una constante k, la media aritmética no varía.
b) El recorrido intercuartílico se define como la diferencia entre el cuarto y segundo cuartil.
c) La mediana es robusta frente a valores extremos, es decir, no se ve afectada por valores extremos.
d) La moda es siempre única.

¿Cuál de las siguientes características no se corresponde con el concepto de mediana?

a) Es el centro de gravedad de la distribución.
b) No se ve afectada por los valores extremos.
c) Deja por debajo el mismo número de datos que por encima.
d) Es el segundo cuartil.