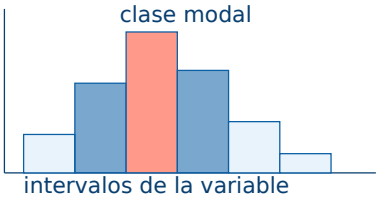
Apuntes - Estadística unidimensional

Curso: 1 Bachillerato Ccss

La estadística unidimensional estudia una sola variable en una población o muestra: notas, sueldos, ventas, estaturas. Sirve para resumir datos con tablas, gráficos y números que indican dónde se centran y cuánto se dispersan.

Actividad de introducción: ¿cómo es «un día normal» en tu clase?

Pide a todos tus compañeros que anoten cuántos minutos tardan en llegar al instituto. Tendrás una lista de 25 o 30 números difíciles de interpretar. Para resumirla, quizá calculas un valor central, por ejemplo el tiempo medio, pero eso no basta: dos clases con un tiempo medio de 20 minutos pueden ser muy distintas si en una todos tardan entre 18 y 22 minutos y en la otra unos viven pegados al instituto y otros tardan una hora. La estadística unidimensional ofrece herramientas para resumir un conjunto de datos de una variable: tablas de frecuencias, gráficos, medidas de centralización (dónde está el centro), medidas de posición (cuartiles y percentiles) y medidas de dispersión (cuánto se separan los datos del centro). Con ellas se puede describir con rigor lo que pasa en una población y comparar dos poblaciones distintas.



Variables y tablas de frecuencias

Definición: tipos de variable y frecuencias

Una variable es **cualitativa** (modo de transporte), **cuantitativa discreta** (número de hijos) o **cuantitativa continua** (sueldo, estatura; se agrupa en intervalos). La **frecuencia absoluta** f_i cuenta las veces que aparece x_i; la **relativa** es $h_i = f_i/n$; las **acumuladas** F_i suman las anteriores.

Ejemplo 1

Notas de 12 alumnos: $5, 6, 7, 5, 8, 9, 5, 6, 7, 6, 10, 6$.

Nota	fi	Fi	hi
5	3	3	0,25
6	4	7	0,333
7	2	9	0,167
8	1	10	0,083

9	1	11	0,083
10	1	12	0,083

Medidas de centralización

Definición: media, mediana y moda

- **Media** $\bar{x} = \dfrac{\sum f_i x_i}{n}$: usa todos los datos y es sensible a extremos.
- **Mediana** Me: valor central de los datos ordenados (o media de los dos centrales).
- **Moda** Mo: valor de mayor frecuencia.

Definición: cuartiles

Los **cuartiles** Q_1, $Q_2 = Me$, Q_3 dividen los datos ordenados en cuatro partes iguales. El **rango intercuartílico** es $Q_3 - Q_1$.

Ejemplo 2

Media, mediana y moda de las notas anteriores.

$\bar{x} = \dfrac{80}{12} \approx 6{,}67$. Ordenadas: $5, 5, 5, 6, 6, 6, 6, 7, 7, 8, 9, 10$; los centrales son 6 y 6, luego $Me = 6$. Moda: 6.

Propiedad

Con valores atípicos (un sueldo muy alto entre sueldos normales), la mediana describe mejor el caso típico que la media.

Medidas de dispersión

Definición: varianza, desviación típica y CV

La **varianza** es $\sigma^2 = \dfrac{\sum f_i (x_i - \bar{x})^2}{n} = \dfrac{\sum f_i x_i^2}{n} - \bar{x}^2$ y la **desviación típica** $\sigma = \sqrt{\sigma^2}$ (en las unidades de los datos). El **coeficiente de variación** $CV = \dfrac{\sigma}{\bar{x}}$ permite comparar dispersiones de variables con distinta media o unidad: a menor CV, mayor homogeneidad.

Método: calcular la desviación típica

1. Calcula la media $\bar{x}$.
2. Calcula $(x_i - \bar{x})^2$ y multiplica por f_i.
3. Suma y divide entre n: es la varianza.
4. Toma la raíz cuadrada.

Ejemplo 3

Desviación típica de $4, 8, 6, 5, 7$.

$\bar{x} = 6$; $\sigma^2 = \dfrac{4+4+0+1+1}{5} = 2$; $\sigma = \sqrt{2} \approx 1{,}41$.

Datos agrupados en intervalos

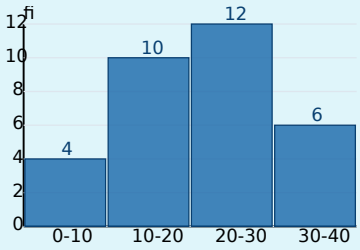
Definición: marca de clase

La **marca de clase** es el punto medio de cada intervalo y lo representa en los cálculos. El **intervalo modal** es el de mayor frecuencia; la **clase mediana** es la primera cuya frecuencia acumulada supera $n/2$.

Ejemplo 4

Intervalos $[0,10)$, $[10,20)$, $[20,30)$, $[30,40)$ con $f_i = 4, 10, 12, 6$ ($n = 32$).

Marcas $5, 15, 25, 35$: $\bar{x} = \dfrac{20 + 150 + 300 + 210}{32} = \dfrac{680}{32} = 21{,}25$. Intervalo modal $[20,30)$. Acumuladas 4, 14, 26, 32; como $n/2 = 16$, la clase mediana es $[20,30)$.



Gráficos estadísticos

Propiedad: qué gráfico usar

- **Diagrama de barras**: variables discretas o cualitativas.
- **Histograma**: datos agrupados; altura proporcional a la frecuencia.
- **Diagrama de sectores**: ángulo $= 360^\circ \cdot h_i$.
- **Diagrama de caja**: mínimo, Q_1, mediana, Q_3 y máximo.

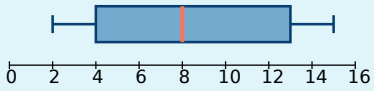
Método: cuartiles y diagrama de caja

1. Ordena los datos y halla la mediana Q_2.
2. Q_1 es la mediana de la mitad inferior y Q_3 la de la superior (sin incluir la mediana si n es impar).
3. Dibuja la caja de Q_1 a Q_3, marca la mediana y traza los bigotes hasta mínimo y máximo.

Ejemplo 5

Cuartiles de $2, 3, 5, 6, 8, 9, 12, 14, 15$.

$Q_2 = 8$; mitad inferior $2, 3, 5, 6$: $Q_1 = 4$; superior $9, 12, 14, 15$: $Q_3 = 13$. Rango intercuartílico 9.



Efecto de transformar los datos

Regla

Si a todos los datos se les suma k, la media aumenta en k y σ no cambia. Si se multiplican por k, la media y σ quedan multiplicadas por $|k|$ y la varianza por k^2.

Ejemplo 6: subida general

Los sueldos (media 1 600 €, $\sigma = 200$ €) suben un 3 %.

Se multiplica por 1,03: media 1648 €, $\sigma = 206$ €. El coeficiente de variación no cambia.

¡Atención!

- Ordena los datos antes de calcular la mediana.
- Con datos agrupados usa la marca de clase, no los extremos.
- En este curso la varianza se divide entre n.
- Para comparar dispersiones con medias muy distintas usa el coeficiente de variación.

Observación: elegir la medida

La media es la mejor medida si los datos son simétricos y sin extremos; la mediana, si hay atípicos o asimetría (rentas, precios de viviendas); la moda, si interesa el valor más común o la variable es cualitativa. Dos conjuntos con la misma media pueden ser muy distintos: acompaña siempre la medida de posición con una de dispersión y redacta una conclusión con el contexto.

Ampliación: cuartiles, percentiles y diagrama de caja

Los cuartiles dividen un conjunto ordenado de datos en cuatro partes con la misma cantidad de observaciones: Q_1 deja a su izquierda el 25 % de los datos, Q_2 es la mediana (50 %) y Q_3 deja a su izquierda el 75 %. Los percentiles generalizan la idea: P_{90} deja por debajo el 90 % de los datos. Los cuartiles permiten construir el diagrama de caja y bigotes, que muestra el centro (mediana), la dispersión central (la caja, de Q_1 a Q_3) y las colas (los bigotes). La longitud de la caja es el **recorrido intercuartílico** $RI = Q_3 - Q_1$, una medida de dispersión que, a diferencia de la desviación típica, resiste a los valores extremos.

Método: cuartiles y valores atípicos

1. Ordena los datos de menor a mayor y localiza la mediana Q_2.
2. Q_1 es la mediana de la mitad inferior y Q_3 la de la mitad superior (sin incluir la mediana si n es impar).
3. Calcula $RI = Q_3 - Q_1$.
4. Los datos fuera de $[Q_1 - 1{,}5\,RI,\ Q_3 + 1{,}5\,RI]$ se consideran atípicos y se marcan con un punto.
5. Dibuja la caja de Q_1 a Q_3, la línea en Q_2 y los bigotes hasta el menor y el mayor dato no atípico.

Ejemplo 7

Los precios (en euros) de ocho menús del día de un barrio son: 9, 10, 10, 11, 12, 13, 14, 30. Estudia la dispersión y los atípicos.

Mediana: $\dfrac{11+12}{2} = 11{,}5$. Mitad inferior (9, 10, 10, 11): $Q_1 = 10$. Mitad superior (12, 13, 14, 30): $Q_3 = 13{,}5$. $RI = 3{,}5$. Límite superior: $13{,}5 + 1{,}5 \cdot 3{,}5 = 18{,}75$. Como $30 > 18{,}75$, el menú de 30 € es un dato atípico (por ejemplo, un restaurante de lujo) y se dibuja aparte. La media (13,625 €) está influida por ese precio, la mediana (11,5 €) no.

Ampliación: datos agrupados, mediana y moda por interpolación

Con datos agrupados en intervalos se pierde información: ya no sabemos el valor exacto de cada observación. Se hace la hipótesis de que los datos de cada intervalo se reparten de forma uniforme, y se sustituye cada intervalo por su marca de clase. La media y la desviación típica se calculan con las marcas, igual que con una tabla discreta. Para la mediana y los cuartiles se localiza primero el intervalo que contiene la posición buscada, usando las frecuencias acumuladas, y se interpola dentro de él.

Regla: mediana y moda en datos agrupados

Si la clase mediana es $[L, U)$ con amplitud a, frecuencia f y frecuencia acumulada

anterior F_{ant}, entonces

$$Me = L + \frac{n/2 - F_{\text{ant}}}{f} \cdot a.$$

Los cuartiles y percentiles se calculan igual, cambiando $n/2$ por $n/4$, $3n/4$ o $kn/100$. La clase modal es la de mayor frecuencia (si las amplitudes son iguales); de forma aproximada, la moda es el punto medio de esa clase.

Ejemplo 8

En una tabla con intervalos de amplitud 10, $n = 60$ y la clase $[20, 30)$ tiene $f = 18$ y frecuencia acumulada anterior 21. Calcula la mediana.

$n/2 = 30$. La mediana está en $[20, 30)$ porque la acumulada anterior (21) es menor que 30 y la acumulada hasta esa clase es $21 + 18 = 39 \geq 30$. Entonces $Me = 20 + \dfrac{30 - 21}{18} \cdot 10 = 20 + 5 = 25$.

Ampliación: elegir la medida adecuada

La media es la medida más usada y la única que tiene en cuenta todos los datos, pero se ve muy influida por valores extremos; por eso, para variables como los salarios o los precios de las viviendas, se prefiere la mediana. La moda es la única posible para variables cualitativas. En cuanto a la dispersión, el recorrido es muy sensible a los extremos, el recorrido intercuartílico no lo es, y la desviación típica es la medida que se combina de forma natural con la media. Para comparar dos variables expresadas en unidades diferentes se emplea el coeficiente de variación *CV*=\sigma/□ar *x*, que se suele expresar como porcentaje. Un CV menor del 10 % indica datos muy homogéneos y un CV alto, por encima del 50 %, que la media representa mal al conjunto.

Errores muy comunes en estadística unidimensional

- Calcular la media de una tabla sin multiplicar por las frecuencias: $\bar{x} \neq (\sum x_i)/k$ si hay frecuencias distintas.
- Olvidar ordenar los datos antes de calcular la mediana o los cuartiles.
- Confundir varianza y desviación típica: la varianza está en unidades al cuadrado.
- Usar la marca de clase como si fuera el dato exacto, sin recordar que es una aproximación.
- Calcular el CV con la varianza en lugar de con la desviación típica.
- Dibujar un histograma con barras de distinta amplitud usando la frecuencia como altura.

¿Sabías que?

La palabra «estadística» procede de «Estado»: los primeros censos se hacían para saber cuántos habitantes y cuántas riquezas tenía un reino. El primer gran estudio

estadístico moderno fue el de John Graunt en 1662, que analizó las listas de defunciones de Londres. En España, el primer censo general moderno es el Censo de Floridablanca de 1787, y hoy el Instituto Nacional de Estadística (INE) publica datos de población, empleo y precios que se resumen con las medidas que has estudiado en este tema.

El consejo de Otto

Si la media y la mediana son muy diferentes, sospecha de valores atípicos y dibuja un diagrama de caja.