Apuntes - Estadística: medidas de localización y dispersión

Curso: 4 Eso

En este tema aprenderás: variables estadísticas y tablas, medidas de localización, medidas de dispersión, tabla de frecuencias: un ejemplo completo, datos agrupados e histogramas y diagrama de caja. Cada idea va en un cuadro de color, con ejemplos resueltos paso a paso.

Actividad de introducción: ¿quién es más regular?

Dos amigas de un club de atletismo de Palencia anotan sus tiempos en cinco carreras de 100 m: Ana corre en $13{,}1;\ 13{,}0;\ 13{,}2;\ 13{,}1;\ 13{,}1$ segundos y Lucía en $12{,}3;\ 13{,}9;\ 12{,}5;\ 13{,}8;\ 13{,}0$. Las dos tienen casi la misma media, unos $13{,}1$ s, pero ¿te parecen igual de regulares? Ana apenas varía y Lucía alterna carreras muy rápidas y muy lentas. La media sola no cuenta toda la historia: necesitas también un número que mida cuánto **se dispersan** los datos alrededor de ella. En este tema aprenderás a resumir una lista de datos con medidas de **centralización** (media, mediana, moda), de **posición** (cuartiles) y de **dispersión** (rango, varianza, desviación típica, coeficiente de variación), y a compararlas con ayuda de gráficos como el histograma o el diagrama de caja.

Variables estadísticas y tablas

Propiedad

La **estadística** estudia datos recogidos de una **población** (o de una muestra). Una variable puede ser **cualitativa** (color, deporte favorito), **cuantitativa discreta** (número de hermanos) o **cuantitativa continua** (estatura, tiempo). Los datos se organizan en una tabla con la **frecuencia absoluta** f_i (cuántas veces aparece cada valor), la **relativa** $h_i = \frac{f_i}{n}$ y la **acumulada**. Si hay muchos valores distintos o la variable es continua, se agrupan en **intervalos** y cada uno se representa por su **marca de clase**, el punto medio: la de $[10, 20)$ es 15.

Medidas de localización

Propiedad: medidas de localización

- **Media:** $\bar{x} = \dfrac{\sum x_i f_i}{n}$. Usa todos los datos y es sensible a los valores extremos.
- **Mediana** Me: el valor central al ordenar los datos (si n es par, la media de los dos centrales). No le afectan los valores atípicos.
- **Moda** Mo: el valor (o intervalo) con mayor frecuencia.

- **Cuartiles:** Q_1, $Q_2 = Me$ y Q_3 dividen los datos ordenados en cuatro partes con el 25 % cada una. Los percentiles dividen en 100 partes.

Ejemplo 1

Datos

4, 6, 7, 7, 8, 9, 9, 9, 10, 12 ($n = 10$). Suma $= 81$, media $= 8{,}1$. Mediana $= \frac{8+9}{2} = 8{,}5$. Moda $= 9$. Q_1 es la mediana de la mitad inferior (4, 6, 7, 7, 8) $= 7$ y Q_3 la de la superior (9, 9, 9, 10, 12) $= 9$.

Medidas de dispersión

Indican cuánto se separan los datos del centro.

Propiedad: medidas de dispersión

- **Rango:** máximo menos mínimo. **Rango intercuartílico:** $IQR = Q_3 - Q_1$.
- **Varianza:** $\sigma^2 = \dfrac{\sum f_i (x_i - \bar{x})^2}{n} = \dfrac{\sum f_i x_i^2}{n} - \bar{x}^2$.
- **Desviación típica:** $\sigma = \sqrt{\sigma^2}$, en las mismas unidades que los datos.
- **Coeficiente de variación:** $CV = \dfrac{\sigma}{\bar{x}}$, sin unidades (suele darse en %). Permite comparar la dispersión de variables con distinta media o unidades.

Regla

En el ejemplo: $\sum x_i^2 = 701$, $\sigma^2 = \frac{701}{10} - 8{,}1^2 = 70{,}1 - 65{,}61 = 4{,}49$, $\sigma \approx 2{,}12$, $IQR = 2$, rango $= 8$, $CV = \frac{2{,}12}{8{,}1} \approx 26{,}2\,\%$.

Propiedades

Si se suma una constante a todos los datos, la media aumenta esa constante y la desviación típica no cambia; si se multiplican todos por k, la media y la desviación típica se multiplican por k.

Cómo se hace: calcular los parámetros de una distribución

1. Ordena los datos y construye la tabla con x_i, f_i, $x_i f_i$ y $x_i^2 f_i$.
2. Calcula la media $\bar{x} = \frac{\sum x_i f_i}{n}$.
3. Calcula la varianza $\sigma^2 = \frac{\sum x_i^2 f_i}{n} - \bar{x}^2$ y la desviación típica.
4. Halla mediana, cuartiles y IQR sobre los datos ordenados.

Ejemplo 2: parámetros de cinco datos

Calcula media, mediana, moda y desviación típica de 2, 3, 3, 5, 7.

Suma $= 20$, $n = 5$, $\bar{x} = 4$. La mediana es 3 y la moda 3. Varianza $\sigma^2 = \frac{(-2)^2+(-1)^2+(-1)^2+1^2+3^2}{5} = \frac{16}{5} = 3{,}2$ y $\sigma \approx 1{,}79$.

Tabla de frecuencias: un ejemplo completo

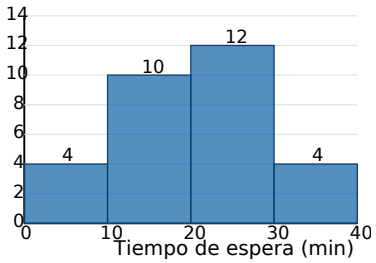
Observación

Notas de 20 alumnos: $x = 3, 4, 5, 6, 7, 8, 9$ con $f = 2, 3, 5, 4, 3, 2, 1$. Entonces $\sum x_i f_i = 113$, $\bar{x} = 5{,}65$; $\sum x_i^2 f_i = 691$, $\sigma^2 = 34{,}55 - 5{,}65^2 = 2{,}6275$, $\sigma \approx 1{,}62$. La moda es 5 y la mediana, entre los datos 10.º y 11.º (5 y 6), vale $5{,}5$.

Datos agrupados e histogramas

Propiedad

Los tiempos de espera (en minutos) de 30 pacientes en un centro de salud: $[0, 10)$: 4; $[10, 20)$: 10; $[20, 30)$: 12; $[30, 40)$: 4. Un **histograma** dibuja un rectángulo sobre cada intervalo con altura proporcional a la frecuencia.



Propiedad

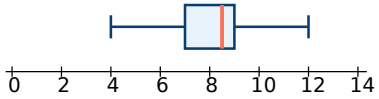
Con las marcas de clase 5, 15, 25, 35: $\bar{x} = \frac{610}{30} \approx 20{,}33$ min; $\sigma^2 = \frac{14\,750}{30} - 20{,}33^2 \approx 78{,}2$, $\sigma \approx 8{,}84$. El **intervalo modal** es $[20, 30)$ y el **intervalo mediano** también, porque $\frac{n}{2} = 15$ se alcanza ahí (acumuladas 4, 14, 26). Con interpolación: $Me = 20 + \frac{15-14}{12} \cdot 10 \approx 20{,}83$.

Diagrama de caja

Propiedad

Resume cinco valores: mínimo, Q_1, mediana, Q_3 y máximo. La caja va de Q_1 a Q_3 (contiene el 50 % central de los datos) y la raya interior marca la mediana. Un dato es **atípico** si queda fuera de $[Q_1 - 1{,}5 \cdot IQR,\ Q_3 + 1{,}5 \cdot IQR]$. Este es el diagrama de

caja de los datos del ejemplo 1:



Ejemplo 3: datos atípicos

En los datos 3, 5, 6, 6, 7, 8, 9, 10, 22, ¿hay algún valor atípico?

$n = 9$, $Me = 7$, $Q_1 = \frac{5+6}{2} = 5{,}5$ y $Q_3 = \frac{9+10}{2} = 9{,}5$, luego $IQR = 4$. El intervalo es $[5{,}5 - 6,\ 9{,}5 + 6] = [-0{,}5,\ 15{,}5]$. Como $22 > 15{,}5$, el dato 22 es atípico.

¡Atención! Errores frecuentes

- Calcular la mediana sin ordenar los datos.
- Dividir entre el número de valores distintos y no entre n al calcular la media con frecuencias.
- Olvidar restar $\bar{x}^2$ en la varianza.
- Comparar desviaciones típicas de variables con medias muy distintas en lugar de usar el coeficiente de variación.

Cómo organizar y comparar distribuciones

Método: construir una tabla de datos agrupados

1. Calcula el rango $R = x_{max} - x_{min}$.
2. Elige entre 5 y 8 intervalos de la misma amplitud, de modo que cubran todos los datos.
3. Cuenta cuántos datos caen en cada intervalo: es la frecuencia absoluta f_i.
4. Calcula la marca de clase de cada intervalo, que es su punto medio.
5. Añade las frecuencias relativas $h_i = \dfrac{f_i}{n}$ y las acumuladas F_i.

Ejemplo 2

Calcula media y desviación típica de los datos agrupados $[0, 10)$: 3; $[10, 20)$: 5; $[20, 30)$: 2.

Marcas $5, 15, 25$ y $n = 10$. $\bar{x} = \dfrac{5 \cdot 3 + 15 \cdot 5 + 25 \cdot 2}{10} = \dfrac{140}{10} = 14$.

$\sum x_i^2 f_i = 25 \cdot 3 + 225 \cdot 5 + 625 \cdot 2 = 2450$, luego $\sigma^2 = \dfrac{2450}{10} - 14^2 = 245 - 196 = 49$ y $\sigma = 7$.

Propiedad: qué medida elegir

- La **media** usa todos los datos y es la mejor opción si no hay valores extremos.
- La **mediana** resiste los datos atípicos y es preferible en distribuciones asimétricas, como los sueldos.
- La **moda** es la única medida que sirve para datos cualitativos (el color favorito, la marca de móvil).
- La **desviación típica** mide la dispersión en las mismas unidades que los datos; el **coeficiente de variación** permite comparar dispersiones de datos con distinta media o unidades.

Ejemplo 3

Dos tiendas venden pantalones. En la tienda A el precio medio es 40 € con $\sigma = 8$ € y en la tienda B es 120 € con $\sigma = 18$ €. ¿Cuál tiene precios más homogéneos?

$CV_A = \dfrac{8}{40} = 20\,\%$ y $CV_B = \dfrac{18}{120} = 15\,\%$. Los precios de la tienda B son relativamente más homogéneos, aunque su desviación típica sea mayor.

¡Atención! Errores habituales en estadística

- No se puede calcular la media de datos cualitativos: solo la moda.
- Con una cantidad par de datos, la mediana es la media de los dos centrales.
- La varianza se mide en unidades al cuadrado; por eso se usa la desviación típica, que vuelve a las unidades originales.
- Al comparar dos distribuciones con medias muy distintas, no basta con comparar desviaciones típicas: hay que calcular el coeficiente de variación.
- Un dato atípico no se elimina sin justificación: se identifica y se analiza.

Interpretar el diagrama de caja

Regla: los cinco números del diagrama de caja

Un diagrama de caja se construye con cinco valores: el mínimo, el primer cuartil Q_1, la mediana, el tercer cuartil Q_3 y el máximo. La caja va de Q_1 a Q_3 y contiene el $50\,\%$ central de los datos; su longitud es el rango intercuartílico $IQR = Q_3 - Q_1$. Los bigotes llegan hasta los datos no atípicos y los puntos aislados fuera de los límites $Q_1 - 1{,}5\,IQR$ y $Q_3 + 1{,}5\,IQR$ se consideran atípicos. Si la mediana está centrada en la caja, la distribución es aproximadamente simétrica.

Ejemplo 4

Los cinco números de unos datos son $2,\ 5,\ 8,\ 10,\ 18$. Describe la distribución.

$Q_1 = 5$, $Me = 8$, $Q_3 = 10$ e $IQR = 5$. La caja, de 5 a 10, contiene la mitad central de los datos. El límite superior es $10 + 1{,}5 \cdot 5 = 17{,}5$ y el inferior $5 - 7{,}5 = -2{,}5$. Como el máximo 18 supera $17{,}5$, es un dato atípico y se dibuja como un punto aislado; el mínimo 2 sí está dentro de los límites.

¿Sabías que?

El belga Adolphe Quetelet fue de los primeros en usar la media y la desviación en el siglo XIX para estudiar la estatura o el peso de las personas y habló del «hombre medio». Florence Nightingale utilizó diagramas estadísticos para convencer al gobierno británico de mejorar los hospitales militares. Y el diagrama de caja lo inventó el estadístico John Tukey en 1970; hoy se usa para comparar de un vistazo desde el rendimiento de equipos deportivos hasta datos científicos.

Observación: transformar los datos

Si a todos los datos se les suma un mismo número, la media aumenta en ese número, pero la desviación típica no cambia, porque los datos siguen igual de separados entre sí. En cambio, si todos los datos se multiplican por un número positivo, la media y la desviación típica quedan multiplicadas por ese número, y el coeficiente de variación no cambia. Estas propiedades permiten, por ejemplo, pasar temperaturas de grados Celsius a Fahrenheit sin repetir todos los cálculos.

El consejo de Otto

Si hay datos extremos que distorsionan, usa la mediana y el rango intercuartílico: son resistentes. Si los datos son estables y simétricos, la media y la desviación típica funcionan muy bien.