UTN
UNIVERSIDAD TECNOLÓGICA NACIONAL

UNIDAD N° 1

ESTADÍSTICA DESCRIPTIVA

Medidas de tendencia central, de dispersión y de posición

UNIDAD N° 1: ESTADÍSTICA DESCRIPTIVA

MEDIDAS PARA DESCRIBIR UN CONJUNTO DE DATOS

En las secciones anteriores hemos visto cómo se construye una tabla de frecuencias.

Conocer las distribuciones de frecuencias es una herramienta central en estadística descriptiva, pero muchas veces resulta útil resumir la información en un solo valor representativo que sintetice la ubicación o comportamiento general de los datos.

Tanto de los datos clasificados como de los gráficos se desprende claramente que hay determinados valores que se presentan más a menudo y otros menos frecuentemente.

En muchas situaciones, las distribuciones de datos (como veremos en los gráficos en la próxima semana) presentan una forma simétrica o campanular. Esto sugiere que los valores más característicos tienden a ubicarse en el centro de la distribución.

Las medidas de tendencia central (como la media, la mediana o la moda) nos permiten resumir los datos en un solo valor representativo.

Las medidas de dispersión, en cambio, indican cuánto varían los datos en torno a ese valor central, es decir, si están muy dispersos o bastante agrupados.

Finalmente, las medidas de posición, como los cuartiles o los percentiles, nos ayudan a ubicar los datos dentro del conjunto, dividiéndolos en partes iguales y permitiéndonos interpretar con mayor precisión cómo se distribuyen.

Encontraremos entonces valores que sirven para describir un conjunto de datos. Si el conjunto en estudio una muestra, dichos valores reciben el nombre de estadísticos o estimadores y se utilizan para estimar características similares en la población, de la cual fue obtenida la muestra. A esas características poblacionales las llamamos parámetros.

MEDIA ARITMÉTICA PARA DATOS NO AGRUPADOS

La media aritmética es una medida de tendencia central que nos permite obtener un valor representativo del conjunto de datos. Se calcula como el cociente entre la suma de todos los valores observados y la cantidad de observaciones. Su fórmula general para una muestra es:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$

La media actúa como un "centro de gravedad" numérico de los datos: si imaginamos una vara numerada sin peso y colocamos pesos idénticos sobre cada número correspondiente a los datos, la vara estaría en equilibrio justo en el punto que marca la media. Esta representación ilustra de manera intuitiva que la media equilibra todas las distancias de los datos.

Ejemplo:

En un examen calificado del 0 al 10; 3 alumnos obtuvieron una calificación de 5 puntos, 5 alumnos obtuvieron una calificación de 4 puntos, y 2 alumnos obtuvieron 3 puntos de calificación. Calcular la calificación media.

Solución:

$$\bar{x} = \frac{5 + 5 + 5 + 4 + 4 + 4 + 4 + 4 + 3 + 3}{10} = \frac{41}{10} = 4{,}1$$

Observamos entonces que la media refleja la calificación promedio de todo el grupo, teniendo en cuenta cuántos estudiantes obtuvieron cada calificación.

La media poblacional se notará con la letra μ y la fórmula de cálculo será:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

En esta fórmula, N representa el total de la población (no solo una muestra), por eso usamos la letra griega μ en lugar de $\bar{x}$.

La media nos da una idea general de qué valor es "típico" o promedio, considerando todos los datos. Sin embargo, esta característica de usar todos los valores también hace que la media sea sensible a los valores extremos o atípicos. Basta con que uno o dos datos sean inusualmente grandes o pequeños para que la media se desplace en esta dirección, perdiendo representatividad. Decimos entonces que la media no es un estadístico robusto.

MEDIA ARITMÉTICA PARA DATOS AGRUPADOS

Cuando trabajamos con datos agrupados, representamos cada intervalo mediante un único valor llamado marca de clase, que corresponde al punto medio del intervalo. Suponemos que todos los datos incluidos en un mismo intervalo se comportan como si tuvieran el valor de la marca de clase, aunque sus valores reales puedan ser diferentes.

Entonces para calcular la media aritmética para datos agrupados utilizamos la siguiente expresión:

$$\bar{x} = \frac{\sum x_i f_i}{n}$$

En esta fórmula:

- x_i es la marca de clase del intervalo i,
- f_i es la frecuencia absoluta del intervalo,
- n es el total de datos.

Para calcular la media de datos que se encuentran agrupados en intervalos , multiplicamos cada marca de clase por su frecuencia, sumamos esos productos, y dividimos por el total de datos.

OTROS TIPOS DE MEDIAS

Media Recortada

La media recortada es una variante de la media aritmética en la que se eliminan previamente un determinado porcentaje de los valores más extremos del conjunto de datos (por ejemplo, el 10 % inferior y el 10 % superior), y luego se calcula la media sobre los datos restantes.

$$\bar{x}_{rec} = \frac{1}{n^*} \sum_{i=r+1}^{n-r} x_{(i)}$$

En esta fórmula:

$x_{(i)}$ representa los datos ordenados de menor a mayor.

r es la cantidad de observaciones eliminadas en cada extremo.

$n^* = n - 2r$ es la cantidad de datos que quedan tras el recorte.

La media recortada surge como una respuesta a una debilidad de la media aritmética: su alta sensibilidad a valores atípicos. En conjuntos de datos reales, especialmente en áreas como economía, educación o ingeniería de datos, es común que haya observaciones extremas que distorsionan el valor promedio. Al eliminar esos extremos, se obtiene una media más robusta, menos influida por observaciones poco representativas.

Volviendo a la imagen de la vara en equilibrio, la media recortada es como si decidiéramos quitar los pesos que están demasiado lejos de los extremos para concentrarnos en dónde se "agrupa" el resto del peso. Esto permite estimar un centro más estable del conjunto, especialmente útil en distribuciones asimétricas o con valores extremos accidentales.

Media Ponderada

La media ponderada es una medida de tendencia central que se utiliza cuando no todos los valores tienen la misma importancia o frecuencia. A cada dato se le asigna un peso o ponderación que indica cuánto influye ese valor en el promedio final.

Ponderar significa asignar más peso o relevancia a ciertos valores en función de algún criterio: cantidad de repeticiones, importancia, duración, nivel de confianza, entre otros.

Es una extensión natural de la media aritmética, en la que cada valor se repite una o más veces, y en lugar de sumarlo varias veces, se lo multiplica por su peso.

Dado un conjunto de valores $x_1, x_2, \cdots, x_n$ y sus respectivos pesos $w_1, w_2, \cdots w_3$ la media ponderada se calcula así:

$$\bar{x}_{pond} = \frac{\sum w_i \cdot x_i}{\sum w_i}$$

Si todos los pesos son iguales, esta fórmula se convierte en la media aritmética.

Donde:

x_i son los valores observados

w_i son los pesos (frecuencias o ponderaciones asignadas)

Ejemplo:

Un estudiante de la TUPaD obtiene las siguientes calificaciones parciales en una materia, cada una con una importancia distinta en la evaluación final:

Evaluación	Nota	Importancia (Peso)
Parcial 1	6	30 %
Parcial 2	8	30 %
Trabajo final	9	40 %

La media ponderada será:

$$\bar{x}_{pond} = \frac{6 \cdot 0{,}30 + 8 \cdot 0{,}30 + 9 \cdot 0{,}40}{0{,}30 + 0{,}30 + 0{,}40} = \frac{1{,}8 + 2{,}4 + 3{,}6}{1} = 7{,}8$$

La nota final del estudiante no es la media aritmética de las tres calificaciones (que sería 7,66), sino 7,8, porque el trabajo final tuvo más peso

.

Media Geométrica

La media geométrica es un tipo especial de promedio que se utiliza cuando los valores de un conjunto se encadenan multiplicativamente, es decir, cuando el resultado de un valor depende del anterior.

Este tipo de situación ocurre, por ejemplo, al analizar porcentajes de crecimiento o disminución sucesiva: rendimientos de una inversión, aumentos de precios, caídas en ventas, entre otros.

En estos casos, el promedio aritmético no refleja adecuadamente el comportamiento global, y por eso se utiliza la media geométrica.

La media geométrica calcula como la raíz enésima del producto de los valores:

$$\bar{x}_{geom} = \bar{G} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

La media geométrica solo puede calcularse si todos los valores del conjunto son positivos.

No se puede aplicar si alguno de los valores es cero (porque el producto total será cero y la raíz también), ni si hay valores negativos (porque no está definida la raíz enésima de un producto negativo en el contexto de números reales).

Por eso, antes de aplicar la media geométrica, es fundamental revisar la naturaleza de los datos y asegurarse de que representen tasas positivas, factores de cambio o porcentajes de crecimiento, donde no hay valores nulos ni negativos.

Ejemplo:

Se quiere calcular la rentabilidad promedio anual de una inversión que tuvo los siguientes rendimientos:

Año 1: gana un 20 %

Año 2: pierde un 10 %

Año 3: gana un 30 %

El capital inicial invertido fue de $10.000

Estos rendimientos actúan de manera acumulativa sobre el capital, año tras año:

Año 1: $\$10.000 \times 1{,}20 = \$12.000$

Año 2: $\$12.000 \times 0{,}90 = \$10.800$

Año 3: $\$10.800 \times 1{,}30 = \$14.040$

Cálculo de la media geométrica:

Los factores de variación fueron:

$$1{,}20;\ 0{,}90;\ 1{,}30$$

Multiplicamos todos los factores:

$$1{,}20 \cdot 0{,}90 \cdot 1{,}30 = 1{,}404$$

Extraemos la raíz cúbica (porque son tres años):

$$\bar{x}_{geom} = \sqrt[3]{1{,}404} \cong 1{,}099$$

Esto equivale a una rentabilidad promedio anual del:

$$(1{,}099 - 1) \cdot 100 = 9{,}9\%$$

Es decir, si la inversión hubiese crecido a un ritmo constante del 9,9 % anual durante los tres años, habría alcanzado el mismo valor final de $14.040.

Lo que equivale a decir que, en promedio, el capital creció un 9,9 % por año, considerando el efecto acumulado de las ganancias y pérdidas sucesivas.

¿Y si usamos la media aritmética?

$$\frac{20\% + (-10\%) + 30\%}{3} = \frac{40\%}{3} \cong 13{,}3\%$$

Este promedio es mayor al real. No tiene en cuenta que una pérdida intermedia impacta negativamente sobre el capital acumulado, por lo que sobreestima el crecimiento.

Media Armónica

La media armónica es una medida de tendencia central que se utiliza cuando los datos expresan tasas, velocidades o razones, es decir, relaciones entre dos magnitudes como “cantidad por unidad de tiempo” o “unidades por cantidad”.

Se utiliza especialmente cuando esas tasas están asociadas a una misma unidad de referencia fija (como una distancia, una tarea o una cantidad total).

A diferencia de la media aritmética, que suma valores absolutos, la media armónica proporciona un promedio más preciso en contextos donde valores bajos tienen un mayor impacto. Esto ocurre, por ejemplo, al calcular:

La velocidad promedio en varios tramos de igual distancia

El consumo de combustible por kilómetro recorrido

El tiempo promedio por unidad producida, cuando se trabaja a distintas eficiencias

En estos casos, se habla de “rendimiento” no en sentido económico, sino como sinónimo de desempeño relativo: cuánto se logra por unidad de algo.

La media armónica pondera estos desempeños de manera inversa, destacando que un valor pequeño puede prolongar mucho un proceso, y por eso recibe más peso en el promedio.

Para un conjunto de n valores positivos $x_1, x_2, \cdots, x_n$ la media armónica se calcula como:

$$\bar{x}_{arm} = \bar{H} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Al igual que la media geométrica, la media armónica solo puede aplicarse cuando todos los valores del conjunto son positivos y diferentes de cero.

Esto se debe a que su fórmula incluye el cálculo del recíproco de cada dato (es decir, $\frac{1}{x_i}$), lo cual no está definido para valores nulos y genera errores para valores negativos en la interpretación del contexto (como tiempos).

Por eso, la media armónica debe utilizarse solo cuando se trata de datos positivos relacionados con tasas, velocidades o rendimientos por unidad fija, donde todos los valores tienen sentido práctico y numérico.

Ejemplo:

Un programador está optimizando un script de análisis de datos. Durante la jornada, ejecuta tres pruebas de rendimiento sobre el mismo conjunto de datos, en distintos momentos, y mide cuánto tarda cada ejecución:

Primer intento: 10 minutos

Segundo intento: 15 minutos

Tercer intento: 20 minutos

El script y el conjunto de datos son siempre los mismos. Solo cambia el rendimiento del sistema (por ejemplo, por carga del procesador o conexión a internet).

Queremos calcular el tiempo promedio por ejecución completa de la prueba.

¿Por qué no usar la media aritmética?

Si simplemente promediamos los tiempos:

$$\frac{10 + 15 + 20}{3} = 15\ minutos$$

Obtenemos un valor intermedio, pero que no representa bien el tiempo real promedio por ejecución, ya que las ejecuciones más lentas consumen más tiempo total y deberían influir más en el resultado.

Aplicando la media armónica:

$$\bar{x}_{arm} = \bar{H} = \frac{3}{\frac{1}{10} + \frac{1}{15} + \frac{1}{20}} = \frac{3}{\frac{13}{60}} = \frac{180}{13} \cong 13{,}85\ minutos$$

El tiempo promedio real por ejecución fue de aproximadamente 13,85 minutos, no 15.

Este valor tiene en cuenta que la tercera ejecución fue más lenta y pesó más en el total. Por eso, la media armónica ofrece una medida más precisa del desempeño general del sistema bajo prueba

Relación entre las medias aritmética, geométrica y armónica

Como regla general, cuando se trabaja con datos positivos y distintos, se cumple la siguiente relación:

$$\bar{x}_{arm} \leq \bar{x}_{geom} \leq \bar{x}_{aritm}$$

Esta desigualdad refleja cómo cada media responde de forma diferente a la distribución de los datos. La elección de una u otra no es arbitraria: depende del tipo de variable, del significado de los valores y del fenómeno específico que se desea analizar.

Entender las condiciones bajo las cuales cada media es adecuada es clave para realizar interpretaciones válidas y representativas.

MODA PARA DATOS NO AGRUPADOS

La moda (Mo o $\hat{x}$) es el valor más frecuente en un conjunto de datos, es decir, aquel que aparece con mayor frecuencia. A diferencia de la media, la moda no requiere cálculos, sino simplemente identificar cuál es el dato que se repite más veces.

La moda puede ser única, múltiple o incluso no existir. Según la cantidad de valores que se repiten con igual máxima frecuencia, podemos clasificar a la distribución de la siguiente manera:

- Si hay una única moda, la distribución es unimodal.
- Si hay dos valores con igual frecuencia máxima, se dice que es bimodal.
- Si hay más de dos modas, se trata de una distribución multimodal.
- Si todos los valores son distintos, entonces no hay moda.

La moda es especialmente útil cuando trabajamos con variables cualitativas o datos categóricos, donde no tiene sentido calcular un promedio numérico.

MODA PARA DATOS AGRUPADOS

Cuando los datos están agrupados en intervalos, no podemos observar directamente cuál es el valor que más se repite, ya que desconocemos los datos individuales. Sin embargo, podemos identificar el intervalo con mayor frecuencia absoluta, al que llamamos clase modal.

Para calcular la moda, aplicamos una fórmula que permite interpolar dentro del intervalo modal, teniendo en cuenta también las frecuencias de las clases anterior y posterior.

$$\hat{x} = M_0 = L_i + \left[\frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})}\right] \times A$$

Donde:

- L_i es el límite inferior del intervalo modal.
- f_i es la frecuencia absoluta del intervalo modal.
- f_{i-1} es la frecuencia absoluta del intervalo anterior al intervalo modal.
- f_{i+1} es la frecuencia absoluta del intervalo posterior al intervalo modal.
- A es la amplitud del intervalo modal.

MEDIANA PARA DATOS NO AGRUPADOS

La mediana (Me o $\tilde{x}$) es el valor que ocupa la posición central cuando los datos están ordenados de menor a mayor. Divide al conjunto en dos partes iguales: la mitad de los valores queda por debajo y la otra mitad por encima.

Es una medida especialmente útil cuando existen valores extremos o atípicos, ya que no se ve influida por ellos como ocurre con la media.

Entonces una vez ordenados los datos en orden de magnitud creciente, la mediana se define como:

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & si\ n\ es\ impar \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right), & si\ n\ es\ par \end{cases}$$

En el caso impar, hay un único valor que queda en el centro de la lista ordenada. En el caso par, no hay un valor central exacto, por lo que se toma el promedio de los dos valores centrales.

Ejemplo: Calcular la media y la mediana del siguiente conjunto de datos:

2,9 ; 3,4 ; 5,1 ; 4,31 ; 15,9

Solución:

$$\bar{x} = \frac{2{,}9 + 3{,}4 + 5{,}1 + 4{,}31 + 15{,}9}{5} = \frac{31{,}61}{5} = 6{,}322$$

Como el número de datos es impar ($n = 5$), la mediana es el valor que ocupa la tercera posición: $\tilde{x} = 4{,}31$

Podemos observar como la media es influida de manera considerable por la presencia de la observación extrema, 15,9; en cambio, la mediana, al depender solo de la posición en el conjunto ordenado, refleja mejor el centro real de la distribución.

MEDIANA PARA DATOS AGRUPADOS

Cuando trabajamos con datos agrupados en intervalos, no podemos identificar directamente el valor que ocupa la posición central, porque los datos individuales están agrupados en clases. Por eso, utilizamos un procedimiento para calcular la mediana dentro del intervalo correspondiente.

Lo primero será identificar el intervalo de la mediana, es decir el intervalo cuya frecuencia relativa acumulada es mayor o igual a 0,5 ($F_r \geq 0{,}5$)

Una vez identificado el intervalo de la mediana, vamos a aplicar la siguiente fórmula para calcularla:

$$\tilde{x} = M_e = L_i + \left[\frac{\frac{n}{2} - F_{i-1}}{f_i}\right] \times A$$

Donde:

- L_i es el límite inferior del intervalo de la mediana.

- f_i es la frecuencia absoluta del intervalo de la mediana.
- F_{i-1} es la frecuencia absoluta acumulada del intervalo anterior al intervalo de la mediana.
- A es la amplitud del intervalo de la mediana.

Aunque el nombre del intervalo es “intervalo de la mediana”, el valor que obtenemos no es la marca de clase, sino un valor más preciso del punto que divide a la distribución en dos mitades.

COMPARACIÓN ENTRE MEDIA, MEDIANA Y MODA. ASIMETRÍA

Las tres medidas de tendencia central (media, mediana y moda) pueden coincidir o diferir entre sí, según cómo se distribuyan los datos. La relación entre ellas nos puede dar una idea general de la forma de esta distribución, especialmente sobre si es simétrica o asimétrica.

Distribución simétrica

- La media, la mediana y la moda coinciden o están muy cercanas.
- Los datos están distribuidos de forma equilibrada alrededor del centro.
- Se dice que la distribución no presenta asimetría.

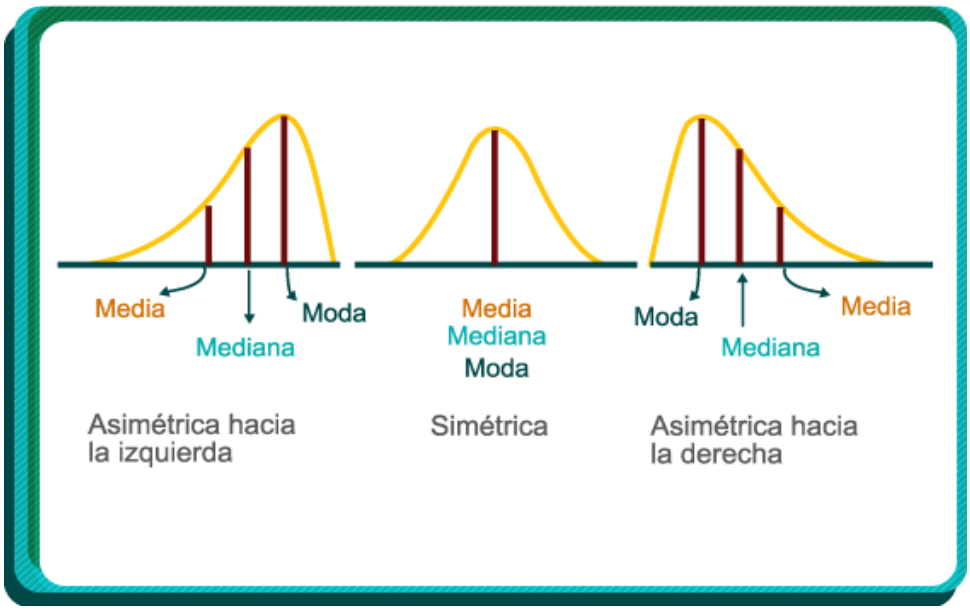
Distribución asimétrica positiva (asimetría a la derecha)

- Media > Mediana > Moda
- Hay valores extremos grandes (outliers) que arrastran la media hacia la derecha.
- El grueso de los datos se concentra en los valores bajos.
- Se dice que la distribución tiene asimetría positiva.

Distribución asimétrica negativa (asimetría a la izquierda)

- Media < Mediana < Moda

- Hay valores extremos pequeños que empujan la media hacia la izquierda.
- La mayoría de los datos se concentra en los valores altos.
- Se dice que la distribución tiene asimetría negativa.



MEDIDAS DE POSICIÓN

CUARTILES

Los cuartiles son medidas de posición que dividen un conjunto de datos ordenados en cuatro partes iguales, dejando el 25 % de los datos en cada tramo.

Son especialmente útiles para comprender la distribución de los valores, más allá de lo que nos indica un promedio.

Podemos pensar los cuartiles como posiciones de referencia que nos permiten conocer cómo se distribuyen los datos alrededor de la mediana.

El primer cuartil (Q_1), es aquel valor que deja el 25% de los datos por debajo y el 75% por encima.

El segundo cuartil (Q_2) es aquel valor que divide al conjunto de datos en dos mitades iguales, 50 % de los datos por debajo y 50 % por encima. Por definición coincide con la mediana.

El tercer cuartil (Q_3) es el valor que deja el 75% de los datos por debajo y el 25% por encima.

Ejemplo: Datos no agrupados

Supongamos el siguiente conjunto de datos ordenados $(n = 12)$:

$$4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10,\ 11,\ 12,\ 13,\ 14,\ 15$$

Q_1: Se encuentra en la posición $\frac{n+1}{4} = \frac{13}{4} = 3{,}25$, este valor nos da la posición, debemos ahora encontrar el valor del primer cuartil para eso Interpolamos entre el 3º y 4º dato:

$$Q_1 = 6 + 0{,}25 \cdot (7 - 6) = 6{,}25$$

Q_2: es la mediana. Promedio entre los valores de las posiciones 6 y 7.:

$$Q_2 = \frac{9 + 10}{2} = 9{,}5$$

Q_3: Se encuentra en la posición $\frac{3(n+1)}{4} = \frac{39}{4} = 9{,}75$, nuevamente este valor nos da la posición, debemos ahora encontrar el valor del tercer cuartil para eso Interpolamos entre el 9º y 10º dato:

$$Q_3 = 12 + 0{,}75 \cdot (13 - 12) = 12{,}75$$

¿Y con datos agrupados?

Cuando los datos están organizados en intervalos de clase, se aplican fórmulas similares a las de la mediana, identificando primero el intervalo que contiene el cuartil y luego aplicando interpolación lineal.

Los cuartiles son especialmente útiles cuando se desea entender la variabilidad y dispersión en distintos tramos del conjunto de datos. Por ejemplo, en estudios de salarios, los cuartiles permiten distinguir a los empleados con ingresos más bajos (debajo de Q_1), los ingresos promedio (alrededor de la mediana o Q_2) y los ingresos altos (por encima de Q_3), facilitando un análisis más equitativo y detallado de la distribución.

PERCENTILES

Los percentiles son medidas de posición que dividen un conjunto de datos ordenados en 100 partes iguales. Cada percentil indica el valor por debajo del cual se encuentra un

determinado porcentaje de los datos.

Son una generalización de los cuartiles:

- El percentil 25 coincide con Q_1
- El percentil 50 coincide con Q_2 (la mediana)
- El percentil 75 coincide con Q_3

Por ejemplo, si se mide el tiempo de respuesta de una API en 1.000 ejecuciones, y el percentil 90 (P_{90}) es 2,5 segundos, eso significa que el 90 % de las respuestas fueron más rápidas que 2,5 segundos.

MEDIDAS DE DISPERSIÓN

Una medida de tendencia central, como la media, no nos dice cuán dispersos o agrupados están los datos alrededor de ese valor. Dos conjuntos pueden tener la misma media, pero comportamientos completamente diferentes si uno tiene valores muy concentrados y el otro, valores muy alejados entre sí.

Por ejemplo, cuando los datos están muy concentrados alrededor de la media, el valor promedio ofrece una buena descripción del conjunto de datos ya que la mayoría de los valores individuales están cerca de la media.

Por otro lado, si los datos están muy dispersos, el promedio puede no representar adecuadamente a la mayoría de los valores. En estos casos, valores muy alejados pueden influir significativamente en la media, y esta puede no coincidir con la percepción central del conjunto.

Por eso, necesitamos medir el grado de variación de los datos. Las medidas de dispersión nos ayudan a complementar la información que brindan la media, la mediana o la moda.

Estudiaremos ahora las principales medidas de dispersión: rango, varianza, desviación estándar y coeficiente de variación.

RANGO

El rango (R) es la medida más simple de dispersión. Se calcula como la diferencia entre el valor máximo y el valor mínimo de un conjunto de datos:

$$R = x_{max} - x_{min}$$

Si bien da una idea rápida de la extensión general de los valores, solo tiene en cuenta

dos datos del conjunto y no muestra cómo se distribuyen el resto. Por eso, se considera una medida rudimentaria.

VARIANZA

La varianza es una medida que nos indica cuánto varían los datos con respecto a la media. Es decir, nos permite saber si los valores están concentrados cerca de la media o alejados de esta.

¿Cómo se calcula la varianza?

Se calcula la media $\bar{x}$.

Se mide cuánto se aleja cada dato respecto a la media: $x_i - \bar{x}$

Esas diferencias pueden ser positivas o negativas, por lo que se elevan al cuadrado para evitar que se cancelen entre sí: $(x_i - \bar{x})^2$

Finalmente, se promedian esos cuadrados, obteniendo una medida general de dispersión.

VARIANZA PARA DATOS NO AGRUPADOS

a) Si los datos provienen de una población:

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

Para el cálculo se utiliza la media poblacional μ y se divide por el total de los datos N.

b) Si los datos provienen de una muestra:

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

Para el cálculo se utiliza la media muestral $\bar{x}$ y se divide por un dato menos de los que hay en la muestra $n - 1$.

¿Por qué se divide por $n - 1$ y no por n en el caso del cálculo de la varianza de una muestra?

Cuando calculamos la varianza de una muestra, usamos la media de ese mismo

conjunto de datos como referencia.

Pero como la media fue calculada a partir de los mismos datos, eso hace que las diferencias respecto a ella tiendan a ser un poco más pequeñas de lo que serían si conociéramos toda la población.

Para que el resultado no quede subestimado, en lugar de dividir por n, se divide por $n-1$.

Esa pequeña corrección permite que la medida sea más justa y se acerque mejor a lo que pasaría si tuviéramos todos los datos posibles.

VARIANCIA PARA DATOS AGRUPADOS

Como vimos anteriormente, cuando los datos están agrupados en intervalos al no conocer cada valor individual usamos la marca de clase como valor representativo de ese grupo de datos que se encuentran dentro del intervalo.

Para calcular la varianza para datos agrupados, también debemos multiplicar cada desvío al cuadrado por su frecuencia absoluta, ya que cada marca de clase representa a varios datos.

La fórmula es:

$$S^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n-1}$$

Donde:

- x_i es la marca de clase de cada intervalo
- f_i es la frecuencia absoluta del intervalo
- $\bar{x}$ es la media muestral
- n es el total de observaciones

Al igual que en la varianza muestral para datos no agrupados, se divide por $n-1$ como forma de corrección.

DESVÍO ESTÁNDAR

El desvío estándar también es una medida que indica cuánto varían los datos respecto de la media, pero expresada en las mismas unidades que la variable original.

Cuando calculamos la varianza, al elevar al cuadrado las diferencias respecto de la media, el resultado queda expresado en unidades cuadradas (por ejemplo, si medimos en días, la varianza estará en días²).

Esto no siempre es conveniente para interpretar los resultados, ya que no usamos unidades cuadradas en la vida cotidiana.

Por eso, tomamos la raíz cuadrada de la varianza y así obtenemos una medida de dispersión más fácil de leer e interpretar, que se conoce como desvío estándar.

$$S = \sqrt{S^2}$$

¿Cómo interpretamos el desvío estándar?

El desvío estándar no solo mide cuánto se alejan los datos de la media en promedio, también nos permite comparar conjuntos de datos y evaluar la homogeneidad o variabilidad interna de una distribución.

Si el desvío estándar es pequeño, los valores están muy concentrados alrededor de la media y si el desvío estándar es grande, los datos están más dispersos y alejados del valor central.

Pero el tamaño del desvío estándar por sí solo no dice mucho si no lo relacionamos con la escala de los datos.

Por ejemplo, un desvío estándar de 2 días en un conjunto con media 5 días es relativamente grande, pero el mismo desvío de 2 días en un conjunto con media 60 días es pequeño en proporción.

Por esto, es conveniente evaluar el desvío estándar no de forma aislada sino relacionado con la media:

¿Qué tan grande es la dispersión respecto al valor promedio?¿Cuánto representa esa variabilidad respecto al total?

Esta comparación entre variabilidad (desvío estándar) y valor central (media) es clave para interpretar los resultados en contexto.

Y justamente eso es lo que nos lleva al próximo concepto: el coeficiente de variación.

COEFICIENTE DE VARIACIÓN

El coeficiente de variación es una medida de dispersión relativa, que expresa cuánto representa la variabilidad con respecto a la media, en forma de porcentaje.

El CV es especialmente útil cuando queremos comparar la dispersión de dos o más conjuntos de datos que tienen medias distintas, comparar datos que están expresados en diferentes unidades (por ejemplo: ingresos en pesos vs. dólares, o longitudes en metros vs. centímetros) o evaluar qué tan consistente es un conjunto de datos en relación con su media.

A diferencia del desvío estándar, el coeficiente de variación no depende de las unidades, decimos entonces que es una “medida adimensional”.

Su fórmula es:

$$CV = \frac{S}{\bar{x}} \times 100$$

Donde:

- S es el desvío estándar
- $\bar{x}$ es la media aritmética
- Multiplicamos por 100 para expresar el resultado en porcentaje (%).

¿Cómo interpretar el resultado del Coeficiente de Variación?

Un CV bajo indica que la dispersión es pequeña en relación con la media y los datos son relativamente homogéneos mientras que un CV alto sugiere que hay gran variabilidad respecto al promedio y los datos son menos consistentes.

Por ejemplo, un conjunto con CV = 8 % es más homogéneo que uno con CV = 26 %, aunque ambos tengan desvíos estándar parecidos.

RANGO INTERCUARTIL

El rango intercuartil (RI) es una medida de dispersión que nos indica qué tan dispersos

están los valores centrales de un conjunto de datos.

Se calcula como la diferencia entre el tercer cuartil Q_3 y el primer cuartil Q_1.

$$RI = Q_3 - Q_1$$

Es decir, el rango intercuartil abarca el 50 % central de los datos, eliminando el efecto de los valores extremos.

El RI es especialmente apropiado cuando la mediana es la medida de tendencia central elegida, ya que ambas se basan en la posición de los datos y no se ven afectadas por valores atípicos.

Por eso, se lo considera la medida de dispersión más adecuada para acompañar a la mediana.