Medidas de variación

Así como se mide la tendencia de los datos por concentrarse alrededor de un valor particular (tendencia central), así también se calcula la variabilidad o dispersión de un conjunto de datos para lograr una mejor descripción de ellos.

Analicemos el caso de cuatro posibles resultados del juego, tiro con dardos:

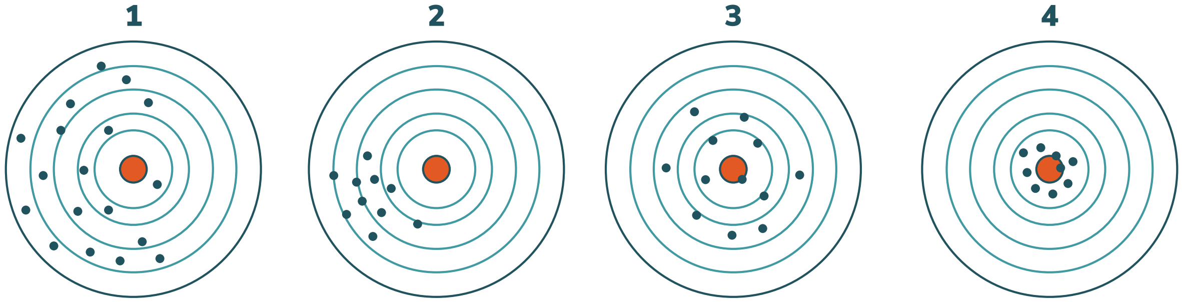

Figura 1. *Representación gráfica del juego con dardos.*

Caso 1: el jugador ni es exacto, ni es preciso.

Caso 2: el jugador tiene precisión, pero le falta exactitud.

Caso 3: el jugador tiene exactitud, pero poca precisión.

Caso 4: el jugador es preciso y exacto.

Al hacer un muestreo para determinar las características de una población ocurren situaciones similares a las mostradas. Por ello, es necesario estudiar la exactitud de la información obtenida con las medidas de centro y determinar la precisión de la información con las medidas de dispersión o variación.

Medidas de dispersión o variación

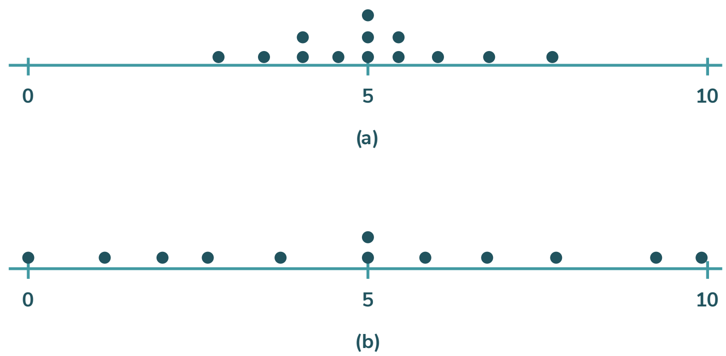

Figura 2. *Diagramas de puntos con valores de centro similares pero diferentes variaciones.*

En la figura 2 la información del gráfico (b) está más dispersa, es menos precisa que en (a).

El rango = Vmáx. – Vmin.

Es una medida de dispersión muy fácil de calcular, pero solo da una medida general de la variabilidad de los datos. Su utilidad se observa al comparar rangos de diferentes muestras de la variable en estudio.

Es poco informativo porque solo tiene en cuenta los valores extremos de la muestra, es muy sensible a la variación de uno de esos valores y al tamaño de la muestra.

Figura 3. *Rangos de diferentes muestras.*

Ejemplo

Los registros diarios de temperatura en °C en las ciudades A y B durante la misma semana fueron:

Ciudad A: 16, 16, 18, 17, 22, 19, 18

$\bar{x}_A = 18$

Mediana: 16, 16, 17, 18, 18, 19, 22

Me = 18

Rango = Vmáx – Vmín = 22 – 16 = 6

Ciudad B: 16, 12, 15, 18, 21, 25, 19

$$\bar{x}_B = 18$$

Mediana: 12, 15, 16, 18, 19, 21, 25

Me = 18.

Rango = Vmáx – Vmín = 25 – 12 = 13

Aun siendo iguales las medidas de centro calculadas, la dispersión de la segunda muestra es algo más que el doble de la primera.

Rango Intercuartílico = Q3 – Q1

Es la diferencia que hay entre el tercer y el primer cuartil.

Da una idea de la variabilidad del 50% de los datos centrales de la muestra y no tiene en cuenta los valores extremos.

Desviación media

La media muestral es una medida de centro; la variación de los datos en forma individual con respecto a ese centro se ve reflejada en las diferencias:

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_n - \bar{x})$$

Las cuales se han denominado como **desviaciones** de la media o simplemente **desviaciones.**

El promedio de estas desviaciones podría ser un indicador de la variación global de la muestra, DM.

Sin embargo, ocurre que siempre:

$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$$

Demostración gráfica

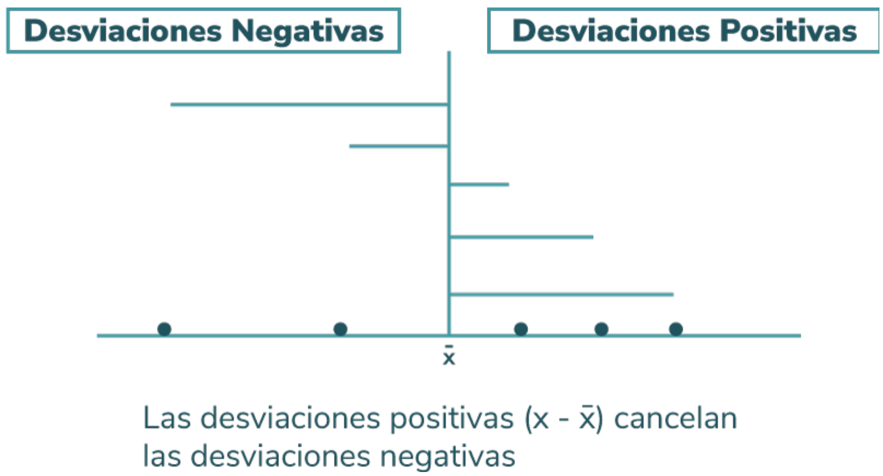

Figura 4. *Representación gráfica de las desviaciones.*

Se define entonces la desviación media:

$$DM = \frac{1}{n}\sum_{i}|x_i - \bar{x}|$$

Tiene en cuenta cada dato y su "distancia" a un valor central como lo es la media. Esto es lo que se conoce como "el error".

Deben eliminarse los signos de las desviaciones negativas antes de promediarlos. ¿Cómo?

Un número al cuadrado, siempre es positivo. Entonces, si:

$$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \dots, (x_n - \bar{x})^2$$

Y se promedia.

La varianza

La varianza muestral, **s²** de un conjunto de **n** mediciones $x_1, \ldots, x_n$ está definida como:

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

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

Desviación estándar

$$s = \sqrt{\text{var}\,ianza} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}}$$

Ejemplo: calcular la varianza y la desviación estándar de las temperaturas en las ciudades A y B, de los ejemplos anteriores.

$$S_A = \sqrt{\text{var}\,i\,anza} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{4.33} = 2.1\,°C$$

$$S_B = \sqrt{\text{var}\,i\,anza} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{16.43} = 4.2\,°C$$

Estas desviaciones estándar son comparables porque las muestras tienen la misma media. La ciudad B muestra mayor variación.

Ejemplo: la producción diaria en litros de las 10 vacas lecheras de un ejercicio anterior fue:
30, 28, 25, 18, 24, 29, 28, 21, 20, 22

$$\sum x_i = 30 + 28 + \ .\ .\ .\ + 22 = 245$$

$$\sum {x_i}^2 = 900 + 784\ +\ .\ .\ .\ + 484 = 6159$$

$$S^2 = \frac{(245)^2 - 10\ x\ 24{,}5^2}{10\ - 1} = 17{,}39\ l^2$$

S = 4,17 litros

Coeficiente de variación

Definido como:

$$Cv = \frac{s}{\overline{x}}$$

Para datos que representen distintas mediciones (muestras) de una misma magnitud, s es un valor promedio del error de medición, por lo tanto, Cv indica la magnitud promedio del error como porcentaje de la cantidad medida.

En datos de una población homogénea, Cv es menor que la unidad.

Cuando Cv es cercana a 1,0 es conveniente investigar posibles fuentes de heterogeneidad en los datos, por ejemplo: *mediciones con distintos instrumentos*, *personas de distinto sexo*, *distintos momentos temporales*, etc.

Tabla 1. *Datos representativos sobre la producción.*

Coeficiente de variación	Precisión
Hasta un 10%	Buena
De 11 a 20%	Aceptable
Superior al 20%	No confiable

Corresponde al investigador determinar si un valor obtenido de una muestra con un Cv mayor al 20% es útil o no en la toma de decisiones.

Ejemplo: en un experimento sicológico con seis sujetos, se utilizó una señal de estimulación de intensidad fija.

Los tiempos de reacción registrados en segundos fueron: 4, 2, 3, 3, 6, 3.

Calcular la media, la varianza, la desviación estándar y Cv para los datos.

Solución

Los cálculos se pueden llevar a cabo en forma esquemática:

Tabla 2. *Esquematización de los datos (1).*

X	4	2	3	3	6	3	$\sum x_i = \quad \bar{x} =$
$(x_i - \bar{x})$							$\sum (x_i - \bar{x}) =$
$(x_i - \bar{x})^2$							$\sum (x_i - \bar{x})^2 =$

Tabla 3. *Esquematización de los datos (2).*

X	4	2	3	3	6	3	$\sum x_i = 21 \quad \bar{x} = \frac{21}{6} = 3.5$
$(x_i - \bar{x})$	0.5	-1.5	-0.5	-0.5	2.5	-0.5	$\sum (x_i - \bar{x}) = 0$
$(x_i - \bar{x})^2$	0.25	2.25	0.25	0.25	6.25	0.25	$\sum (x_i - \bar{x})^2 = 9.50$

La medida muestral:

$$\bar{x} = 3{,}5\ segundos$$

La varianza muestral:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{9{,}5}{5} = 1{,}9\ segundos^2$$

La desviación estándar = 1,4 segundos.

El coeficiente de variación = 1,4/3,5 = 0,394.

La imprecisión se debe al tamaño de la muestra.

Ejemplo: los precios promedios por kilo en pie de ganado mayor y menor en la feria de ganados en un año cualquiera, fueron:

Tabla 4. *Promedios del ganado.*

Calidad primera categoría	Precio promedio anual	Desviación típica	
Macho cruzado	5731	499	8.71 %
Hembra cruzada	4740	388	8.19 %
Porcino	9444	796	8.43 %

Se puede concluir que el precio del ganado mayor y menor tuvo la misma variación a través del año, para la primera categoría. La información es confiable.

Referencias bibliográficas

Díaz Cadavid, A. y Gutiérrez Arias, A. (1991). *Estadística General*. Universidad de Antioquia.

Wayne, W. D. (1991). *Bioestadística: base para el análisis de las ciencias de la salud*. Limusa S.A.