Medidas de Dispersión

Estadística IV° Medio

Profesor Alberto Alvaradejo Ojeda

Dispersión es el grado de variación o diseminación de los datos.

Dos conjuntos de datos pueden diferir tanto en tendencia central como en dispersión o dos conjuntos de datos pueden tener las mismas medidas de tendencia central, pero diferir mucho en términos de dispersión.

Ejemplo:

$$1)\ 2, 2, 2, 2, 2 \quad \overline{x} = 2$$
$$2)\ 1, 1, 2, 3, 3 \quad \overline{x} = 2$$

Los estadígrafos de dispersión nos indican si la distribución o conjunto de datos forma grupos homogéneos o heterogéneos.

Las medidas de dispersión a estudiar son: **rango**, **desviación media**, **rango intercuartilico**, **varianza y desviación estándar**.

Rango

Indica el número de valores que toma la variable. El rango es la diferencia entre el valor máximo y el valor mínimo de un conjunto de datos.

$$R = x_{máx} - x_{mín}$$

Si los datos están agrupados en una tabla de frecuencias, el recorrido es la diferencia entre el límite superior del último intervalo y el límite inferior del primer intervalo.

$$R = LS_{máx} - LI_{mín}$$

Ejemplo:

1) Sea el siguiente conjunto de datos

12, 15, 17, 23, 25, 28

$x_{máx} = 28$ $x_{mín} = 12$ $R = x_{máx} - x_{mín}$ = 28 – 12 = 16

2) En datos agrupados

Peso $(Kg.)$	f_i
$55,0 - 63,0$	5
$63,1 - 71,1$	15
$71,2 - 79,2$	12
$79,3 - 87,3$	5
$87,4 - 95,4$	3
Total	40

$R = L_{máx} - L_{mín}$ = 95,4 – 55 = 40,4

Desviación Media

Es la media aritmética de los valores absolutos de las desviaciones de todos los datos respecto a la media aritmética. Su símbolo es ***DM***.

Es distancia promedio que están los datos del la media aritmética

a) Desviación media para datos no agrupados

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

Ejemplo: Obtener la desviación media para los datos 5, 7, 8, 10 y 16

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

$$\overline{x} = \frac{5+7+8+10+16}{5} = 9,2$$

$$DM = \frac{|5-9,2| + |7-9,2| + |8-9,2| + |10-9,2| + |16-9,2|}{5}$$

$$DM = \frac{15,2}{5} \qquad DM = 3,04$$

b) Desviación media para datos agrupados

$$DM = \frac{\sum_{i=1}^{m} |\, x_i - \overline{x} \,|\, f_i}{n}$$

Donde $\boldsymbol{x}_i$ es la marca de clase

Ejemplo: Determinar la **desviación media** de los siguientes datos agrupados

Pesos $(Kg.)$	f_i
$60-62$	5
$63-65$	18
$66-68$	42
$69-71$	27
$72-74$	8
Total	100

Pesos $(Kg.)$	x_i	f_i	$x_i \cdot f_i$	$x_i - \overline{x}$	$\|x_i - \overline{x}\|\, f_i$
$60-62$	61	5	305	$6,45$	$32,25$
$63-65$	64	18	1152	$3,45$	$62,10$
$66-68$	67	42	2814	$0,45$	$18,90$
$69-71$	70	27	1890	$2,55$	$68,85$
$72-74$	73	8	584	$5,55$	$44,40$
Total		100	6745		$226,5$

$$\overline{x} = \frac{6745}{100} = 67,45$$

$$DM = |61-67,5| \cdot 5 + |64-67,5| \cdot 18 + |67-67,5| \cdot 42 + |40-67,5| \cdot 27 + |73-67,5| \cdot 8 / 100$$

$$DM = \frac{226,5}{100} = 2,265$$

Rango intercuartiílico

Una medida que no es afectada por los valores extremos es el **rango intercuartílico** (**RIC**). Esta medida de variabilidad es la diferencia entre el tercer cuartil Q_3 y el primer cuartil Q_1. En otras palabras, el rango intercuartílico es el rango en que se encuentra el 50% central de los datos.

$$IQR = Q_3 - Q_1$$

Varianza y Desviación Estándar

Dos medidas de **dispersión** que se utilizan con frecuencia y que **sí** toman en consideración la forma en que se distribuyen los valores son la **varianza** y su raíz cuadrada, la **desviación estándar**.

Estas medidas establecen la forma en que los valores fluctúan con respecto a la media.

Varianza

La varianza se define como el promedio aritmético de las diferencias entre cada uno de los valores del conjunto de datos y la media aritmética del conjunto elevadas al cuadrado.

Su símbolo es S^2 si estamos trabajando con una muestra

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

donde x_i representa los datos de la muestra.

Ejemplo: Determinar la varianza del siguiente conjunto de datos: 25, 12, 23, 28, 17, 15

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

$$\overline{x} = \frac{25+12+23+28+17+15}{6} = 20$$

$$S^2 = \frac{(25-20)^2 + (12-20)^2 + (23-20)^2 + (28-20)^2 + (17-20)^2 + (15-20)^2}{6-1}$$

$$S^2 = \frac{196}{5} \qquad \Rightarrow \qquad S^2 = 39,2 \text{ (en unidades al cuadrado)}$$

b) Varianza para datos agrupados

$$S^2 = \frac{\sum_{i=1}^{m} (x_i - \overline{x})^2 f_i}{n-1}$$

donde x_i es la marca de clase.

Ejemplo: Considere la tabla con los datos de los edades de 26 personas

Edades (años)	f_i
$15 - 20$	2
$21 - 26$	7
$27 - 32$	8
$33 - 38$	5
$39 - 44$	4
Total	26

Edades (años)	x_i	f_i	$x_i \cdot f_i$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 f_i$
$15 - 20$	$17,5$	2	$35,0$	$155,2516$	$310,5032$
$21 - 26$	$23,5$	7	$164,5$	$41,7316$	$292,1212$
$27 - 32$	$29,5$	8	$236,0$	$0,2116$	$1,6928$
$33 - 38$	$35,5$	5	$177,5$	$30,6916$	$153,458$
$39 - 44$	$41,5$	4	$166,0$	$133,1716$	$532,6864$
Total		26	$779,0$		$1290,4616$

$$\overline{x} = \frac{779,0}{26} = 29,96 \text{ años}$$

$$S^2 = \frac{1290,4616}{25} = 51,618 \,(\text{ en años}^2\,)$$

Desviación Típica o Desviación Estándar

Es la raíz cuadrada positiva de la Varianza. Su símbolo es si se está trabajando ***S*** con una muestra.

a) Desviación estándar para datos no agrupados

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

Donde $\boldsymbol{x}_i$ representa los datos de la muestra.

Ejemplo: para el conjunto de datos: 25, 12, 23, 28, 17, 15; donde se obtuvo una varianza de $\mathbf{S^2 = 39{,}2}$ su desviación estándar es $\boldsymbol{S} = \sqrt{\mathbf{39{,}2}} = \mathbf{6{,}26}$ unidades.

b) Desviación estándar para datos agrupados

$$S = \sqrt{\frac{\sum_{i=1}^{m} (x_i - \overline{x})^2 f_i}{n-1}}$$

Donde $\boldsymbol{x}_i$ es la marca de clase

Ejemplo: Para el ejemplo de los datos tabulados sobre las edades de 26 personas se obtuvo como varianza $\boldsymbol{S}\uparrow\mathbf{2} = \mathbf{51{,}618}$; luego su desviación estándar será:

$\boldsymbol{S} = \sqrt{\mathbf{51{,}618}} = \mathbf{7{,}18}$ ***a*ñ*os***

¿Qué indican la Varianza y la Desviación Estándar?

La varianza y la desviación estándar miden la dispersión "***promedio***" en torno a la media aritmética, es decir, cómo fluctúan las observaciones mayores por encima de la media aritmética y cómo se distribuyen las observaciones menores por debajo de ella.

La **varianza A**l calcularla se obtienen unidades al cuadrado cm^2, $pulgadas^2$, mm^2, $(edades)^2$, $(horas)^2$, etc.

La **desviación estándar**, cuyo valor está dado en las unidades originales: cm, pulgadas, mm, edades, horas, etc.

En los ejemplos anteriores:

a) Para la muestra de datos: 25,12,23,28,17,15 se obtuvo por desviación estándar S=6,26 (unidades). Esto indica que la mayor parte de los datos de esta muestra se agrupan dentro de 6,26 unidades por encima y por debajo de la media aritmética, es decir:

   i. 20 - 6,26 = 13,74
   ii. 20 + 6,26 = 26,26

b) Para el caso de los datos tabulados correspondientes a las edades de 26 personas, se obtuvo una desviación estándar de ***S=7,18*** años. La mayor parte de los datos están agrupados entre:

   i. 29,26 - 7,18 = 22,78 años
   ii. 29,26 + 7,18 =37,14 años

Edades (años)	f_i
$15-20$	2
$21-26$	**7**
$27-32$	**8**
$33-38$	**5**
$39-44$	4
Total	26

Criterio de Homogeneidad

Una distribución se considera **homogénea**, si la desviación estándar se encuentra entre la quinta y la cuarta parte del rango. Si no es así, entonces se considera que la muestra es **heterogénea**.

a) Para la muestra de datos:
25, 12, 23, 28, 17, 15

R = 28 – 12 = 16
S = 6,26

$$\left(\frac{R}{5},\frac{R}{4}\right)=\left(3{,}2;4{,}0\right)$$

$$S\notin\left(3{,}2;4{,}0\right)$$

La muestra es heterogénea

b) Para el caso de los datos tabulados de las edades de26 personas:

Edades (años)	f_i
15 – 20	2
21 – 26	7
27 – 32	8
33 – 38	5
39 – 44	4
Total	26

R = 44 - 15 = 29
S = 7,18 años

$$\left(\frac{R}{5},\frac{R}{4}\right)=\left(6;7{,}5\right)$$

$$S\in\left(6;7{,}5\right)$$

La muestra es homogénea

Observaciones

1) Cuanto más separados o dispersos estén los datos, es decir, para muestras heterogéneas, mayores serán el rango, la varianza y la desviación estándar.

2) Si los datos están más concentrados, es decir, para muestras homogéneas, menores serán el rango, la varianza y la desviación estándar.

3) Si todas las observaciones son iguales (de manera que no haya variación en los datos), el rango, la varianza y la desviación estándar serán iguales a cero.

Coeficiente de variación

En algunas ocasiones se requiere un estadístico descriptivo que indique cuán grande es la desviación estándar en relación con la media. Esta medida es el **coeficiente de variación** y se representa como porcentaje.

$$\left(\frac{\text{Desviación estándar}}{\text{Media}} \times 100\right)\%$$

Ejemplo: En los datos de los tamaños de los cinco grupos de estudiantes, se encontró una media muestral de 44 y una desviación estándar muestral de 8.

El coeficiente de variación es [(8/44) 100]% 18.2%.

Expresado en palabras, el coeficiente de variación indica que la desviación estándar muestral es 18.2% del valor de la media muestral.

Ejercicios

1. Considere una muestra con los datos 10, 20, 12, 17 y 16. Calcule el rango y el rango intercuarílico.

2. Considere una muestra que tiene como valores 10, 20, 12, 17 y 16. Calcule la varianza y la desviación estándar.

3. Considere una muestra con valores 27, 25, 0, 15, 30, 34, 28 y 25. Calcule el rango, el rango intercuartílico, la varianza y la desviación estándar.

4. Las puntuaciones obtenidas por un jugador de boliche en seis juegos fueron 182, 168, 184, 190, 170 y 174. Use estos datos como una muestra y calcule los estadísticos descriptivos siguientes:
   a) Rango
   b) Desviación estándar
   c) Varianza
   d) Coeficiente de variación

5. A home theater in a box es la manera más sencilla y económica de tener sonido envolvente en un centro de entretenimiento en casa. A continuación se presenta una muestra de precios (Consumer Report Buying Guide 2004). Los precios corresponden a modelos con y sin reproductor de DVD.

Modelos con reproductor de DVD	Precio	Modelos sin reproductor de DVD	Precio
Sony HT-1800DP	$450	Pioneer HTP-230	$300
Pioneer HTD-330DV	300	Sony HT-DDW750	300
Sony HT-C800DP	400	Kenwood HTB-306	360
Panasonic SC-HT900	500	RCA RT-2600	290
Panasonic SC-MTI	400	Kenwood HTB-206	300

a) Calcule el precio medio de los modelos con reproductor de DVD y el precio medio de los modelos sin reproductor de DVD. ¿Cuánto es lo que se paga de más por tener un reproductor de DVD en casa?

b) Calcule el rango, la varianza y la desviación estándar de las dos muestras. ¿Qué le dice esta información acerca de los precios de los modelos con y sin reproductor de DVD?

6. Las tarifas de arriendo de automóviles por día en siete ciudades del este de Estados Unidos son las siguientes (*The Wall Street Journal* 16 de enero de 2004).

Ciudad	Tarifa por día
Boston	$43
Atlanta	35
Miami	34
New York	58
Orlando	30
Pittsburgh	30
Washington, D.C.	36

a) Calcule la media, la varianza y la desviación estándar de estas tarifas.
b) En una muestra similar de siete ciudades del oeste la media muestral de las tarifas fue de $38 por día. La varianza y la desviación estándar fueron 12.3 y 3.5 cada una. Analice la diferencia entre las tarifas de las ciudades del este y del oeste.