

MEDIDAS DE DISPERSION

Las medidas que hasta ahora conocemos, medias, moda, percentiles, etc., tienen todas ellas la propiedad de ubicarse siempre entre los dos valores extremos de los datos, mínimo y máximo, pues indican posición, bien sea central, o bien sea extrema como por ejemplo el percentil 5, o el percentil 95.

Las medidas que van a ser estudiadas en este capítulo no gozan de esta propiedad, y persiguen como objetivo describir la homogeneidad o heterogeneidad de los datos.

Las medidas de tendencia central son insuficientes para describir el comportamiento de los datos, pues no proporcionan información acerca de cuán cerca o cuán lejos se encuentran estos datos, con relación a ese valor central.

Así por ejemplo el trío de datos {8, 9, 10} y {1, 10, 16} tienen ambos media 9; pero resulta obvio, que en el primero de ellos existe una menor desviación con respecto a este valor central, que en el segundo.

Medir la variabilidad resulta muy importante en diversas situaciones prácticas, pues a través de su medición se podrán comparar conjuntos de datos, y establecer cuando existe una mayor concentración de ellos en la región central.

Así por ejemplo, en estudios sociales las medidas de dispersión proporcionan la información requerida para analizar como es la distribución de los ingresos dentro de la sociedad; en los estudios de calidad industrial, estas mismas medidas de dispersión se utilizan para medir la precisión de las máquinas utilizadas en el proceso de producción.

Antes de comenzar a analizar las medidas de dispersión, se recomienda revisar las propiedades de la media aritmética. (Véase Medidas de Tendencia Central. Arvelo)

1 Varianza y Desviación Típica: Cuando se tiene un conjunto de datos $\{x_1, x_2, x_3, \dots, x_n\}$ sin agrupar, se define como desvío de cada dato con relación a su media aritmética, a la diferencia: $d_i = x_i - \bar{X}$.

Aunque los desvíos pueden ser calculados respecto de cualquier valor, salvo que se indique lo contrario, se sobreentiende que estos se calculan respecto de $\bar{X}$, y como sabemos, la suma de sus cuadrados es mínima, cuando estos desvíos se calculan respecto de la media aritmética.

Cuanto mayor sea este desvío, más alejado se encuentra el dato x_i de $\bar{X}$.

Estos desvíos pueden ser positivos o negativos, según el dato x_i se encuentre a la derecha o a la izquierda de $\bar{X}$, y su suma algebraica da siempre cero.

En consecuencia, no se puede tomar al promedio de los desvíos como medida de dispersión, debido a que este promedio sería siempre nulo.

El problema del signo de los desvíos puede ser resuelto de dos maneras, o bien tomando su valor absoluto, o como segunda opción elevándolos al cuadrado.

Para definir la varianza, se adopta esta segunda alternativa.

La varianza de un conjunto de datos cuantitativos $\{x_1, x_2, x_3, \dots, x_n\}$ sin agrupar, se define como la media aritmética del cuadrado de sus desvíos.

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

La expresión anterior es una definición, y por lo tanto debe ser aceptada como tal, sin demostración.

Sin embargo, debido a que la varianza es la más importante de las medidas de dispersión, es importante hacer las siguientes aclaratorias:

1°) La varianza es una medida de dispersión que representa exclusivamente lo que establece la definición: Media aritmética del cuadrado de los desvíos.

Cuanto más desviado esté un dato de $\bar{X}$, mayor será su cuadrado, y en consecuencia mayor será varianza.

La varianza no puede ser interpretada como algo diferente a lo que la definición misma establece.

2°) La varianza viene expresada en unidades de los datos al cuadrado, y así por ejemplo, si los datos están en centímetros, la varianza está en centímetros cuadrados.

Esto ocasiona que el orden de magnitud de la varianza sea completamente diferente al de los datos originales, y que además no sean comparables con ellos pues vienen expresados en unidades diferentes.

Por estos dos motivos, en muchas oportunidades interesa regresar a las unidades originales, y allí es cuando aparece el concepto de “**Desviación Típica**”, que se define simplemente como la raíz cuadrada de la varianza.

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

La desviación típica o “estándar”, viene en las mismas unidades de los datos, y constituye junto con la varianza las más importantes medidas de dispersión.

3°) Otro comentario importante con relación a la varianza, es el que se refiere a su denominador, si es “n” o “n-1”, pues a lo largo de toda la bibliografía estadística, existe una gran confusión con relación a este punto.

Cuando una medida se calcula tomando en consideración a toda una población recibe el nombre de “Parámetro Poblacional”; mientras que cuando se calcula sobre una muestra se llama “Estadígrafo”, “Estadístico Muestral”, o simplemente “estadístico”.

La nomenclatura más utilizada en “Inferencia Estadística” consiste en designar a los parámetros poblacionales con letras griegas, y a los estadísticos con letras latinas.

Así por ejemplo, si se tienen “N” datos cuantitativos que constituyen una población y sobre ellos se quiere calcular su media, la misma vendría

representada por: $\mu = \frac{\sum_{i=1}^{i=N} x_i}{N}$, y se llamaría “Media Poblacional”

Si de esa población de “N” elementos, se toma al azar una muestra de “n” y se

calcula su media, ésta se designaría por : $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$.

Uno de los problemas importantes que se estudia en “Inferencia Estadística” es el de estimación, que trata sobre la metodología a seguir para inferir el valor desconocido de un parámetro poblacional a partir del estadístico muestral.

Cuando esta nomenclatura se aplica sobre la varianza, se obtiene que la

poblacional designada por σ^2 ,viene dada por : $\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$; mientras que la

muestral : $S^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}$.

Sin embargo, en Inferencia Estadística se demuestra que especialmente para

muestras pequeñas ($n < 30$), un mejor estimador¹ de σ^2 es: $S_c^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}$ que recibe el nombre de “cuasi varianza muestral” .

Como la diferencia en la estimación, entre uno y otro es realmente muy pequeña especialmente para muestras grandes y además S^2 es mejor estimador, muchos autores definen de una vez a S_c^2 como “varianza muestral” y omiten definir al otro. La “Estadística Descriptiva” no hace distinción entre población y muestra, su único objetivo es describir el comportamiento de los datos, y por este motivo usa a S^2 con divisor “n” como definición de “Varianza de los datos”, sin pronunciarse sobre si es muestral o poblacional.

Propiedades de la Varianza

Propiedad N°1°: $S^2 \geq 0$.

Resulta obvio que por ser la varianza la media de cuadrados de los desvíos, sea siempre una cantidad positiva, pues los cuadrados siempre lo son.

Es importante destacar que el caso $S^2 = 0$ implica que todos los datos son iguales, es decir que no existe variabilidad, y recíprocamente cuando todos los datos son iguales entonces $S^2 = 0$.

Esta propiedad es común para todas las medidas de dispersión, es decir:

$$\text{Medida de Dispersión} = 0 \Leftrightarrow x_1 = x_2 = x_3 = \dots = x_n$$

Propiedad N°2°: Cuando los datos son sometidos a una transformación lineal $Y = a + b X$, entonces $S_Y^2 = b^2 S_X^2$

¹ Es un estimador “Inssegado”. Véanse textos de “Inferencia Estadística” .

Demostración : Supongamos que se tiene un conjunto de datos $\{x_1, x_2, x_3, \dots, x_n\}$ y se le somete a la transformación $Y = a + bX$, dando lugar a unos nuevos datos $\{y_1, y_2, y_3, \dots, y_n\}$.

Por definición : $S_Y^2 = \frac{\sum_{i=1}^n (y_i - \bar{Y})^2}{n}$; $y_i = a + bx_i$; por la transformación $\bar{Y} = a + b\bar{X}$; por propiedad N 5 de la media

Por lo tanto: $S_Y^2 = \frac{\sum_{i=1}^n (a + bx_i - a - b\bar{X})^2}{n} = \frac{\sum_{i=1}^n b^2(x_i - \bar{X})^2}{n} = b^2 S_X^2$.

Como corolario de esta propiedad N° 2 se deducen las siguientes:

2.a) Si se le suma a cada dato una constante, la varianza no se altera.

Es el caso : $b = 1$.

2.b) Si cada dato es multiplicado por una constante, la varianza queda multiplicada por el cuadrado de dicha constante. Es el caso $a = 0$.

2.c) Si se somete a los datos a una transformación lineal, la desviación típica que multiplicada por la pendiente de la transformación: $S_Y = b S_X$.

Ejemplo 1 Los siguientes datos representan el peso de 4 personas expresados en kilogramos: 64 , 70 , 66 y 80.

a) Calcular la varianza y la desviación típica.

b) Transformar los datos en libras, y recalcular su varianza y su desviación típica.

Solución: a) Para calcular la varianza, es necesario calcular primero la media.

$$\bar{X} = \frac{64 + 70 + 66 + 80}{4} = 70 \text{ Kgs.}$$

$$S_X^2 = \frac{(64 - 70)^2 + (70 - 70)^2 + (66 - 70)^2 + (80 - 70)^2}{4} = 38 \Rightarrow S = \sqrt{38} = 6,16 \text{ Kgs.}$$

b) El factor de conversión de kilogramos a libras es 2,2 , de manera que si el peso de estas 4 personas se expresara en libras los datos quedarían transformados en: 140,80 , 154,00 , 145,20 y 176,00 .

Para la varianza de los nuevos datos, se puede repetir el procedimiento anterior.

$$\bar{Y} = \frac{140,80 + 154,00 + 145,20 + 176,00}{4} = 154,00$$

$$S_Y^2 = \frac{(140,80 - 154)^2 + (154 - 154)^2 + (145,20 - 154)^2 + (176 - 154)^2}{4} = 183,92$$

$$S_Y = \sqrt{183,92} = 13,56$$

Gracias a la propiedad N° 2 de la varianza, este procedimiento es innecesario, pues los datos "Y" en libras son los datos "X" en kilogramos sometidos a la transformación: $Y = 2,2 X$, y por lo tanto:

$$S_Y^2 = b^2 S_X^2 = (2,2)^2 38 = 183,92 ; S_Y = b S_X = (2,2) 6,16 = 13,56$$

Propiedad N°3 : La varianza también puede ser calculada mediante la expresión:

$$S^2 = \frac{\sum_{i=1}^{i=n} x_i^2}{n} - \bar{X}^2$$

Esta es una expresión alternativa, cuyo cálculo suele ser más sencillo que el explicado con anterioridad a partir de la definición.

Esta expresión recibe el nombre de “cálculo de la varianza por momentos”, por una razón que se explicará en el capítulo siguiente, y establece que la varianza es igual a la media de los cuadrados menos el cuadrado de la media.

Demostración: Partiendo de la definición, y desarrollando el cuadrado, se tiene:

$$S^2 = \frac{\sum_{i=1}^{i=n} (x_i - \bar{X})^2}{n} = \frac{\sum_{i=1}^{i=n} (x_i^2 - 2x_i \bar{X} + \bar{X}^2)}{n} = \frac{\sum_{i=1}^{i=n} x_i^2 - 2\bar{X} \sum_{i=1}^{i=n} x_i + \sum_{i=1}^{i=n} \bar{X}^2}{n}$$

Como es $\bar{X}$ una constante, según las propiedades de las sumatorias se tiene que

$\sum_{i=1}^{i=n} \bar{X}^2 = n \bar{X}^2$, y además $\sum_{i=1}^{i=n} x_i = n \bar{X}$. Sustituyendo:

$$S^2 = \frac{\sum_{i=1}^{i=n} x_i^2 - 2\bar{X} \sum_{i=1}^{i=n} x_i + \sum_{i=1}^{i=n} \bar{X}^2}{n} = \frac{\sum_{i=1}^{i=n} x_i^2 - n \bar{X}^2}{n} = \frac{\sum_{i=1}^{i=n} x_i^2}{n} - \bar{X}^2$$

Como consecuencia de esta propiedad se tiene que una fórmula alternativa para

calcular la desviación típica es: $S = \sqrt{\frac{\sum_{i=1}^{i=n} x_i^2}{n} - \bar{X}^2}$

Ejemplo 2 Calcular la varianza y la desviación típica de los pesos en kilogramos 64, 70, 66 y 80 del Ejemplo 7.1, aplicando la fórmula por momentos.

Solución: Hay que calcular la media $\bar{X} = 70$

$$S^2 = \frac{64^2 + 70^2 + 66^2 + 80^2}{4} - (70)^2 = 38 \Rightarrow S = \sqrt{38} = 6,16$$

El cálculo de la varianza y de la desviación típica por momentos suele más breve que por definición; sin embargo se corre el riesgo de perder de vista lo que se está calculando, pues no se necesitan calcular los desvíos.

Propiedad N° 4: Desigualdad de Tchebychev. Dado un conjunto de datos

$\{x_1, x_2, x_3, \dots, x_n\}$ con media $\bar{X}$; la proporción de datos que se desvían de $\bar{X}$ en

una cantidad menor que ε , es por lo menos $1 - \frac{S^2}{\varepsilon^2}$.

Esta propiedad es general, y una vez demostrada veremos que es mucha utilidad práctica, especialmente cuando se aplica en ciertos casos particulares como por ejemplo, en la curva normal.

Demostración: Supongamos que existen “m” datos que se desvían de $\bar{X}$ en una cantidad menor que ε . Llamemos “C” al conjunto de esos “m” datos.

Se tiene: $m \leq n$, y además $|x_i - \bar{X}| < \varepsilon$, cuando $x_i \in C$

$p = \frac{m}{n}$ = Proporción de datos que pertenecen al conjunto “C”, y que por tanto

difieren de $\bar{X}$ en una cantidad menor que ε .

$$\text{Por definición : } S^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n} \Rightarrow n S^2 = \sum_{i=1}^n (x_i - \bar{X})^2$$

La suma $\sum_{i=1}^n (x_i - \bar{X})^2$ puede ser descompuesta en dos partes, sobre los datos que pertenecen al conjunto “C”, y sobre los que no pertenecen a él.

$$\text{Por tanto : } \sum_{i=1}^n (x_i - \bar{X})^2 = \sum_{x_i \in C} (x_i - \bar{X})^2 + \sum_{x_i \notin C} (x_i - \bar{X})^2$$

$$\text{Como: } \sum_{x_i \notin C} (x_i - \bar{X})^2 \geq 0 \Rightarrow n S^2 = \sum_{i=1}^n (x_i - \bar{X})^2 \geq \sum_{x_i \notin C} (x_i - \bar{X})^2$$

Existen $(n - m)$ datos $\{x_i\}$ no pertenecientes al conjunto “C”, y para ellos se

$$\text{verifica: } \forall x_i \notin C \Rightarrow |x_i - \bar{X}| \geq \varepsilon \Rightarrow (x_i - \bar{X})^2 \geq \varepsilon^2 \Rightarrow \sum_{x_i \notin C} (x_i - \bar{X})^2 \geq (n - m) \varepsilon^2$$

$$\text{Por lo tanto : } n S^2 \geq \sum_{x_i \notin C} (x_i - \bar{X})^2 \geq (n - m) \varepsilon^2 \Rightarrow \frac{S^2}{\varepsilon^2} \geq \frac{n - m}{n} = 1 - p$$

y en conclusión: $p \geq 1 - \frac{S^2}{\varepsilon^2}$ tal como se quería demostrar.

Frecuentemente se toma “ ε ” como un múltiplo de “S”, de la forma: $\varepsilon = k S$, y en este caso la desigualdad de Tchebychev establece que la proporción de datos que difiere su media en una cantidad menor que “k veces” la desviación típica, es por

lo menos $1 - \frac{1}{k^2}$, es decir: $p \geq 1 - \frac{1}{k^2}$

Es evidente que cuando $k=1$, esta desigualdad carece de utilidad práctica alguna, pues en este caso se concluye en: $p \geq 0$, lo que es obvio.

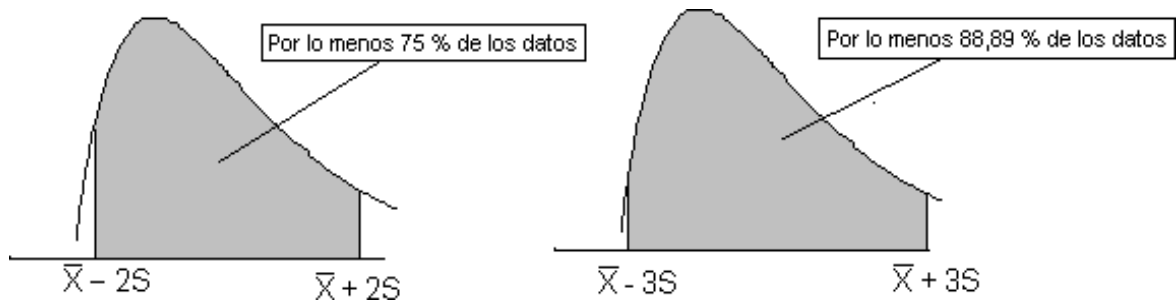
Sin embargo, para valores de “k” mayores que la unidad se obtienen conclusiones importantes, y así por ejemplo para $k = 2$ concluye que en un conjunto de datos cualquiera, la proporción de datos que difiere de su media en menos de dos veces

la desviación típica es por lo menos $1 - \frac{1}{4} = \frac{3}{4}$, es decir $p \geq 75\%$.

Cuando se toma $k = 3$, la conclusión es que la proporción de datos que difiere de su media en menos de tres desviaciones típicas es por lo menos de $1 - \frac{1}{9} = \frac{8}{9}$

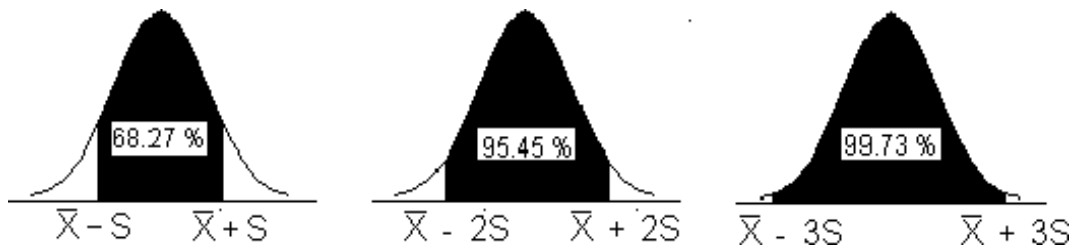
lo que equivale al 88,89 %.

Para cualquier distribución de frecuencias distribuciones , y según lo explicado anteriormente por la desigualdad de Tchevishev , en el intervalo $(\bar{X} \pm 2 S)$ deben caer por lo menos el 75% , y en el intervalo $(\bar{X} \pm 3 S)$ por lo menos el 88,89 % de los datos .



En el caso de la distribución normal , esta proporción de datos ha sido determinada exactamente , y los resultados obtenidos son de amplia utilidad en muchas aplicaciones.

Para una normal, en el intervalo $(\bar{X} \pm S)$ deben caer el 68,27% de los datos muestrales , en el intervalo $(\bar{X} \pm 2 S)$ deben caer el 95,45% de los datos muestrales , y en el intervalo $(\bar{X} \pm 3 S)$ deben caer el 99,73% de los datos muestrales .



Propiedad N° 5. Para el caso de datos agrupados, el cálculo de la varianza se

$$\text{hace mediante la expresión : } S^2 = \frac{\sum_{i=1}^{i=k} (L_i^* - \bar{X})^2 f_i}{\sum_{i=1}^{i=k} f_i} = \frac{\sum_{i=1}^{i=k} (L_i^*)^2 f_i}{\sum_{i=1}^{i=k} f_i} - \bar{X}^2$$

La expresiones anteriores son sencillas de justificar , pues constituyen adaptaciones de la definición de varianza, y de su cálculo por momentos, tomando en cuenta que para el caso de datos agrupados, la marca de clase L_i^* representa a todos los datos que caen dentro del intervalo.

Otra expresión útil para el cálculo, es la que resulta de sustituir en la fórmula de la derecha para el cálculo de S^2 por momentos, la media $\bar{X}$ por su correspondiente para datos agrupados en función de las marcas de clase, y por lo tanto:

$$S^2 = \frac{\sum_{i=1}^k (L_i^*)^2 f_i}{\sum_{i=1}^k f_i} - \left(\frac{\sum_{i=1}^k L_i^* f_i}{\sum_{i=1}^k f_i} \right)^2$$

Ejemplo 3 Calcular la varianza y la desviación típica de la siguiente distribución de frecuencias.

Intervalo	0 a 10	10 a 20	20 a 30	30 a 40	40 a 50	50 a 60	60 a 70
frecuencia	8	34	76	60	31	28	13

Solución : Hay que organizar los cálculos en la siguiente tabla :

Límites reales	f_i = frecuencia	L_i^* = Marca de clase	$L_i^* f_i$	$(L_i^*)^2 f_i$
0 a 10	8	5	40	200
10 a 20	34	15	510	7650
20 a 30	76	25	1900	47500
30 a 40	60	35	2100	73500
40 a 50	31	45	1395	62775
50 a 60	28	55	1540	84700
60 a 70	13	65	845	54925
	331		8330	331250

$$S^2 = \frac{331250}{331} - \left(\frac{8330}{331} \right)^2 = 214,78 \Rightarrow S = \sqrt{214,78} = 14,66$$

Método abreviado para el cálculo de la varianza en tablas de frecuencia para datos agrupados de igual amplitud

Los cálculos requeridos en el Ejemplo 7.3 pueden resultar complicados si no se tiene una buena calculadora.

Un procedimiento abreviado para efectuar los cálculos consiste en definir unas marcas de clase artificiales designadas por U_i^* .

Al intervalo de mayor frecuencia o clase modal se le da una marca de clase artificial $U_m^* = 0$, a los anteriores marcas de clases artificiales -1, -2, etc., y a los posteriores +1, +2, +3, etc.

Con este artificio, la tabla queda :

Límites reales	f_i = frecuencia	U_i^* = Marca de clase artificial	$U_i^* f_i$	$(U_i^*)^2 f_i$
0 a 10	8	-2	- 16	32
10 a 20	34	-1	- 34	34
20 a 30	76	0	0	0
30 a 40	60	1	60	60

40 a 50	31	2	62	124
50 a 60	28	3	84	252
60 a 70	13	4	52	208
Sumatorias	250		208	710

Posteriormente se calcula la varianza de los datos artificiales, mediante la

$$\text{expresión : } S_U = \frac{\sum_{i=1}^K (U_i^*)^2 f_i}{\sum_{i=1}^K f_i} - \frac{\left(\sum_{i=1}^K U_i^* f_i \right)^2}{\left(\sum_{i=1}^K f_i \right)^2} = \frac{710}{250} - \frac{208^2}{250^2} = 2,1478$$

Finalmente : $S^2 = c^2 S_U^2$, donde $c = \text{Amplitud} = 10 \Rightarrow S^2 = (10)^2 2,1478 = 214,78$ que coincide con el resultado anterior.

Justificación del método abreviado : Se ha definido una función lineal de los datos, que los transforma en unos datos artificiales más sencillos de trabajar.

Esta transformación es de la forma : $U = \frac{X - L_m^*}{c}$; donde teóricamente L^* pudiera

ser cualquier origen , pero que por conveniencia se toma la marca de clase del intervalo de mayor frecuencia, pues así la mayor frecuencia queda multiplicada por cero simplificando aún más los cálculos.

Por efecto de la transformación, las marcas de clase anteriores se convierten en -1, -2 , (siempre que la amplitud “c” sea igual para todos los intervalos) , y las marcas de clase posteriores se convierten en +1 , +2 , etc.

La expresión para calcular S_U^2 es la correspondiente al cálculo de la varianza por momentos para los datos artificiales “U”.

La relación entre los datos originales “X” y los artificiales “U” es lineal pues : $X = L_m^* + c U \Rightarrow$ Por la propiedad N° 2 de la varianza : $S_X^2 = c^2 S_U^2$, lo que constituye la justificación del método abreviado de cálculo.

2 La Media de las desviaciones absolutas : En virtud de que los desvíos respecto de $\bar{X}$ se compensan, otra manera de omitir su signo es considerar sus valores absolutos , en lugar de elevarlos al cuadrado como se hace en el cálculo de la varianza..

Cuando se promedian los desvíos absolutos, aparece otra medida de dispersión, que se conoce con el nombre de “Media de las desviaciones absolutas”, o “Desviación Media”, y que representa lo que en promedio se desvían los datos de $\bar{X}$, sin considerar su signo.

$$D.M = \frac{\sum_{i=1}^n |x_i - \bar{X}|}{n}$$

Ejemplo 4 Los siguientes datos representan la duración de cinco cuñas de T.V expresadas en segundos: 38 , 46 , 42 , 28 , 26 . Calcular la media de las desviaciones absolutas.

Solución : Se calcula su media $\bar{X} = \frac{38 + 46 + 42 + 28 + 26}{5} = 36$, y posteriormente

los desvíos: $38-36 = 2$, $46-36 = 10$, $42-36 = 6$, $28-36 = -8$ y $26-36 = -10$

La suma algebraica de los desvíos resulta ser cero, tal como debe ocurrir siempre, y para calcular la desviación media se promedian los desvíos absolutos:

$$D.M = \frac{2 + 10 + 6 + |-8| + |-10|}{5} = 7,20$$

Este resultado significa que en promedio, la duración de las cuñas se alejan en forma absoluta de su media 36 segundos, en 7,20 segundos .

La desviación media absoluta tiene las siguientes propiedades:

1º) Al igual que cualquier otra medida de dispersión $D.M \geq 0$, y $D.M = 0$ es condición necesaria y suficiente para que todos los datos sean iguales.

2º) En el caso de una distribución normal con desviación típica poblacional “ σ ”, existe una relación directa entre la D.M y “ σ ”, dada por la expresión² :

$$D.M = \sqrt{\frac{2}{\pi}} \sigma$$

A partir de esta propiedad se obtiene una fórmula aproximada entre la desviación media absoluta y la desviación típica, según la cual para curvas de frecuencias simétricas o con poco grado de asimetría, se verifica : $D.M \approx \frac{4}{5} S$

3º) Para datos agrupados, la marca de clase sustituye al verdadero valor de cada dato, y la desviación media absoluta se calcula mediante la expresión:

$$D.M = \frac{\sum_{i=1}^{i=k} |L_i^* - \bar{X}| f_i}{\sum_{i=1}^{i=k} f_i}$$

Ejemplo 5 Calcular la D.M para los datos del Ejemplo 3 , y verificar que se cumple la propiedad N° 2 .

Solución: La media es $\bar{X} = \frac{8.330}{250} = 33,32$, y se organizan los cálculos

Límites reales	f_i = frecuencia	L_i^*	$ L_i^* - \bar{X} $	$ L_i^* - \bar{X} f_i$
0 a 10	8	5	28,32	226,56
10 a 20	34	15	18,31	622,54
20 a 30	76	25	8,32	632,32
30 a 40	60	35	1,68	100,80
40 a 50	31	45	11,68	362,08
50 a 60	28	55	21,68	607,04
60 a 70	13	65	31,68	411,84
Sumatorias	250		208	2.963,18

² Para una demostración de esta propiedad, véase la obra de este mismo autor: “ Capacidad de los Procesos Industriales” , U.C.A.B 1998 .

$$D.M = \frac{2.963,18}{250} = 11,85$$

$$S = 14,66 \Rightarrow \text{Por la propiedad N}^\circ 3: D.M \approx \frac{4}{5} (14,86) = 11,89$$

3 La media de las desviaciones absolutas respecto de la mediana: La media de las desviaciones absolutas puede ser calculada también respecto de la mediana, dando lugar así a otra medida de dispersión, conocida también “Desviación Media respecto de la Mediana”, y que se define de la siguiente manera:

$$D.M_{Med} = \frac{\sum_{i=1}^{i=n} |x_i - \text{Mediana}|}{n}$$

Como una de las propiedades de la Mediana, estable que la suma de las desviaciones absolutas es mínima, cuando estas se calculan respecto de la ella, entonces se puede garantizar que : $D.M_{Med} \leq D.M$

$$\text{Para datos agrupados : } D.M_{Med} = \frac{\sum_{i=1}^{i=k} |L_i^* - \text{Mediana}| f_i}{\sum_{i=1}^{i=k} f_i}$$

Ejemplo 6 Calcular la $D.M_{Med}$ para los datos del Ejemplo .4

Solución : La mediana entre { 38 , 46 , 42 , 28 , 26} es 38, y las desviaciones respecto de ella son {0, 8, 4 , -10 y -12} respectivamente.

$$\text{Por lo tanto: } D.M_{Med} = \frac{0 + 8 + 4 + |-10| + |-12|}{5} = 6,80 < D.M = 7,20$$

4 El Rango o Recorrido: Esta medida de dispersión se define como la diferencia entre el mayor y el menor valor de los datos.

$$R = \text{Máximo Valor} - \text{Mínimo Valor}$$

El uso del rango o recorrido, es muy difundido en las aplicaciones industriales de la Estadística, pues dada su sencillez de cálculo se utiliza en la construcción de las gráficas de control³ conocidas como ($\bar{X}$, R) cuyo objetivo es detectar a tiempo la aparición de causas especiales dentro de un proceso de producción.

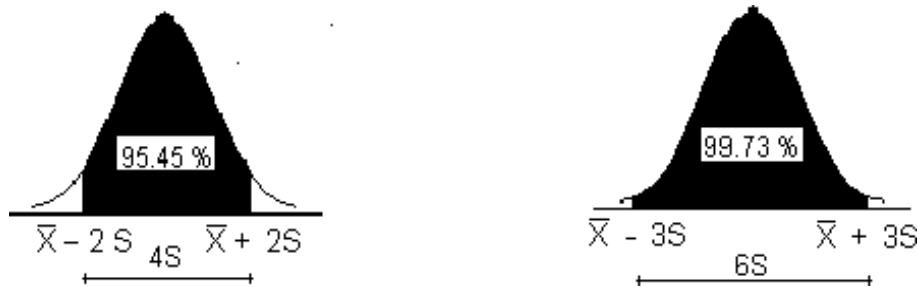
En el caso de la distribución normal, existe una relación aproximada entre el rango y la desviación típica, y así por ejemplo cuando :

$$n=100: S \cong \frac{R}{4} ; \quad \text{y cuando } n=1.000: S \cong \frac{R}{6}$$

estas relaciones empíricas son consecuencia de la propiedad ya estudiada para la distribución normal, según la cual en el intervalo ($\bar{X} \pm 2 S$) deben caer el 95,45% de los datos, y en el intervalo ($\bar{X} \pm 3 S$) deben caer el 99,73% .

³ Para una mayor información sobre estas gráficas, véase la misma referencia citada en la nota 2.

Cuando $n=100$, el rango debe ser aproximadamente 4 veces “S” que corresponde a la longitud del intervalo, y cuando $n= 1.000$, el rango debe ser aproximadamente 6 veces “S” , que es la longitud del intervalo $\bar{X} \pm 3S$.



Ejemplo 7 Calcular el rango para los datos del ejemplo 7.4

Solución : En este conjunto de datos el mayor valor es 46 y el menor 26 , en consecuencia: $R = 46 - 26 = 20$.

VII. 5 Otras Medidas absolutas de dispersión: Además de las ya estudiadas, existen otras medidas de menor aplicación práctica, tales como:

a) El Rango intercuartílico: Se define como la diferencia entre el tercer y primer cuartil, y representa la amplitud del intervalo 50% central.

$$R_q = Q_3 - Q_1$$

b) El Rango semi – intercuartílico : Es la mitad del rango intercuartílico, y representa el radio del intervalo 50% central .

$$Q = \frac{Q_3 - Q_1}{2} .$$

Con el rango semi-intercuartílico, el intervalo 50 % central puede ser escrito como:

$$\frac{Q_1 + Q_3}{2} \pm \frac{Q_3 - Q_1}{2} = \frac{Q_1 + Q_3}{2} \pm Q$$

Para distribuciones acampanadas: $Q \cong \frac{2}{3} S$

c) El Rango Percentílico : Se define como la diferencia entre el Percentil 90 y el Percentil 10, y representa la longitud del intervalo 80 % central .

$$R_p = P_{90} - P_{10}$$

Todas estas medidas de dispersión al igual que las anteriores son absolutas, pues no toman en cuenta el orden de magnitud de los datos, y vienen en sus mismas unidades, a excepción de la varianza que viene en unidades al cuadrado.

6 El Coeficiente de Variación o Dispersión Relativa: Las medidas absolutas de dispersión, al no tomar en cuenta el orden de magnitud de los datos, no proporcionan una información completa sobre su variabilidad, pues no es lo mismo por ejemplo, una desviación típica de 100 en unos datos que sean del orden de cientos, que esa misma desviación típica de 100 en unos datos que sean

del orden de millones. Resulta obvio que en el primer caso existe una variabilidad mucho mayor que en el segundo, a pesar de que el valor absoluto de la desviación típica sea el mismo.

Otro problema que tienen las medidas absolutas de dispersión es el de las unidades, pues esto impide hacer comparaciones entre conjuntos de datos que tengan diferente naturaleza.

Así por ejemplo, si se quisiera saber cuál variable tiene un comportamiento más homogéneo, el peso o la estatura de un conjunto de personas, no es posible comparar las desviaciones típicas entre esas ellas, por venir expresadas en diferentes unidades.

Para solucionar este par de inconvenientes que presentan las medidas absolutas de dispersión, se utiliza al coeficiente de variación o dispersión relativa, definido por:

$$C. V = \frac{S}{\bar{X}} 100\%$$

Propiedades del Coeficiente de Variación

1º) Es un porcentaje de razón entre la desviación típica y la media, de manera que representa cuantas veces es la desviación típica con relación a la media.

Así por ejemplo, un C.V = 50% significa que la desviación típica es la mitad de la media, lo que revela una alta variabilidad.

Valores del C.V menores al 10 % revelan poca variabilidad de los datos ; y así por ejemplo, para en el Control de Calidad de la preparación de mezclas, es frecuente exigir un C.V menor al 5% entre las muestras tomadas en distintos puntos de la mezcla, a fin de garantizar su homogeneidad.

2º) El C.V es un número abstracto, es decir sin unidades, pues tanto S como $\bar{X}$ vienen en las mismas unidades de los datos, y al hacer la división se simplifican.

Esta propiedad permite utilizar al C.V para hacer comparaciones entre varios conjuntos de datos, y concluir que cuanto más pequeño sea su valor, más homogéneo es el comportamiento.

3º) El C.V no se altera cuando los datos son multiplicados por una constante, pues en virtud de las propiedades de $\bar{X}$ y de "S" ambos quedan multiplicados por esa constante, sin alterar al cociente.

Esta propiedad trae como consecuencia que el C.V sea invariante frente a cambios de unidades, como, por ejemplo, pasar de libras a kilogramos o de pies a centímetros, etc.

Ejemplo 8 Calcular el C.V para los datos del ejemplo 3

Solución : $S = 14,66$, $\bar{X} = \frac{8.330}{250} = 33,32 \Rightarrow C.V = \frac{14,66}{33,32} 100\% = 44 \%$

Ejemplo 9 : Al analizar la estatura y el peso de un conjunto de personas, se obtuvo, para el peso una media de 68,50 Kgs. con una desviación típica de 9,20

kgs.; y para la estatura una media de 165 cms., con una desviación típica de 11,30 cms. ¿Cuál de las dos variables tiene un comportamiento más homogéneo?

Solución: Para el peso: $C.V = \frac{9,20}{68,50} 100\% = 13,43 \%$

Para la estatura: $C.V = \frac{11,30}{165} 100\% = 6,85 \%$

Se concluye que la estatura tiene un comportamiento más homogéneo.

Una de las limitaciones que tiene el coeficiente de variación, es que sólo puede ser utilizado cuando los datos corresponden a mediciones sobre una escala de razón, y por lo tanto existe el cero absoluto.

Cuando existen datos positivos y negativos, la media puede resultar igual a cero, negativa o muy próxima a cero, en cuyo caso este coeficiente de variación carece de sentido como medida de dispersión.

EJERCICIOS RESUELTOS

Ejemplo 10 Al examinar el diámetro de 300 tornillos, en milímetros, se encuentra:

Díámetro	2,00 a 2,49	2,50 a 2,99	3,00 a 3,49	3,50 a 3,99	4,00 a 4,49	4,50 a 4,99
Frecuencia	17	51	93	86	41	12

a) Calcule la varianza, la desviación típica, el coeficiente de variación, la media de desviaciones absolutas, la media de desviaciones absolutas respecto de la mediana, el rango intercuartílico y el rango percentílico.

b) Calcule el porcentaje de observaciones que caen en los intervalos $\bar{X} \pm \bar{S}$, $\bar{X} \pm 2S$ y $\bar{X} \pm 3S$.

Solución : Para calcular la varianza, se organizan los cálculos:

L_i^* = Marca de clase	f_i = Frecuencia	$L_i^* f_i$	$(L_i^*)^2 f_i$
2,245	17	38,1650	85,6804
2,745	51	139,9950	384,2863
3,245	93	301,7850	979,2923
3,745	86	322,0700	1.206,1522
4,245	41	174,0450	738,8210
4,745	12	56,9400	270,1803
Sumatorias	300	1.033,000	3.664,4125

$$S = \frac{\sqrt{\frac{3.664,4125}{300} - \left(\frac{1.033,00}{300}\right)^2}}{1} = 0,3582 \Rightarrow S = \sqrt{0,3582} = 0,5985$$

Utilizando el método abreviado de cálculo:

U_i^* = Marca de clase artificial	f_i = Frecuencia	$U_i^* f_i$	$(U_i^*)^2 f_i$
-2	17	-34	68
-1	51	-51	51
0	93	0	0
1	86	86	86

2	41	82	164
3	12	36	108
Sumatorias	300	119	477

$$S_U^2 = \frac{G_{119} K J}{300 H 300} = 1,4327$$

$$c = \text{Amplitud Real} = 0,50 \Rightarrow S^2 = c^2 S_U^2 = (0,50)^2 0,1,4327 = 0,3582$$

$$\text{Para calcular el coeficiente de variación, se necesita: } \bar{X} = \frac{1033}{300} = 3,4433$$

$$\text{Por el método abreviado: } U = \frac{119}{300} = 0,3967$$

$$\bar{X} = L_m^* + c \bar{U} \Rightarrow \bar{X} = 3,245 + (0,50) 0,3967 = 3,4433$$

$$\text{y por lo tanto: } C.V = \frac{S}{\bar{X}} 100\% = \frac{0,5985}{3,4433} 100\% = 17,38 \%$$

Para calcular la desviación media:

L_i^* = Marca de clase	f_i = Frecuencia	$ L_i^* - \bar{X} $	$ L_i^* - \bar{X} f_i$
2,245	17	1,1983	20,3711
2,745	51	0,6983	35,6133
3,245	93	0,1983	18,4419
3,745	86	0,3017	25,9462
4,245	41	0,8017	32,8697
4,745	12	1,3017	15,6204
Sumatorias	300		148,8626

$$\text{Por lo tanto } D.M = \frac{148,8626}{300} = 0,4962$$

Si se quisiera obtener un cálculo rápido pero aproximado, se pudiera aplicar la propiedad para distribuciones acampanadas como esta, según la cual:

$$D.M \approx \frac{4}{5} S = \frac{4}{5} (0,5985) = 0,4788$$

Para las restantes medidas de dispersión se necesitan los cuartiles, y los percentiles 10 y 90, a partir de las frecuencias acumuladas:

Diámetro	< 2,495	< 2,995	< 3,495	< 3,995	4,495	4,995
Frecuencia	17	68	161	247	288	300

$$Q_1 = 2,995 + \frac{\frac{300}{4} - 68}{93} 0,50 = 3,0326 ; Q_3 = 3,495 + \frac{\frac{3}{4}(300) - 161}{86} 0,50 = 3,8671$$

$$Q_2 = \text{Med} = 2,995 + \frac{\frac{1}{2}(300) - 68}{93} 0,50 = 3,4359$$

$$P_{10} = 2,495 + \frac{10}{100} \left(\frac{300}{51} - 1 \right) \cdot 0,50 = 2,6225$$

$$P_{90} = 3,995 + \frac{90}{100} \left(\frac{300}{41} - 1 \right) \cdot 0,50 = 4,2755$$

Para calcular la desviación media absoluta, respecto de la mediana, hay que organizar los cálculos en una tabla similar a la de la D.M , pero calculando los desvíos absolutos respecto de la mediana.

L_i^* = Marca de clase	f_i = Frecuencia	$ L_i^* - Med $	$ L_i^* - Med \cdot f_i$
2,245	17	1,1909	20,2453
2,745	51	0,6909	35,2359
3,245	93	0,1909	17,7537
3,745	86	0,3091	26,5826
4,245	41	0,8091	33,1731
4,745	12	1,3091	15,7092
Sumatorias	300		148,6998

$$D.M_{Med} = \frac{148,6998}{300} = 0,4957$$

El rango intercuartílico: $R_q = Q_3 - Q_1 = 3,8671 - 3,0326 = 0,8345$

El rango percentílico: $R_p = P_{90} - P_{10} = 4,2755 - 2,6225 = 1,6530$

Estos rangos representan la amplitud de los intervalos 50% central, y 80% central respectivamente.

Para calcular el porcentaje de valores que cae en los intervalos $\bar{X} \pm S$, $\bar{X} \pm 2S$ y $\bar{X} \pm 3S$, hay que determinar la posición percentil que le corresponde a cada uno de los extremos de los intervalos, para luego restarlos.

Para $\bar{X} \pm S$ se tiene : $3,4433 \pm 0,5985 = [2,8448 ; 4,0418]$

El % por debajo de 3,9418 : $p_2 = \frac{100}{300} \left[\frac{4,0418 - 3,995}{0,5} \right] = 83,61\%$

El % por debajo de 2,8448 : $p_1 = \frac{100}{300} \left[\frac{2,8448 - 2,495}{0,5} \right] = 17,56\%$

Dentro de $\bar{X} \pm S$, caen entonces $83,61\% - 17,56\% = 66,05\%$, que es aproximadamente el porcentaje teórico de 68,27 % dado por la normal.

Para los intervalos $\bar{X} \pm 2S$ y $\bar{X} \pm 3S$, se procede de manera análoga, y se obtiene 95,45 % y 100 % respectivamente.

Ejemplo 11 Dada la siguiente tabla de frecuencias, para datos sin agrupar:

Valor	0	1	2	3	4	5	6	7
frecuencia	8	14	25	13	20	12	6	2

a) Calcule su varianza, la desviación típica, el coeficiente de variación, el rango, la media de las desviaciones absolutas, la media de las desviaciones absolutas respecto de la mediana, el rango intercuartílico.

b) Determine el porcentaje de observaciones que caen en el intervalo $\bar{X} \pm S$.

Solución : A diferencia del ejemplo anterior, en este caso los datos no están agrupados, y por lo tanto para calcular la varianza no se toma a la marca de clase, sino al verdadero valor del dato.

$X_i =$ Valor del dato	$f_i =$ Frecuencia	$X_i f_i$	$(X_i)^2 f_i$
0	8	0	0
1	14	14	14
2	25	50	100
3	13	39	117
4	20	80	320
5	12	60	300
6	6	36	216
7	2	14	98
Sumatorias	100	293	1165

$$s^2 = \frac{1165}{100} - \left(\frac{293}{100} \right)^2$$

293

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}} = \sqrt{\frac{3,07}{100}} = 1,75 ; \quad \bar{X} = \frac{\sum X_i}{n} = \frac{172}{100} = 1,72$$

$$C.V = \frac{1,75}{1,72} 100\% = 101,74\% ; \quad R = 7 - 0 = 7$$

X_i = Valor del dato	f_i = Frecuencia	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
0	8	-1,72	2,94
1	14	-0,72	0,52
2	25	0,28	0,08
3	13	1,28	1,64
4	20	2,28	5,18
5	12	3,28	10,76
6	6	4,28	18,32
7	2	5,28	27,89
Sumatorias	100		147,42

$$D.M = \frac{147,42}{100} = 1,4742$$

Para hallar la mediana, y demás cuartiles hay que proceder según lo explicado en el Ejemplo 6.13 donde se obtuvo $Q_1 = 2$, y siguiendo la misma metodología

$$Q_2 = \text{Med} = 3, \quad Q_3 = 4$$

X_i = Valor del dato	f_i = Frecuencia	$ X_i - \text{Med} $	$ X_i - \text{Med} \cdot f_i$
0	8	3	24
1	14	2	28
2	25	1	25
3	13	0	0
4	20	1	20
5	12	2	24
6	6	3	18
7	2	4	8
Sumatorias	100		147

$$D.M_{Med} = \frac{147}{100} = 1,47 ; R_q = Q_3 - Q_1 = 4 - 2 = 2$$

En cuanto al porcentaje de datos comprendidos en el intervalo $\bar{X} \pm S$, tenemos que éste es : $2,93 \pm 1,75 = [1,18 ; 4,68]$, y en él caen todos los valores comprendidos entre 2 y 4 ambos inclusive, que representan el 58 % de los datos.

Preguntas de Revisión

- 1°) Si se tienen dos conjuntos de datos expresados en las mismas unidades, ¿puede decirse que el que tenga mayor varianza presenta una mayor dispersión?
- 2°) Si en un conjunto de datos todos los valores son negativos, ¿puede alguna de las medidas absolutas de dispersión ser negativa? .
- 3°) ¿Cuál es la diferencia entre las medidas absolutas y las medidas relativas de dispersión? ¿Cuál de las dos mide mejor la variabilidad?
- 4°) Analice las modificaciones que sería necesario realizar, si se quisiera aplicar el método abreviado de cálculo para la varianza, en una tabla de frecuencias con intervalos de diferente amplitud.
- 5°) En un conjunto de datos, el menor es 30 , el rango es 10 , y la media 35. ¿Puede el coeficiente de variación ser igual a 40%? .
- 6°) El valor más bajo de un conjunto de datos es 75, y el más alto 110. ¿Puede ser la desviación típica 39,50?.
- 7°) ¿ Qué representa el rango percentílico? .
- 8°) Suponga que un conjunto de datos es sometido a una transformación no lineal, como por ejemplo: $Y = a + bX^2$. Explique el procedimiento a seguir para calcular la varianza de los datos transformados.
- 9°) Si un conjunto de datos es sometido a una transformación lineal: $Y = a + bX$. ¿Se altera el coeficiente de variación? .
- 10°) ¿Por qué la desigualdad de Tchebychev no proporciona información alguna cuando se quiere hallar el porcentaje de datos que difieren de la media en menos de $\frac{1}{2}$ la desviación típica?
- 11°) Según la desigualdad de Tchebychev ¿cuál es el porcentaje mínimo de datos que difieren de la media en menos de 1,50 veces la desviación típica?.
- 12°) ¿Puede un coeficiente de variación ser mayor al 100% ?.

13°) Suponga que sobre un conjunto de datos $\{x_1, x_2, x_3, \dots, x_n\}$ con media " $\bar{X}$ ", y desviación típica " S ", se define la transformación: $z_i = \frac{x_i - \bar{X}}{S}$ dando lugar a otro

conjunto de datos $\{z_1, z_2, z_3, \dots, z_n\}$ llamados "datos tipificados".

¿Cuál es la media y la varianza de los datos tipificados?

14°) En un conjunto de datos en donde la mitad sea igual a un valor, y la otra mitad sea igual a otro valor diferente, ¿puede afirmarse que la media de las desviaciones absolutas es igual a la desviación típica?

15°) Si $\{x_1, x_2, x_3, \dots, x_n\}$ es una muestra que tiene media muestral $\bar{X}$, y proviene de una población con media poblacional " μ ". ¿Cuál de las siguientes dos

expresiones es menor: $\sum_{i=1}^n (x_i - \mu)^2$ ó $\sum_{i=1}^n (x_i - \bar{X})^2$? . Justifique su respuesta.

16°) Explique por qué en el método de abreviado de cálculo para la varianza, se toma como referencia cero al punto medio de la clase modal. ¿Pudiera tomarse a otra marca de clase sin modificar el método?