CAPITULO CUATRO

MEDIDAS DE DISPERSION, ASIMETRIA Y CURTOSIS

El conocimiento de las medidas de centralización no es suficiente para caracterizar completamente a una distribución por ejemplo: si las edades medias de dos grupos de personas fueran iguales, esto no implica que las edades en ambos grupos sean las mismas y esta igualdad en las medias persistirá aún cuando en un grupo todos tengan las mismas edades y en el otro grupo solo sean unos cuantos los que tienen mayores edades.

Para caracterizar completamente una distribución, es necesario conocer cómo están distribuidos los valores de la variable alrededor de un promedio.

Son medidas de dispersión; cuantifican el grado de concentración o de dispersión de los datos alrededor de un "promedio".

¿Por qué estudiar la dispersión?

- Una medida de dispersión se aplica para evaluar la confiabilidad del promedio que se está utilizando.
- Permite apreciar cuán dispersas están dos o más distribuciones.

Ejemplo:

Observemos los siguientes tres conjuntos de datos:

1 2 3 4 5, 5 10 15 20 25, 10 20 30 40 50

En el primero, cuya media es 3, notamos que los datos están muy concentrados alrededor de su media.

En el segundo, cuya media es 15, los datos están menos concentrados alrededor de su media.

En el tercero, cuya media es 30, los datos están más distantes, más dispersos alrededor de su media.

4.1 EL RECORRIDO (AMPLITUD TOTAL O RANGO) (R)

Es la distancia entre los valores máximo y mínimo de la variable de una población o muestra.

Cálculo Datos no agrupados	Cálculo Datos agrupados
$\boxed{R = \text{Obs. Mayor} - \text{Obs. Menor}}$	**Método 1** $\boxed{R = M_s - M_i}$ M_s: Marca de clase superior. M_i : Marca de clase interior. **Método 2** $\boxed{R = L_s - L_i}$ L_s :Limite superior del intervalo más alto. L_i :Limite interior del intervalo más bajo.

4.2 LA DESVIACION MEDIA

Es el promedio de los valores absolutos las desviaciones con respecto a la media aritmética, mediana u otra medida de tendencia central. Denominada también como desviación promedio, mide el promedio en donde los valores de una población, o muestra, varían con respecto a su media.

Cálculo Datos no agrupados	Cálculo Datos agrupados

RESPECTO A LA MEDIA

Para una muestra:

$$\boxed{DM = \frac{|x_1 - \bar{x}| + \ldots + |x_n - \bar{x}|}{n}}$$

donde:

x_i : observaciones en la muestra.
$\bar{x}$: media muestral.
n : tamaño de la muestra.

RESPECTO A LA MEDIA

Para una muestra:

$$\boxed{\mathrm{DM} = \frac{|\mathrm{x}_1 - \bar{\mathrm{x}}|\mathrm{f}_1 + \ldots + |\mathrm{x}_m - \bar{\mathrm{x}}|\mathrm{f}_m}{\mathrm{n}}}$$

donde:

x_i : marca de clase del intervalo i, donde j varia de 1 a m.
$\bar{x}$: media muestral.
f_i : frecuencia intervalo i.
n : tamaño de la muestra .

RESPECTO A LA MEDIANA

Para una muestra:

$$\boxed{DM = \frac{|x_1 - k| + \ldots + |x_n - k|}{n}}$$

donde:

x_i : observaciones en la muestra.
k : cualquier medida de tendencia central de la muestra (mediana)
n : tamaño de la muestra.

RESPECTO A LA MEDIANA

Para una muestra:

$$\boxed{\mathrm{DM} = \frac{|\mathrm{x}_1 - \mathrm{k}|\mathrm{f}_1 + \ldots + |\mathrm{x}_m - \mathrm{k}|\mathrm{f}_m}{\mathrm{n}}}$$

donde:

x_i : marca de clase del intervalo i, donde i varía de 1 a m.
k : cualquier medida T.C. muestral, tal como la mediana.
f_i : frecuencia intervalo i.
n : tamaño de la muestra.

Características:

El valor de la desviación media depende del valor de la variable en cada unidad de la población o muestra.
Se puede calcular alrededor de la media aritmética, mediana o cualquier otra
Medida de tendencia central.

4.3 LA VARIANZA

La varianza es una forma especial de desviación promedio alrededor de la media. Indica la variación de las observaciones en torno a su media.
Para una población se denota por la letra griega σ^2 y para una muestra por s^2.

Cálculo Datos no agrupados	Cálculo Datos agrupados

Para una muestra:

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

x_i : observaciones en la muestra.
$\bar{X}$: media muestral.
n : tamaño muestra.

Para una muestra:

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

x_i : marca de clase del intervalo i, donde i varía de 1 a m.
$\bar{X}$: media muestral.
f_i : frecuencia intervalo i.
n : tamaño muestra.

Para una población:

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

x_i : observaciones en la población.
μ : media poblacional.
N : tamaño de la población.

Para una población:

$$\boxed{\sigma^2 = \frac{\sum (X_i - \mu)^2 f_i}{N}}$$

X_i: observaciones en la población.
f_i : frecuencia de clase.
N: tamaño de la población.

Características:

Suma de cuadrados y reglas elementales: constante, aditiva, multiplicativa.

Reglas Elementales

CONSTANTE:

$$x_i = k \Rightarrow \overline{x} = k \Rightarrow x_i - \overline{x} = k - k = 0 \Rightarrow SC = 0$$

$$V(k) = 0$$

ADITIVA:

$$\begin{aligned}
&x_i + k \Rightarrow \overline{x + k} = \overline{x} + k \\
&\qquad \Rightarrow \left(x_i + k - \left(\overline{x + k}\right)\right) = \left(x_i + k - \overline{x} - k\right) = \left(x_i - \overline{x}\right) \\
&\qquad \Rightarrow \left(x_i + k - \left(\overline{x + k}\right)\right)^2 = \left(x_i - \overline{x}\right)^2
\end{aligned}$$

$$V(x_i + k) = V(x_i)$$

MULTIPLICATIVA:

$$\begin{aligned}
&cx_i \Rightarrow \overline{cx} = c.\overline{x} \\
&\qquad \Rightarrow \left(cx_i - \overline{cx}\right) = c\left(x_i - \overline{x}\right) \\
&\qquad \Rightarrow \left(cx_i - \overline{cx}\right)^2 = cx^2\left(x_i - \overline{x}\right)^2
\end{aligned}$$

$$\begin{aligned}
V(cx_i) &= c^2 V(x_i) \\
E\left(cx_i - \overline{cx}\right) &= cE\left(x_i - \overline{x}\right)^2 \\
V(x_i) &= E\left(x_i + \overline{x}\right)^2 \\
V(x_i + k) &= E\left(\left(x_i + k\right) - \left(\overline{x_i + k}\right)\right)^2
\end{aligned}$$

4.4 LA DESVIACION ESTANDAR

La Desviación Estándar es la raíz cuadrada positiva de la varianza, es decir, σ para una población y S, para una muestra.

Cálculo Datos no agrupados	Cálculo Datos agrupados

Para una muestra:

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

donde:

x_i: observaciones en la muestra.
$\bar{X}$: media muestral.
n : tamaño muestra.

Para una muestra:

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

donde:

x_i : Marca de Clase del intervalo i, donde i varía de 1 a m.
$\bar{X}$: media muestral.
f_i : frecuencia intervalo i.
n : tamaño población.

Para una población:

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

donde:
X_i: observaciones en la población.
μ: media poblacional.
N: tamaño de la población.

Para una población:

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2 f_i}{N}}$$

donde:
X_i: observaciones en la población.
f: frecuencia.
N: tamaño de la población.

Características:

Al igual que la varianza las características o propiedades de la desviación estándar se corresponden con las Reglas Elementales: constante, aditiva y multiplicativa.

4.5 EL COEFICIENTE DE VARIACION

Es un número abstracto que, denotado por CV, se obtiene como cociente entre la desviación estándar y su media aritmética.

Cálculo para Datos no agrupados y agrupados

$$\boxed{CV = \frac{\sigma}{\mu} 100 \quad para\, una\ población}$$

$$\boxed{CV = \frac{S}{\overline{x}} 100 \quad para\, una\, muestra}$$

donde:

σ : desviación estándar poblacional.
S : desviación estándar muestral.
μ : media aritmética poblacional.
$\overline{x}$: media aritmética muestral.

COEFICIENTE DE VARIACION DE LA DESVIACION MEDIANA

$$\boxed{CV_{DM} = \frac{DM}{Me} 100}$$

donde:

CV_{DM} : coeficiente de variación de la desviación mediana.
DM : desviación mediana.
Me : mediana.

Características:

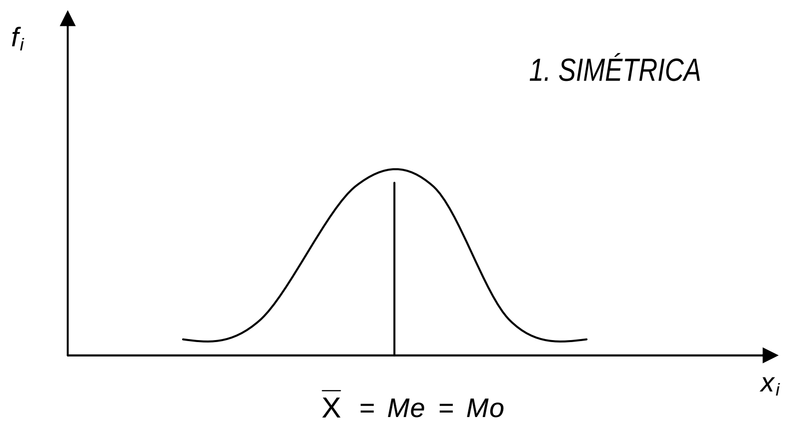
El coeficiente de variación es muy útil especialmente cuando se aplica a muestras homogéneas.

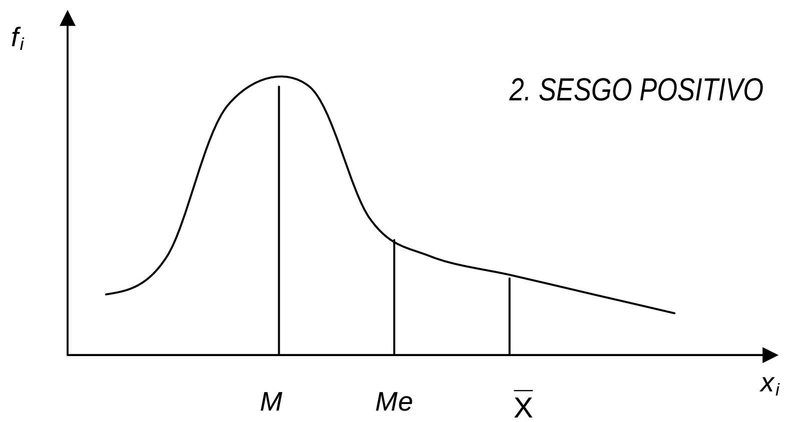
4.6 ASIMETRIA O SESGO

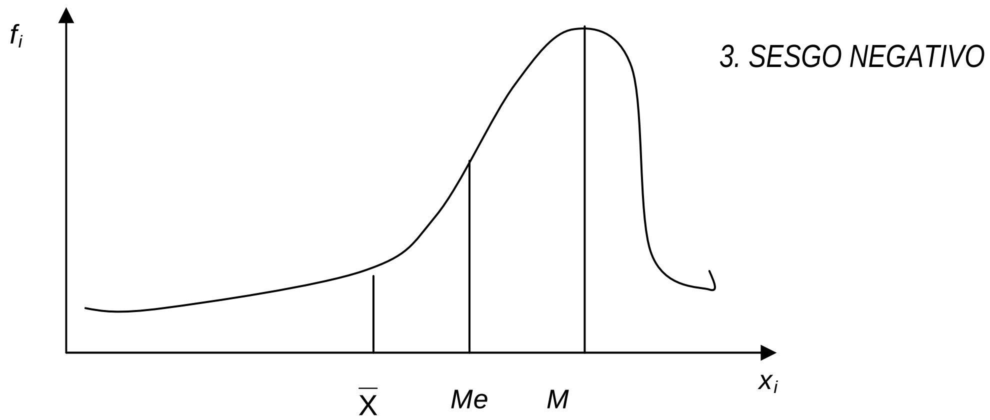
El grado de asimetría de la distribución de frecuencias constituye uno de sus caracteres de mayor importancia. En la práctica casi nunca se encuentran polígonos de frecuencias o histogramas completamente simétricos, por lo cual, el grado en el cual la distribución es asimétrica constituye su **sesgo**.

Si una distribución de frecuencias es simétrica, no tiene sesgo, es decir, el sesgo es nulo. Si una o mas observaciones son grandes, la media de la distribución se vuelve mayor que la Me o la Mo, en tales casos se dice que la distribución tiene sesgo positivo. Si una o más observaciones muy pequeñas se encuentran presentes, la media es la menor de los tres promedios y se dice que la distribución tiene sesgo negativo. Obsérvese el siguiente diagrama:

Diagrama







Karl Pearson desarrolló una medida para desarrollar el sesgo de una distribución denominada **coeficiente de asimetría** (C.A.).

$$\boxed{C.A. = \frac{3(media - mediana)}{desviación\ estándar}}$$

Ejemplo: Las duraciones de estándar en el piso de cancerología de un hospital, se organizaron en una distribución de frecuencias. La duración media fue de 28 días, la mediana 25 días, y la duración modal 23 días. Se calculó una desviación estándar de 4.2 días.

1. ¿Es la distribución simétrica con sesgo positivo o sesgo negativo?
2. ¿Cuál es el coeficiente de asimetría? Interprételo.

Solución:

1. Es asimétrica con sesgo positivo porque la media es la mayor de los tres promedios.
2. Lo calculamos de la siguiente manera:

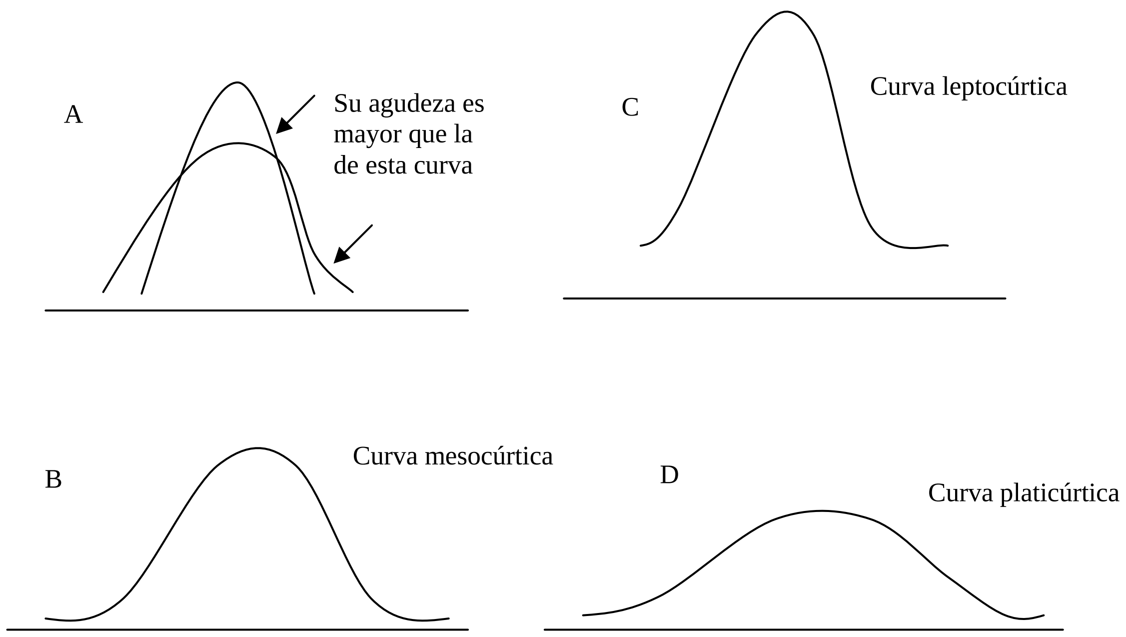
$$C.A = \frac{3(media - mediana)}{desviación\ estándar} = \frac{3(28-25)}{4.2} = 2.14$$

Interpretando esto, el coeficiente de asimetría por lo general se encuentra entre –3 y +3. En tal caso + 2.14 indica un grado importante de asimetría con sesgo positivo. En apariencia unos cuantos pacientes cancerosos permanecen en el hospital durante largo tiempo, provocando que la media sea mayor que la mediana o la moda.

4.7 CURTOSIS

Permite medir el grado de la agudeza de una distribución, es decir, para saber cuán agudo o plano es un polígono de frecuencias.
Observemos los tipos de curtosis, en las siguientes gráficas:

En la figura A se observa que ambas curvas son simétricas y tienen la misma media, mientras que una de las curvas es más cúrtica. La figura B se le denomina mesocúrtica (intermedio con punta). La figura C se le denomina leptocúrtica (delgada con punta) y la figura D se le denomina platicúrtica (aplanado con punta).



El coeficiente de curtosis de un grupo de datos, es una medida del apuntamiento o aplastamiento de su polígono de frecuencias, se define como:

$$\boxed{k = \frac{(0,5)(C_{75} - C_{25})}{C_{90} - C_{10}}}$$

en donde C_{75} es el percentil 75, etc.

Cuando el coeficiente de curtosis tiende a 0,5; esto es, si las diferencias C_{75}–C_{25} y C_{90}-C_{10}, son aproximadamente iguales, la curva se llama **leptocúrtica.**

Si el coeficiente de curtosis tiende a 0, esto es, cuando la diferencia C_{75}–C_{25} es pequeña, respecto de C_{90}-C_{10}, la curva se llama **platicúrtica.**

Si el coeficiente de curtosis es aproximadamente 0,25; esto es, si C_{90}-C_{10} es aproximadamente el doble de C_{75}–C_{25} , la curva se llama **mesocúrtica.**

4.8 APLICACIONES DE MEDIDAS DE DISPERSION

En base al ejercicio N° 1, se tiene:

d. Desviación Media

Designaremos por DM_A : desviación media - Empresa A.

DM_B : desviación media - Empresa B

Para datos agrupados, la desviación media se define como:

$$DM = \frac{|X_1 - \overline{X}|\, f_1 + \dots + |X_m - \overline{X}| \,.\, f_m}{n}$$

Entonces calcularemos las desviaciones con respecto a la media aritmética en valor absoluto y luego, las multiplicaremos por sus respectivas frecuencias.

Empresa A

Salarios (S/.)	Marcas de Clase X_i	Frecuencia f_i	$X_i - \overline{X}_A$	$\lvert X_i - \overline{X}_A \rvert$	$\lvert X_i - \overline{X}_A \rvert \,.\, f_i$
500 – 1 000	750	1	-1 360	1 360	1 360
1 000 – 1 500	1 250	3	-860	860	2 580
1 500 – 2 000	1 750	8	-360	360	2 880
2 000 – 2 500	2 250	5	140	140	700
2 500 – 3 000	2 750	6	640	640	3 840
3 000 – 3 500	3 250	2	1.140	1 140	2 280
Total		n_A = 25			13 640

Donde: $\overline{X}_A$ = S/. 2 110

Por lo tanto:

$$DM_A = \frac{13\ 640}{25} = 545.6$$

Existe una desviación promedio de 545.6 de los sueldos percibidos por los empleados de la Empresa A, alrededor de la media aritmética $\overline{X}_A$= S/. 2 110

Empresa B

Salarios (S/.)	Marcas de Clase X_1	Frecuencia f_i	$X_i - \overline{X}_B$	$\lvert Xi - \overline{X}_B \rvert$	$\lvert X_i - \overline{X}_B \rvert \cdot f_i$
500 – 1 000	750	5	-1 360	1 360	6 800
1 000 – 1 500	1 250	1	-860	860	860
1 500 – 2 000	1 750	3	-360	360	1 080
2 000 – 2 500	2 250	7	140	140	980
2 500 – 3 000	2 750	5	640	640	3 200
3 000 – 3 500	3 250	4	1 140	1 140	4 560
Total		$n_B = 25$			17 480

Donde: X_B = S/. 2.110

Entonces:

$$DM_B = \frac{17\ 480}{25} = 699{,}2$$

Hay una desviación promedio igual a 699.2 de los salarios percibidos por los empleados de la Empresa B, alrededor de la media aritmética X_B = S/. 2110

Puesto que la DM_B es mayor que la DM_A, se concluye que los salarios de los empleados de la Empresa B están más dispersos alrededor de su media aritmética que los salarios de los empleados de la Empresa A.

e. Recorrido (Amplitud de clase o Rango)

Denotaremos por Recorrido $_A$: recorrido de salarios - Empresa A

Recorrido $_B$: recorrido de salarios - Empresa B.

Para datos agrupados, hay dos formas de calcular el recorrido.

1ª. Forma : Recorrido = límite superior de la clase más alta- límite inferior de la clase más baja.

2ª Forma : Recorrido = Marca de clase superior - Marca de clase inferior.

Entonces para la:

Empresa A

1ª Forma : Recorrido $_A$ = 3.500 - 500 = S/. 3 000

2ª Forma : Recorrido $_A$ = 3.250 - 750 = S/. 2 500

Empresa B

1ª Forma : Recorrido $_B$ = 3.500 - 500 = S/. 3 000
2ª Forma : Recorrido $_B$ = 3.250 - 750 = S/. 2 500
Para ambas Empresas, A y B, los sueldos de los empleados oscilan entre S/. 500 y S/. 3 500. Si eliminamos los valores extremos, tendríamos que los salarios de los empleados para las dos empresas, varían entre S/. 750 y S/. 3 250.

f. Varianza

Designaremos por S^2_A : varianza de salarios - Empresa A

S^2_B : varianza de salarios - Empresa B

La fórmula de la varianza para datos agrupados es:

$$S^2 = \frac{(X_1 - \overline{X})^2 \cdot f_1 + \ldots + (X_m - \overline{X})^2 \cdot f_m}{n-1}$$

Entonces, se deben calcular las desviaciones al cuadrado con respecto a la media aritmética y luego, multiplicarlas por las frecuencias correspondientes.

Empresa A

Salarios (S/.)	Marcas de Clase X_1	Frecuencia f_i	$(X_i - \overline{X}_A)$	$(Xi - \overline{X}_A)^2$	$(X_i - \overline{X}_A)^2 \cdot f_i$
500 – 1 000	750	1	-1 360	1 849 600	1 849 600
1 000 – 1 500	1 250	3	-860	739 600	2 218 800
1 500 – 2 000	1 750	8	-360	129 600	1 03 800
2 000 – 2 500	2 250	5	140	19 600	98 000
2 500 – 3 000	2 750	6	640	409 600	2 457 600
3 000 – 3 500	3 250	2	1 140	1 299 600	2 599 200
Total		n_A = 25			10 260 000

Por lo tanto,

$$S^2_A = \frac{10\ 260\ 000}{24} = 427\ 500\ (S/.)^2$$

La varianza de los salarios de los empleados de la Empresa A es de 427.500 $(S/.)^2$.

Empresa B

Salarios (S/.)	**Marcas de Clase X_1**	**Frecuencia f_i**	**$(X_i - \overline{X}_B)$**	**$(Xi - \overline{X}_B)^2$**	**$(X_i - \overline{X}_B)^2 . f_i$**
500 – 1 000	750	5	-1 360	1 849 600	9 248 000
1 000 – 1 500	1 250	1	-860	739 600	739 600
1 500 – 2 000	1 750	3	-360	129 600	388 800
2 000 – 2 500	2 250	7	140	19 600	137 200
2 500 – 3 000	2 750	5	640	409 600	2 048 000
3 000 – 3 500	3 250	4	1 140	1 299 600	5 198 400
Total					17 760 000

Así, $S^2_B = \dfrac{17\,760\,000}{24} = 740\,000\ (S/.)^2$

En la Empresa B, la varianza de los salarios es de 740.000 $(S/.)^2$; la cual es mayor que en la Empresa A.

g. Desviación Estándar

Sean S_A : desviación estándar de los salarios - Empresa A.

S_B : desviación estándar de los salarios - Empresa B.

Como la desviación estándar es la raíz cuadrada positiva de la varianza, es decir:

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

Tenemos que:

$$S_A = +\sqrt{S_A^2} = +\sqrt{427\,500} = S/.\ 653.8$$

$$S_B = +\sqrt{S_B^2} = +\sqrt{740\,000} = S/.\ 860.2$$

La desviación estándar de los salarios de la Empresa B es mayor que la desviación estándar de los sueldos de la Empresa A.

h. Coeficiente de Variación

Si denotamos por CV_A : Coef. de variación - Empresa A.

CV_B : Coef. de variación - Empresa B

Sabemos que el Coeficiente de Variación, se calcula como:

$$CV = \frac{S}{\overline{X}} 100$$

Entonces, a partir de los resultados obtenidos en (a) y en (f), sabemos que

$\overline{X}_A$ = S/. 2.110 y S_A = S/. 653.8

$\overline{X}_B$ = S/. 2.110 y S_B = S/. 860.2

Reemplazando en la fórmula, obtenemos:

$$CV_A = \frac{653.8}{2.110} \cdot 100 = 0{,}3098 \cdot 100$$

$$= 30.98$$

$$CV_B = \frac{860.2}{2.110} \cdot 100 = 0{,}408 \cdot 100$$

$$= 40{,}8$$

A partir de estos resultados, puede apreciarse que si bien el ingreso promedio de los empleados en ambas empresas son iguales, vemos que hay mayor dispersión en salarios que perciben en la Empresa B.

¿Que es un Error Estándar?

Para la inferencia estadística, digamos una prueba estadística y de estimación, se necesita estimar los parámetros de la población. La estimación implica la determinación, con un error posible debido al muestreo, del valor desconocido de un parámetro de la población, tal como la proporción que tiene una cualidad específica o el valor medio m de una cierta medida numérica. Para expresar la exactitud de las estimaciones de las características de la población, se debe también calcular los errores estándar de las estimaciones. Éstas son las medidas de exactitud que determinan los errores posibles que se presentan del hecho de que las estimaciones están basadas en muestras escogidas al azar de la población entera, y no en un censo completo de la población.

El error estándar es un estadístico que indica la exactitud de una estimación. Es decir, nos dice cuan diferente la estimación (como) es del parámetro de la población (como m).

Por lo tanto, esta es la desviación estándar de una distribución muestral para un estimador como.

Los siguientes son una colección de errores estándar para la extensamente usada estadística:

Error Estándar para la Media is: $S/n^{\frac{1}{2}}$.

Como cualquiera esperaría, el error estándar disminuye mientras que el tamaño de la muestra aumenta. Sin embargo la desviación estándar de la estimación disminuye por un factor del $n^{\frac{1}{2}}$ no n. Por ejemplo, si usted desea reducir el error en 50%, el tamaño de la muestra debe ser 4 veces n, lo cual es costoso. Por lo tanto, como alternativa a incrementar el tamaño de la muestra, se puede reducir el error obteniendo los datos de "calidad" el cual proporciona una estimación más exacta.

EJERCICIOS

1.- Cinco representantes de servicio de clientes de una empresa electrónica, trabajaron durante las ventas del viernes. Las cantidades respectivas de videograbadoras que vendieron durante las primera cuatro horas de servicio son: 5,8,4,10 y 3.

a. ¿Cuál es la amplitud total de los datos?
b. ¿Cuál es la media aritmética?
c. ¿Cuál es la desviación media?
d. Interprete la amplitud total.

2.- El departamento de estadística de una universidad ofrece ocho cursos de estadística básica. Las siguientes son las cantidades de estudiantes inscritos en tales cursos:34,46,52,29,41,38,36 y 28.

a. ¿Cuál es la amplitud total?
b. ¿Cuál es la media aritmética de las cantidades de estudiantes inscritos en los cursos?
c. ¿Cuál es la desviación promedio?
d. Interprete la amplitud total

3.- Una empresa de equipos instala abridores automáticos para puertas de garaje. La siguiente lista indica el número de minutos necesarios para tal instalación en una muestra de 10 puertas:28,32,24,46,44,40,54,38,32y 42.

a. ¿Cuál es la amplitud total?
b. ¿Cuál es la media aritmética?
c. ¿Cuál es la desviación media?
d. Interprete esta desviación promedio

4.- Una muestra de ocho compañías en la industria aerospacial fueron entrevistadas acerca de sus rendimientos sobre la inversión de un cierto año. Los resultados son en porcentaje:
10.6,12.6,14.8,18.2,12.0,14.8,12.2y 15.6

a. ¿Cuál es la amplitud total de los rendimientos?
b. ¿Cuál es la media aritmética de los mismos?
c. ¿Cuál es la desviación media?