UNIDAD 4: MEDIDAS DE TENDENCIA CENTRAL

Objetivo terminal:
Calcular e interpretar medidas de tendencia central para un conjunto de datos estadísticos.

Objetivos específicos:

1. Mencionar las características particulares donde se aplica cada medida de tendencia central.
2. Calcular diversas medidas de tendencia central para un conjunto de datos agrupados ó no agrupados.
3. Interpretar las diversas medidas calculadas.

Introducción

Son medidas estadísticas que se usan para describir como se puede resumir la localización de los datos. Ubican e identifican el punto alrededor del cual se centran los datos. Las medidas de tendencia central nos indican hacia donde se inclinan o se agrupan más los datos. Las más utilizadas son: la media, la mediana y la moda.

El propósito de las medidas de tendencia central son:

1. Mostrar en qué lugar se ubica el elemento promedio o típica del grupo.

2. Sirve como un método para comparar o interpretar cualquier valor en relación con el puntaje central o típico.

3. Sirve como un método para comparar el valor adquirido por una misma variable en dos diferentes ocasiones.

4. Sirve como un método para comparar los resultados medios obtenidos por dos o más grupos.

La Media

La **media** o **media aritmetica**, usualmente llamada **promedio**, se obtiene sumando todos los valores de los datos y divide el resultado entre la cantidad de datos. Si los datos proceden de una muestra la media se representa con una x testada (x) y si provienen de la poblacion se representan con la letra griega miu (μ).

Media aritmetica para datos no agrupados muestrales

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

Media aritmetica para datos no agrupados poblacionales

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

Media aritmetica para datos agrupados

$$\bar{x} = \frac{\sum_{clases} f * Xc}{\sum_{clases} fi}$$

Donde

X: promedio muestral (estadistico).
μ: promedio poblacional (parametro).
∑: signo de sumatoria.
N = numero de datos de la poblacion.
n: numero de datos de la muestra.
fi: frecuencia absoluta.
Xc: Marca de clase o punto medio.

Ejemplo de como se emplea la media o promedio con el siguiente ejemplo para datos no agrupados:

a) A continuación se presenta una muestra de las puntuaciones en un examen de un curso de estadística:

70	90	95	74	58	70	98	72	75	85
95	74	80	85	90	65	90	75	90	69

Podemos calcular el promedio de las puntuaciones para conocer cuántos estudiantes obtuvieron puntuaciones por encima y por debajo del promedio.

Primero, sumamos todos los valores de los datos y el resultado lo divide entre el total de datos o tamaño de la muestra. Al sumar todas las puntuaciones en el ejemplo anterior obtendrás un total de 1600, que dividido por 20(total de datos), es igual a 80. Si empleamos la fórmula obtenemos:

$$\bar{x} = \frac{\sum x}{n} \qquad \bar{x} = \frac{1600}{20} = 80$$

La media para datos agrupados, ejemplo:
Ejemplo: Para los gastos diarios en periódicos del hotel agrupados en una tabla de frecuencia:

Intervalo de clase	Fi	Xc	Fi * Xc
5.2 - 6.0	3	5.6	16.8
6.1 – 6.9	5	6.5	32.5
7.0 – 7.8	9	7.4	66.6
7.9 – 8.7	7	8.3	58.1
8.8 – 9.6	5	9.2	46.0
9.7 – 10.5	3	10.1	30.3
Total	**32**		**250.4**

El promedio aritmetico es:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i X_i}{n} = \frac{250.4}{32} = 7.825$$

1ro) Se construye la tabla de distribución de frecuencias.
2do) se obtiene el total de la frecuencia absoluta de clase por el punto medio.
3ro) El resultado obtenido se divide entre el tamaño de la muestra.

Propiedades de la Media:

1ª) La suma de las desviaciones de los valores o datos de una variable X, respecto a su media aritmética es cero.

Ventajas e inconvenientes:

- La media aritmética viene expresada en las mismas unidades que la variable.
- En su cálculo intervienen todos los valores de la distribución.
- Es el centro de gravedad de toda la distribución, representando a todos los valores observados.
- Es única.
- Su principal inconveniente es que se ve afectada por los valores extremadamente grandes o pequeños de la distribución.

La Mediana

La segunda medida de tendencia central que analizaremos es la mediana, en ocasiones se le llama media posicional, porque queda exactamente en la mitad de un grupo de datos, luego de que los datos se han colocado de forma ordenada. En este caso la mitad (50%) de los datos estará por encima de la mediana y la otra mitad (50%) estará por debajo de ella. La mediana es el valor intermedio cuando los valores de los datos se han ordenado.

La Mediana (Me) para datos no agrupados:

1. Primero se ordenan los datos.
2. Luego se calcula la pocision de la mediana con la siguiente formula: (n+1)÷2 donde, n es el número de datos.

    a) Por ejemplo, se tiene una muestra de tamaño 5 con los siguientes valores: 46, 54, 42, 48 y 32.

Primer paso, ordenar los datos: 32 42 46 48 54

Como la cantidad de datos es impar (5 datos), la mediana es el valor del dato que se encuentra ubicado en la posición (5+1)÷2=3, la mediana es:

Me = **46**.

b) Se ha obtenido una muestra con los valores de datos: 27, 25, 27, 30, 20 y 26. ¿cómo se determina la mediana en este caso?.

Primer paso, ordenar los datos de forma ascendente: 20 25 **26** **27** 27 30

Como el número de datos es par (6), la mediana es el promedio de los datos que se encuentran en las posiciones (6+1) ÷1 = 3.5. Por lo tanto la mediana es:

$$\mathbf{Me} = \frac{26+27}{2} = \mathbf{26.5}$$

Para Datos Agrupados.

$$Me = L_i + \frac{\left(\frac{n}{2} - F_{i\text{-}1}\right)}{f_i} \quad \text{(i)}$$

Donde:
Li: Limite inferior real de la clase que contiene la mediana.
n: tamaño de la muestra.
Fi-1 = AFA: Frecuencia acumulada anterior a la clase que contiene la mediana.
Fi: frecuencia de clase absoluta de la clase mediana.

Para identificar la clase mediana se divide n/2 y la primera clase que contenga una frecuencia acumulada mayor que n/2.

n = 32, entonces n/2 = 32/2 = 16. Buscar la primera frecuencia acumulada mayor que 16, esa sera la clase mediana.

Intervalo de clase	fi	Xc	fi * Xc	fa	Limites reales
5.2 - 6.0	3	5.6	16.8	3	5.15 – 6.05
6.1 – 6.9	5	6.5	32.5	8	6.05 – 6.95
7.0 – 7.8	**9**	**7.4**	**66.6**	**17**	**6.95 – 7.85**
7.9 – 8.7	7	8.3	58.1	24	7.85 – 8.75
8.8 – 9.6	5	9.2	46.0	29	8.75 – 9.65
9.7 – 10.5	3	10.1	30.3	32	9.65 – 10.55
Total	**32**		**250.4**		

Ahora se aplica la formula:

Me = (6.95 + (((32/2 – 8)/9)*(0.9)) = 6.95 + (16 – 8) / 9)*(0.9)

Me = (6.95 +(8/9)*(0.9)) = 6.95 +(0.88*0.9)

Me = 6.95 + 0.79

Me = 7.75 ≈ 7.8

Ventajas e inconvenientes :

- Es la medida más representativa en el caso de variables que solo admitan la escala ordinal.
- Es fácil de calcular.
- En la mediana solo influyen los valores centrales y es insensible a los valores extremos u “outliers ”.
- En su determinación no intervienen todos los valores de la variable.

La Moda (Mo)

La moda es el dato que más se repite o el dato que ocurre con mayor frecuencia**..** Un grupo de datos puede no tener moda, tener **una moda (unimodal), dos modas (bimodal) o más de dos modas (multimodal).**

Veamos los siguientes ejemplos:

a) Se tiene una muestra con valores 20, 23, 24, 25, 25, 26 y 30.
Mo = 25 es unimodal

b) Se tiene una muestra con valores 20, 20, 23, 24, 25, 25, 26 y 30.
Mo= 20 y 25, se dice que es **bimodal**.

c) Se tiene una muestra con valores 20, 23, 24, 25, 25, 26, 30 y 30.
Mo= 20, 25 y 30, se dice que es **multimodal**.

En los datos agrupados la Mo es la marca de clase de la clase que contenga la mayor frecuencia absoluta.

Intervalo de clase	fi	Xc	fi * Xc	fa	Limites reales
5.2 - 6.0	3	5.6	16.8	3	5.15 – 6.05
6.1 – 6.9	5	6.5	32.5	8	6.05 – 6.95
7.0 – 7.8	**9**	**7.4**	**66.6**	**17**	**6.95 – 7.85**
7.9 – 8.7	7	8.3	58.1	24	7.85 – 8.75
8.8 – 9.6	5	9.2	46.0	29	8.75 – 9.65
9.7 – 10.5	3	10.1	30.3	32	9.65 – 10.55
Total	**32**		**250.4**		

Mo = 7.4

Tambien se puede calcular a traves de la formula:

$$M_o = L_{ir} + \left[\frac{d_1}{d_1 + d_2}\right] i$$

, donde

Lir: limite inferior verdadero de la clase modal.

$d_1=(f_i - f_{i-1})$ y $d_1=(f_i - f_{i+1})$

fi es la frecuencia absoluta de la clase modal.
fi-1 es la frecuencia de clase absoluta anterior a la clase modal
fi+1 es la frecuencia de clase absoluta posterior a la de la clase modal.
i es el intervalo de clase.

La clase modal es aquella que contiene la mayor frecuencia absoluta.

Intervalo de clase	fi	Xc	fi * Xc	fa	Limites reales
5.2 - 6.0	3	5.6	16.8	3	5.15 – 6.05
6.1 – 6.9	5	6.5	32.5	8	6.05 – 6.95
7.0 – 7.8	**9**	**7.4**	**66.6**	**17**	**6.95 – 7.85**
7.9 – 8.7	7	8.3	58.1	24	7.85 – 8.75
8.8 – 9.6	5	9.2	46.0	29	8.75 – 9.65
9.7 – 10.5	3	10.1	30.3	32	9.65 – 10.55
Total	**32**		**250.4**		

d1 = 9 – 4 = 4

d2 = 9 – 7 = 2

Mo = 6.95 + (4 / 4 + 2) * 0.9 = 6.95 + (4 / 6) * 0.9 = 6.95 + 0.66 * 0.9

Mo = 6.95 + 0.59

Mo = 7.55 ≈ 7.6

Es mejor utilizar la formula para el calculo de la moda.

Ventajas e inconvenientes:
- Su cálculo es sencillo.
- Es de fácil interpretación.
- Es la única medida de posición central que puede obtenerse en las variables de tipo cualitativo.
- En su determinación no intervienen todos lo valores de la distribución.

Cuartiles

Los cuartiles dividen los datos en cuatro partes. Cada una de las partes representa una cuarta parte, o el 25% de las observaciones. Los cuartiles son percentiles específicos.

Los cuartiles se definen de la siguiente manera

Q1 = primer cuartil, o percentil 25.
Q2 = segundo cuartil, o percentil 50 (La mediana).
Q3 = tercer cuartil, o percentil 75.

Ejemplo: a continuación se presenta un conjunto de datos con los siguientes valores; 10, 5, 12, 8, 14, 11, 15, 20, 18, 30 y 25.

Primero, ordenamos los datos

5 8 11 12 14 15 18 20 25 30

Segundo, determinamos (i) para cada cuartil:

Q1 = primer cuartil, o percentil 25.

Q3 = tercer cuartil, o percentil 75.

Calcular posicion de los Cuartiles:

Q1 = primer cuartil, o percentil 25

$$i \approx \left(\frac{25}{100}\right)10 \quad \text{o bien} \quad Q_1 = \frac{n+1}{4}$$

i = (10+1)/4 = (11)/4 = 2.75

Como(i) no es un número entero, se redondea al próximo entero mayor que 2.5, o sea 3. Al referirnos a los datos vemos que el primer cuartil está ubicado en la posición 3 de los datos que este caso es 11. El primer cuartil en los datos se divide de la siguiente forma:

5 8 11 | 12 14 15 18 20 25 30
$Q_1 = 11$ |

Tercer cuartil: Q3 = tercer cuartil, o percentil 75

$$i \approx \left(\frac{75}{100}\right)10 \quad \text{o bien} \quad Q_3 = \frac{3(n+1)}{4}$$

i = 3(10+1)/4 = 3(11)/4 = 33/4 = 8.25

Como (i) no es un número entero, se trunca al entero anterior que 8.25, o sea 8. Al referirnos a los datos , vemos que el tercer cuartil está ubicado en posición 8 de los datos que en este caso es el **20.** Finalmente, los cuartiles en este caso se presentan de la siguiente forma:

5 8 11 | 12 14 15 18 | 20 25 30
Q1=11 | | **Q3=20**

Cuartiles para datos Agrupados.

$$Q_1 = L_{i1} + \frac{\left(\frac{n}{4} - F_{i-1}\right)}{f_i}(i) \qquad Q_3 = L_{i1} + \frac{\left(\frac{3n}{4} - F_{i-1}\right)}{f_i}(i)$$

Para identificar las clases del primer y tercer cuartil, se utiliza la formula n/4 para el primer cuartil y 3n/4 para el tercer cuartil. La clase que contenga la primera clase con frecuencia acumulada mayor que n/4 , esa es la clase del primer cuartil y (3*n)/4 para el tercer cuartil. Luego se aplica la formula.

Lir es el limite inferior real de la clase cuartilica.
n es el tamaño de la muestra.
fi es la frecuencia de clase cuartilica.
fi-1 es la frecuencia de clase anterior a la clase cuartilica.
i es el tamaño del intervalo.

Usos de los cuartiles:

1. Para indicar el porcentaje igual o menor que el valor de un cuartil.

2. Para describir el 50% central de las observaciones

3. Elaboración del gráfico de caja.

UNIDAD 5 : MEDIDAS DE DISPERSIÓN

Objetivo terminal:

Calcular e interpretar medidas de dispersión para un conjunto de datos estadísticos.

Objetivos específicos:

- Entender la importancia de analizar la dispersión de un grupo de datos.
- Calcular diversas medidas de dispersión para un conjunto de datos agrupados ó no agrupados.
- Interpretar diversas medidas de dispersión para un conjunto de datos agrupados ó no agrupados.

Introducción

Miden la variabilidad de un conjunto de datos. Las medidas mas utilizadas son: Rango, Varianza, Desviación estándar, Coeficiente de variación, Intervalo cuartilar.

Rango

Es la diferencia entre el valor más grande y el más pequeño del conjunto de datos.

Rango para datos no agrupados.

Rango = Valor máximo - Valor mínimo
R = 64 – 12 = 52

Rango para datos agrupados:

R = límite superior de la última clase - límite inferior de la primera clase
R = 10.5 – 5.2 = 5.3

Varianza

Es la medida que cuantifica la variabilidad de los datos respecto al valor de la media.

La varianza para la muestra se representa mediante una s al cuadrado:

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

Donde:
x_i valores de la variable, x_1 , x_2, etc.
n es de la muestra
$\bar{x}$ es la media aritmética

Usos de la Varianza:

- En inferencia estadística.
- Para calcular la desviación estándar.
- Para calcular el tamaño de muestra.

Ejemplo: Para los datos siguientes calcular la varianza.

22	38	35	56	45	33	28	36	45	55	20	38
46	27	45	23	64	21	34	22	29	36	12	54
45	37	53	26	35	32	21	43	39	28	28	

Se debe calcular primero la media.

$X = (22 + 38 + 35 + 56 + 45 + 33 + 28 + 36 + 45 + 55 + 20 + 38 + 46 + 27 + 45 + 23 + 64 + 21 + 34 + 22 + 29 + 36 + 12 + 54 + 45 + 37 + 53 + 26 + 35 + 32 + 21 + 43 + 39 + 28 + 28) / 35$

$X = 1250.99 / 35$

$X = 35.74$

$S^2 = ((22 – 35.74)^2 + (38 – 35.74)^2 + (35 – 35.74)^2 + (56 – 35.74)^2 + (45 – 35.74)^2 + (33 – 35.74)^2 + (28 – 35.74)^2 + (36 – 35.74)^2 + (45 – 35.74)^2 + (55 – 35.74)^2 + (20 – 35.74)^2 + (38 – 35.74)^2 + (46 – 35.74)^2 + (27 – 35.74)^2 + (45 – 35.74)^2 + (23 – 35.74)^2 + (64 – 35.74)^2 + (21 – 35.74)^2 + (34 – 35.74)^2 + (22 – 35.74)^2 + (29 – 35.74)^2 + (36 – 35.74)^2 + (12 – 35.74)^2 + (54 – 35.74)^2 + (45 – 35.74)^2 + (37 – 35.74)^2 + (53 – 35.74)^2 + (26 – 35.74)^2 + (35 – 35.74)^2 + (32 – 35.74)^2 + (21 – 35.74)^2 + (43 – 35.74)^2 + (39 – 35.74)^2 + (28 – 35.74)^2 + (28 – 35.74)^2) / (35 – 1)$

$S^2 = 145$

Desviación Estándar

Es la raíz cuadrada positiva de la varianza. Mide la variabilidad de los datos en las unidades en que se midieron originalmente. Los símbolos son: s, si es una muestra y ; σ si es una población.

$s = \sqrt{s^2}$

$s = \sqrt{145}$

$s = 12.04$

Características de la desviación estándar:

1. Siempre es un valor positivo.
2. Está influenciada por todos los valores de la muestra o población.
3. Mayor influencia ejercen los valores extremos debido a que son elevados al cuadrado en el cálculo.
4. Sirve para definir la dispersión de los datos alrededor de la media.

La desviación estándar para datos agrupados.

$s = \sqrt{(\sum(fi * Xc^2) - (\sum(fiXc)^2) / n) / n - 1}$

Proceso:

- Primero se eleva el punto medio al cuadrado y luego se multiplica por la frecuencia absoluta de clase.
- Se obtiene el total de la frecuencia absoluta por la marca de clase y se eleva al cuadrado, este resultado se divide entre el tamaño de la muestran.
- Se resta el primer resultado del segundo y se divide ente n – 1.

Para obtener la varianza se eleva la desviación estándar al cuadrado.

Ej.

Intervalo de clase	Fi	Xc	Fi * Xc	Xc^2	Fi * Xc^2
5.2 - 6.0	3	5.6	16.8	31.36	94.08
6.1 – 6.9	5	6.5	32.5	42.25	211.25
7.0 – 7.8	9	7.4	66.6	54.76	492.84
7.9 – 8.7	7	8.3	58.1	68.89	482.23
8.8 – 9.6	5	9.2	46.0	84.64	423.20
9.7 – 10.5	3	10.1	30.3	102.01	306.30
Total	**32**		**250.4**		**2009.63**

$s = \sqrt{2009.63 – 250.4^{2} / 32) / 32 – 1}$

$s = \sqrt{2009.63 –62700.16/ 32) / 31}$

$s = \sqrt{2009.63 – 1959.38) / 31}$

$s = \sqrt{50.25 / 31}$

$s = \sqrt{1.62}$

s = 1.27

Varianza es: $s^2 = 1.62$

Coeficiente De Variabilidad

Medida de variabilidad relativa: Se usa para comparar la variabilidad entre dos o más muestras medidas en las mismas unidades o no. Los datos que se expresan en porcentaje en la cual se compara la desviación estándar con el respectivo valor del promedio de los datos:

$$C.V. = \left(\frac{S}{\overline{x}}\right) x100$$

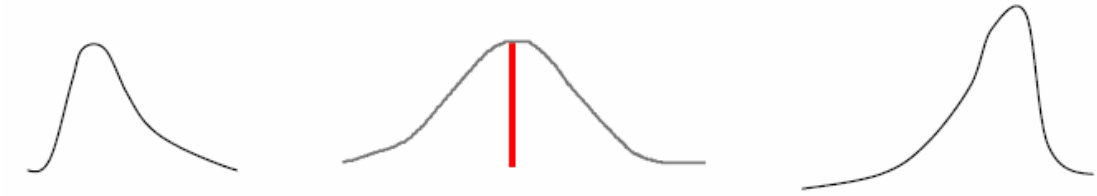
Grado de variabilidad de los datos	*Coeficiente de variabilidad*
Con variabilidad baja	Menos de 10%
Con variabilidad moderada	De 10% a 30%
Con alta variabilidad	Más de 30%

Medida De Forma: Asimetría O Sesgo

Evalúa el grado de distorsión o inclinación que adopta la distribución de los datos respecto a su valor promedio tomado como centro de gravedad. El coeficiente de asimetría de Pearson es:

$$A_K = \frac{3(\overline{X} - M_e)}{S}$$

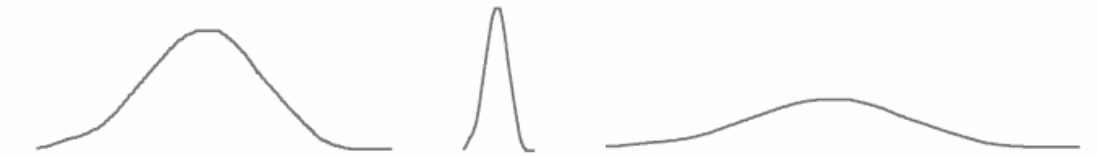
Grado de Asimetría	*Valor del Sesgo*
Simetría Perfecta	**Cero**. El promedio es igual a la mediana
Sesgo Positivo	**Positivo**. Promedio mayor que la mediana
Sesgo Negativo	**Negativo**. Promedio menor que mediana



Asimetría Positiva (Promedio>Mediana) | Simétrica Promedio=Mediana | Asimetría Negativa Promedio<Mediana

Medida De Forma: Curtosis

Grado de Apuntamiento	*Valor de la Curtosis*
Mesocurtica (Distribución normal)	0.263
Leptocúrtica (Elevada)	Mayor a 0.263 ó se aproxima a 0.5
Platicúrtica (Aplanada)	Menor a 0.263 ó se aproxima a 0



K_u=0.263 **Mesocúrtica** | K_u>0.263 **Leptocúrtica** | K_u<0.263 **Platicúrtica**

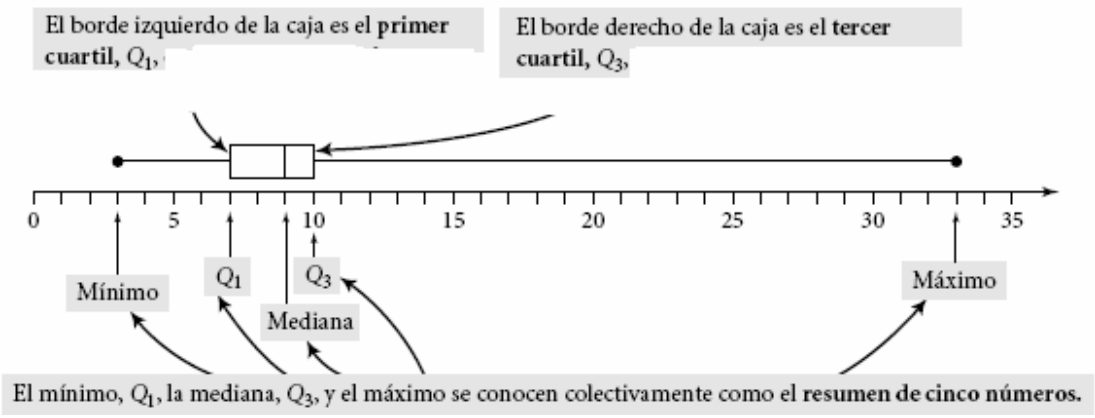
Amplitud cuartílica.
AC = tercer cuartil – primer cuartil

Desviación cuartílica.

DC = (tercer cuartil - primer cuartil) / 2

La Grafica de caja y brazo y el Resumen de 5 números.

Una buena descripción de un conjunto de datos incluye una medida de la tendencia central, junto con información sobre la forma y la dispersión de los datos. Una gráfica de caja es una herramienta útil para mostrar la forma y la dispersión de los datos.



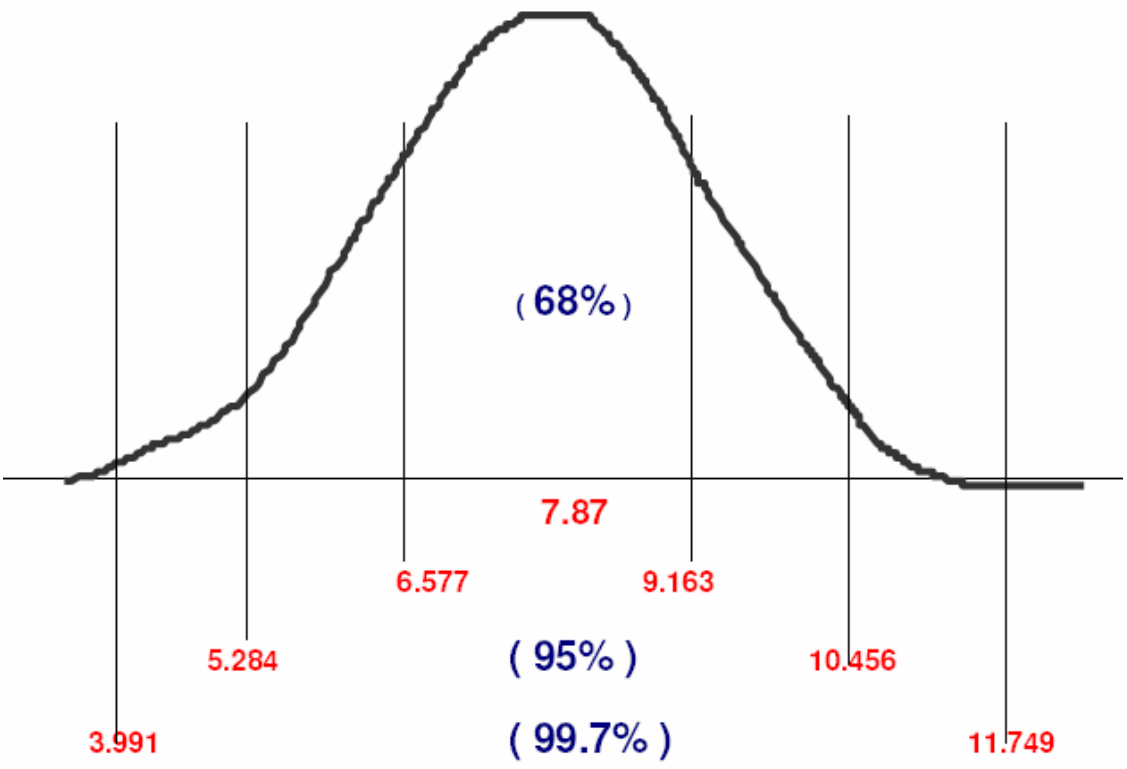
Los segmentos que se salen de la "caja" se llaman "bigotes" ("whiskers"). Una gráfica de caja divide los datos en cuatro partes iguales. El bigote izquierdo, la parte izquierda de la caja, la parte derecha de la caja, y el bigote derecho representan cada uno un cuarto de los datos y la mediana se coloca en el centro de la caja.

El resumen de cinco números da los valores de los puntos claves de una gráfica de caja. Los cinco números son el valor mínimo, el primer cuartil, la mediana, el tercer cuartil, y el valor máximo, respectivamente. Los datos superiores al valor máximo y los inferiores al valor mínimo se denominan Datos Atípicos.

Regla Empírica

Cuando la distribución de frecuencia es **simétrica**:
Por ejemplo: Si el Promedio es 7.87 y Desviación estándar 1.293 podremos afirmar que:

- El 68% se encuentran (+- 1s) mas menos una desviación estándar.
- El 95% se encuentran (+- 2s) mas menos dos desviación estándar.
- 99.7% se encuentran (+- 3s) mas menos tres desviación estándar.


(68%)
7.87
6.577
9.163
5.284
(95%)
10.456
3.991
(99.7%)
11.749

UNIDAD 6: ANÁLISIS DE CORRELACIÓN SIMPLE

Objetivo terminal:

Utilizar las técnicas del análisis de correlación simple para determinar si existe relación entre dos variables.

Objetivos específicos:

- Elaborar un diagrama de dispersión como recurso gráfico para observar la correlación entre dos variables.
- Diferenciar la relación lineal de la no lineal.
- Calcular e interpretar el coeficiente de correlación.

Introducción

En ocasiones nos puede interesar estudiar si existe o no algún tipo de relación entre dos variables aleatorias. A través de este análisis se trata de determinar el grado de relación o correspondencia entre dos conjuntos de valores denominados variables. Cuando la relación tiene un valor positivo significa que a valores altos en una variable corresponden valores altos en la otra variable. Y la relación con signo negativo significa que las variables están relacionadas de manera inversa de modo que cuando el valor aumenta en una, disminuye en la otra.

Las variables estudiadas asumen los nombres de: **variable dependiente representada por Y y la variable independiente representada por x.**

Conceptos:

Análisis de correlación: se usa un gupo de técnicas estadísticas para medir la fuerza de la relación (correlación) entre dos variables.

Diagrama de dispersión: gráfica que describe la relación entre las dos variables de interés.

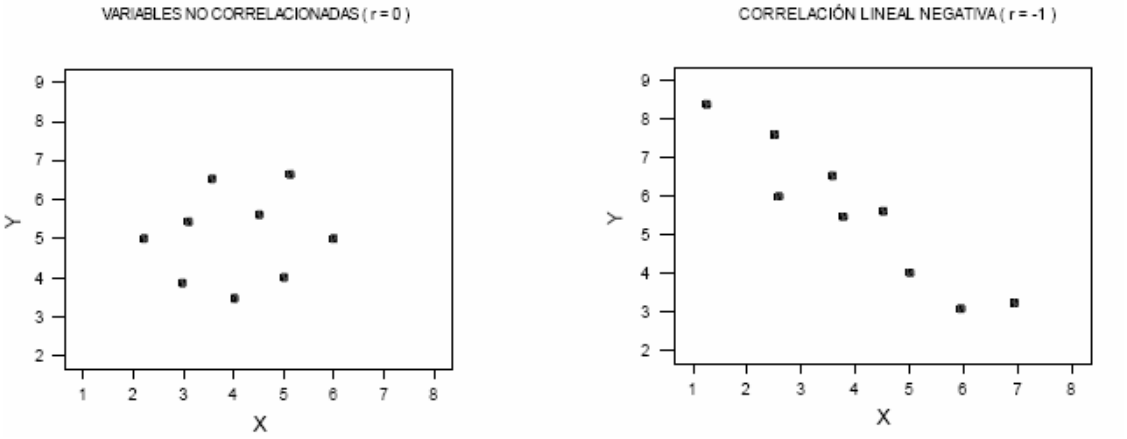
Variable dependiente: la variable que se pronostica o estima.

Variable independiente: la variable que proporciona la base para la estimación. Es la variable predictora.

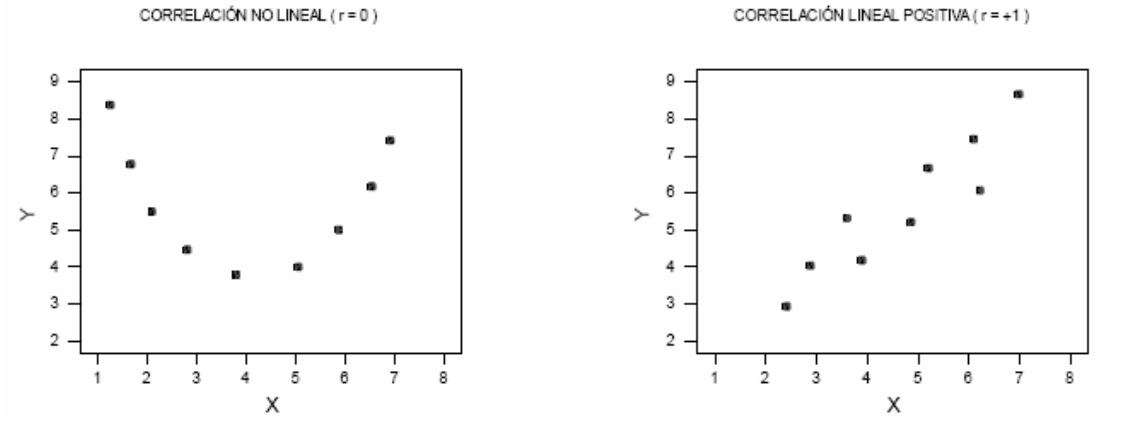
El coeficiente de determinación, r2 es la proporción de la variación total en la variable dependiente Y que está explicada por o se debe a la variación en la variable independiente X. El coeficiente de determinación es el cuadrado del coeficiente de correlación, y toma valores de 0 a 1.

El coeficiente de correlación (r) es una medida de la intensidad de la relación entre dos variables. Requiere datos con escala de intervalo o de razón (variables), y puede tomar valores entre -1.00 y 1.00.

$$-1 \le r = \frac{Cov(X,Y)}{s_X s_Y} = \frac{\sum_{t=1}^{n}(X_t - \overline{X}) * (Y_t - \overline{Y})}{\sqrt{\sum_{t=1}^{n}(X_t - \overline{X})^2} * \sqrt{\sum_{t=1}^{n}(Y_t - \overline{Y})^2}} \le +1$$



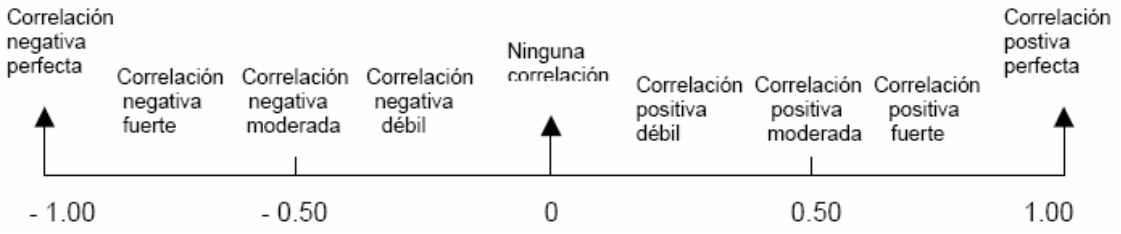
Valores de -1.00 o 1.00 indican correlación fuerte y perfecta. Los valores cercanos a 0.0 indican correlación débil. Valores negativos indican una relación inversa y valores positivos indican una relación directa.



Como se observa en los diagramas anteriores, el valor de r se aproxima a +1 cuando la correlación tiende a ser lineal directa (mayores valores de X significan mayores valores de Y), y se aproxima a –1 cuando la correlación tiende a ser lineal inversa.

Es importante notar que la existencia de correlación entre variables no implica causalidad. Si no hay correlación de ningún tipo entre dos v.a., entonces tampoco habrá correlación lineal, por lo que r = 0. Sin embargo, el que ocurra r = 0 sólo nos dice que no hay correlación lineal, pero puede que la haya de otro tipo.

El siguiente diagrama resume el análisis del coeficiente de correlación entre dos variable:



Definición y características del concepto de Regresión Lineal

En aquellos casos en que el coeficiente de regresión lineal sea "cercano" a +1 o a –1, tiene sentido considerar la ecuación de la recta que "mejor se ajuste" a la nube de puntos (recta de mínimos cuadrados). Uno de los principales usos de dicha recta será el de predecir o estimar los valores de Y que obtendríamos para distintos valores de X. Estos conceptos quedarán representados en lo que llamamos **diagrama de dispersión**.

Análisis de regresión

Propósito: determinar la ecuación de regresión; se usa para predecir el valor de la variable dependiente (Y) basado en la variable independiente (X).

Procedimiento: seleccionar una muestra de la población y enumerar los datos por pares para cada observación; dibujar un diagrama de dispersión para visualizar la relación; determinar la ecuación de regresión.

La ecuación de regresión: Y'= a + bX, donde:

Y' es el valor promedio pronosticado de Y para cualquier valor de X.

a es la intercepción en Y, o el valor estimado de Y cuando X = 0, es decir, el valor del punto en que la recta cruza, corta el eje de las coordenadas (y).

x es cualquier valor de x que desee utilizarse para predecir su correspondiente valor en y.

b es la pendiente de la recta, o cambio promedio en Y' por cada cambio de una unidad en X se usa el principio de mínimos cuadrados para obtener a y b:

$$b = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2}$$

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n}$$

Definición del Coeficiente de Determinación

Denominamos coeficiente de determinación R2 como el coeficiente que nos indica el porcentaje del ajuste que se ha conseguido con el modelo lineal, es decir el porcentaje de la variación de Y que se explica a través del modelo lineal que se ha estimado, es decir a través del comportamiento de X. A mayor porcentaje mejor es nuestro modelo para predecir el comportamiento de la variable Y.

También se puede entender este coeficiente de determinación como el porcentaje de varianza explicada por la recta de regresión y su valor siempre estará entre 0 y 1 y siempre es igual al cuadrado del coeficiente de correlación (r).

$R^2 = r^2$

Es una medida de la proximidad o de ajuste de la recta de regresión a la nube de puntos. También se le denomina **bondad del ajuste**.

1– R^2 nos indica qué porcentaje de las variaciones no se explica a través del modelo de regresión, es como si fuera la varianza inexplicada que es la varianza de los residuos.

Procedimiento para el análisis de correlación y regresión Lineal

1. Identificar la variable dependiente y la variable independiente.
2. Construir el diagrama de dispersión. Los datos de la variable independiente x se colocan en el eje de las X y los de la variable dependiente en el eje de las Y.
3. Calcular el coeficiente de correlación lineal.
4. Calcular la ecuación de mejor ajuste de los mínimos cuadrados.
5. Trazar la línea de mejor ajuste.

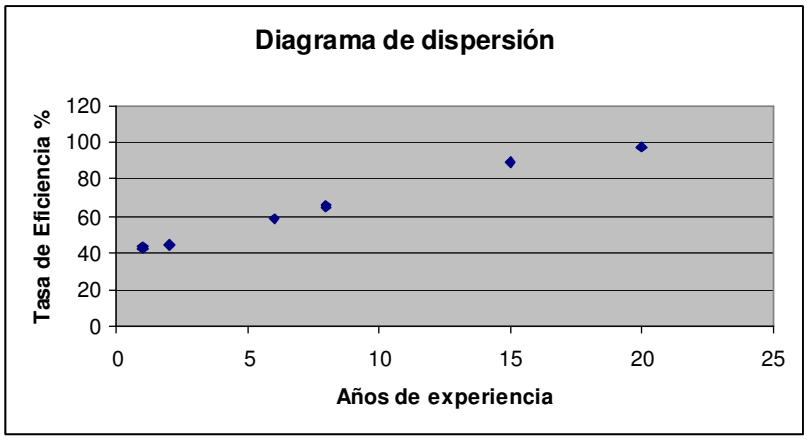
Ejemplo: el siguiente conjunto de datos: Se llevó a cabo un proyecto de investigación para determinar si existe alguna relación entre los años de servicio en un hospital y la eficiencia de las enfermeras. Se recogieron los datos siguientes. Se desea predecir la eficiencia del empleado.

Enfermera	Años de servicio	Tasa de eficiencia(%)
1	1	43
2	20	97
3	6	59
4	8	66
5	2	44
6	1	42
7	15	89
8	8	65

1. Primero identificamos la variable dependiente y la independiente.

Se puede decir que la **variable dependiente es la tasa de eficiencia** por que depende de los años de servicio (experiencia). Por lo tanto la variable **independiente son los años de experiencia**.

2. Se traza el diagrama de dispersión. Para ello los valores de la variable dependiente se colocan en el eje de las Y y los valores de la variable independiente en el eje de las X. Luego se coloca un punto de intersección entre los valores de los datos ordenados, al grafico de resultado se le conoce como diagrama de dispersión.



3. Se calcula el coeficiente de relación.

Empleado	Años de servicio X	Tasa de eficiencia (%) Y	XY	X^2	Y^2
1	1	43	43	1	1849
2	20	97	1940	400	9409
3	6	59	354	36	3481
4	8	66	528	64	4356
5	2	44	88	4	1936
6	1	42	42	1	1764
7	15	89	1335	225	7921
8	8	65	520	64	4225
Total	**61**	**505**	**4850**	**795**	**34941**

$r = [n(\sum xy)-(\sum x)*(\sum y)] / \sqrt{[(n(\sum x^2))-(\sum x)^2][(n(\sum y^2))-(\sum y)^2]}$

$r = [8*(4850)-(61)*(505)] / \sqrt{[(8*795)-(61)^2][(8*34941)-(505)^2]}$

$r = [38800-30805] / \sqrt{[6360-3721][279528-255025]}$

$r = 7995 / \sqrt{[2639][24503]}$

$r = 7995 / \sqrt{64663417}$

r = 7995 / 8041.357

r = 0.994235, lo que tiene a indicar que existe una correlación positiva intensa.

R^2 = r*r

R^2 = 0.994235 * 0.994235

R^2 = 0.98850 * 100

R^2 = 98.850 %,

Es el porcentaje de la variación de Y(tasa de eficiencia) que se explica a través del modelo lineal que se ha estimado, es decir a través del comportamiento de X (años de servicio) .

1– R^2

1 – 0.98850= 0.0115

≈ 1.15 %,

Esto nos indica qué porcentaje de las variaciones no se explica a través del modelo de regresión.

4. Calcular la ecuación de mejor ajuste de los minimos cuadrados. Primero se calcula b y luego a y se escribe la ecuación de mejor ajuste.

$$b = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2}$$

$$a = \frac{\Sigma Y}{n} - b\,\frac{\Sigma X}{n}$$

b = [8 *(4850)-(61)*(505)] / [(8*795)-$(61)^2$] ;

b = [38800-30805] / [6360-3721]

b = 7995 / 2639

b = 3.0295567

a = [505 / 8] - [3.0295567 * (61/8)];

a =63.125 - [3.0295567 * 7.625]

a =63.125 - 23.10037

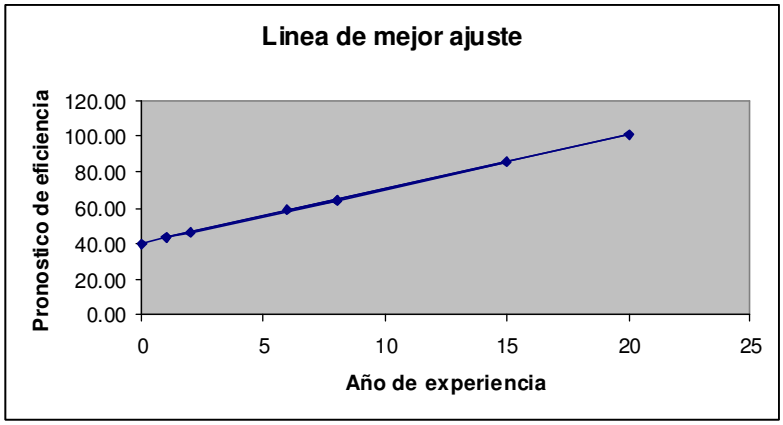
a =40.02463

La ecuación de regresión: Y'= a + bX, donde:

Y'= 40.02463 + 3.0295567 X

6. Trazar la línea de mejor ajuste, para ello se debe hacer un pronóstico de los valores de x en la ecuación.

Años de servicio X	Pronóstico Y' = a + bx
0	40.025
1	43.054
20	100.616
6	58.202
8	64.261
2	46.084
1	43.054
15	85.468
8	64.261
61	



Error estándar de la estimación.

$$s = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

El **error estándar de estimación**, es el mismo concepto que la desviación estándar, aunque ésta mide la dispersión alrededor de la media y el error estándar mide la dispersión alrededor de la línea de regresión.

$s = \sqrt{((43 - 43.05)^2 + (97 - 100.62)^2 + (59 - 58.2)^2 + (66 - 64.26)^2 + (44 - 46.08)^2 + (42 - 43.05)^2 + (89 - 85.47)^2 + (65 - 64.26)^2) / 8 - 2}$

$s = \sqrt{[(43 - 43.05)^2 + (97 - 100.62)^2 + (59 - 58.2)^2 + (66 - 64.26)^2 + (44 - 46.08)^2 + (42 - 43.05)^2 + (89 - 85.47)^2 + (65 - 64.26)^2] / 6}$

$s = \sqrt{((0.5)^2 + (-3.62)^2 + (0.8)^2 + (1.74)^2 + (-2.08)^2 + (-1.05)^2 + (3.53)^2 + (0.74)^2)/6}$

$s = \sqrt{(0.25 + 13.1044 + 0.64 + 3.0276 + 4.3264 + 1.1025 + (12.461 + 0.5476)/6}$

$s = \sqrt{35.4595/6}$

$s = \sqrt{5.909917}$

s = 2.431032, esta es la dispersión de los datos con respecto a la línea de mejor ajuste.