CBS

Colegio Bautista Shalom

Estadística II
Quinto PAE
Segundo Bimestre

Contenidos

Alégrate, joven, en tu juventud, y tome placer tu corazón en los días de tu adolescencia; y anda en los caminos de tu corazón y en la vista de tus ojos; pero sabe, que sobre todas estas cosas te juzgará Dios. Eclesiastés 11:9

- ✓ COEFICIENTE DE VARIACIÓN DE PEARSON (Cvx).
- ✓ VARIABLE NORMALIZADA O TIPIFICADA.
- ✓ UNIDADES TIPIFICADAS.

MEDIDAS DE FORMA

- ✓ ASIMETRÍA.
- ✓ TIPOS DE ASIMETRÍA.
- ✓ MEDIDAS DE ASIMETRÍA.
- ✓ CURTOSIS O APUNTAMIENTO.
  - TIPOS DE CURTOSIS.
  - MEDIDAS DE CURTOSIS.

LA CORRELACIÓN

- ✓ CORRELACIÓN LINEAL.
- ✓ CORRELACIÓN LINEAL SIMPLE.

AJUSTE DE CURVAS

- ✓ REGRESIÓN POR MÍNIMOS CUADRADOS.
- ✓ EMPLEANDO LA HOJA ELECTRÓNICA DE EXCEL.

NOTA: conforme avances en tu aprendizaje tu catedrático(a) te indicará la actividad o ejercicio a realizar. Sigue sus instrucciones.

COEFICIENTE DE VARIACIÓN *DE PEARSON* (Cvx)

Indica la relación existente entre la desviación típica de una muestra y su media.

$$CV = \frac{S}{x}$$

Al dividir la desviación típica por la media se convierte en un valor excento de unidad de medida. Si comparamos la dispersión en varios conjuntos de observaciones tendrá menor dispersión aquella que tenga menor coeficiente de variación. El principal inconveniente, es que al ser un coeficiente inversamente proporcional a la media aritmética, cuando está tome valores cercanos a cero, el coeficiente tenderá a infinito.

Ejemplo: calcula la varianza, desviación típica y la dispersión relativa de esta distribución.

Sea x el número de habitaciones que tienen los 8 pisos que forman un bloque de vecinos.

X	ni
2	2
3	2
5	1
6	3
	N= 8

$$\bar{x} = \frac{\sum_{i-1}^{n} x_i n_j}{N} = \frac{2 \times 2 + 3 \times 2 + 5 \times 1 + 6 \times 3}{8} = 4.125 \ \ habitaciones.$$

$$S^2 = \frac{\sum_{i-1}^{r} x_i {n_j}^2}{N} - \overline{x^2} = \frac{2^2 \times 2 + 3^2 \times 2 + 5^2 \times 1 + 6^2 \times 3}{8} - (4.125)^2 = 2.86 \ \ habitaciones$$

$$S_x = \sqrt{{S_x}^2} = \sqrt{2.86} = 1.69 \ habitaciones$$

$$CV = \frac{S}{\overline{x}} = \frac{1.69}{4.125} = 0.41$$

VARIABLE NORMALIZADA O TIPIFICADA

En análisis de datos, centrar y reducir las variables (normalizar) permite comparaciones independientes de la unidad de medida.

- ✓ *Centrar* una variable consiste en sustraer su media a cada uno de sus valores inicial.
- ✓ *Reducir* una variable consiste en dividir todos sus valores por su desviación típica.

Una variable centrada reducida tiene:

- Una media nula,
- Una desviación típica igual a uno.

Así obtenemos:

- Datos independientes de la unidad, o de la escala escogida,
- Variables que tienen misma dispersión y misma media.

Podemos entonces comparar más fácilmente las variaciones. Centrar reducir las variables es muy útil en análisis de datos:

- Esto equivale a un *cambio de unidad*, y no tiene incidencia sobre los perfiles de variación.

- Los valores de los coeficientes de correlación entre variables centradas reducidas permanecen idénticos a lo que eran antes de la operación de centrado reducción.

Supongamos que X es un valor procedente de una muestra (o población) con media $\bar{a}$ o μ y desviación típica s o σ.

En tal caso, el valor de x en unidades tipificadas representado por z se define de la siguiente manera.

UNIDADES TIPIFICADAS

$$z = \frac{x - \bar{a}}{s} \qquad z = \frac{x - \mu}{\sigma}$$

Las unidades tipificadas muestran el número de desviaciones típicas en que un valor dado se sitúa por encima o debajo de la media de su muestra o población. Se usan también para comparar valores de diferentes muestras o poblaciones. Por ejemplo: un alumno A saca una puntuación de 85 en un examen cuyas puntuaciones tienen una media de 79 con una desviación típica de 8. Un alumno B saca 74 en un examen cuyas puntuaciones tienen una media de 70 y desviación típica de 5. ¿Cuál de los dos alumnos obtuvo una puntuación mejor? La respuesta, desde el punto de vista de la "unidad tipificada", se obtiene así:

Las puntuaciones tipificadas de los alumnos A y B son respectivamente:

$$z_a = \frac{85 - 79}{8} = \frac{6}{8} = 0,75 \qquad z_b = \frac{74 - 70}{5} = \frac{4}{5} = 0,8$$

Así el alumno B lo hizo mejor que el A, aunque su puntuación de 74 es inferior a 85.

MEDIDAS DE FORMA

Las medidas de forma permiten comprobar si una distribución de frecuencia tiene características especiales como simetría, asimetría, nivel de concentración de datos y nivel de apuntamiento que la clasifiquen en un tipo particular de distribución.

ASIMETRÍA

Es una medida de forma de una distribución que permite identificar y describir la manera como los datos tiende a reunirse de acuerdo con la frecuencia con que se hallen dentro de la distribución. Permite identificar las características de la distribución de datos sin necesidad de generar el gráfico.

TIPOS DE ASIMETRÍA

La asimetría presenta las siguientes formas:

Asimetría Negativa o a la Izquierda. Se da cuando en una distribución la minoría de los datos está en la parte izquierda de la media. Este tipo de distribución presenta un alargamiento o sesgo hacia la izquierda, es decir, la distribución de los datos tiene a la izquierda una cola más larga que a la derecha. También se dice que una distribución es simétrica a la izquierda o tiene sesgo negativo cuando el valor de la media aritmética es menor que la mediana y éste valor de la mediana a su vez es menor que la moda, en símbolos:

$$\bar{x} < Md < Mo$$

Nota: sesgo es el grado de asimetría de una distribución, es decir, cuánto se aparta de la simetría.

Simétrica. Se da cuando en una distribución se distribuyen aproximadamente la misma cantidad de los datos a ambos lados de la media aritmética. No tiene alargamiento o sesgo. Se representa por una curva normal en forma de campana llamada campana de Gauss (matemático Alemán 1777-1855) o también conocida como de Laplace (1749-1827). También se dice que una distribución es simétrica cuando su media aritmética, su mediana y su moda son iguales, en símbolos:

$\bar{x}$ Md=Mo

Asimetría Positiva o a la Derecha. Se da cuando en una distribución la minoría de los datos está en la parte derecha de la media aritmética. Este tipo de distribución presenta un alargamiento o sesgo hacia la derecha, es decir, la distribución de los datos tiene a la derecha una cola más larga que a la izquierda. También se dice que una distribución es simétrica a la derecha o tiene sesgo positivo cuando el valor de la media aritmética es mayor que la mediana y éste a valor de la mediana a su vez es mayor que la moda, en símbolos:

$$\bar{x} > Md > Mo$$

MEDIDAS DE ASIMETRÍA

Coeficiente de Karl Pearson:

$$As = \frac{3(\bar{x} - Md)}{s}$$

Donde:

$\bar{x}$= media aritmética.
Md = Mediana.
s = desviación típica o estándar.

Nota:

El Coeficiente de Pearson varía entre -3 y 3.
Si As < 0 ? la distribución será asimétrica negativa.
Si As = 0 ? la distribución será simétrica.
Si As > 0 ? la distribución será asimétrica positiva.

Medida de Yule Bowley o Medida Cuartílica:

$$As = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

Donde:

Q_1= Cuartil uno; Q_2= Cuartil dos = Mediana; Q_3= Cuartil tres**.**

Nota:

La Medida de Bowley varía entre -1 y 1
Si As < 0 ? la distribución será asimétrica negativa.
Si As = 0 ? la distribución será simétrica.
Si As > 0 ? la distribución será asimétrica positiva.

Medida de Fisher:

Para datos sin agrupar se emplea la siguiente fórmula:

$$\mathrm{As} = \frac{\sum(\mathrm{x_i} - \bar{\mathrm{x}})^3}{\mathrm{n}\sigma^3}$$

Para datos agrupados en tablas de frecuencias se emplea la siguiente fórmula:

$$\mathrm{As} = \frac{\sum \mathrm{f}(\mathrm{x_i} - \bar{\mathrm{x}})^3}{\mathrm{n}\sigma^3}$$

Para datos agrupados en intervalos se emplea la siguiente fórmula:

$$As = \frac{\sum f(xm - \bar{x})^3}{n\sigma^3}$$

Donde:

x_i= cada uno de los valores; n = número de datos; $\bar{x}$= media aritmética; f = frecuencia absoluta
σ^3 = cubo de la desviación estándar poblacional; xm = marca de clase.

Nota:

Si As < 0 Indica que existe presencia de la minoría de datos en la parte izquierda de la media, aunque en algunos casos no necesariamente indicará que la distribución sea asimétrica negativa.
Si As = 0 la distribución será simétrica.
Si As > 0 Indica que existe presencia de la minoría de datos en la parte derecha de la media, aunque en algunos casos no necesariamente indicará que la distribución sea asimétrica positiva.

Por ejemplo: calcular el Coeficiente de Pearson, Medida Cuartílica y la Medida de Fisher dada la siguiente distribución: 6, 9, 9, 12, 12, 12, 15 y 17.

Solución: calculando la media aritmética se obtiene:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{6+9+9+12+12+12+15+17}{8} = \frac{92}{8} = 11{,}5$$

Para calcular los cuartiles se ordena los datos de menor a mayor:

6	9	9	12	12	12	15	17
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8

Fórmula:

$$Q_k = X_{\left[\frac{n\cdot k+2}{4}\right]}$$

Calculando el cuartil uno se obtiene:

$$Q_1 = X_{\left[\frac{n+2}{4}\right]} = X_{\left[\frac{8+2}{4}\right]} = X_{\left[\frac{10}{4}\right]} = X_{2,5} = \frac{x_2 + x_3}{2} = \frac{9+9}{2} = 9$$

Calculando el cuartil dos se obtiene:

$$Q_2 = X_{\left[\frac{n\cdot 2+2}{4}\right]} = X_{\left[\frac{2n+2}{4}\right]} = X_{\left[\frac{2\cdot 8+2}{4}\right]} = X_{\left[\frac{16+2}{4}\right]} = X_{4,5} = \frac{x_4 + x_5}{2} = \frac{12+12}{2} = 12$$

Calculando el cuartil tres se obtiene:

$$Q_3 = X_{\left[\frac{3n+2}{4}\right]} = X_{\left[\frac{3\cdot 8+2}{4}\right]} = X_{\left[\frac{24+2}{4}\right]} = x_{\frac{26}{4}} = X_{6,5} = \frac{x_6 + x_7}{2} = \frac{12+15}{2} = 13{,}5$$

Calculando la desviación estándar muestral se obtiene:

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

$$s = \sqrt{\frac{(6-11{,}5)^2 + (9-11{,}5)^2 + (9-11{,}5)^2 + (12-11{,}5)^2 + (12-11{,}5)^2 + (12-11{,}5)^2 + (15-11{,}5)^2 + (17-11{,}5)^2}{8-1}}$$

$$s = 3{,}505$$

Calculando el Coeficiente de Pearson se obtiene:

$$As = \frac{3(\bar{x} - Md)}{s} = \frac{3(11{,}5-12)}{3{,}505} = \frac{-1{,}5}{3{,}505} = -0{,}428$$

Calculando la Medida de Bowley se obtiene:

$$As = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1} = \frac{9 + 13{,}5 - 3 \cdot 12}{13{,}5 - 9} = -0{,}333$$

Calculando la desviación estándar poblacional se obtiene:

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

$$\sigma = \sqrt{\frac{(6-11{,}5)^2 + (9-11{,}5)^2 + (9-11{,}5)^2 + (12-11{,}5)^2 + (12-11{,}5)^2 + (12-11{,}5)^2 + (15-11{,}5)^2 + (17-11{,}5)^2}{8}}$$

$$\sigma = 3{,}279$$

Calculando la Medida de Fisher se obtiene:

Datos	$(x_i - \bar{x})^3$
6	-166,375
9	-15,625
9	-15,625
12	0,125
12	0,125
12	0,125
15	42,875
17	166,375
Total	12

$$As = \frac{\sum(x_i - \bar{x})^3}{n\sigma^3} = \frac{12}{8(3{,}279)^3} = 0{,}035$$

Los cálculos en ***Excel*** se muestran en la siguiente figura:

	A	B	C	D	E	F	G
1	Datos	$(x_i - \bar{x})^3$					
2	6	-166,375					
3	9	-15,625					
4	9	-15,625					
5	12	0,125					
6	12	0,125					
7	12	0,125					
8	15	42,875					
9	17	166,375					
10	Total	12	=SUMA(B2:B9)				
11	n	8	=CONTAR(A2:A9)				
12	Media aritmética	11,5	=PROMEDIO(A2:A9)				
13	Desviación estándar	3,5050983	=DESVEST.M(A2:A9)				
14	Desviación poblacional	3,2787193	=DESVEST.P(A2:A9)				
15	Cuartil 1	9	=CUARTIL.INC(A2:A9;1)				
16	Cuartil 2	12	=CUARTIL.INC(A2:A9;2)				
17	Cuartil 3	13,5	=CUARTIL.INC(A2:A9;3)+0,25*(A8-A7)				
18	Coeficiente de Pearson						
19	$As = \frac{3(\bar{x} - Md)}{s}$	-0,427948	=3*(B12-B16)/B13				
20							
21	Medida de Bowley						
22	$As = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$	-0,333333	=(B15+B17-2*B16)/(B17-B15)				
23							
24	Medida de Fisher						
25	$As = \frac{\sum(x_i - \bar{x})^3}{n\sigma^3}$	0,0425577	=B10/(B11*B14^3)				
26							
27	Coeficiente de Asimetría en Excel		0,0530788	=COEFICIENTE.ASIMETRIA(A2:A9)			

Nota. El COEFICIENTE.ASIMETRIA (A2:A9) es un valor que tiene consideraciones semejantes a la Medida de Fisher.

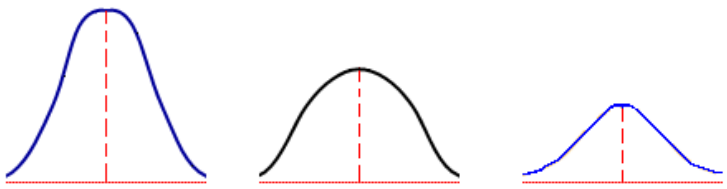
CURTOSIS O APUNTAMIENTO

La curtosis mide el grado de agudeza o achatamiento de una distribución con relación a la distribución normal, es decir, mide cuán puntiaguda es una distribución.

TIPOS DE CURTOSIS

La curtosis determina el grado de concentración que presentan los valores en la región central de la distribución. Así puede ser:

Leptocúrtica. Existe una gran concentración.
Mesocúrtica. Existe una concentración normal.
Platicúrtica. Existe una baja concentración.



MEDIDAS DE CURTOSIS

Medida de Fisher:

Para datos sin agrupar se emplea la siguiente fórmula:

$$\alpha = \frac{\sum(x_i - \bar{x})^4}{n\sigma^4}$$

Para datos agrupados en tablas de frecuencias se emplea la siguiente fórmula:

$$\alpha = \frac{\sum f(x_i - \bar{x})^4}{n\sigma^4}$$

Para datos agrupados en intervalos se emplea la siguiente fórmula:

$$\alpha = \frac{\sum f(xm - \bar{x})^4}{n\sigma^4}$$

Donde: $\mathbf{x_i}$ = cada uno de los valores; n = número de datos; $\bar{\mathbf{x}}$ = media aritmética; $\boldsymbol{\sigma^4}$ = Cuádruplo de la desviación estándar poblacional; f = frecuencia absoluta; xm = marca de clase.

Nota:

Si a < 3 ? la distribución es platicútica.
Si a = 3 ? la distribución es normal o mesocúrtica.
Si a > 3 ? la distribución es leptocúrtica.

Medida basada en Cuartiles y Percentiles:

$$\kappa = \frac{\text{Desviación cuartílica}}{\text{Amplitud cuartílica}} = \frac{\frac{Q_3 - Q_1}{2}}{P_{90} - P_{10}} = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$

κ(letra griega minúscula kappa) = Coeficiente percentil de curtosis.

Nota:

Si κ< 0,263 ? la distribución es platicúrtica

Si κ= 0,263 ? la distribución es normal o mesocúrtica

Si κ> 0,263 ? la distribución es leptocúrtica.

Esta medida no es muy utilizada.

Ejemplo: determinar qué tipo de curtosis tiene la siguiente distribución: 6, 9, 9, 12, 12, 12, 15 y 17. Emplear la medida de Fisher y el coeficiente percentil de curtosis.

Solución: Calculando la media aritmética se obtiene:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{6+9+9+12+12+12+15+17}{8} = \frac{92}{8} = 11{,}5$$

Calculando la desviación estándar poblacional se obtiene:

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

$$\sigma = \sqrt{\frac{(6-11{,}5)^2 + (9-11{,}5)^2 + (9-11{,}5)^2 + (12-11{,}5)^2 + (12-11{,}5)^2 + (12-11{,}5)^2 + (15-11{,}5)^2 + (17-11{,}5)^2}{8}}$$

$$\sigma = 3{,}279$$

Calculando la Medida de Fisher se obtiene:

Datos	$(x_i - \bar{x})^4$
6	915,0625
9	39,0625
9	39,0625
12	0,0625
12	0,0625
12	0,0625
15	150,0625
17	915,0625
Total	2058,5

$$\alpha = \frac{\sum(x_i - \bar{x})^4}{n\sigma^4} = \frac{2058,5}{8 \cdot (3,279)^4} = 2,23$$

Para calcular los cuartiles y percentiles se ordena los datos de menor a mayor:

6	9	9	12	12	12	15	17
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8

Calculando el cuartil uno se obtiene:

$$Q_k = X_{\left[\frac{n \cdot k+2}{4}\right]} \qquad Q_1 = X_{\left[\frac{n+2}{4}\right]} = X_{\left[\frac{8+2}{4}\right]} = X_{\left[\frac{10}{4}\right]} = X_{2,5} = \frac{x_2 + x_3}{2} = \frac{9+9}{2} = 9$$

Calculando el cuartil tres se obtiene:

$$Q_k = X_{\left[\frac{n \cdot k+2}{4}\right]} \qquad Q_3 = X_{\left[\frac{3n+2}{4}\right]} = X_{\left[\frac{3 \cdot 8+2}{4}\right]} = X_{\left[\frac{24+2}{4}\right]} = x_{\frac{26}{4}} = X_{6,5} = \frac{x_6 + x_7}{2} = \frac{12+15}{2} = 13,5$$

Calculando el percentil 90 se tiene:

$$P_k = X_{\left[\frac{n \cdot k+50}{100}\right]} \qquad P_{90} = X_{\left[\frac{n \cdot 90+50}{100}\right]} = X_{\left[\frac{8 \cdot 90+50}{100}\right]} = X_{\left[\frac{770}{100}\right]} = X_{7,7} = \frac{x_7 + x_8}{2} = \frac{15+17}{2} = 16$$

Calculando el percentil 10 se tiene:

$$P_k = X_{\left[\frac{n \cdot k+50}{100}\right]} \qquad P_{10} = X_{\left[\frac{n \cdot 10+50}{100}\right]} = X_{\left[\frac{8 \cdot 10+50}{100}\right]} = X_{\left[\frac{130}{100}\right]} = X_{1,3} = x_1 = 6$$

Calculando el coeficiente percentil de curtosis se obtiene:

$$\kappa = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} = \frac{13,5 - 9}{2(16 - 6)} = 0,225$$

Como a= 2,23 y $\kappa = 0,225$ la distribución es platicúrtica.

Los cálculos en **Excel** se muestran en la siguiente figura:

	A	B	C	D	E	F
1	Datos	$(x_i - \bar{x})^4$				
2	6	915,06250				
3	9	39,06250				
4	9	39,06250				
5	12	0,06250				
6	12	0,06250				
7	12	0,06250				
8	15	150,06250				
9	17	915,06250				
10	Total	2058,5	=SUMA(B2:B9)			
11	n	8	=CONTAR(A2:A9)			
12	Media aritmética	11,5	=PROMEDIO(A2:A9)			
13	Desviación poblacional	3,2787193	=DESVEST.P(A2:A9)			
14	Cuartil 1	9	=CUARTIL.INC(A2:A9;1)			
15	Cuartil 3	13,5	=CUARTIL.INC(A2:A9;3)+0,25*(A8-A7)			
16	$\alpha = \frac{\sum(x_i - \bar{x})^4}{n\sigma^4}$	2,226609	=B10/(B11*B13^4)			
17						
18	Perceltil 10	7,4	=PERCENTIL.INC(A2:A9;0,1)-0,25*(A3-A2)			
19	Perceltil 90	16,350	=PERCENTIL.INC(A2:A9;0,9)+0,25*(A8-A7)			
20	$\kappa = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$	0,2500	=(B15-B14)/(2*(B19-B18))			
21						
22						
23	**Curtosis en Excel**	-0,224121	=CURTOSIS(A2:A9)			
24	Valor semejante a la α	2,7758789	=B23+3			

EJERCICIO REPASANDO, PRACTICANDO Y APRENDIENDO:

El número de días necesarios por 10 equipos de trabajadores para terminar 10 instalaciones de iguales características han sido: 21, 32, 15, 59, 60, 61, 64, 60, 71, y 80 días.
Calcular la media, mediana, moda, varianza y desviación típica.

La media: suma de todos los valores de una variable dividida entre el número total de datos de los que se dispone:

$$\bar{X} = \frac{21 + 32 + 15 + 59 + 60 + 61 + 64 + 60 + 71 + 80}{10} = 52.3 \textbf{ dias}$$

La mediana: es el valor que deja a la mitad de los datos por encima de dicho valor y a la otra mitad por debajo.
Si ordenamos los datos de mayor a menor observamos la secuencia:

15, 21, 32, 59, 60, 60,61, 64, 71, 80

Como quiera que en este ejemplo el número de observaciones sea par (10 individuos), los dos valores que se encuentran en el medio son 60 y 60. Si realizamos el cálculo de la media de estos dos valores nos dará a su vez **60,** que es el valor de **la mediana.**

La moda: el valor de la variable que presenta una mayor frecuencia es **60.**

La varianza S^2: es la media de los cuadrados de las diferencias entre cada valor de la variable y la media aritmética de la distribución.

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

$$S^2 = \frac{(15-52.3)^2 + (21-53,2)^2 \ldots (80-53,2)^2}{10} = 427,61$$

La desviación típica S: es la raíz cuadrada de la varianza.

$S = \sqrt{S^2}$

$$S = \sqrt{427,61} = 20.67$$

El rango: diferencia entre el valor de las observaciones mayor y el menor: **80 - 15 = 65 días.**

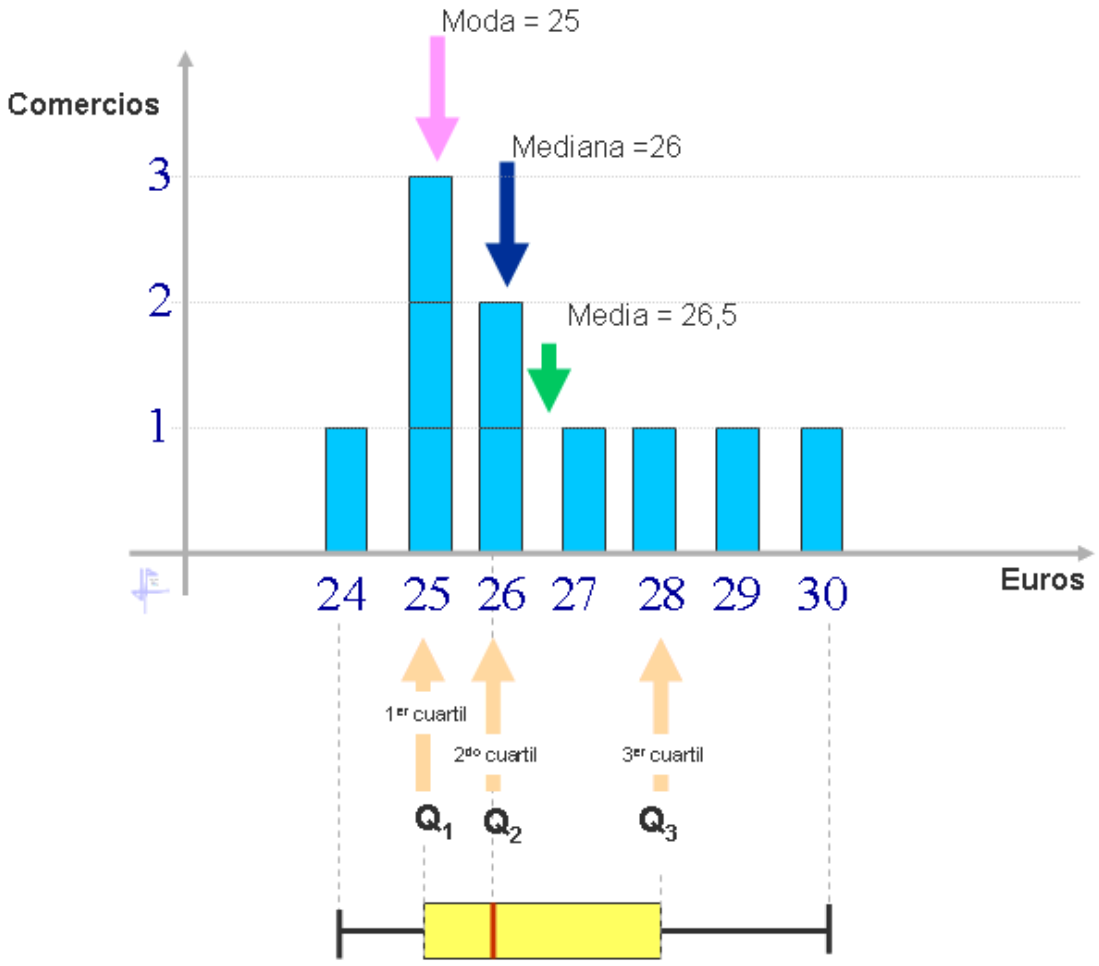
El coeficiente de variación: cociente entre la desviación típica y el valor absoluto de la media aritmética.

$$CV = \frac{20,67}{52,3} = 0,39$$

SIGAMOS RECORDANDO, PRACTICANDO Y APRENDIENDO:

El precio de un interruptor en 10 comercios de electricidad: 25, 25, 26, 24, 30, 25, 29, 28, 26, y 27 Euros.

Hallar la media, moda, mediana, (abrir la calculadora estadística, más abajo) diagrama de barras y el diagrama de caja.



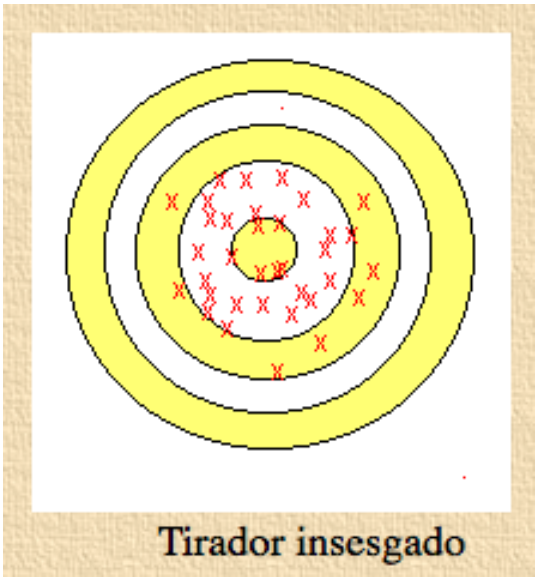
El diagrama de cajas: caja desde Q_1 a Q_3 (50% de los datos), bigotes el recorrido].

SESGO

Otra propiedad razonable que podemos pedir al estimador de un parámetro θ es que, en promedio, sus valores coincidan con θ. Cuando sucede esto decimos que el estimador es centrado o insesgado.

Un símil coloquial que suele aplicarse a la estimación puntual es considerarla como un ejercicio de tiro a una diana. En este sentido, el centro de la diana sería el parámetro a estimar (θ).

De manera los disparos de un tirador insesgado estarían centrados alrededor del centro de la diana. Mientras que los disparos de un tirador sesgado estarían sistemáticamente desviados de la diana (como sucedería si el cañón de nuestra arma no estuviese recto).



Podemos fijarnos que en la diana del tirador insesgado, el centro de masas de los disparos coincide con el centro de la diana (que representa el verdadero valor del parámetro). Como ya vimos anteriormente, el concepto de centro de masas está relacionado con la esperanza de una variable aleatoria y, precisamente así, obtenemos la definición formal de estimador insesgado: un estimador *T* de un parámetro θ diremos que es centrado o insesgado si su esperanza es precisamente θ.

$$T \text{ estimador insesgado de } \theta \quad \Leftrightarrow \quad E(T) = \theta$$

Si, al contrario, tenemos un estimador U sesgado, la desviación respecto al verdadero valor a estimar se mide por el sesgo:

$$T \text{ estimador de } \theta; \qquad \text{sesgo}\,(T) = E(T) - \theta$$

De manera que el sesgo de un estimador puede ser:

- ✓ Positivo: si producen, en promedio, estimaciones por exceso.
- ✓ Cero: si es un estimador centrado o insesgado.
- ✓ Negativo: si producen, en promedio, estimaciones por defecto.
- ✓ Un ejemplo de estimador insesgado es la media aritmética, que es un estimador insesgado de la esperanza de una variable aleatoria.

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \qquad \text{donde} \quad E(X_i) = \mu \qquad \forall\, i = 1, \ldots, n$$

$$E(\overline{X}_n) = E\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\, n\mu = \mu$$

Un ejemplo de estimador sesgado es la varianza muestral, que es un estimador sesgado de la varianza poblacional.

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 \quad \text{con} \quad E(X_i) = \mu \quad \text{y} \quad V(X_i) = \sigma^2$$

Sabemos que $E(X_i^2) = V(X_i) + (E(X_i))^2 = \sigma^2 + \mu^2$ y que $E(\overline{X}_n^2) = V(\overline{X}_n) + (E(\overline{X}_n))^2 = \frac{\sigma^2}{n} + \mu^2$

Por tanto:

$$E(S^2) = \frac{1}{n}\left[\sum_{i=1}^{n} E(X_i^2)\right] - \left[E(\overline{X}_n^2)\right] = \frac{1}{n}\left[n(\sigma^2 + \mu^2)\right] - \left(\frac{\sigma^2}{n} + \mu^2\right) =$$

$$= (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{n-1}{n}\sigma^2$$

Con lo que el sesgo de este estimador será:

$$\text{Sesgo}\ (S^2) = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

Es decir, el sesgo es negativo y, por tanto, la varianza muestral es un estimador de la varianza poblacional sesgado por defecto. Por dicha razón, suele utilizarse la llamada *varianza muestral corregida* como estimador de la varianza poblacional:

$$\hat{S}_n^2 = \frac{n}{n-1} S_n^2$$

que se comprueba trivialmente que sí es un estimador insesgado.

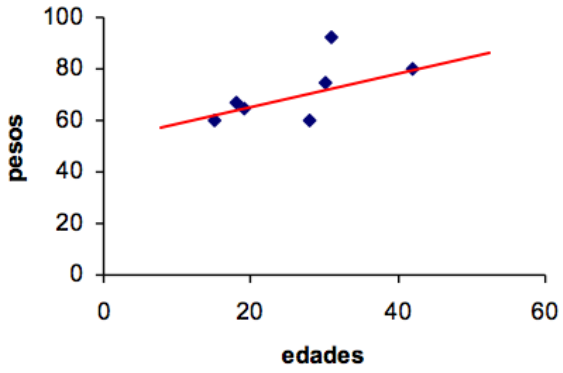
LA CORRELACIÓN

La correlación es la forma numérica en la que la estadística ha podido evaluar la relación de dos o más variables, es decir, mide la dependencia de una variable con respecto de otra variable independiente. Para poder entender esta relación tendremos que analizarlo en forma gráfica:

edad	*peso*
15	**60**
30	**75**
18	**67**
42	**80**
28	**60**
19	**65**
31	**92**

Si tenemos los datos que se presentan en la tabla y consideramos que la edad determina el peso de las personas entonces podremos observar la siguiente gráfica:

Donde los puntos representan cada uno de los pares ordenados y la línea podría ser una recta que represente la tendencia de los datos, que en otras palabras podría decirse que se observa que a mayor edad mayor peso.



La correlación se puede explicar con la pendiente de esa recta

estimada y de esta forma nos podemos dar cuenta que también existe el caso en el que al crecer la variable independiente decrezca la variable dependiente. En aquellas rectas estimadas cuya pendiente sea cero entonces podremos decir que no existe correlación.

Así en estadística podremos calcular la correlación para **datos no agrupados** con la siguiente formula.

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i * \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

En donde:

R = coeficiente de correlación
N = número de pares ordenados
X = variable independiente
Y = variable independiente

Ejemplo:

Edad (x)	*Peso (y)*	*X²*	*Y²*	*X* Y*
15	60	225	3600	900
30	75	900	5625	2250
18	67	324	4489	1206
42	80	1764	6400	3360
28	60	784	3600	1680
19	65	361	4225	1235
31	92	961	8464	2852
183	***499***	***5319***	***36403***	***13483***

Supóngase que deseamos obtener la correlación de los datos de la tabla anterior:

Ahora podemos observar que:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i * \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} = \frac{7 * 13483 - (183 * 499)}{\sqrt{\left[7 * 5319 - (183)^2\right]\left[7 * 36403 - (499)^2\right]}} = 0.65638606$$

Se debe aclarar que el coeficiente de correlación sólo puede variar de la siguiente manera: $-1 \leq r \leq 1$ y que para entenderlo mejor se debe obtener el **coeficiente de determinación** que se obtiene con "r" cuadrada, ya que este representa el porcentaje que se explica "y" mediante los datos de "x".

En nuestro ejemplo decimos que la correlación es casi perfecta, ya que, está muy cerca de 1 y que el porcentaje de datos que explican a "y" es $(0.65638606)^2 = 0.430842$ o sea el 43.08 %

En el caso de que fueran **datos agrupados** tendremos lo siguiente:

Primero tendremos que pensar que se genera una matriz, ya que, ahora estamos juntando dos tablas de distribución de frecuencias y por ello nuestros cálculos serán más laboriosos, por lo que les recomiendo el uso de una hoja de cálculo o al menos una calculadora con regresión para datos agrupados.

De cualquier forma aquí también estamos evaluando numéricamente si existe relación entre dos variables y lo haremos con la siguiente ecuación.

En donde podemos encontrar **k** como el número de clases para la variable "y" y **l** para el número de clases de "x".

$$r = \frac{n\sum_{j=1}^{k}\sum_{i=1}^{l} f\, x_i y_i - \sum_{i=1}^{l} f_x x_i * \sum_{i=1}^{k} f_y y_i}{\sqrt{\left[n\sum_{i=1}^{l} f_x x_i^2 - \left(\sum_{i=1}^{l} f_x x_i\right)^2\right]\left[n\sum_{i=1}^{k} f_y y_i^2 - \left(\sum_{i=1}^{k} f_y y_i\right)^2\right]}}$$

También podemos observar que hay varios tipos de "f" es decir, la que se encuentra sola (sin subíndice) que nos habla de las frecuencias celdares (cada una de las frecuencias que se encuentran en la intersección entre una columna y un renglón) y las "f" con subíndices que representan las frecuencias de cada una de las variables.

Para entender el uso de esta fórmula usaremos un ejemplo:

Los resultados que se presentan en la siguiente tabla representan los pesos y las estaturas de 48 alumnos entrevistados.

Marcas de clase de "x"

marcas de clase de "Y"	1.445	1.545	1.645	1.745	1.845	1.945	Σf_y	$\Sigma f_x y$	$\Sigma f_x y^{\wedge 2}$
44.5		3	1				4	178	7921
54.5		5	9	5			19	1035.5	56434.75
64.5		1	2	4	1	1	9	580.5	37442.25
74.5				5	1	1	7	521.5	38851.75
84.5				2	2	1	5	422.5	35701.25
94.5				1	3		4	378	35721
Σf_x	0	9	12	17	7	3	48	3116	212072
$\Sigma f_x x$	0	13.905	19.74	29.665	12.915	5.835	82.06		
$\Sigma f_x x^{\wedge 2}$	0	21.483225	32.4723	51.765425	23.828175	11.349075	140.8982		

$\Sigma \Sigma f\, x\, y$ 5380.77

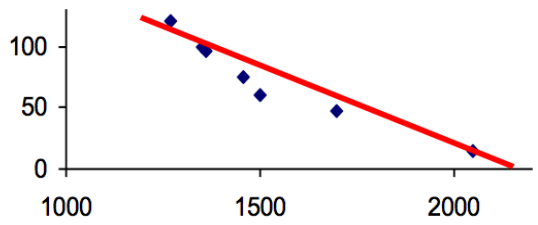
Correlación= 0.695

La sustitución de la fórmula es la siguiente:

$$r = \frac{n\sum_{j=1}^{k}\sum_{i=1}^{l} f\, x_i y_i - \sum_{i=1}^{l} f_x x_i * \sum_{i=1}^{k} f_y y_i}{\sqrt{\left[n\sum_{i=1}^{l} f_x x_i^2 - \left(\sum_{i=1}^{l} f_x x_i\right)^2\right]\left[n\sum_{i=1}^{k} f_y y_i^2 - \left(\sum_{i=1}^{k} f_y y_i\right)^2\right]}} =$$

$$\frac{48 * 5380.77 - (82.06 * 3116)}{\sqrt{((48 * 140.8982) - 82.06^2) * ((48 * 212072) - 3116^2)}} = 0.695$$

Al interpretar nuestro resultado podemos concluir que si existe relación entre el peso y la estatura, es decir, que a mayor estatura mayor peso. En muchas ocasiones el resultado de la correlación es negativo y lo que debemos pensar es que la relación de las variables involucradas en el cálculo es inverso es decir que en la medida que crece la variable independiente la variable dependiente decrece (imagen derecha).



CORRELACIÓN LINEAL

Un análisis de correlación nos permite cuantificar el grado de asociación lineal entre variables continuas, indica la fuerza y dirección de la relación lineal entre dos o más variables. Cuando exista dicha relación se podrá proceder a la obtención del modelo de regresión (simple o múltiple) que veremos posteriormente (Pérez, 2014).

Existen *diferentes tipos de correlación*, la correlación simple, la correlación múltiple y la correlación parcial. Utilizaremos la correlación simple cuando contemos con una sola variable predictora para explicar una respuesta, y los coeficientes de correlación parcial y múltiple cuando tengamos varios predictores.

CORRELACIÓN LINEAL SIMPLE

Utilizamos la correlación lineal simple para estudiar el grado de variación conjunta entre dos o más variables. Queremos detectar si la variación de una de las variables tiene conexión con la variación de la otra, esperamos que, si una variable de desvía de la media, la otra variable se desvíe de la media de manera similar.

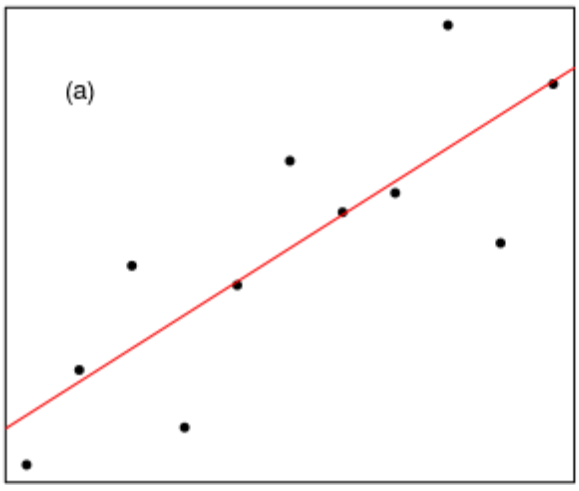
Una *relación lineal positiva* entre dos variables indica que los valores de las dos variables varían de forma parecida: los sujetos que puntúan alto en una variable tienden a puntuar alto en la otra y los que puntúan bajo en la primera tienden a puntuar bajo en la segunda, existe una relación directa entre ambas variables.

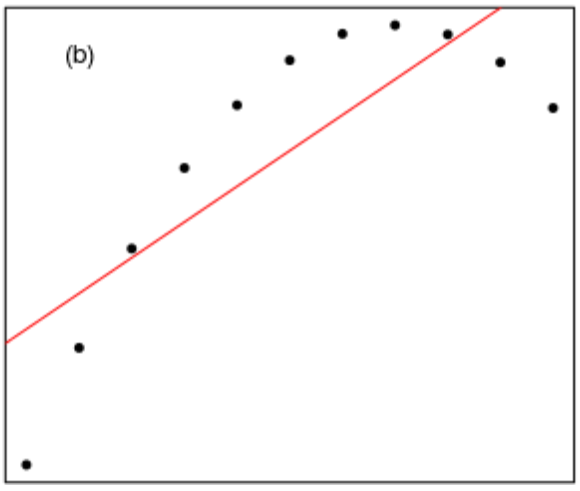
Una *relación lineal negativa* significa que los valores de las dos variables tienen una relación inversa: valores pequeños de una variable van asociados ahora a valores grandes de la otra y, equivalentemente, valores grandes de una se asocian a valores pequeños de la otra.

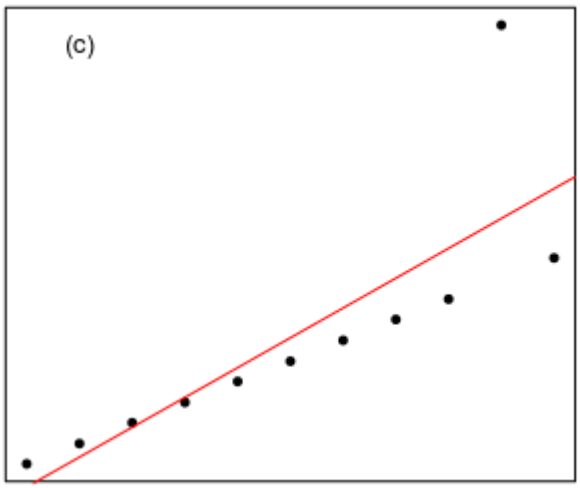
La forma más directa e intuitiva de formarnos una primera impresión sobre el tipo de relación existente entre dos variables es a través de un *diagrama de dispersión*. Se trata de un gráfico en el que una de las variables, XX, se coloca en el eje de abscisas, la otra, YY, en el de ordenadas y los pares (xi,yi)xiyi se representan como una *nube de puntos*. La forma de la nube de puntos nos informa sobre el tipo de relación existente entre las variables.

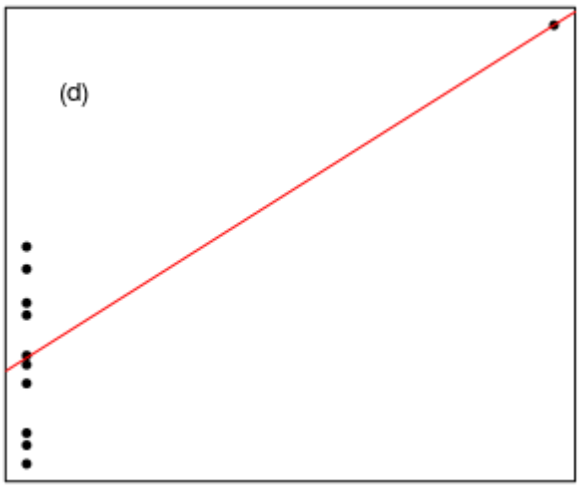
Una regla fundamental es que cuanta mayor correlación haya entre dos variables en la representación bidimensional, más próximos a la recta estarán los valores.

Veamos un ejemplo: en el siguiente gráfico mostramos cuatro diagramas de dispersión que reflejan cuatro tipos de relación diferentes (Ferrari & Head, 2010).









Para todos estos conjuntos de datos la recta de regresión es la misma:

$$\hat{y} = 3 + 0.5 \times x$$

con los coeficientes significativos con un nivel de significación:

$$< 0.01, \text{ y además todos tienen la misma } R^2 = 0.67 \text{ y } \hat{\sigma} = 1.24.$$

Sin embargo, solamente podemos escribir mediante un modelo lineal los datos del gráfico (a). El gráfico (b) muestra un conjunto de datos es claramente no lineal y sería mejor ajustarlo mediante una función cuadrática.

El gráfico (c) muestra un conjunto de datos que tiene un punto que distorsiona los coeficientes de la recta ajustada. Por último, el gráfico muestra un conjunto de datos totalmente inapropiado para un ajuste lineal, la recta ajustada está determinada esencialmente por la observación extrema (Ali S. Hadi, 2006).

Tras haber realizado una representación de los datos, una buena manera de cuantificar la relación a entre dos variables es mediante la **covarianza:**

$$r = Cov(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{N - 1},$$

donde *N* es el número de observaciones.

Sin embargo, la covarianza no es una medida útil para comparar rectas de regresión de variables distintas, o comparar el grado de asociación lineal entre distintos pares de variables, ya que depende de las escalas de medida de las variables. La solución está en estandarizarla y es de aquí de donde surgen llamados *coeficientes de correlación*.

COEFICIENTES DE CORRELACIÓN

El más importante de los coeficientes de correlación es el ***Coeficiente de Pearson***, que explicaremos en mayor profundidad, pero también están la ***Rho de Spearman*** y la ***Tau de Kendall***. Veamos sus **propiedades generales**:

- ✓ Todos los coeficientes varían entre -1 y 1.
- ✓ Si el coeficiente de correlación es -1 existe correlación negativa, es decir, a medida que una variable aumenta, la otra disminuye. Cuando el coeficiente es 1 hay correlación positiva, cuando aumenta una variable, también aumenta la otra.
- ✓ Un valor cercano o igual a cero indica poca o nula relación lineal entre las variables.
- ✓ Se utilizan como una medida de la fuerza de asociación: valores ±0.10.1 representan pequeñas asociación, ±0.30.3 asociación mediana, ±0.50.5asociación moderada, ±0.70.7 gran asociación y ±0.90.9 asociación muy alta.

Las principales **diferencias entre los coeficientes** son:

- ✓ La correlación de ***Pearson*** funciona bien con variables cuantitativas y que sigan bien la distribución normal.
- ✓ La correlación de ***Spearman*** se utiliza para datos ordinales o de intervalo que **no** satisfacen la condición de normalidad. (usualmente tiene valores muy parecidos a la de Pearson).
- ✓ La correlación de ***Kendall*** es una medida no paramétrica para el estudio de la correlación. Debemos utilizar este coeficiente en vez de la de Spearman cuando tengamos un conjunto de datos pequeño y muchas puntuaciones estén en el mismo nivel.

AJUSTE DE CURVAS

Básicamente el ajuste de curvas se utiliza cuando se tiene una serie de datos calculados y se desea conocer valores intermedios no conocidos, o también en aquellos casos en que se desee una versión simplificada de una función que se ajuste a un número de valores concretos, y posteriormente usar la función simplificada para derivar nuevos valores.

> "Ajustar una curva implica ajustar una función g(x) a un conjunto de datos (xi,yi), *i=1,2,...,L.* g(x) puede ser un polinomio, una función lineal o combinación de funciones conocidas"[1]

Hay básicamente dos métodos para lograr ajuste de curvas:

1. Si los datos no son muy exactos o tienen asociado un error (ruido) entonces la mejor manera es establecer una sola curva que represente la tendencia general de los datos observados. Se conoce como REGRESIÓN LINEAL, cuyo método más sencillo es la REGRESIÓN POR MÍNIMOS CUADRADOS.
2. Si los datos que se tienen son precisos se trazan una o varias curvas que pasan por cada uno de los puntos de datos. A esto se le llama INTERPOLACIÓN, la cual puede ser lineal o curvilínea.

REGRESIÓN POR MÍNIMOS CUADRADOS

En este método se pretende trazar la recta que más se acerque al conjunto de datos dado, a la cual se le llama "línea (recta) de regresión", expresada matemáticamente como:

$$Y = C_1x + C_2 + Error \qquad (2.1)$$

Los valores de C_1, C_2 y el *Error,* se pueden calcular de las siguientes maneras:

Empleando las fórmulas matemáticas:

$$C_1 = \frac{n\sum XiYi - \sum Xi \sum Yi}{n\sum Xi^2 - \left(\sum Xi\right)^2}, C_2 = \frac{\sum Yi}{n} - C_1\frac{\sum Xi}{n} \text{ y } Error = \sum (Yi - C_2 - C_1Xi)^2 \qquad (2.2)$$

El error se da en términos de la suma de los cuadrados de la diferencia entre el valor muestral y el valor calculado con la recta de regresión.

Ejemplo: encontrar la línea de regresión que se ajusta mejor a los siguientes datos.

x	1.0	1.5	2.0	2.5	3.0
y	2.0	3.2	4.1	4.9	5.9

La siguiente tabla muestra los valores calculados necesarios para hallar los valores de los coeficientes de la recta buscada:

n	*Xi*	*Yi*	*Xi*Yi*	*Xi²*	*Error*
1	1,00	2,00	2,00	1,00	0,0144
2	1,50	3,20	4,80	2,25	0,0169
3	2,00	4,10	8,20	4,00	0,0064
4	2,50	4,90	12,25	6,25	0,0049
5	3,00	5,90	17,70	9,00	0,0004
TOTAL	10,00	20,10	44,95	22,50	0,043

Reemplazando en las ecuaciones (2.2) tenemos:

$$C_1 = \frac{5*44.95 - 10.00*20.10}{5*22.50 - 10^2} = 1.90\,,\; C_2 = \frac{20.10}{5} - 1.90\frac{10.0}{5} = 0.22\,,\; Error = 0.043$$

Con estos valores al reemplazarlos en la ecuación (2.1) tenemos: Y = 1.90X + 0.22.

Ahora bien,…

EMPLEANDO LA HOJA ELECTRÓNICA DE EXCEL

La hoja electrónica trae incorporada la función **ESTIMACION.LINEAL**, con la cual se puede realizar el cálculo de la recta de regresión. Se procede de la siguiente manera: en una hoja de cálculo se crea una tabla con los valores de las variables independiente y dependiente.

Se resalta un bloque de al menos dos celdas, en la misma fila, en donde van a quedar los resultados de los coeficientes **C_1** y **C_2** buscados, luego se digita la fórmula: ***ESTIMACION.LINEAL(Rango_Y;Rango_X).*** **Rango_Y** nos indica el rango de celdas en donde se encuentran los valores de la variable dependiente; **Rango_X** contiene los valores de la variable independiente.

Luego, se presionan simultáneamente las teclas: ***Shift + Ctrl + Enter.***

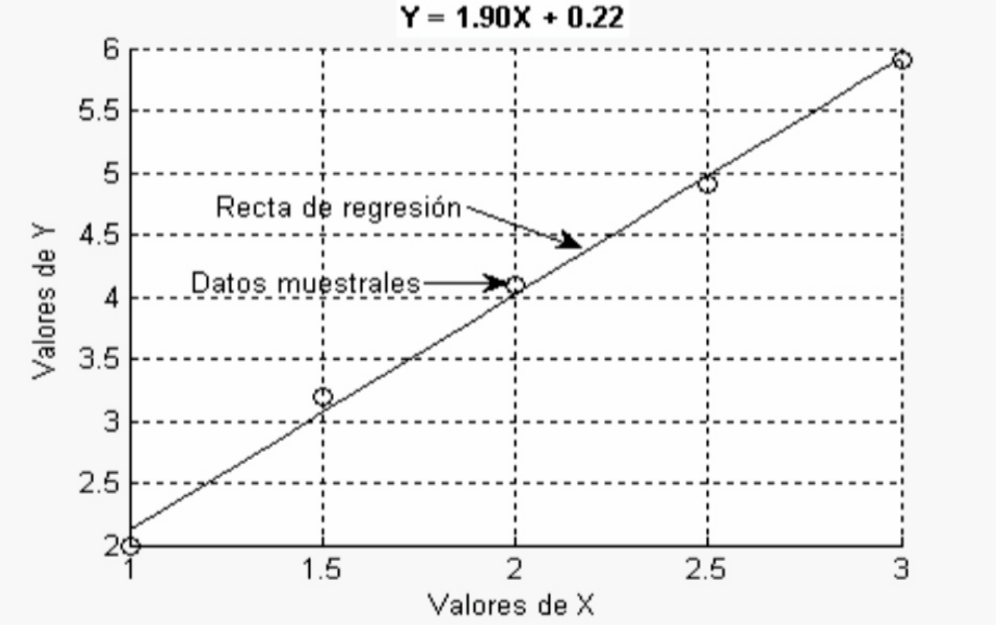
En la primera celda queda el valor de ***C_1*** y en la segunda el valor de ***C_2,*** con los cuales se forma la recta de regresión.

Para trazar la recta de regresión es necesario recalcular unos nuevos puntos de datos con espaciado más fino que los datos originales y luego evaluar la recta de regresión en estos puntos. Los comandos en MatLab que hacen esta tarea son:

```
x1=[x(1):0.1:x(length(x))]  %Se generan puntos de datos en el intervalo.
y1=polyval(C,x1)            %Se evalúa la recta de regresión en los nuevos puntos.
plot(x1,y1)                 %Dibuja la recta de regresión
hold on                     %Impide borrar la ventana gráfica
plot(x,y,'o')               %Dibuja los datos muestrales
```

Los tres últimos comando se pueden reemplazar por el comando: plot(x,y,'o',x1,y1).

Gráficamente los datos se observarían como se muestra en la figura siguiente:



Ejemplo 2: en una muestra se obtuvieron los siguientes resultados:

x	0.1	0.4	0.5	0.7	0.7	0.9
y	0.61	0.92	0.99	1.52	1.47	2.03

Hallar la recta de regresión y estimar el valor para x=1.3? Los cálculos a realizar se muestran en la siguiente tabla:

n	x	y	*Xi*Yi*	*Xi²*	*Error*
1	0,1	0,61	0,061	0,01	0,021722034
2	0,4	0,92	0,368	0,16	0,00518157
3	0,5	0,99	0,495	0,25	0,031840412
4	0,7	1,52	1,064	0,49	1,82307E-06
5	0,7	1,47	1,029	0,49	0,002636844
6	0,9	2,03	1,827	0,81	0,024254448
Total	3,3	7,54	4,844	2,21	0,085637131

Evaluar	1,3	***2,58008***
	C1	***C2***
	1,764557	0,28616

Los valores de C1 y C2 se obtuvieron utilizando las ecuaciones (2.3).

Podemos plantear la recta de regresión como Y = 1.7645X + 0.2862, y al evaluarla en X=1.3 da como resultado: 2.58008.

De la misma manera usando MatLab el proceso sería:

```
x=[0.1 0.4 0.5 0.7 0.7 0.9];
y=[0.61 0.92 0.99 1.52 1.47 2.03];
C=polyfit(x,y,1);
```

Que dan como resultado: C = [1.76455696202532 0.28616033755274] lo que nos permite construir la recta de regresión: Y=1.76455696202532X+ 0.28616033755274. Al evaluarla en X=1.3 digitamos el comando:

polyval(C,1.3)

Dando como respuesta: 2.58008438818565, que es el valor buscado.

INFORMACIÓN (INCLUÍDA EN ESTE DOCUMENTO EDUCATIVO) TOMADA DE:

Libros:

1. BENALCÁZAR, Marco, (2002), Unidades para Producir Medios Instruccionales en Educación, SUÁREZ, Mario Ed. Graficolor, Ibarra, Ecuador.
2. DAZA, Jorge, (2006), Estadística Aplicada con Microsoft Excel, Grupo Editorial Megabyte, Lima, Perú.
3. SUÁREZ, Mario, (2004), Interaprendizaje Holístico de Matemática, Ed. Gráficas Planeta, Ibarra, Ecuador.
4. SUÁREZ, Mario, (2011), Interaprendizaje de Estadística Básica TAPIA, Fausto Ibarra, Ecuador.

Sitios web:

1. http://www.eumed.net/cursecon/dic/oc/pearson.htm
2. https://es.wikipedia.org/wiki/Unidad_tipificada
3. https://es.wikipedia.org/wiki/Unidad_tipificada
   http://www.monografias.com/trabajos87/medidas-forma-asimetria-curtosis/medidas-forma-asimetria-curtosis.shtml#ixzz4bYGElSKb
4. http://halweb.uc3m.es/esp/Personal/personas/amalonso/esp/bstat-tema2p.pdf
5. http://www.um.es/ae/FEIR/40/
6. http://www.ub.edu/stat/GrupsInnovacio/Statmedia/demo/Temas/Capitulo7/B0C7m1t5.htm
7. http://www.tuveras.com/estadistica/estadistica02.htm