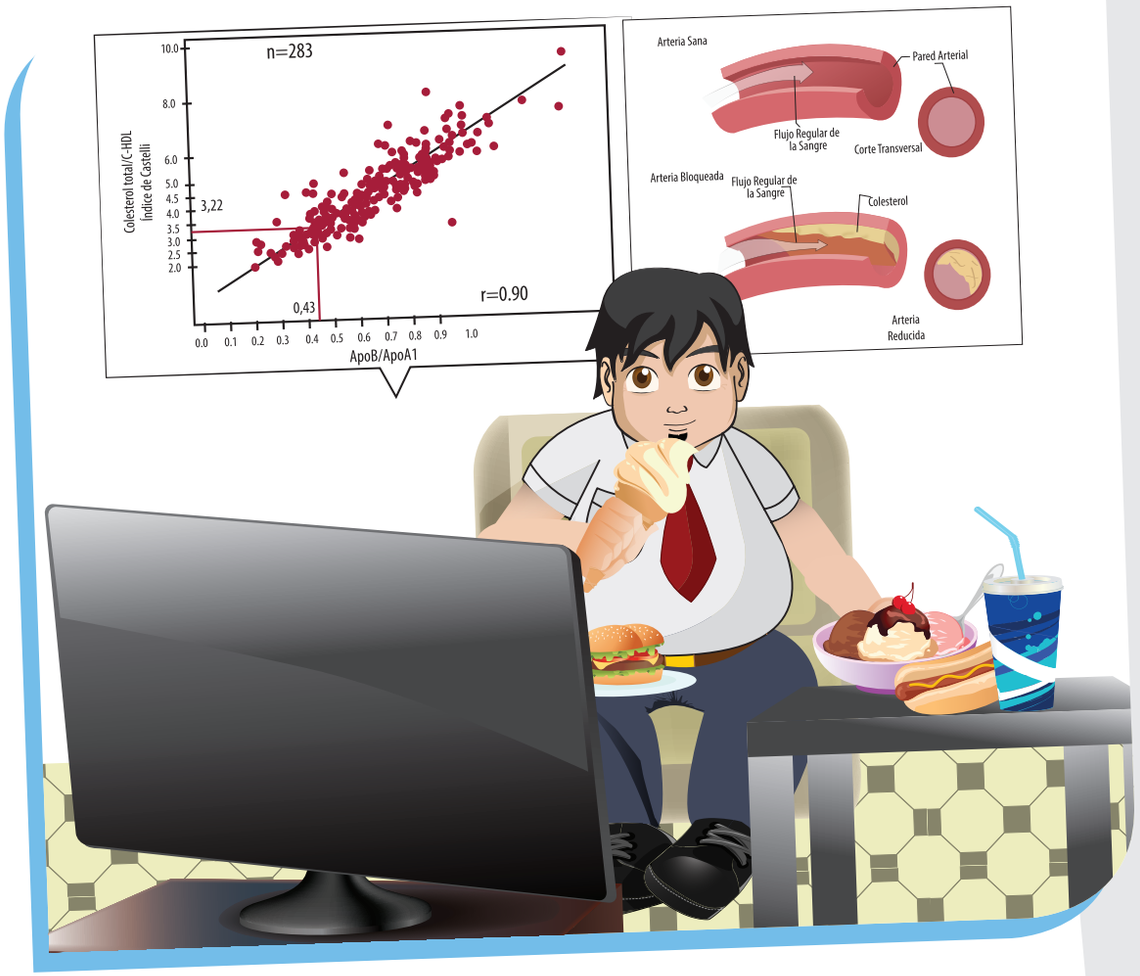
Guía 3

Utilicemos la correlación y la regresión para estudiar datos



Indicadores de desempeño

Conceptual:

- Define las características de la correlación y la regresión.

Procedimental:

- Emplea la correlación y la regresión para analizar un conjunto de datos.

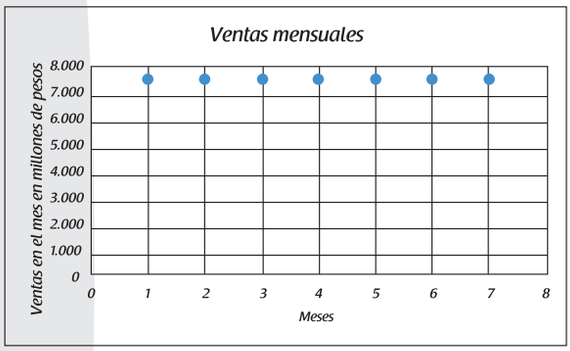
Actitudinal:

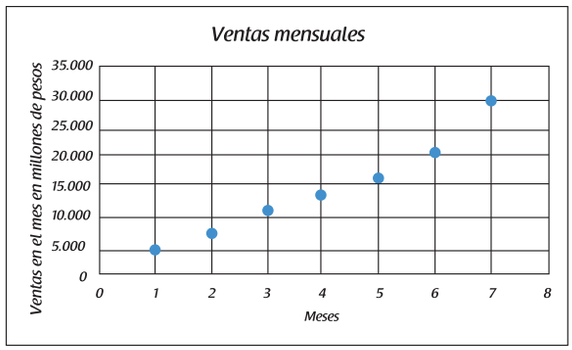
- Muestra responsabilidad en las decisiones que se toman al analizar un conjunto de datos.

A Vivencia

TRABAJO EN PAREJAS

1. Determinamos las tablas que permitieron elaborar las gráficas que se muestran a continuación:





   a. ¿Es posible establecer una relación entre las variables?
   b. Escribimos una posible fórmula algebraica que represente la relación entre las variables.

2. Elaboramos los gráficos de dispersión de los siguientes conjuntos de datos:

Mes	1	2	3	4	5	6
Demandas por discriminación laboral	13	16	21	28	48	61

Mes	1	2	3	4	5	6
Demandas por discriminación por discapacidad	2	6	12	20	30	42

Mes	1	2	3	4	5	6
Demandas por maltrato físico y verbal	2	4	8	16	32	64

   a. ¿Es posible establecer una relación entre las variables?
   b. Escribimos una posible fórmula algebraica que represente la relación entre las variables.

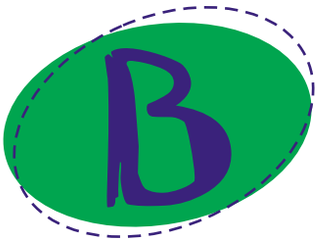
TRABAJO EN EQUIPO

3. Discutimos con nuestros compañeros sobre la forma en la que se relacionan las variables.

4. Analizamos las tendencias de las gráficas elaboradas a partir de la tabla de datos:

   a. ¿Es posible predecir el dato del séptimo mes?
   b. Si continúa la tendencia es posible determinar que en todos los casos existe un incremento de la violación de los derechos humanos según los datos. ¿Cuál es la posición del grupo con respecto a esta afirmación? Justificamos nuestra respuesta.
   c. ¿Cuál tipo de demanda demuestra el derecho más vulnerado?

5. Convocamos a nuestro profesor para que revise nuestro trabajo.



Fundamentación Científica

TRABAJO EN EQUIPO

1. En subgrupos de tres personas, realizamos la siguiente lectura y elaboramos en nuestros cuadernos un mapa conceptual con las ideas principales:

Regresión

El término regresión se lo adjudican al biólogo y estadístico inglés, **SIR FRANCIS GALTON**, quien lo introdujo en 1889. Empleó este concepto para indicar la relación que existía entre la estatura de los niños y la estatura de su padre en una muestra. Encontró, que si los padres son altos, los hijos también o si los padres son bajos, los hijos también. Pero ocurrían casos como en los que el padre es muy alto o muy bajo, y el hijo tiene una estatura media de la población, es decir, se presentó una regresión, de modo que los hijos retroceden hacia la media de la de sus padres, por lo tanto, están muy alejados. Hoy en día, este término no se utiliza en ese sentido.

Actualmente, la regresión s**e considera como la relación o dependencia entre dos características cuantitativas**, o más de una, consideradas sobre la misma población que es objeto de estudio, por ejemplo, la talla y el peso.

Los casos que estudiaremos a continuación muestran el análisis de la relación entre dos variables:

Caso 1: Si ambas variables están realmente relacionadas entre sí o son independientes.

Caso 2: Si existe dependencia, se requiere del **"grado de relación"**, así como del "tipo" de relación entre ambas.

Caso 3: Es posible predecir la variable que es considerada como dependiente a partir de los valores de la otra que es considerada independiente, y si es así, con qué precisión.

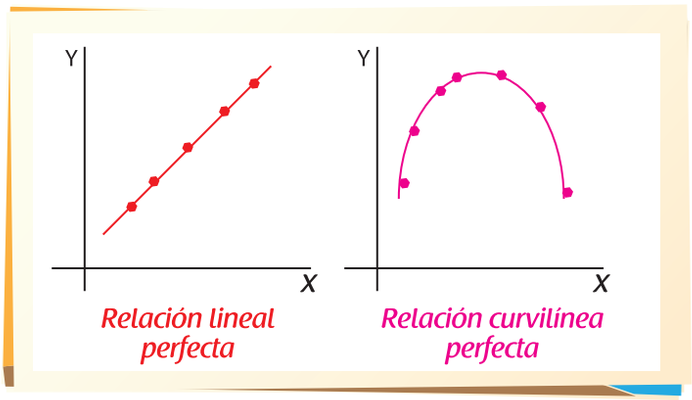
La existencia de una regresión se establece al determinar cuál es la variable aleatoria independiente, que se denominará X, y la variable aleatoria dependiente, que se denominará Y. Luego, se analizan los datos en parejas a partir de una muestra, los cuales se representan como las parejas ordenadas (x_i, y_i).

Estas se representan en un plano cartesiano, dando lugar al **diagrama de dispersión o nube de puntos**. La gráfica nos permite identificar la forma y dispersión de los puntos para establecer la manera en la que las variables se relacionan. Cuando examinamos un diagrama de dispersión, debemos estudiar el patrón de los puntos graficados. Si existe un patrón, es importante examinar su dirección: Si los datos van hacia arriba, esto sugiere que cuando una variable aumenta, la otra también lo hace; en cambio, si los datos van hacia abajo, esto sugiere que cuando una variable aumenta, la otra disminuye. También ubicamos los puntos mas alejados como valores extremos.

En primer lugar, debemos distinguir entre **dependencia funcional** y **dependencia estocástica.**

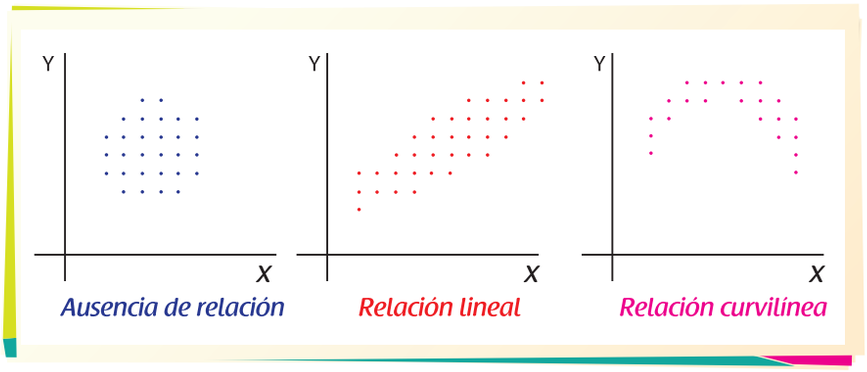
Dependencia funcional

La relación entre los datos de las variables es perfecta $y = f(x)$, es decir, los puntos del diagrama de dispersión corresponden a los valores de la función $y = f(x)$.



Dependencia estocástica

Es una relación menos rigurosa entre las variables, la cual se acerca a una dependencia funcional. La relación entre X e Y se establece con un margen de error residual.



En la dependencia estocástica, se distinguen dos tipos de técnicas:

1. Análisis de regresión: Este contesta a las siguientes preguntas: ¿Cuál es el tipo de dependencia entre las dos variables? ¿Pueden estimarse los valores de Y a partir de los de X? Y ¿Con qué precisión?

2. Análisis de correlación: Este contesta a las siguientes preguntas: ¿Existe dependencia estocástica entre las variables? y ¿Cuál es el grado de dicha dependencia?

Tipos de regresión

Si las dos variables X e Y se relacionan según un modelo de **línea recta**, entonces nos estamos refiriendo a una regresión **lineal simple**: $Y = a + bX$.

Cuando las variables X e Y se relacionan según una **línea curva**, hablamos de **regresión no lineal o curvilínea**. Aquí podemos distinguir entre **regresión parabólica, exponencial, potencial, etc.** Cuando tenemos más de una variable independiente (X_1, X_2,..., X_p), y una sola variable dependiente Y, la denominamos **regresión múltiple.**

Correlación

En un conjunto de datos es posible establecer inferencias sobre **la correlación** (o relación) entre dos variables. El **coeficiente de correlación lineal r** es una medida numérica de la fuerza de la relación entre dos variables que representan datos cuantitativos. Calculamos el valor de r haciendo uso de datos muestrales apareados y luego utilizamos este valor para concluir si existe o no una relación entre las dos variables.

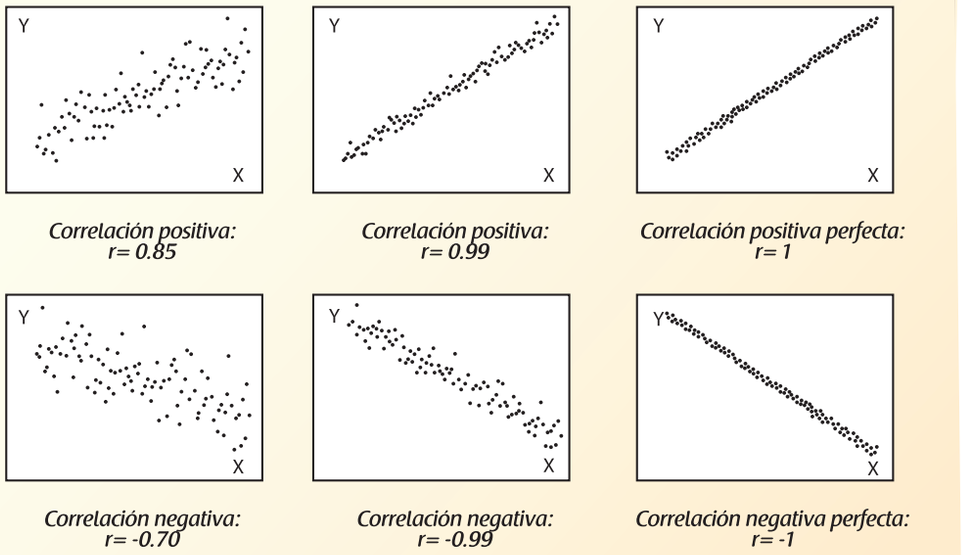
La correlación que abordaremos es la lineal. Esto significa que los datos se comportan como una línea recta. Para establecer el valor de r se requiere:

1. La muestra de datos (x, y), la cual es aleatoria y tiene datos cuantitativos.

2. El examen visual del diagrama de dispersión debe confirmar que los puntos se acercan al patrón de una línea recta.

3. Es necesario eliminar cualquier valor extremo, si se sabe que se trata de un error. Los efectos de cualquier otro valor extremo deben tomarse en cuenta calculando r con y sin el valor extremo incluido.

Los posibles valores de r se asocian a los siguientes diagramas de dispersión:



La fórmula para calcular el coeficiente de correlación r es:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\ \sqrt{n(\sum y^2) - (\sum y)^2}}$$

n	representa el número de pares de datos presentes.
$\sum$	denota la suma de los elementos indicados.
$\sum x$	denota la suma de todos los valores de x.
$\sum x^2$	indica que cada valor de x debe elevarse al cuadrado y después deben sumarse esos cuadrados.
$(\sum x)^2$	indica que los valores de x deben sumarse y el total elevarse al cuadrado. Es sumamente importante evitar confundirse entre $\sum x^2$ y $(\sum x)^2$
$\sum xy$	indica que cada valor de x debe multiplicarse primero por su valor y correspondiente. Después de obtener todos estos productos, se calcula su suma.
r	representa el coeficiente de correlación lineal de una **muestra.**
p	la letra griega rho se usa para representar el coeficiente de correlación lineal de una **población.**

El valor de r siempre debe estar entre -1 y +1. Si r se acerca a 0, concluimos que no existe una correlación lineal entre x y y, pero si r se acerca a -1 o +1, concluimos que hay una correlación lineal entre x y y. En la siguiente tabla se muestran las apreciaciones que existen con los valores de r:

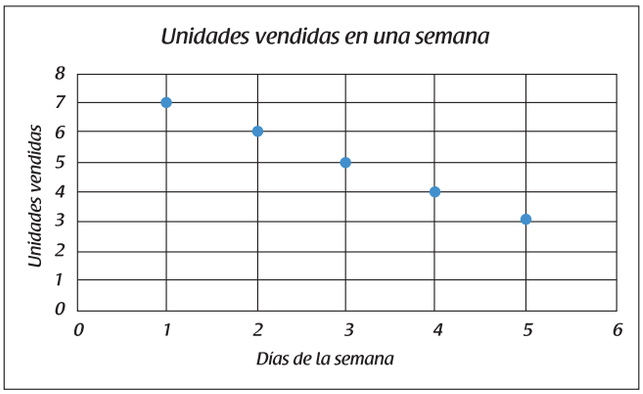
Valores de r	Tipo y grado de correlación
-1	Negativa perfecta
$1 < r \leq -0.8$	Negativa fuerte
$-0.8 < r < -0.5$	Negativa moderada
$-0.5 \leq r < 0$	Negativa débil
0	No existe
$0 < r \leq 0.5$	Positiva débil
$0.5 < r < 0.8$	Positiva moderada
$0.8 \leq r < 1$	Positiva fuerte
1	Positiva perfecta

Ejemplo 1:

Los siguientes datos son parte del registro de las unidades de carros vendidos en una semana en un concesionario:

Días de la semana	4	3	1	2	5
Unidades vendidas	4	5	7	6	3

Al realizar la gráfica de dispersión, encontramos que el patrón se asemeja a una línea recta. Entonces, esta cumple con los requisitos establecidos:



Realizamos los cálculos para determinar el nivel de correlación de los datos:

x	y	$x \cdot y$	x^2	y^2
4	4	16	16	16
3	5	15	9	25
1	7	7	1	49
2	6	12	4	36
5	3	15	25	9
15 ↑ $\sum x$	25 ↑ $\sum y$	65 ↑ $\sum xy$	55 ↑ $\sum x^2$	135 ↑ $\sum y^2$

Ahora, determinamos el cálculo de r con la fórmula:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\ \sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r = \frac{5(65) - 15 \cdot 25}{\sqrt[2]{5(55) - 15^2}\sqrt[2]{5(135) - 25^2}}$$

$$r = \frac{-50}{\sqrt[2]{50}\sqrt[2]{50}}$$

$$r = -1$$

Entonces, es una correlación negativa significativa.

Ecuación de la regresión lineal

Cuando los datos se organizan en una línea recta, es posible establecer una tendencia a través de una ecuación lineal.

Recordemos que la función es $y = mx + b$
m se calcula con la siguiente fórmula:

$$m = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum {x_i}^2 - (\sum x_i)^2}$$

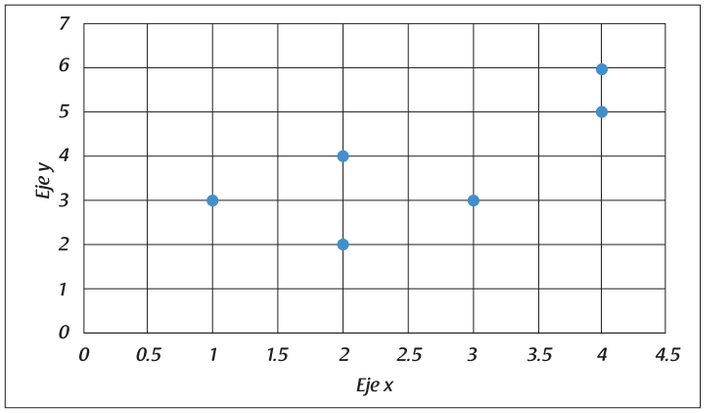
Donde $b = \bar{y} - m\bar{x}$. Además, $\bar{y} = \frac{\sum yi}{n}$ y $\bar{x} = \frac{\sum x_i}{n}$

Ejemplo 2:

Los siguientes datos se tomaron aleatoriamente de las variables x y y. Los ajustamos a una línea recta:

x	1	2	2	3	4	4
y	3	2	4	3	5	6

Paso 1: Elaboramos la gráfica de dispersión:



Paso 2: Realizamos los cálculos para determinar el valor de m:

	x	y	$x \cdot y$	x^2
	1	3	3	1
	2	2	4	4
	2	4	8	4
	3	3	9	9
	4	5	20	16
	4	6	24	16
Total	16	23	68	50

Aplicamos la fórmula para obtener el valor de m:

$$m = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum {x_i}^2 - (\sum x_i)^2}$$

Entonces, al reemplazar:

$$m = \frac{6 \cdot 68 - 16 \cdot 23}{6 \cdot 50 - 256}$$

$$m = \frac{408 - 368}{300 - 256}$$

$$m = \frac{40}{44} = \frac{10}{11}$$

$$m = 0.909$$

Paso 3: Calculamos el valor de b:

$$b = \frac{\sum yi}{n} - m\frac{\sum x_i}{n}$$

$$b = \frac{23}{6} - \frac{10}{11}\left(\frac{16}{6}\right)$$

$$b = \frac{31}{22}$$

$$b = 1.4091$$

Paso 4: Escribimos la ecuación de la recta:

$$y = 0.909x + 1.4091$$

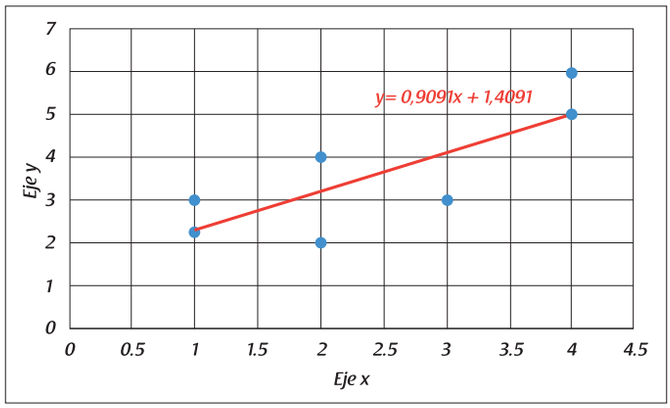
Paso 5: Representamos la recta y buscamos las coordenadas de los valores extremos de x. Estos valores son $x = 1$ y $x = 4$.

Entonces, sus coordenadas en y correspondientes son:

$$y(1) = 0.909\ (1) + 1.4091 = 2.3181$$

$$y(4) = 0.909\ (4) + 1.4091 = 5.0451$$

Realizamos el trazo de la línea:

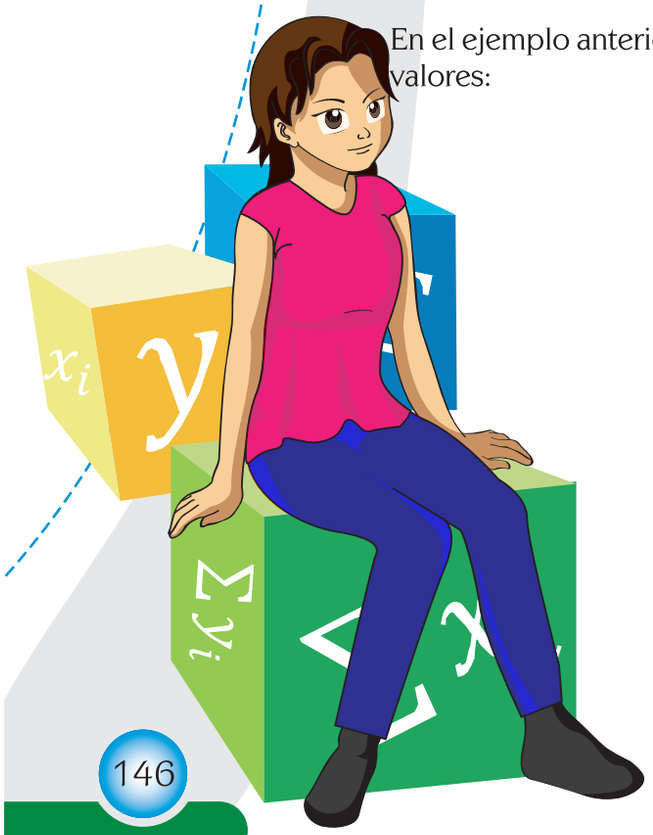


Como observamos, en la gráfica existe una distancia entre los datos reales y los puntos de la recta, lo que nos indica que existe un **error de estimación a la ecuación lineal**. Este valor sirve para determinar la validez de la ecuación. Para calcular el error se usa la siguiente fórmula:

$$S_{yx} = \sqrt[2]{\frac{\sum y^2 - b(\sum y) - m(\sum xy)}{n-2}}$$

Ejemplo 3:

En el ejemplo anterior, la ecuación lineal es $y = 0.909x + 1.4091$. Calculamos los valores:



	x	y	xy	y^2
	1	3	3	9
	2	2	4	4
	2	4	8	16
	3	3	9	9
	4	5	20	25
	4	6	24	36
Total	16	23	68	529

Al aplicar la fórmula:

$$S_{yx}=\sqrt[2]{\frac{529-1.4091(23)-0.909(68)}{7-2}}$$

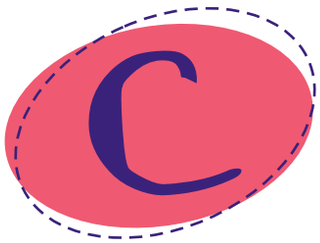
$$S_{yx}=\sqrt[2]{\frac{534.7787}{5}}$$

$$S_{yx}=\sqrt[2]{56.95574}$$

$$S_{yx}=9.325006166$$

El error estándar de estimación es de 9.325006166, que representa la variabilidad alrededor de la recta de regresión. Esto equivale al 93%, lo cual indica que la relación lineal es muy fuerte.

2. Invitamos a nuestro docente a que revise nuestro trabajo y le solicitamos amablemente que nos aclare las dudas.



Ejercitación

TRABAJO EN PAREJAS

1. Buscamos el número de correlación entre las variables estocásticas. Podemos utilizar la calculadora del CRA o la ayuda de una hoja de Excel:

   a. Los siguientes datos muestran el número de cigarrillos que consumen las personas en un día:

Número de cigarrillos	2	7	8	10	8	9
Número de personas que consumen	4	8	6	7	5	6

   b. Los siguientes datos son de las personas que se presentan en el hospital entre las 4am y 5am, algunas de las cuales van por accidentes de tránsito:

Personas en el hospital	12	5	7	9	3	10
Personas por accidentes de tránsito	10	5	6	7	4	9

c. Los siguientes datos muestran cuántas personas compran gaseosa de 1.5 litros entre las 12 am y 2:00 pm en la tienda del municipio de Nocaima:

Compran gaseosa	20	15	10	5
Compran gaseosa de 1.5 l	3	6	9	7

2. Establecemos la ecuación lineal que se ajusta al conjunto de los datos dados en el ejercicio anterior. Podemos utilizar la calculadora del CRA.

3. Determinamos el error de estimación del conjunto de datos dados en el ejercicio anterior. Recomendamos utilizar la calculadora del CRA.

4. Determinamos las decisiones que se pueden tomar con relación al análisis de cada una de las situaciones. ¿Es posible predecirlas?

5. Socializamos las respuestas de esta actividad. Solicitamos a nuestro profesor que dirija el ejercicio y aclare nuestras inquietudes al respecto.

Aplicación

TRABAJO EN EQUIPO

1. Leemos con atención cada una de las siguientes situaciones. Para resolverlas, podemos utilizar la calculadora del CRA o apoyarnos en una hoja de Excel:

   a. En la siguiente tabla se muestran los datos recogidos de 30 familias del casco rural de Manizales sobre sus ingresos y consumo de productos:

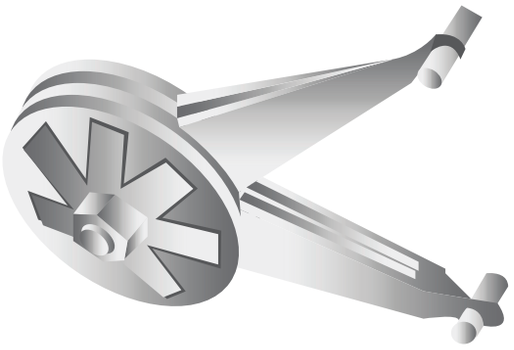
Número de familias	Ingreso en diez miles	Consumo en diez miles
1	119	154
2	85	121
3	97	125
4	120	140
5	92	130
6	105	141
7	110	134
8	198	130
9	98	134
10	81	115
11	81	79
12	103	125
13	105	144
14	100	137
15	68	65
16	78	98
17	56	67
18	67	134
19	108	147
20	116	144
21	86	133
22	108	120
23	36	45
24	67	126
25	78	145
26	76	35
27	87	120
28	70	81
29	60	72
30	70	65

a. Establecemos la correlación entre las variables.
b. Determinamos la ecuación de la regresión lineal.
c. Generamos una explicación sobre la relación entre las variables de consumo e ingresos por familia.
d. ¿Cuál sería un plan de control de gastos coherente con los ingresos familiares para obtener una regresión lineal perfecta?

e. ¿Es posible tomar una decisión coherente con la ecuación y los datos reales sobre el control de gastos e ingresos?

2. En unas pruebas de rebote de un amortiguador de fricción se obtuvieron para distintas alturas de caída (variable independiente) y las alturas de rebote (variable dependiente) los siguientes datos:

Prueba	1	2	3	4	5	6	7
Distancia de caída dm	1.2	2.4	3.2	4.3	5.7	7	9
Distancia de altura mm	1.9	1.3	2.1	2.7	2	3.6	3.3



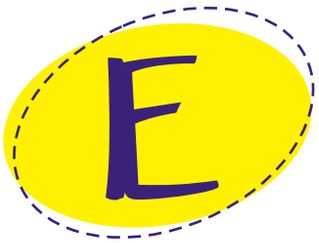
a. Determinamos la recta de regresión.
b. Ajustamos los datos elaborando una nueva tabla y señalamos las diferencias entre el valor real y el valor que da la fórmula.
c. Establecemos la correlación entre las variables.
d. ¿Es posible determinar la calidad del amortiguador con esta información?
e. ¿Es una toma de decisión beneficiosa la implementación de este amortiguador?

3. La siguiente tabla muestra las calificaciones de 9 estudiantes de grado noveno en las asignaturas de matemáticas y estadística:

Estudiante	Matematicas	Estadística
Diana	10	8
Carolina	6	6
María	7	10
Pedro	9	9
Matías	8	10
Brayan	5	8
Ana	4	7
Lucía	7	9
Edwin	8	9

a. Calculamos la correlación entre las variables.
b. Determinamos la regresión lineal.
c. Dibujamos los diagramas de nubes.
d. ¿Es posible tomar una decisión de plan de mejoramiento en alguna de las dos áreas a partir de los resultados reales y posibles con la ecuación? Justificamos nuestra respuesta.

4. Invitamos al docente a revisar nuestro trabajo.



Complementación

TRABAJO EN PAREJAS

1. La siguiente información complementa las formas de realizar los cálculos para determinar el grado de correlación entre las variables y las formas de calcular la regresión lineal. Anotamos la información en el cuaderno y establecemos las ventajas que tiene cada una de las fórmulas:

Un método para calcular la ecuación de la regresión lineal es a través de los **mínimos cuadrados**, en el cual se da un ajuste de los datos muestrales a una recta representativa de la relación de dependencia entre las dos variables. Como sabemos, la ecuación de una recta es $y = a + bx$, en donde a y b son coeficientes constantes; y, x e y representan los valores de las variables independiente y dependiente, respectivamente. Utilizando esta fórmula, vamos a obtener la recta óptima en la cual las distancias entre los puntos de la distribución a la recta sean mínimas.

Para hallar la recta de regresión lineal mediante el método de los mínimos cuadrados utilizaremos las siguientes fórmulas:

$$y = a + bx$$

$$a = \frac{\sum y - b\sum x}{n}$$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - [\sum x]^2}$$

Igualmente, se puede calcular la forma de determinar **el coeficiente de correlación** r entre las variables, el que también representa una medida en el que tanto los valores de x y de y se determinan mutuamente.
El valor de r varía siempre entre -1 y 1, de hecho, r tiene siempre el mismo signo de la pendiente de la recta, entre más cercanos se encuentren los puntos a la recta, el coeficiente r se acercará más al valor de -1 ó 1.
Si $r = 0$, no existe correlación entre las variables.

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2\,(n\sum y^2 - (\sum y)^2}}$$

Otra forma es el coeficiente de **correlación de Karl Pearson** o coeficiente de correlación producto-momento. Se calcula aplicando la siguiente ecuación:

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)\,(\sum y^2)}}$$

Donde $\boldsymbol{x}$ es la diferencia entre los valores de la variable independiente y el promedio de las variables independientes, $x = X - \bar{X}$. Por otro lado, $\boldsymbol{y}$ es la diferencia entre los valores de la variable dependiente y el promedio de las variables dependientes, $y = Y - \bar{Y}$.

Ejemplo 1:

Los siguientes son los datos de temperatura de dos días diferentes de una ciudad de Colombia. A partir de ellos, determinamos el tipo de correlación que existe entre las variables mediante el coeficiente de PEARSON:

							Suma	Promedio
X	18	17	15	16	9	10	85	14.2
Y	13	15	14	9	10	12	73	12.2

Para calcular el coeficiente de correlación de Pearson se requiere lo siguiente:

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy	y^2
18	13	3.8	0.8	14.7	3.2	0.7
17	15	2.8	2.8	8.0	8.0	8.0
15	14	0.8	1.8	0.7	1.5	3.4
16	9	1.8	-3.2	3.4	-5.8	10.0
9	10	-5.2	-2.2	26.7	11.2	4.7
10	12	-4.2	-0.2	17.4	0.7	0.0
				70.8	18.8	26.8

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{18.8}{\sqrt{(70.8)(26.8)}} = \frac{18.8}{43.5} = 0.4316$$

Esto indica que la correlacion es moderada.

2. Utilizamos las fórmulas para calcular los coeficientes de correlación y los comparamos con el obtenido en el ejemplo 1 de esta sesión.

3. Hacemos uso de las fórmulas para determinar la ecuación de la regresión lineal y revisamos con el ejemplo 1 de esta sesión si es la misma ecuación.

4. Presentamos nuestros cuadernos al profesor para que revise lo desarrollado y valore nuestro desempeño.

Evaluación por competencias

INFORMACIÓN PARA CONTESTAR LAS PREGUNTAS 1, 2 Y 3

La tabla muestra el promedio de las pérdidas de peso, observadas en 9 grupos de 25 pájaros, después de 6 días de ser sometidos a distintos grados de humedades relativas:

Pérdida de peso (mg)	8.98	8.14	6.67	6.08	5.90	5.83	4.68	4.20	3.72
% Humedad relativa	0	12	29.5	43	53	62,5	75.5	85	93

1. La variable independiente es:

A. Número de pájaros.
B. Días de pérdida de peso.
C. Peso.
D. Humedad.

1

2. El coeficiente de correlación de Pearson es:

A. 0.50716983
B. -0.13389158
C. 0.3924568
D. -0.32260598

2

3. El tipo de correlación entre las variables es:

A. Positiva moderada.
B. Negativa moderada.
C. Positiva fuerte.
D. Negativa fuerte.

3

INFORMACIÓN PARA CONTESTAR LAS PREGUNTAS 4 Y 5

La tabla muestra los datos recogidos sobre la altura y el peso de 10 hombres entre los 20 y 30 años:

Estatura en cm	Peso en Kg
152	56
157	61
162	67
173	72
178	89
182	83
188	75
165	90
170	67
167	77

4. Dibujo el diagrama de nubes.

5. ¿La regresión lineal que representa el conjunto de datos es $y = 0\,6188x - 31{,}117$?

   A. SÍ
   B. NO

Glosario

- **Correlación:** Es un valor numérico que indica la fuerza de relación entre las variables.
- **Dependencia estocástica:** Los puntos del diagrama de dispersión corresponden a los datos y presentan un residuo.
- **Dependencia funcional:** Los puntos del diagrama de dispersión corresponden a los valores de la función $y = f(x)$.
- **Error estándar de estimación:** Mide la dispersión de los valores observados alrededor de la línea de regresión. Está representado por el símbolo S_{yx}.
- **Grado de relación:** La medida indica el grado de dependencia que existe entre las variables involucradas. Existen varios modos de calcular esta relación.
- **Nube de puntos:** Representa cada uno de los puntos de los datos de las variables. La forma de su distribución permite identificar un patrón.
- **Variable:** Es toda característica de algún fenómeno susceptible de medición y que puede tomar diferentes valores: Peso, estatura, ingresos, productividad, etc.
- **Variable dependiente:** Es aquella cuyos valores van a estar determinados por el valor que se le asigne a la variable.
- **Variable independiente:** Es aquella que puede controlar el investigador y a la que se le puede asignar cualquier valor.