UNIDAD II. REGRESIÓN Y CORRELACIÓN LINEAL

¡Hola!

Tienes en tus manos el material de estudio para aprender más sobre otra aplicación de la estadística: la regresión y correlación lineal simple.

Prepárate para sumergirte en el mundo de la estadística y sus aplicaciones. Recuerda que es importante comprometerse con tu aprendizaje para tener éxito.

Espero lo disfrutes.

¡Comencemos!

En nuestro entorno conocemos variables y las utilizamos todo el tiempo, tanto que, muchas veces no les damos importancia; pero ¿te has preguntado si estas variables influyen unas con otras y si estos comportamientos se pueden matematizar?

La respuesta no siempre es simple, pues depende del experimento y la información que te arroja. Recuerda que las variables se miden de unidades experimentales. Cuando estudiamos una sola variable, por ejemplo: el color de ojos de los estudiantes, decimos que los datos son **univariados,** lo que significa obtener un solo dato de la unidad experimental (un estudiante). Estos datos se pueden presentar u obtener medidas estadísticas si son cuantitativos, e interpretarlas. Ahora bien, si de los mismos estudiantes (la unidad experimental) podemos medir otra variable: la graduación de sus lentes, se llaman datos **bivariados**. Lo interesante de los datos bivariados es que podemos compararlos y responder la pregunta que nos planteamos al inicio: ¿Una variable puede influir en la otra?

Aunque se pueden comparar datos bivariados cualitativos, para este curso nos centraremos en las comparaciones de solo datos **bivariados cuantitativos.** Por otro lado, aunque el comportamiento de las variables puede ajustarse con diversas funciones (cuadrática, exponencial, polinomial, logarítmica, etc) nos centraremos en lo más simple matemáticamente hablando: la línea recta.

En esta sesión estudiaremos la asociación entre dos variables, así como su representación gráfica, para posteriormente en la próxima sesión estudiar un método que se utiliza para estimar (explicar) una variable: regresión lineal simple. En el esquema siguiente se muestran los temas que abordaremos en la unidad II.

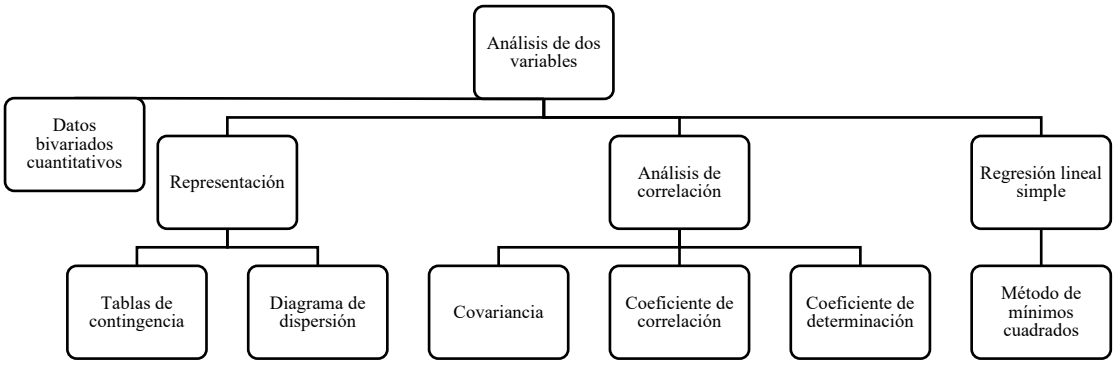


Figura. 1. Esquema de la unidad II.

1. REPRESENTACIÓN DEL ANÁLISIS DE DOS VARIABLES.

Cuando las variables son cuantitativas, y se relacionan entre sí, podemos representarlas como pares ordenados (x,y) para analizarlas y graficarlas como en el curso de Geometría Analítica o en Cálculo, dónde seguramente aprendiste que en las funciones existe una variable dependiente y una independiente.

En estadística, cuando analizamos datos bivariados, seleccionamos cualquiera de las dos variables como dependiente o independiente, pero el análisis lo hacemos según nuestra elección. Así, una vez clasificadas nuestras variables, construimos una tabla de contingencia y/o un diagrama de dispersión, donde colocamos a la variable independiente en el eje x y a la variable dependiente en el eje y.

1.1 Tabla de contingencia

La tabla de contingencia se utiliza para clasificar el número de observaciones respecto a dos características o variables de interés. En una columna se coloca a la variable X y en otra a la variable Y. Para el análisis de dos variables se requiere que dichas variables sean cuantitativas.

Ejemplo 1.1 Construcción de tabla de contingencia

A Alicia le gusta leer y un día se preguntó si existía una relación entre el número de páginas de un libro con su precio. De tal forma que representó los datos en la tabla 1. Represéntalos en una tabla de contingencia.

Título	Núm de págs	Precios
El sueño del celta	464	$269
El asedio	736	$289
La peste	232	$188
Cartas a una joven matemática	238	$219
Kafka en la orilla	574	$165
Harry Potter	192	$149
Bajo la misma estrella	208	$113

Tabla 1. Libros de Alicia

Solución

Primero elegimos quien será la variable independiente y quien será la dependiente.

X= número de páginas Y= Precio.

Es decir, el precio depende del número de páginas del libro.

Posteriormente armamos la tabla de contingencia:

X	Y
464	269
736	289
232	188
238	219
574	165
192	149
208	113

Tabla 2 Representación en tabla de contingencia del ejercicio 1.1

1.2 Diagrama de dispersión

Un diagrama de dispersión es una gráfica de puntos representados en el plano cartesiano. Cada punto indica un par de valores (x, y). Este diagrama permite observar cómo se relacionan dos variables; generalmente, lo que se busca al usar un diagrama de este tipo es determinar si los puntos siguen una línea recta y si ésta tiene pendiente positiva o negativa.

Ejemplo 1.2 Construcción de un diagrama de dispersión.

Realiza un diagrama de dispersión con la tabla 1.

Solución

Primero debemos pensar la escala, que por los datos que tenemos conviene que sea de 100 cada línea en el eje de las Xs y de 50 cada línea en el eje de las Ys.

Luego ubicamos las coordenadas de cada punto y dibujamos los puntos, para que quede como sigue:

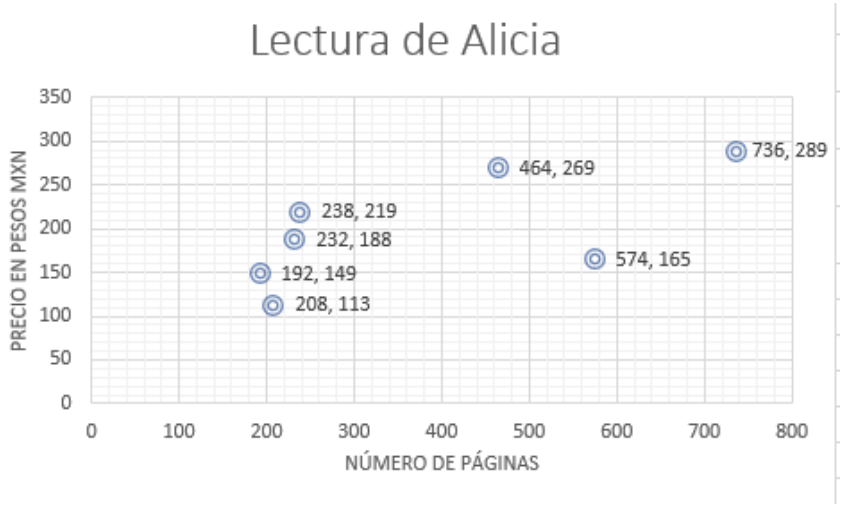


Figura 2. Diagrama de dispersión de los libros de Alicia

Como se observa en la gráfica, podría haber una relación lineal y ésta sería positiva, ya que la posible recta sería creciente (pendiente positiva).

Ejemplo 1.3 Deporte en la Escuela de Bachilleres

En el curso de educación física de la Escuela de Bachilleres UAQ, plantel San Juan del Río se tomaron varias notas. La siguiente muestra es el número de "lagartijas" y "sentadillas" hechas por 10 estudiantes seleccionados al azar:

(27, 30) (22, 26) (15, 25) (35, 42) (30, 38) (52, 40) (35, 32) (55, 54) (40, 50) (40, 43)
Realiza una tabla de contingencia y un diagrama de dispersión.

Solución:

Primero elegimos las variables

X- número de lagartijas Y – número de sentadillas.

Representando los datos en la tabla:

x	27	22	15	35	30	52	35	55	40	40
y	30	26	25	42	38	40	32	54	50	43

Tabla 3. Representación en tabla de contingencia de lagartijas y sentadillas

Observa como el formato de la tabla es diferente, y no importa, también puede ser así.

Por último, realizamos el diagrama de dispersión, por el tipo de datos conviene una escala de 10 por línea tanto en X como en Y.

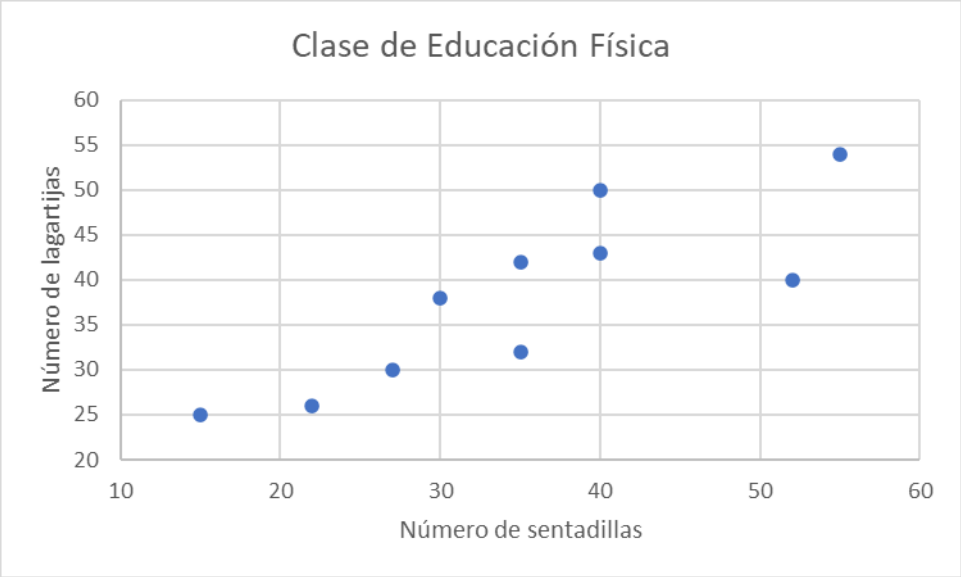


Figura 3. Diagrama de dispersión de la clase de educación Física

Observa que la tendencia de los datos es creciente y que la escala no necesariamente inicia en cero.

2. ANÁLISIS DE CORRELACIÓN

Muchas veces, los diagramas de dispersión no muestran claramente si existe una relación lineal entre dos variables, por lo que es necesario medir el grado de asociación de ellas, es decir, calcular un valor numérico que indique el tipo de relación que hay entre ellas.

Los siguientes son los tipos de asociación lineal que puede haber entre dos variables:

- Directa, si la línea recta es creciente (pendiente positiva).
- Inversa, si la línea recta es decreciente (pendiente negativa).
- Inexistente, que es, obviamente, cuando no hay relación entre las variables, si la línea recta es horizontal o vertical (pendiente igual a cero o inexistente)

2.1 Covarianza

La covarianza es una medida descriptiva que permite determinar el tipo de asociación lineal entre dos variables.

La covarianza muestral se obtiene mediante la fórmula:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Donde

$x_i = valor\ i-ésimo\ de\ la\ variable\ x$

$\bar{x} = media\ muestral\ de\ la\ variable\ x$

$y_i = valor\ i-ésimo\ de\ la\ variable\ y$

$\bar{y} = media\ muestral\ de\ la\ variable\ y$

$n = tamaño\ de\ la\ muestra.$

IMPORTANTE: También se puede calcular la covarianza poblacional σ_x, la formula varia muy poco, nosotros solo trabajaremos la muestral.

2.1.1 Interpretación de la covarianza

La interpretación de este valor es muy sencilla:

Si S_{xy}=0 no existe relación entre variables

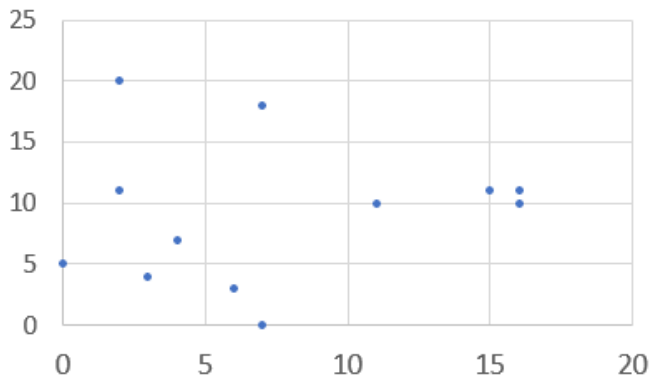


Figura 4. relación inexistente

Si S_{xy} es negativo hay una relación inversa entre las variables. Mientras X aumenta, Y disminuye.

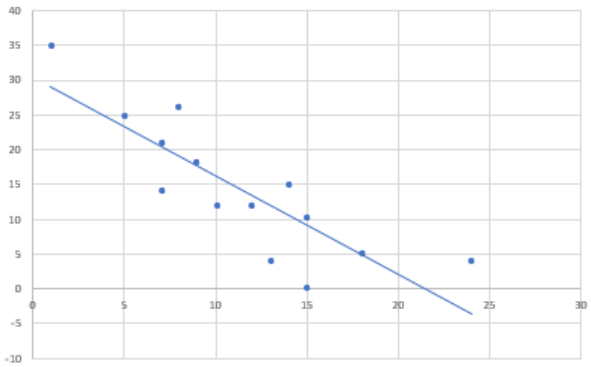


Figura 5. relación inversa

Si S_{xy} es positivo hay una relación directa entre las variables. Mientras X aumenta, Y aumenta.

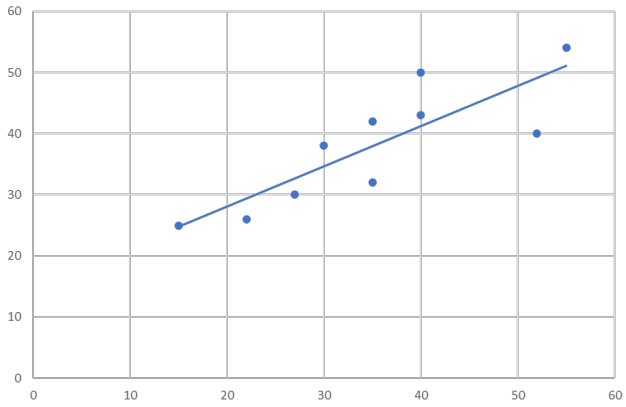


Figura 6. relación directa

Ejemplo 2.1 Cálculo de la covarianza

Una vez que Alicia dibujó el diagrama de dispersión para relacionar el número de páginas y el precio de los libros, calculó la covarianza entre estas variables para determinar numéricamente, cómo es la relación lineal, si ésta existe, entre ellas.

Solución

Para poder hacer mejor el cálculo se debe realizar una tabla que poco a poco vamos llenando con cálculos más sencillos.

Paso 1. Las primeras columnas corresponden a la tabla de contingencia.

Paso 2. Calcula el valor de la media de X y la media de Y

$$\bar{x} = \frac{464 + 736 + 232 + 238 + 574 + 192 + 208}{8} \qquad \bar{x} = 377.7$$

$$\bar{y} = \frac{269 + 289 + 188 + 219 + 165 + 149 + 113}{8} \qquad \bar{y} = 198.9$$

Paso 3. En la tercera columna calcula el valor de la resta de cada x menos la media de x.

Paso 4. En la cuarta columna calcula el valor de la resta de cada y menos la media de y.

Paso 5. En la quinta columna multiplica las columnas tres y cuatro.

Paso 6. Suma los valores de la columna cinco.

X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$
464	269	86.29	70.14	6052.327
736	289	358.29	90.14	32296.898
232	188	-145.71	-10.86	1582.041
238	219	-139.71	20.14	-2814.245
574	165	196.29	-33.86	-6645.673
192	149	-185.71	-49.86	9259.184
208	113	-169.71	-85.86	14571.184
			$\sum(X-\bar{X})(Y-\bar{Y})$	54301.7143

Tabla 4. Tabla de contingencia para calcular la covarianza

Realizando los cálculos tenemos que:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} \qquad S_{xy} = \frac{54\,301.7143}{7-1} \qquad S_{xy} = +9050.28571$$

Interpretación: existe una relación directa entre las variables, a mayor número de páginas, mayor es el costo de los libros de Alicia.

2.2 Coeficiente de correlación

Aunque la covarianza indica el tipo de relación lineal que hay entre dos variables, no podemos saber la fortaleza de esta relación. Para eso debemos calcular otro valor.

El coeficiente de correlación se utiliza para medir la magnitud de la relación lineal entre dos variables, es decir, indica cuán fuerte o débil es una relación lineal. Se denota con la letra r y también se le conoce como r de Pearson, en honor a Karl Pearson. Se calcula de esta forma:

$$r = \frac{S_{xy}}{S_x S_y}$$

Donde

$S_{xy} = covarianza\ muestral\ entre\ x, y$

$S_x = desviación\ estándar\ muestral\ de\ la\ variable\ x$

$S_{xy} = desviación\ estándar\ muestral\ de\ la\ variable\ y$

Para su interpretación el rango de valores varia entre -1 y 1. Los valores intermedios se interpretan de forma intuitiva, para hacerlo te puedes guiar del siguiente diagrama:

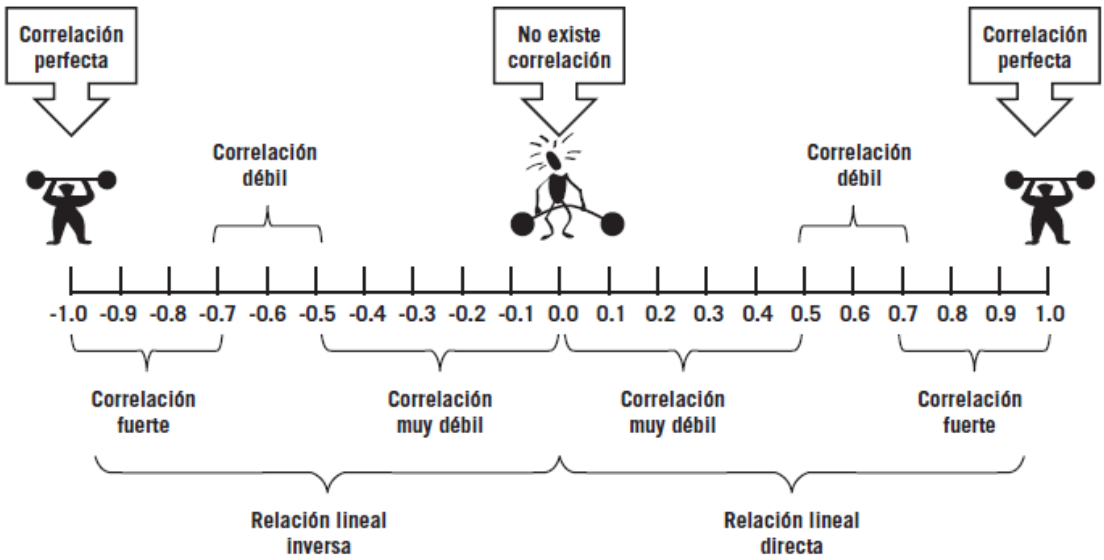


Figura 7. Interpretación del coeficiente de correlación

De tal forma que tenemos los siguientes casos:

- Correlación positiva perfecta, r=1

Figura 8. Gráfica de una correlación positiva perfecta

- Correlación negativa perfecta, r= -1

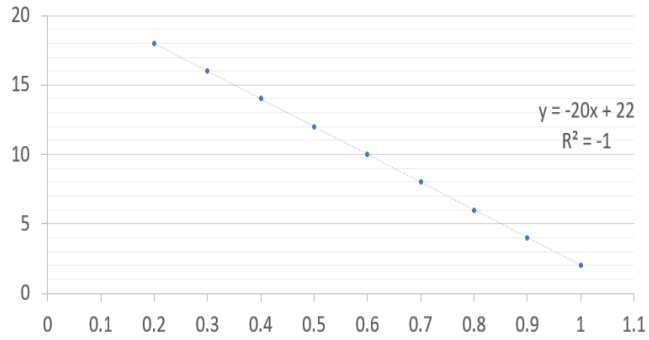


Figura 9. Gráfica de una correlación negativa perfecta

- Correlación inexistente, r=0

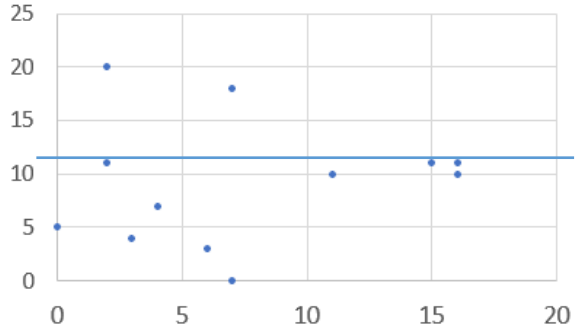


Figura 10. Gráfica de una correlación inexistente

- Correlación positiva fuerte, r= 0.8

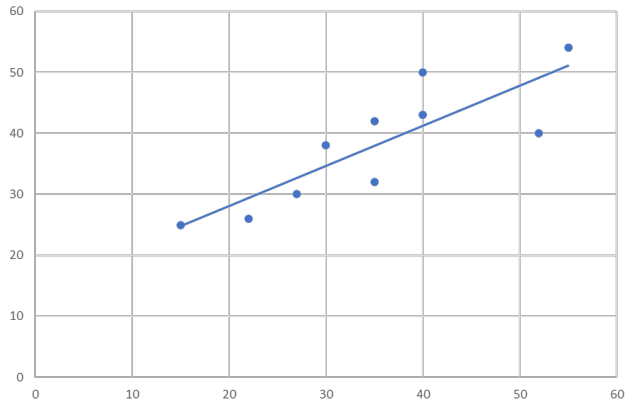


Figura 11. Gráfica de una correlación positiva fuerte

Ejemplo 2.2 Cálculo de la correlación

Ahora, Alicia calcula el coeficiente de correlación entre número de páginas y el precio de los libros para establecer cuán fuerte (o débil) es la relación lineal entre las variables: Completar la tabla para calcular S_x y S_y. Recuerda la fórmula de desviación estándar:

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Paso 1. Calcular para cada x, $(x_i - \bar{x})^2$ en la columna seis.

Paso 2. Calcular para cada y, $(y_i - \bar{y})^2$ en la columna siete.

X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
464	269	86.29	70.14	6052.327	7445.22	4920.02
736	289	358.29	90.14	32296.898	128368.65	8125.73
232	188	-145.71	-10.86	1582.041	21232.65	117.88
238	219	-139.71	20.14	-2814.245	19520.08	405.73
574	165	196.29	-33.86	-6645.673	38528.08	1146.31
192	149	-185.71	-49.86	9259.184	34489.80	2485.73
208	113	-169.71	-85.86	14571.184	28802.94	7371.45
			$\sum(X-\bar{X})(Y-\bar{Y})$	54301.7143	278387.43	24572.86

Tabla 5. Tabla de contingencia para calcular el coeficiente de correlación

Así tenemos que

$$S_x = \sqrt{\frac{278387.43}{7-1}} \qquad S_x = \sqrt{46397.90}$$

$$S_x = 215.4$$

$$S_y = \sqrt{\frac{24572.86}{7-1}} \qquad S_y = \sqrt{4095.47}$$

$$S_y = 63.99$$

Y del ejemplo 2.1

$$S_{xy} = 9050.28571$$

Por lo tanto

$$r = \frac{9050.28571}{(215.4)(63.99)}$$

$$r = 0.6565$$

Interpretación: El coeficiente de correlación es de 0.65 por lo que tenemos una correlación entre la cantidad de páginas y el costo de un libro directa débil. Es decir, a mayor número de páginas mayor costo.

2.3 Coeficiente de determinación

Cuando lo que interesa es analizar una relación de causalidad entre dos variables, primero debemos definir cuál de ellas es la variable Y, variable dependiente, y cuál es la variable X, variable independiente. La variable dependiente Y es la que se busca explicar; en términos estadísticos, es la que se busca estimar o pronosticar. A su vez, la variable independiente X es la que brinda información para explicar Y y recibe el nombre de variable de predicción.

Para saber si una variable X es "buena" para explicar la variable Y se calcula el **coeficiente de determinación,** que representaremos con r^2 y que tiene las características siguientes:

- Es el cuadrado del coeficiente de correlación.
- Su rango de valores está entre 0 a 1.
- No da ninguna información sobre la dirección de la relación entre las variables.
- Se puede expresar en porcentaje.

Cuanto más cerca esté de 1, la variable independiente X será una buena variable para explicar Y. Es decir, es un factor determinante para Y. En contraparte, conforme r^2 se acerca a 0, indica que X no es un factor significativo para explicar Y.

Ejemplo 2.3 Cálculo del coeficiente de determinación

Por último, Alicia desea calcular el coeficiente de determinación para sus variables.

Solución

$$r^2 = 0.6565^2$$

$$r^2 = 0.43$$

$$r^2 = 43\%$$

Interpretación. Aunque el número de páginas de un libro puede ser uno de los factores que influyan en su costo, es lógico pensar que hay otros factores como el autor, el tipo de libro, la demanda, etc. El coeficiente de determinación indica que la variable "Núm. de páginas" explica el 43% de la variabilidad del "Precio del libro". En otras palabras, para poder explicar completamente el precio de un libro debemos considerar otros factores, no sólo el número de páginas.

Actividad 1.1

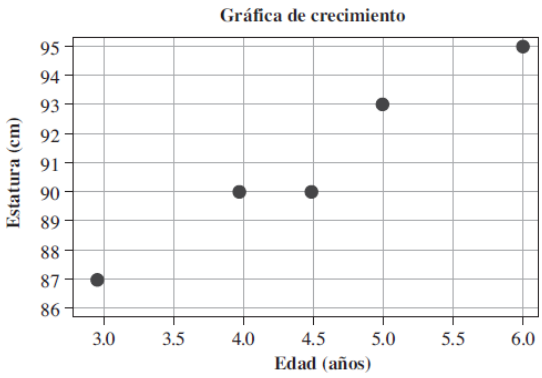
Tiempo estimado: 60 minutos

En cada uno de los ejercicios contesta lo que se te pide de forma limpia y ordenada. Realiza la gráficas, los procedimiento y los cálculos en hojas aparte.

1. ¿Da resultados estudiar para un examen?
   a. Trace un diagrama de dispersión del número de horas estudiadas, x, comparado con la calificación de examen recibida, y.

x	2	5	1	4	2
y	80	80	70	90	60

   b. Explique lo que pueda concluir con base en el patrón de datos que se muestran en el diagrama de dispersión.
2. Por lo general, los pediatras usan gráficas de crecimiento para observar el crecimiento de un niño. Considere la gráfica de crecimiento que sigue:



   a. ¿Cuáles son las dos variables mostradas en la gráfica?
   b. ¿Qué información representa el par ordenado (3,87)?
   c. Describa la forma en que el pediatra podría usar esta gráfica y qué tipos de conclusiones podrían basarse en la información mostrada por la gráfica.
3. Del "Ejemplo 1.3 Deporte en la Escuela de Bachilleres" calcula la covarianza, el coeficiente de correlación y el coeficiente de determinación. Interprétalos

¡Llegamos al final de la lección!

¡Lo has hecho bien!

Como podrás darte cuenta la estadística puede ayudar mucho cuando se busca analizar modelos con matemáticas, y además las cosas pueden ir mucho más complejas, pero para este curso, lo que veremos a continuación está bien para tu nivel de aprendizaje. Ojalá algún día puedas estudiar estadística con métodos más complejos y te acuerdes de la simpleza de este tema.

Ahora estás list@ para contestar el Quiz.

Si quieres saber más, y entender mejor, puedes ver este video:

Psico Facil. (12 de Junio de 2019). *CORRELACION PEARSON Y SPEARMAN FACIL + Tutorial SPSS*. [Obtenido de video] Recuperado de https://www.youtube.com/watch?v=AWsVevNvURw

Píldoras matemáticas. (17 de Enero de 2019). 05 Covarianza-significado. [Obtenido de video] Recuperado de https://www.youtube.com/watch?v=XW-yuLXX4PY&t=271s

Bibliografía

Banegas, A. L. (2012). *Probabilidad y estadística. Enfoque por competencias.* México: INTERAMERICANA EDITORES S.A. DE C.V.

Robert Johson, P. K. (2008). *Estadítica elemental: Lo esencial.* México: Cengage Learning Editores, S.A. de C.V.