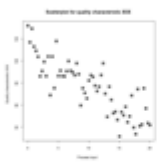
Probabilidad y Estadística

12

Actividades de Aprendizaje



Conceptos y definiciones de esta clase:

- Regresión y Correlación de variables
- Regresión Lineal
- El diagrama de Dispersión
- El Método de Mínimos Cuadrados
- Error de la estimación

Regresión y Correlación de variables

El análisis de correlación de variables nos permite analizar en forma conjunta dos o más variables, para luego inferir resultados sobre una de ellas a partir de la otra (y otras). Esto es algo con lo que estamos muy familiarizados cuando leemos en los titulares de los periódicos algo como sigue:

"*La ingesta de 2 g de canela durante 12 semanas reduce significativamente la HbA1c, SBP y DBP, entre los pacientes con diabetes de tipo 2"*

El titular anterior nos está indicando que hay una correlación de algún tipo entre:

- la cantidad de canela consumida
- el tiempo durante el cual se mantiene la ingesta
- los niveles de la hemoglobina glicosilada (HbA1) y otros parámetros de salud

Este tipo de relaciones entre parámetros disímiles son naturales en nosotros, y más habituales de lo que podemos pensar en primera instancia. Analicemos, a modo de ejemplo, alguna situación en la que uno se ha visto desfavorecido, por ejemplo, acabamos de salir de un examen y nos encontramos con la noticia de que nos ha ido mal en el mismo. Inmediatamente comienza en nosotros un proceso natural de, en primer lugar, análisis de las acciones llevadas a cabo, horas de estudio, notas anteriores, fuentes consultadas, etc. Al mismo tiempo, comenzamos a relevar a nuestro alrededor a quienes consideramos nuestros pares frente a la misma situación (o sea, nuestros compañeros de examen), y los indagamos sobre los mismos parámetros, con preguntas al estilo:

- ¿vos cuánto estudiaste?
- ¿hace mucho que venís preparando la materia?
- ¿estudiaste sólo?
- ¿te preparaste con alguien?
- ¿hiciste todos los ejercicios?

En resumen, estamos intentando encontrar una **correlación** entre lo que consideramos parámetros habituales y pertinentes con respecto a la consigna (rendir correctamente el examen) y su cuantificación (horas de estudio, cantidad de ejercicios, etc.), para finalmente intentar establecer, aún sin un método

preestablecido, la relación entre todo lo anteriormente mencionado y el resultado final. De esta manera, nuestras conclusiones suelen ser del siguiente tipo:

- Este examen hay que comenzar a prepararlo por lo menos tres semanas antes
- Tengo que estudiar, como mínimo, cuatro horas por día
- Hay que hacer casi todos los ejercicios del práctico para pasar el escrito

Es así como, de manera natural, intentamos encontrar respuestas a nuestras acciones y sus resultados de manera precisa y predictible. No nos sorprende pensar, entonces, que, a mayor experiencia, se corresponderán mejores y más precisos pronósticos. Esto es materia recurrente en muchas profesiones en las cuales, aún sin demasiados conocimientos matemáticos por parte del agente estimador, la confección de presupuestos son un acto diario. Pensemos en un pintor, un albañil, un carpintero, un tapicero, y cualquier trabajador que deba realizar tanto presupuestos de costos como de tiempos.

Y como ejemplo final, pensemos en las relaciones establecidas con las que contamos a diario para administrar nuestras finanzas hogareñas, como por ejemplo la relación entre los viajes que realizo y el gasto mensual en combustibles.
Llegamos a una primera conclusión entonces, que es que no nos es difícil encontrar relaciones entre parámetros, sino que la dificultad radica más bien en establecer cuál es el tipo de relación.

1. Regresión Lineal

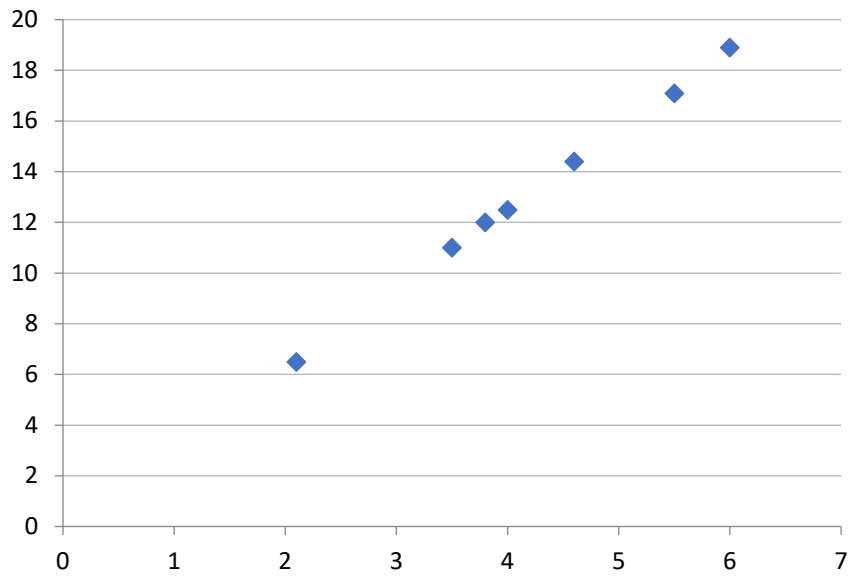
1.1 El diagrama de Dispersión

Vamos a comenzar realizando un análisis de correlación entre variables. Una manera de hacerlo es mediante un sencillo diagrama de dispersión, en el cual volcaremos los datos previamente reunidos mediante cálculos, encuestas, o bien en forma experimental.

Para este ejemplo, tomaremos las medidas de diversas circunferencias y sus diámetros, obtenidas mediante observación directa y posteriormente incluidas en la siguiente tabla:

Nro. de observación (n)	Diámetro en cm (x)	Longitud en cm (y)
1	2.10	6.50
2	5.50	17.10
3	4.00	12.50
4	3.80	12.00
5	6.00	18.90
6	3.50	11.00
7	4.60	14.40

A continuación, graficamos los pares de datos en un sistema de ejes, haciendo x al diámetro e y a la longitud de la circunferencia, ambas expresadas en centímetros.



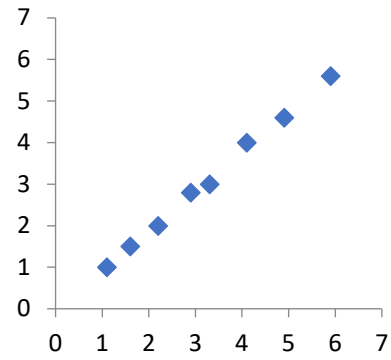
Importante:

La decisión sobre cuál de las variables es x y cuál es y se basa, fundamentalmente, en determinar qué variable va a depender de la otra, o sea qué variable utilizaremos para obtener datos sobre la otra.

En estadística, y (la variable dependiente) es también llamada ***variable de respuesta***, y x (la variable independiente) es llamada ***variable predictora o variable explicativa***.

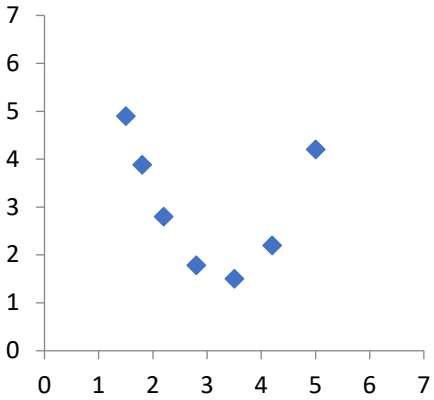
Veamos ahora que realizando graficas que denominaremos diagramas de dispersión nos brindaran de inmediato, información sobre si existe relación entre ambas variables y de qué tipo es ésta.

Veamos algunos ejemplos:



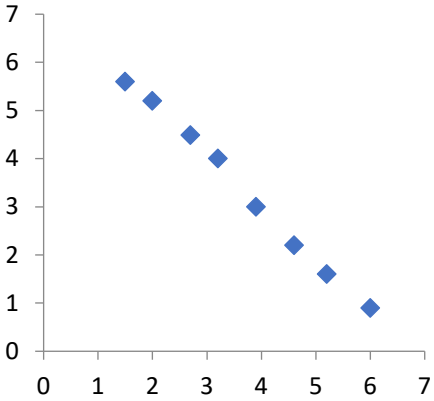
Existe correlación lineal positiva

$r \cong 1$



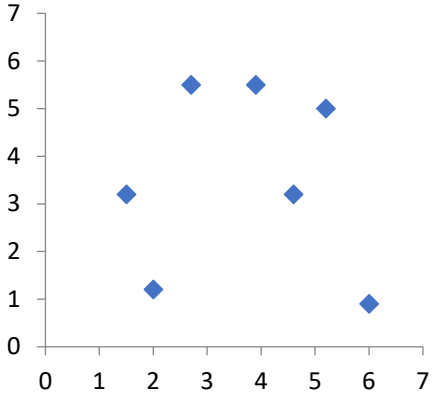
Existe correlación no lineal

$r = 0$



Existe correlación lineal negativa

$r \cong -1$



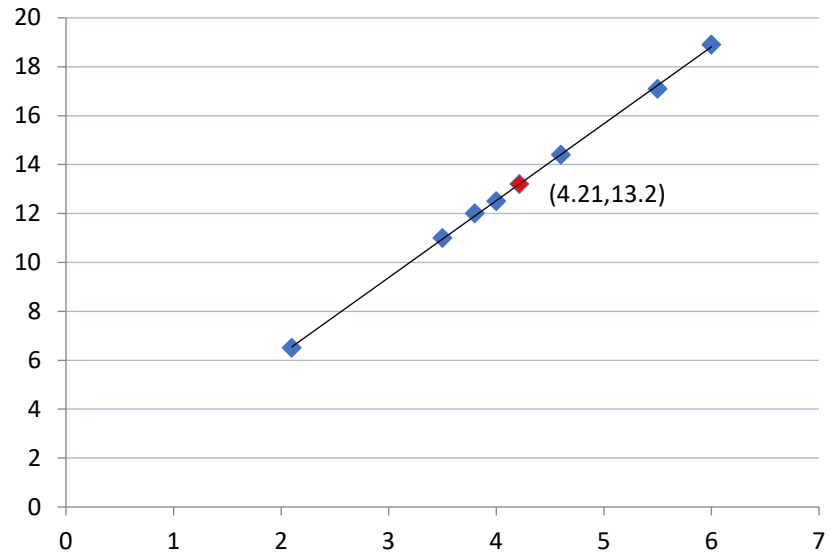
No existe correlación entre los datos

$r = 0$

El valor r establecido para cada diagrama es el valor de correlación, que se encuentra siempre entre los valores -1 a +1. Valores cercanos a estos extremos indican alta correlación entre las variables analizadas. Valores cercanos al 0 indican baja o nula correlación.

1.2 La Recta de Regresión

Cuando la relación entre las dos variables queda representada por una línea recta, se establece la denominada recta de regresión, que es aquella que se ajustan de la mejor manera a los puntos del diagrama de dispersión. Para nuestro caso, la recta de regresión quedaría graficada de la siguiente manera:



En este diagrama, el punto rojo identificado con sus coordenadas establece el centro de gravedad de la dispersión, que se obtiene hallando las medias aritméticas de las variables.

Centro de gravedad: $(\bar{x}, \bar{y})$

Calculamos para nuestro caso:

$$\bar{x} = \frac{\sum_1^7 x_i}{7} = \frac{2.1 + 5.5 + 4.0 + 3.8 + 6.0 + 3.5 + 4.6}{7} = \frac{29.5}{7} \cong 4.21$$

$$\bar{y} = \frac{\sum_1^7 y_i}{7} = \frac{6.5 + 17.1 + 12.5 + 12.0 + 18.9 + 11.0 + 14.4}{7} = \frac{92.4}{7} = 13.2$$

Con lo cual, el punto buscado es $(4.21,13.2)$

Nota:

La recta de regresión siempre pasa por el centro de gravedad $(\bar{x}, \bar{y})$.

1.3 El Método de Mínimos Cuadrados

Este método nos permite establecer una manera de encontrar la ecuación de la recta de regresión establecida en el punto anterior, de tal manera que la misma presente el mejor ajuste para todos los puntos del diagrama de dispersión y a la vez minimice el error. Podríamos simplemente buscar aquella recta en la que los errores sean los mínimos posibles, pero este procedimiento tiene el defecto de que un gran error por exceso se equilibraría con muchos pequeños errores por defecto, lo cual nos daría una recta que no sería la adecuada. De la misma manera, si tratáramos con los errores absolutos, impidiendo de esta manera que se anularan entre sí, estaríamos encontrando una recta más adecuada, pero que aún no haría foco en la magnitud de los errores cometidos.

¿Cuál es, entonces, una manera apropiada de encontrar los puntos de la recta?

Bueno, una posible solución consiste en elevar al cuadrado los desvíos de las ordenadas de los puntos a considerar en nuestra recta con respecto a los puntos de las mediciones. De esta manera lograremos dos objetivos:

1. Amplificar la magnitud de los errores cometidos.
2. Hacer que todos los errores den positivo.

Para ello, primero diferenciaremos las ordenadas de los puntos originales de las de los puntos de la recta con la siguiente notación:

y: son las ordenadas de los puntos originales
$\hat{y}$: son las ordenadas de los puntos de la recta

De esta manera, la obtención de los desvíos, para cada par de puntos, se obtendrá calculando la diferencia $y - \hat{y}$ y elevándola al cuadrado.

Entonces, hasta el momento tenemos:

- Un conjunto de puntos (x, y) correspondientes a las muestras.
- Un diagrama de dispersión que se asemeja a una línea recta.
- Una estrategia para obtener los puntos de una recta que se ajuste de la mejor manera a los puntos del diagrama.

Como lo que queremos es minimizar, entonces lo que debemos hacer es encontrar una fórmula que vincule los datos con la pendiente y la ordenada al origen de la recta deseada, de tal manera que podamos escribir su ecuación. Esto se obtiene derivando la fórmula de los errores cuadráticos para posteriormente despejar los elementos de la recta mencionada. No vamos a realizar todo este desarrollo en este apunte y nos limitaremos a escribir las fórmulas de dichos elementos. Así pues, tenemos:

- La ecuación de la recta de estimación de mejor ajuste

$$y = a + bx$$

Siendo b la pendiente, que se obtiene con la fórmula:

$$b = \frac{\sum_1^n x_i y_i - n\bar{x}\bar{y}}{\sum_1^n x_i^2 - n\bar{x}^2}$$

y "a" es la ordenada al origen, cuya fórmula es:

$$a = \bar{y} - b\bar{x}$$

Es importante aclarar que las fórmulas que anteceden pueden encontrarse escritas de diversas formas en los textos sobre el tema, pero con todas ellas se obtiene el mismo resultado.
A continuación, vamos a tomar los datos de nuestro ejemplo (Tabla 6.1) y calcularemos la pendiente y ordenada al origen con las fórmulas vistas. Para ello, ampliaremos la tabla agregando las columnas que necesitamos para nuestros cálculos y las sumas de cada columna.

Nro de observación (n)	Diámetro en cm (x)	Longitud en cm (y)	$x.y$	x^2
1	2.10	6.50	13.65	4.41
2	5.50	17.10	94.50	30.25
3	4.00	12.50	50.00	16.00
4	3.80	12.00	45.60	14.44
5	6.00	18.90	113.40	36.00
6	3.50	11.00	38.50	12.25
7	4.60	14.40	66.24	21.16
Totales	29.50	92.40	421.44	134.51

Las medias de cada variable ya las habíamos calculado, siendo:

$$\bar{x} \cong 4.21$$

$$\bar{y} = 13.2$$

Calculemos ahora los parámetros de la recta:

$$b = \frac{\sum_1^n x_i y_i - n\bar{x}\bar{y}}{\sum_1^n x_i^2 - n\bar{x}^2} = \frac{421.44 - 7.(4.21).(13.2)}{134.51 - 7.(4.21)^2} = \frac{32.04}{10.19} \cong 3.1446999$$

$$a = \bar{y} - b\bar{x} = 13.2 - (3.144).(4.21) \cong -0.052664$$

Por último, mostramos la ecuación de la recta de regresión obtenida por este método:

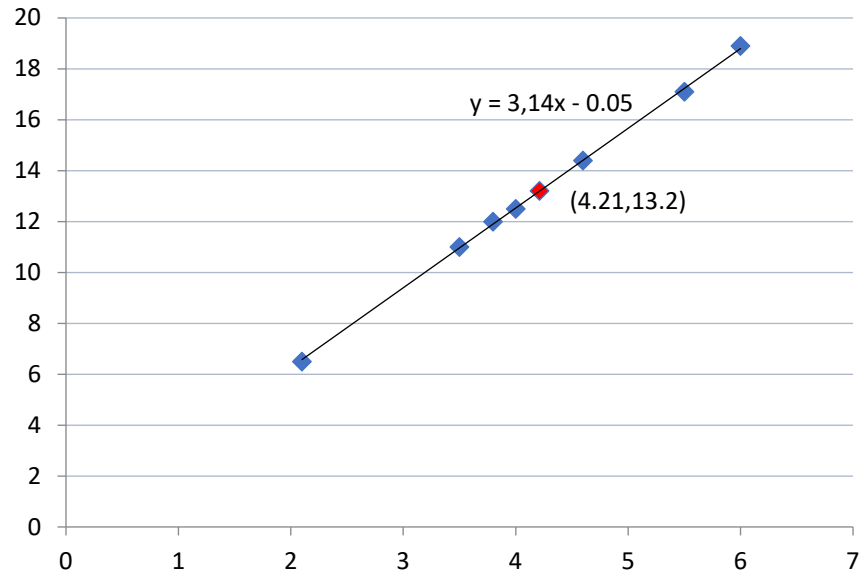
$$y = a + bx = -0.05 + 3.144\,x$$

No es de extrañar que tenga pendiente 3.144 puesto que es la relación existente entre el diámetro de una circunferencia y su longitud.

Nota:

La diferencia entre los valores mostrados en esta página y los que surgen de los mismos cálculos obtenidos con una calculadora se debe a que se han tomado muchos más decimales que los dos mostrados en este texto

Incorporaremos la recta obtenida a nuestra gráfica:



1.4 Error de la estimación

Para continuar nuestro análisis de la regresión, calcularemos a continuación qué tan confiable es la ecuación hallada, lo cual haremos mediante el error estándar de la estimación, que puede calcularse con la siguiente fórmula:

$$S_e = \sqrt{\frac{\sum_1^n (\hat{y}_i - y_i)^2}{n-2}}$$

Ahora bien, el uso de la anterior fórmula implica agregar nuevas columnas con los valores obtenidos a través de la fórmula de la recta de estimación, para luego realizar los cálculos. Existe también la posibilidad de utilizar un método abreviado que simplifique nuestros cálculos, a saber:

$$S_e = \sqrt{\frac{\sum_1^n {y_i}^2 - a\sum_1^n y_i - b\sum_1^n x_i y_i}{n-2}}$$

De esta manera, evitamos tener que incurrir en numerosos cálculos, puesto que ya contamos en nuestra tabla con los utilizados por esta fórmula. De todas maneras, existe la desventaja de que estaremos utilizando datos obtenidos en el paso anterior, que, en el caso de estar equivocados, nos harán incurrir en nuevos errores.

Calculemos para nuestro ejemplo de las circunferencias, ampliando una vez más la tabla para incorporar el cálculo de y^2:

Nro. de observación (n)	Diámetro en cm (x)	Longitud en cm (y)	$x.y$	x^2	y^2
1	2.10	6.50	13.65	4.41	42.25
2	5.50	17.10	94.50	30.25	292.41
3	4.00	12.50	50.00	16.00	156.25
4	3.80	12.00	45.60	14.44	144.00
5	6.00	18.90	113.40	36.00	357.21
6	3.50	11.00	38.50	12.25	121.00
7	4.60	14.40	66.24	21.16	207.36
Totales	29.50	92.40	421.44	134.51	1320.48

$$S_e = \sqrt{\frac{\sum_1^n {y_i}^2 - a\sum_1^n y_i - b\sum_1^n x_i y_i}{n-2}} =$$

$$= \sqrt{\frac{1320.48 - (-0.05)(92.40) - (3.14)(421.44)}{7-2}} =$$

$$= \sqrt{\frac{0.0438}{5}} = \sqrt{0.0876} = 0.093$$

Insistimos una vez más en utilizar una buena cantidad de decimales para estos cálculos.

Para el estudiante:

Se propone, antes de seguir, que vuelque los datos del ejercicio en una hoja de cálculo, y arme las columnas necesarias para calcular el error mediante la otra fórmula y comprobar que se obtiene idéntico resultado