INSTITUTO de ENSEÑANZAS a DISTANCIA de ANDALUCÍA

PAU

Mayores de 25 años

Contenidos

Matemáticas Aplicadas a las Ciencias Sociales
Sucesiones y estadística: Estadística Bidimensional

1. ¿Qué es la Estadística Bidimensional?

Cuando realizamos un estudio estadístico de una población o de una muestra de ella, podemos ceñirnos a observar un solo carácter de cada individuo (como hacíamos en el tema anterior), o bien de cada uno de los individuos nos interesen dos o más caracteres, dando lugar a las **variables estadísticas bidimensionales** o multidimensionales.

Por ejemplo, podemos estudiar en una población las variables: "Número de horas durmiendo" y "Nº de horas viendo la televisión", pero también la relación y la dependencia existente entre ellas.

Composición de elaboración propia
a partir de imágenes de Dominio Público

En este tema estudiaremos precisamente eso, la relación existente entre dos variables estadísticas. Esto nos permitirá hacer predicciones sobre futuros comportamientos en función de la relación existente entre ellas.

Importante

Una **Variable Estadística Bidimensional (X,Y)** es el resultado del estudio de dos caracteres X e Y en los elementos de una población.

Para cada elemento de estudio obtenemos un par de valores que notaremos (x_i,y_i), donde x_i es el valor para el factor X, e y_i para el factor Y.

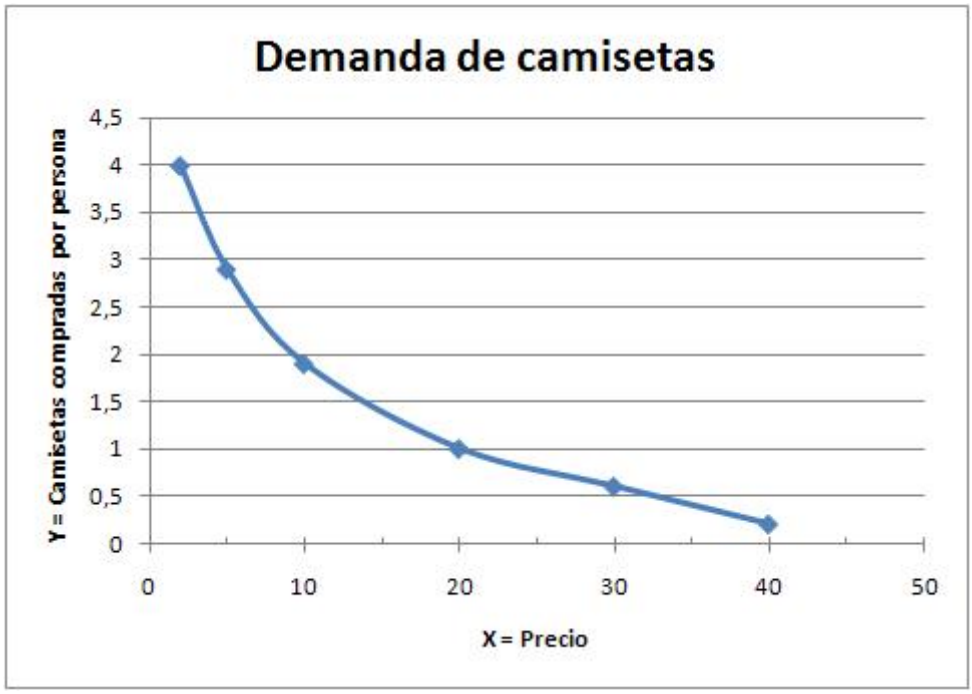
Comprueba lo aprendido

Al entrar en una tienda de ropa, ves en un cartel que venden una camiseta que te gusta por 20 €. A ese precio, puede que te la acabes comprando. Si costara solo 10 €, seguramente te comprarías esa y puede que otro modelo del mismo precio. ¿Y si las vendieran por 3 €? Es probable que compres más de dos... por si acaso.

Esta es básicamente la Ley de la Demanda: la modificación del precio causa variaciones en la demanda de un artículo. Como tenemos dos factores X = precio del artículo, Y = número de artículos demandados, podemos establecer una variable estadística bidimensional

bidimensional.

Cuando tenemos suficientes datos, podemos hacer una primera estimación de la dependencia entre ambas variables. En la siguiente gráfica hemos representado los datos obtenidos al estudiar esta variable bidimensional.



a) Cuando aumenta el precio de la camiseta, ¿qué ocurre con la demanda?

- Disminuye.
- Aumenta.
- No varía.

¡CORRECTO!

Recuerda que el precio está en el eje horizontal y la demanda en el vertical.

Recuerda que el precio está en el eje horizontal y la demanda en el vertical.

Solution

1. Opción correcta
2. Incorrecto
3. Incorrecto

b) Veamos si lo hace en la misma proporción. Por ejemplo, si aumento el precio el doble, ¿la demanda cae a la mitad?

Sugerencia

- Sí.
- No.

Si vendemos a 20 € nos comprarían 1 camiseta, pero por 40 € no se vendería ni media.

¡MUY BIEN! En este caso podríamos decir que hay una **dependencia negativa** (porque al aumentar una variable, la otra disminuye) y **aleatoria** (porque la dependencia no es exacta).

Solution

1. Incorrecto
2. Opción correcta

2. Opción correcta

2. Organización y representación de datos

Como has visto en el apartado anterior, tan solo con unos cuantos datos ya se pueden establecer relaciones entre dos variables. Pero lo normal para obtener resultados fiables es contar con una gran cantidad de datos estadísticos. En estos casos no es cómodo hacer una **tabla simple** como las que hemos utilizado en los últimos ejemplos, en los que solo había seis o siete datos. Vamos a ver cómo organizar la información en una **tabla de doble entrada** cuando tenemos muchos pares de datos.

Por cierto, aunque en principio por tabla de doble entrada no te venga nada a la mente, si te paras un poquito a pensar en la siguiente imagen, verás como llevas años utilizándolas. Observa como interaccionan las filas con las columnas, los elementos centrales van surgiendo de combinar las colores con las figuras, obteniendo de esta forma figuras coloreadas:

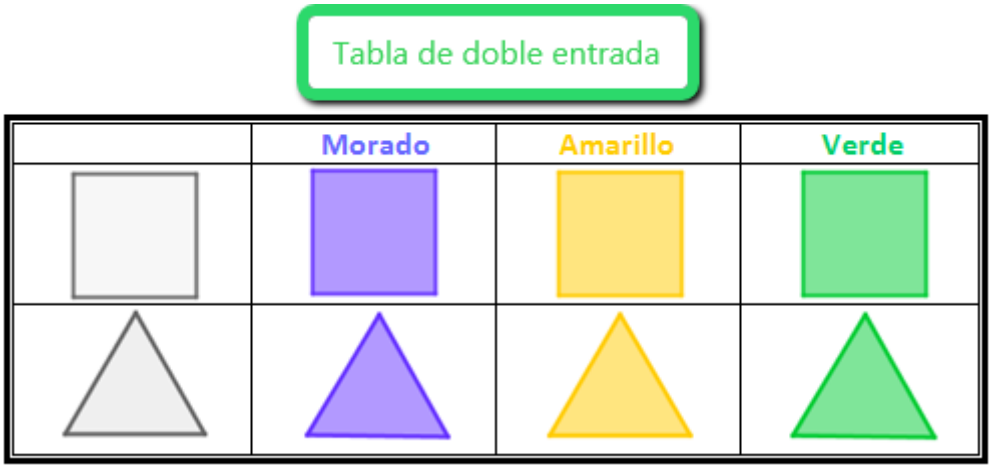


Imagen de elaboración propia

Además, en este apartado verás cómo podemos resumir y representar la información que nos proporcionan estas tablas, en unos ejes de coordenados. Esta representación llamada nube de puntos, nos dará una idea de la posible dependencia que puede existir entre las dos variables que estamos estudiando.

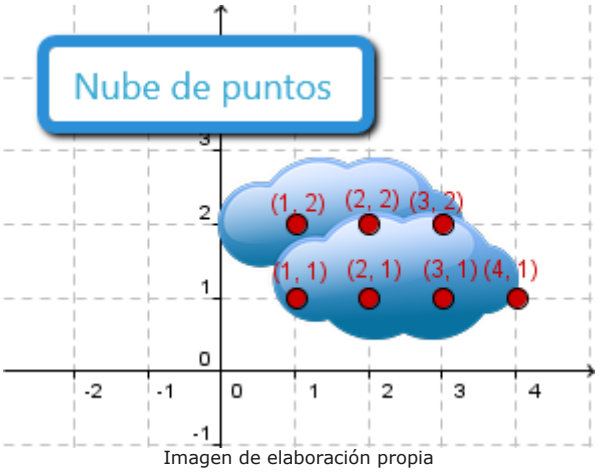


Imagen de elaboración propia

Tablas de doble entrada

Este tipo de tablas, también llamadas tablas de contingencia, brindan información estadística de dos variables relacionadas entre sí, independientemente de si son cualitativas o cuantitativas.

Son útiles en casos en los cuales un experimento es dependiente del otro. Si haces memoria y recuerdas el ejemplo de las variables, X="nº de horas que dedicamos a dormir" e Y="nº de horas que dedicamos a ver la televisión", ¿sería conveniente colocarlas en una tabla de doble entrada?, ¿podríamos detectar una relación entre ambas? La experiencia nos dice que sí.

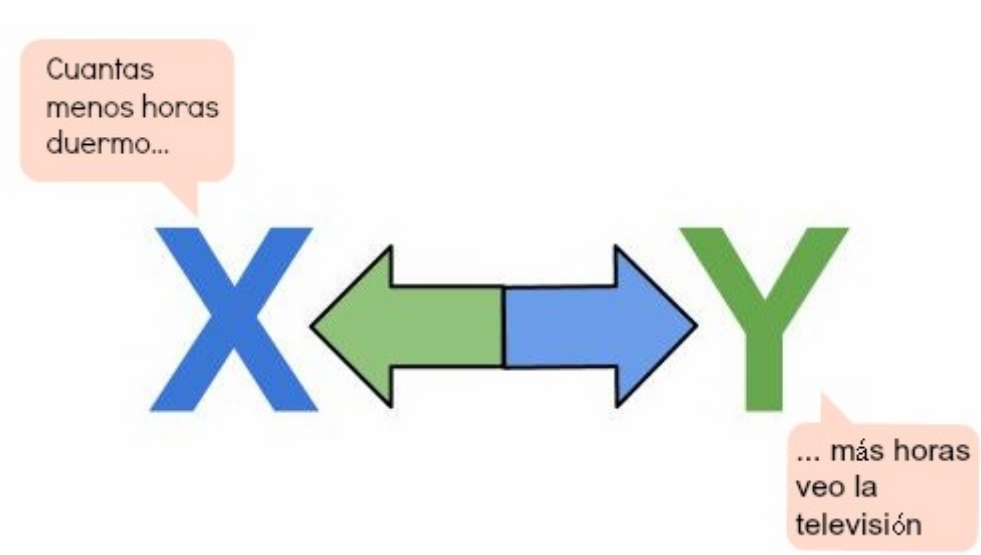


Imagen de elaboración propia

En la siguiente presentación verás con un ejemplo cómo se crean estas tablas:

Presentación en Slideshare de **Patricia_Perez basada en otra de Saúl Valverde**

Recuerda que las tablas de doble entrada son útiles en casos en los que tenemos gran cantidad de datos o en los que los pares de datos pueden aparecer repetidos. En caso contrario, hacemos uso de una tabla simple.

Mira el siguiente ejemplo:

Estamos estudiando los datos de un grupo de alumnos en las asignaturas de Matemáticas (X) y Geografía (Y) y pretendemos recogerlos en una tabla. Los resultados han sido los siguientes: (6,7), (7,8), (9,8), (6,6), (4,3), (7,7), (3,4), (4,5) y (5,5), donde cada par corresponden a las notas de Matemáticas y Geografía de cada alumno, la forma más sencilla de ordenarlos sería:

Matemáticas (X)	**6**	**7**	**9**	**6**	**4**	**7**	**3**	**4**	**5**
Geografía (Y)	7	8	8	6	3	7	4	5	5

Ejercicio resuelto

Vamos a trabajar con la tabla de la presentación anterior para sacar algunas conclusiones. Recuerda que X="número de días por mes en los que se supera el límite permitido de concentración de NO_2", e Y="número de días por mes en los que se supera el límite permitido de concentración de ozono".

		Y					
		$y_1=0$	$y_2=1$	$y_3=2$	$y_4=3$	$y_5=4$	n_i
X	$x_1=0$	7	1	2	2	0	12
	$x_2=1$	4	3	1	1	5	14
	$x_3=2$	3	0	0	2	0	5
	$x_4=3$	0	2	2	1	0	5
	n_j	14	6	5	6	5	36

a) ¿Cuántos meses tuvieron 2 días con niveles excesivos de NO_2, pero ninguno con nivel excesivo de ozono?

Mostrar retroalimentación

Respuesta: 3 meses. Tenemos que fijarnos en la fila de $x_3=2$ y en la columna de $y_1=0$.

b) ¿Cuántos meses tuvieron solo un día de exceso de concentración de NO_2 en aire?

Mostrar retroalimentación

Respuesta: 14 meses. Como sólo nos piden información de X, tendremos que sumar todas las casillas que corresponden a $x_2=1$, que coincide con la suma parcial que tenemos en la última columna.

Cuando tenemos los datos de una variable **agrupados por intervalos**, las frecuencias corresponden al número de observaciones que hay en cada intervalo.

Comprueba lo aprendido

En una de las estaciones meteorológicas del Alto Guadalquivir se han recogido medidas de temperatura media (ºC) y precipitaciones medias (l/m^2) cada mes. Los datos de los años 2007 y 2008 son los siguientes:

(7,5 ; 7,7)	(10,2 ; 56,5)	(11 ; 28,6)	(12,8 ; 93,5)	(17,6 ; 76,5)	(22,5 ; 5)
(27,5 ; 0,2)	(26,3 ; 1,9)	(22,2 ; 44,4)	(16,7 ; 29,9)	(10,2 ; 61,3)	(7,5 ; 6,1)
(8,48 ; 62,51)	(10,89 ; 38,34)	(11,72 ; 16,46)	(14,71 ; 122,7)	(16,55 ; 64,66)	(23,87 ; 5,38)
(26,86 ; 10,54)	(27,16 ; 0,14)	(20,63 ; 48,58)	(16,18 ; 58,55)	(8,57 ; 67,71)	(6,46 ; 48,56)

El primer par significa que en Enero de 2007 la media de temperatura fue de 7,5 ºC y la media de precipitaciones fue de 7,7 l/m^2.

Con estos datos, completa la tabla de doble entrada en la que las variables son X = "Temperatura media mensual" e Y = "Precipitaciones medias mensuales". Fíjate que en este caso las variables se han agrupado por intervalos. En la primera casilla tendrás que contar el número de meses en los que la temperatura media está entre 0 y 10 grados, y las precipitaciones entre 0 y 30 l/m^2, que son los pares (7,5 ; 7,7) y (7,5 ; 6,1), por lo que en esa casilla pondremos un 2.

		Y					
		[0-30)	[30-60)	[60-90)	[90-120)	[120-150]	n_i
X	[0-10)	2	1	2	0	0	5
	[10-20)						
	[20-30]						
	n_j						24

Enviar

Distribuciones marginales

Al analizar una distribución bidimensional, uno puede centrar su estudio en el comportamiento de una de las variables, con independencia de cómo se comporta la otra. Estaríamos así en el análisis de una distribución marginal.

Importante

En una variable bidimensional (X,Y), cada una de las variables por separado (X) e (Y) constituyen variables unidimensionales estadísticas. A estas variables se les conoce como marginales

como marginales.

Las distribuciones marginales de las variables estadísticas X e Y se obtienen a partir de la tabla de doble entrada considerando una sola variable.

Para la distribución marginal de X tomamos la primera y última columnas de la tabla de doble entrada.

Para la distribución marginal de Y tomamos la primera y última filas de la tabla de doble entrada.

De esta forma, podemos calcular sus medias y sus desviaciones típicas, como hacíamos en el tema anterior, utilizando la siguiente notación:

	Distribución marginal X	Distribución marginal Y
Media	$\bar{x} = \sum_{i=1}^{n} \frac{x_i \cdot n_i}{N}$	$\bar{y} = \sum_{j=1}^{m} \frac{y_j \cdot n_j}{N}$
Varianza	$\sigma_x^2 = \sum_{i=1}^{n} \frac{x_i^2 \cdot n_i}{N} - \bar{x}^2$	$\sigma_y^2 = \sum_{j=1}^{m} \frac{y_j^2 \cdot n_j}{N} - \bar{y}^2$

Imagen de elaboración propia

Representación de datos

Fotografía en Flickr de Ian Sane bajo CC

Si te fijas en los pasos que hemos dado hasta este momento, verás que es un proceso muy lógico:

1. Nos planteamos una pregunta sobre la relación entre dos parámetros.

2. Tomamos suficientes datos de ambos parámetros sobre la población que nos interesa.

3. Organizamos estos datos en una tabla simple o de doble entrada.

El siguiente paso será visualizar estos datos en una **gráfica**, de modo que nos resulte más fácil dar respuesta a nuestra pregunta inicial.

CASO I: Datos ordenados en una tabla de doble entrada

Como este caso no lo vamos a trabajar, te recomendamos que visites el apartado de Curiosidades en el apéndice si quieres saber algo más sobre esta cuestión. Eso sí, te adelantamos que se hablarán de Histogramas tridimensionales y Diagramas de dispersión o de burbujas.

CASO II: Datos recogidos en una tabla simple

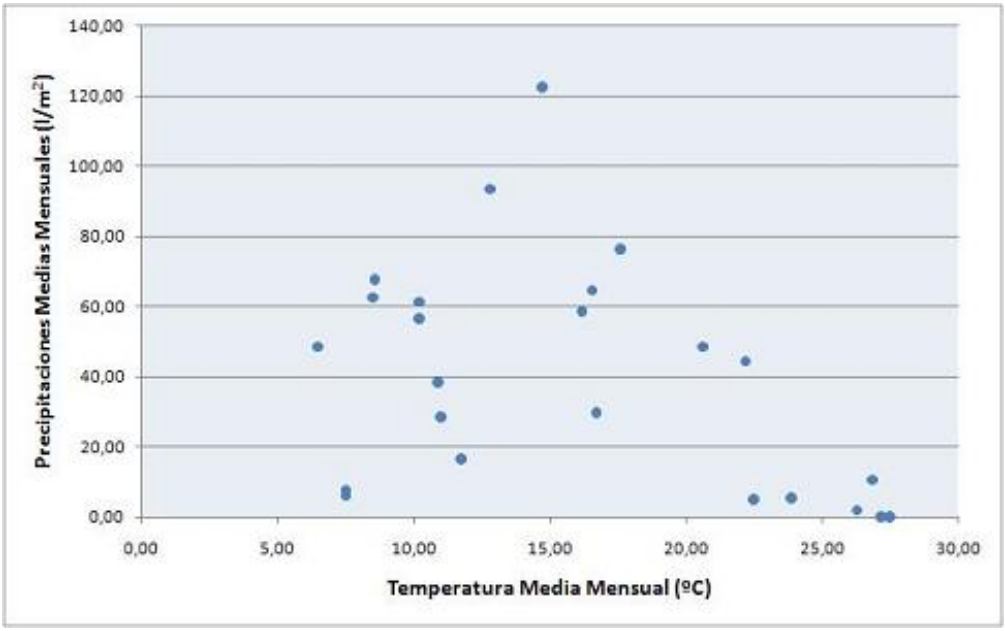
Este es el caso sobre el que vamos a trabajar. Además, la representación te resultará familiar de la unidad 2, ya que tenemos unos ejes cartesianos y puntos con dos coordenadas x e y.

Por ejemplo: vamos a representar los valores de Temperatura y Precipitaciones medias mensuales en una determinada estación climatológica que tenías en la autoevaluación del apartado anterior.

(7,5 ; 7,7)	(10,2 ; 56,5)	(11 ; 28,6)	(12,8 ; 93,5)	(17,6 ; 76,5)	(22,5 ; 5)
(27,5 ; 0,2)	(26,3 ; 1,9)	(22,2 ; 44,4)	(16,7 ; 29,9)	(10,2 ; 61,3)	(7,5 ; 6,1)
(8,48 ; 62,51)	(10,89 ; 38,34)	(11,72 ; 16,46)	(14,71 ; 122,7)	(16,55 ; 64,66)	(23,87 ; 5,38)
(26,86 ; 10,54)	(27,16 ; 0,14)	(20,63 ; 48,58)	(16,18 ; 58,55)	(8,57 ; 67,71)	(6,46 ; 48,56)

Diagrama de dispersión o Nube de puntos:

Al igual que el Diagrama de Burbujas, se representa sobre un par de ejes cartesianos. En este caso, cada punto representa un par de datos de la Variable Estadística Bidimensional.

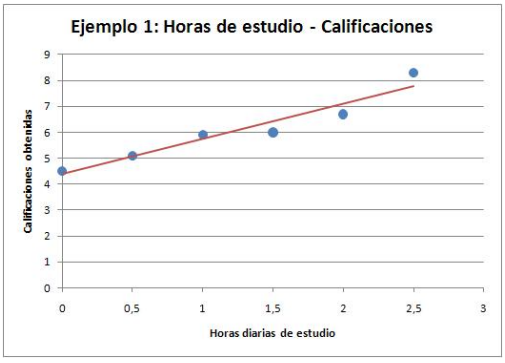


Dependencia

Las Nubes de Puntos también nos ayudan a ver la dependencia entre las variables. Si recuerdas, en el primer apartado vimos que la dependencia podía ser:

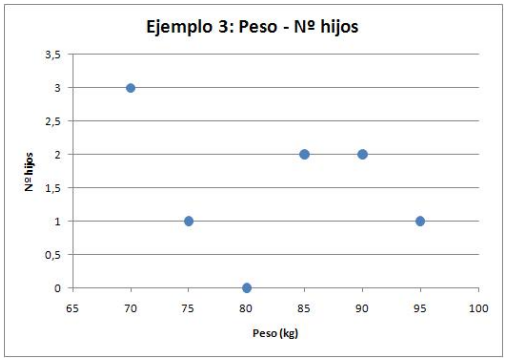
- **Dependencia positiva**: Al aumentar la variable X, también aumenta la Y.
- **Dependencia negativa**: Al aumentar la variable X, disminuye la Y.
- **Sin dependencia**: No se observa ninguna relación entre las dos variables.
- **Dependencia funcional**: Podemos encontrar una relación exacta entre ambas variables que siempre se cumple. Por ejemplo, si estudias la relación entre el número de cajas de leche y el número de litros que se compra de una marca, tenemos una dependencia funcional, porque cada caja tiene siempre el mismo número de litros. Puede ser más o menos fuerte dependiendo de que el diagrama de dispersión tienda a acercarse más o menos a la representación de la función. Nos interesará conocer si es positiva o negativa, así como si es lineal o curvilínea.
- **Dependencia aleatoria**: No hay una regla exacta que determine la relación entre ambas variables, como en el ejemplo anterior.

Mira las siguientes gráficas correspondientes a diferentes ejemplos. Verás que es mucho más fácil ver así la dependencia:

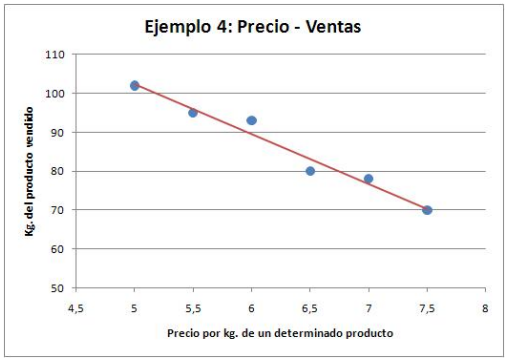


Dependencia positiva aleatoria

Dependencia positiva funcional



Sin dependencia



Dependencia negativa aleatoria

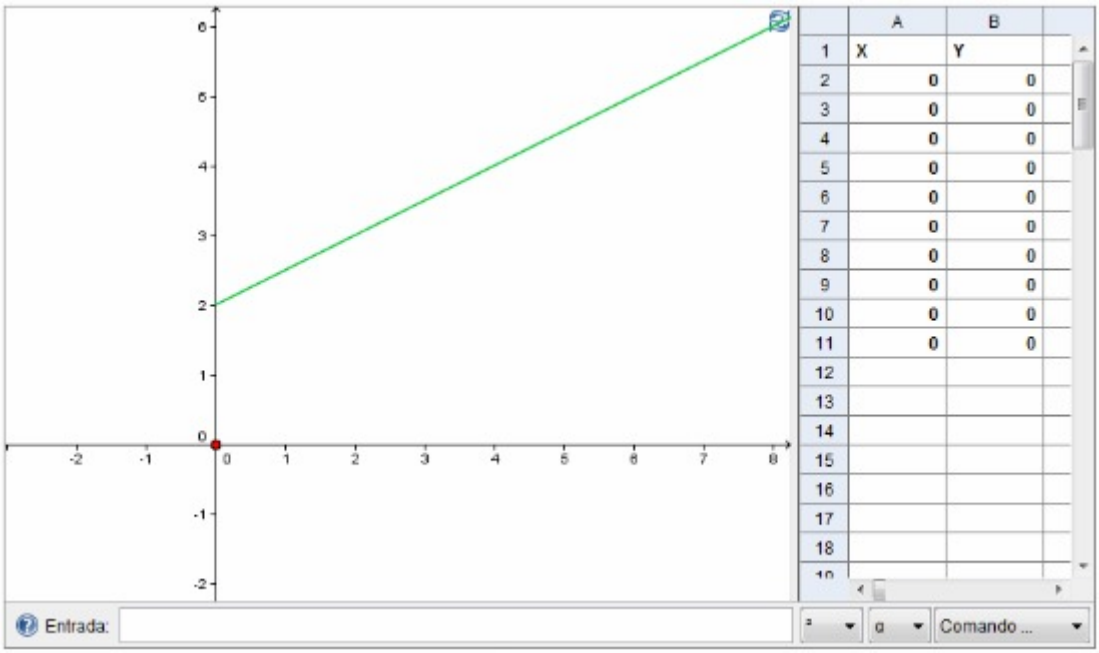
Comprueba lo aprendido

Una compañía química está estudiando el uso de un fertilizante líquido que han inventado en una determinada planta. Para ello miden dos variables: X = "cantidad diaria de fertilizante que se aporta a la planta (en ml)" e Y = "crecimiento de la planta al cabo de 10 días (en cm)".

Los resultados son los siguientes pares de datos:

X	1	2	3	4	5	6	7	8	9	10
Y	2	2.9	3	4	4	4.5	5	4.8	4.9	5.2

Si haces clic en la siguiente imagen podrás representar estos puntos en los ejes de coordenadas. Para ello, escribe cada par en las columnas X e Y. ¡Ojo!, para escribir los números decimales debes utilizar el punto, no la coma.



a) A la vista de la gráfica, ¿crees que existe dependencia funcional?

- Sí.
- No.

¿Crees que hay una relación exacta entre las dos variables?

Como puedes observar, los puntos no se adaptan a una recta o una parábola exactamente. Por tanto no hay dependencia funcional.

Solution

1. Incorrecto
2. Opción correcta

b) ¿Cómo clasificarías la dependencia según la nube de puntos?

- Positiva.
- Negativa.
- Sin dependencia.

Muy bien. Al aumentar la cantidad de fertilizante, también aumenta el tamaño de la planta.

Observa si la nube de puntos crece o decrece.

Observa si la nube de puntos crece o decrece.

Solution

1. Opción correcta
2. Incorrecto
3. Incorrecto

c) La línea verde de la gráfica representa los resultados que se obtienen con el fertilizante de la competencia, ¿con cuál crees que se obtienen mejores resultados?

- Con el de la compañía.
- Con el de la competencia.

Observa que los resultados de la compañía quedan por debajo de los de la competencia.

Muy bien. Nuestra nube de puntos está por debajo de los resultados de la competencia.

Solution

1. Incorrecto
2. Opción correcta

Comprueba lo aprendido

Determina el tipo de dependencia que existe en las siguientes variables estadísticas bidimensionales, con tan solo observar los datos que aparecen en las tablas.

a) Al alumnado de una clase se le pregunta sobre las horas que dedican diariamente a estudiar y las calificaciones de matemáticas obtenidas. Los resultados medios son los siguientes:

X = Horas de estudio	0	0,5	1	1,5	2	2,5
Y = Calificación obtenida	4,5	5,1	5,9	6	6,7	8,3

- Dependencia positiva funcional.
- Dependencia positiva aleatoria.
- Sin dependencia.

Si fuera funcional, podríamos determinar exactamente la calificación que obtendría un alumno que estudiara 3 horas ¿crees que sería posible con estos datos?

¡CORRECTO! Al aumentar el número de horas, aumenta la calificación (positiva), pero no podemos apreciar ninguna relación exacta (aleatoria).

Comprueba si al aumentar el número de horas, aumenta o disminuye la calificación.

Solution

1. Incorrecto
2. Opción correcta
3. Incorrecto

b) En una empresa pagan las horas extra según la siguiente tabla, donde X = número de horas extra e Y = sueldo recibido.

X = Nº de horas	1	2	3	4	5
Y = Sueldo	15	30	45	60	75

- Dependencia positiva funcional.
- Dependencia positiva aleatoria.
- Sin dependencia.

¡CORRECTO! Podemos decir que cada hora siempre se cobra a 15€, por lo tanto la relación se puede determinar exactamente.

¿No crees que podrías determinar exactamente la relación que hay entre las horas de trabajo y el sueldo que percibe?

Comprueba si al aumentar el número de horas extra, también aumenta o disminuye el sueldo.

Solution

1. Opción correcta
2. Incorrecto
3. Incorrecto

c) Queremos saber si existe alguna relación entre el peso de un hombre y el número de hijos que tiene. Para ello, después de preguntar a una población, hemos obtenido los siguientes datos:

X = Peso	70	75	80	85	90	95
Y = Nº de hijos	3	1	0	2	2	1

- Dependencia positiva funcional.
- Dependencia positiva aleatoria.
- Sin dependencia.

¿Al aumentar el peso siempre aumenta el número de hijos?

¿Al aumentar el peso siempre aumenta el número de hijos?

¡CORRECTO! Como ya podías imaginar, no existe ninguna relación entre el peso de un hombre y el número de hijos que tiene.

Solution

1. Incorrecto
2. Incorrecto
3. Opción correcta

3. Análisis de datos

Últimamente, y de manera cada vez más frecuente, habrás escuchado en los informativos noticias que nos hablan de **previsiones** sobre la economía de una cierta región, el comercio o la evolución de la población mundial a un medio o largo plazo. Seguro que te has planteado: ¿cómo se puede pronosticar un acontecimiento que aún está por ocurrir?

Por ejemplo, algunas de las noticias que has podido oír referentes a nuestra historia sobre el cambio climático, han sido, entre otras, las siguientes:

- La población mundial en el año 2050 aumentará en más de mil millones de personas, y con ello las emisiones de CO_2, según datos de la ONU.
- La Agencia Estatal de Meteorología prevé que la temperatura en España aumentará hasta en 6 ºC en 2100.
- El nivel del mar podría aumentar hasta un metro para 2100 según científicos australianos.

En este tema veremos que, si entre dos variables de una determinada situación existe cierta relación, podremos hacer previsiones sobre el comportamiento futuro de estas.

3.1. Estudio de la Correlación

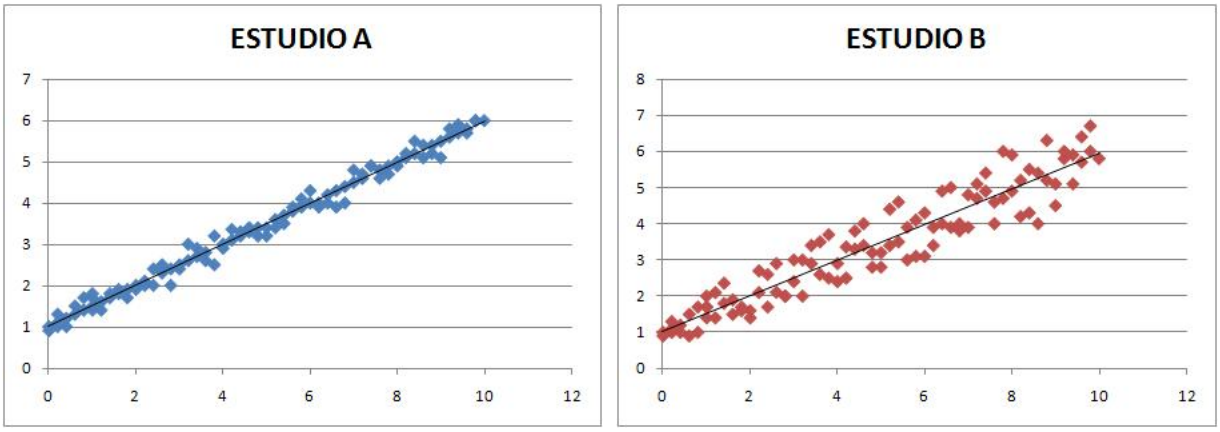
¿Recuerdas las nubes de puntos del apartado anterior? Con ellas podíamos determinar si había algún tipo de relación o dependencia entre dos variables, y en ese caso decidir si la relación era positiva o negativa (al aumentar la primera, aumentaba o disminuía respectivamente la segunda).

Fotografía en Flickr de Kevin Dooley bajo CC

Esta relación la definiremos como **correlación**. Y esta puede ser:

- **Correlación funcional**: si existe una relación funcional entre las variables X e Y. Es decir, podemos calcular los valores de Y a partir de los de X, con una función.
- **Correlación positiva o directa**: existe cierta relación entre ambas variables, y al aumentar los valores de X también aumentan los de Y.
- **Correlación negativa o inversa**: existe cierta relación entre las variables, pero al aumentar los valores de X disminuyen los de Y.
- **Correlación nula**: no existe ningún tipo de relación entre ambas.

Observa ahora las dos siguientes gráficas obtenidas al hacer el mismo estudio sobre dos poblaciones diferentes.



Como puedes ver, en ambas la correlación es positiva. Pero, ¿crees que en los dos casos existe la misma dependencia entre las variables? Por lo que se puede apreciar en las gráficas, en el Estudio A la dependencia parece ser más fuerte que en el Estudio B. Por tanto, debe existir alguna forma para **medir** la correlación.

A continuación, definiremos dos parámetros, la **covarianza** y el **coeficiente de correlación lineal**, que nos servirán para establecer esta medida.

Covarianza

Al igual que teníamos en el tema anterior medidas que nos ayudaban a interpretar los datos de una distribución unidimensional, en las bidimensionales tenemos la **covarianza**, que nos permite saber si la relación entre las variables es directa o inversa, y si dicha relación puede ser lineal o no.

Importante

La **covarianza** de una variable bidimensional (X,Y), que representaremos por σ_{XY}, es una medida estadística que se calcula usando una de las expresiones:

una medida estadística que se calcula usando una de las expresiones:

Tablas simples

$$\sigma_{XY} = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y}$$

Tablas de doble entrada

$$\sigma_{XY} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} x_i \cdot y_j \cdot n_{ij}}{N} - \bar{x} \cdot \bar{y}$$

donde N indica el tamaño de la muestra.

Interpretación: el signo de la covarianza nos permitirá saber el tipo de correlación:

- Si la covarianza es positiva, la correlación será directa.
- Si la covarianza es negativa, la correlación será inversa.

En la siguiente presentación puedes ver cómo calcularemos la covarianza a partir de una tabla simple.

Presentación en Slideshare de Saúl Valverde

Si aún así las fórmulas te parecen muy complicadas, no te preocupes que seguro que con estos dos ejemplos lo vas a entender. En el primero tenemos una tabla simple, con únicamente cinco datos y en el segundo vamos a tener una tabla de doble entrada:

Ejercicio resuelto

En un estudio sociológico se está analizando el nivel de estudios de la población y el salario mensual de estos. Los datos obtenidos se reflejan en esta tabla:

Nivel de estudios	1	2	3	4	5
Salario medio (€)	700	940	1.120	1.300	2.180

donde 1=Sin titulación, 2=Estudios secundarios, 3=Técnicos de grado medio, 4= Bachillerato y 5=Técnicos superiores o licenciados.

¿Cuál es el valor de la covarianza?

Mostrar retroalimentación

En este caso se trata de datos simples (sin frecuencias) por lo que aplicamos la fórmula:

$$\sigma_{XY} = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{N} - \overline{x} \cdot \overline{y}$$

Empezamos calculando las medias de las variables X e Y (marginales):

$$\overline{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

$$\overline{y} = \frac{700+940+1.120+1.300+2.180}{5} = \frac{6.240}{5} = 1.248$$

Ahora calculamos la suma de los productos $x_i \cdot y_i$:

$$\sum_{i=1}^{5} x_i \cdot y_i = 1 \cdot 700 + 2 \cdot 940 + 3 \cdot 1.120 + 4 \cdot 1.300 + 5 \cdot 2.180 = 22.040$$

Por último, sustituimos en la fórmula inicial de la covarianza y obtenemos:

$$\sigma_{XY} = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{N} - \overline{x} \cdot \overline{y} = \frac{22.040}{5} - 3 \cdot 1.248 = 664$$

A la salida de un restaurante se hace una encuesta en la que se pregunta el número de persona que vienen a comer juntas, X, y la calificación de 1 a 4 que le pondrían al restaurante, Y. Los datos recogidos se han ordenado en la tabla:

X \ Y	1	2	3	4
1	0	1	0	0
2	2	3	4	1
3	2	3	6	1
4	0	0	2	0

¿Cuál es la covarianza?

Mostrar retroalimentación

Ve pasando la presentación para ver el cálculo.

Coeficiente de correlación lineal

También llamado coeficiente de correlación de Pearson es el parámetro que nos va a decir si la correlación es débil o fuerte, además de indicarnos también si es directa o inversa dependiendo de su signo.

Importante

Para calcularlo, necesitamos conocer el valor de las desviaciones típicas marginales de cada variable σ_x y σ_y, ya que su expresión viene dada por:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

El valor del coeficiente de correlación lineal r siempre será un número comprendido entre -1 y 1 $(-1 \leq r \leq 1)$. Su signo nos indicará el sentido de la correlación (positiva o negativa) y mientras más próximo esté su valor a 1 o -1, más fuerte será la correlación.

Interpretación: Según el valor de r, la correlación entre las dos variables será:

- r = 0: No existe correlación (correlación nula).
- r = 1 : La correlación es perfecta y positiva (correlación funcional positiva).
- r = -1: La correlación es perfecta y negativa (correlación funcional negativa).
- r próximo a 1: La correlación es fuerte y positiva.
- r próximo a -1: La correlación es fuerte pero negativa.
- r próximo a 0: La correlación es débil.

Veamos cómo calcularla con el ejemplo anterior.

Presentación en Slideshare por Saúl Valverde

Comprueba lo aprendido

Vamos a recuperar los Estudios A y B del comienzo de este apartado. Vas a calcular el coeficiente de correlación de Pearson para comprobar si el resultado se corresponde con lo que habíamos supuesto por la gráfica.

Para realizar los cálculos te daremos la suma de los totales de cada columna, como en los ejemplos anteriores. El valor de N=104.

ESTUDIO	x_i	v_i	x_iv_i	x_i^2	v_i^2

A					
TOTALES	502,2	354,56	2165,43	3336,04	1437,21

ESTUDIO B	x_i	y_i	x_iy_i	x_i^2	y_i^2
TOTALES	502,2	353,83	2157,92	3336,04	1451,59

Observación: por cuestión de redondeo, puede que los resultados que te ofrecemos a continuación no coincidan exactamente con los tuyos. Elige los que más se aproximen.

a) ¿Cuáles de los siguientes valores corresponden a las medias marginales de X e Y para ambos estudios?

- Estudio A: Media X = 4,83
  Estudio B: Media Y = 4,53
- Estudio A: Media Y = 3,41
  Estudio B: Media X = 4,83

Mostrar retroalimentación

Solution

1. Incorrecto
2. Correcto

b) ¿Qué valores corresponden a las desviaciones típicas marginales de X e Y para esos dos estudios?

- Estudio A: Desviación típica de X = 2,96
  Estudio B: Desviación típica de Y = 2,15
- Estudio A: Desviación típica de Y = 1,48
  Estudio B: Desviación típica de X = 2,96

Mostrar retroalimentación

Solution

1. Incorrecto
2. Correcto

c) ¿Cuáles de las siguientes covarianzas corresponden a la de esos estudios?

- Estudio A: Covarianza = 4,36
  Estudio B: Covarianza = 4,32
- Estudio A: Covarianza = 4,58
  Estudio B: Covarianza = 4,52

Mostrar retroalimentación

Solution

1. Correcto
2. Incorrecto

d) Calcula los correspondientes Coeficientes de Correlación de Pearson.

- [ ] Estudio A: r = 0,99
  Estudio B: r = 0,95

- [ ] Estudio A: r = 0,97
  Estudio B: r = 0,91

Mostrar retroalimentación

Solution

1. Correcto
2. Incorrecto

e) Por último, a la vista de los resultados, podemos afirmar que:

- [ ] Ambos estudios tienen correlación positiva, y la del Estudio A es más fuerte que la del Estudio B.

- [ ] Ambos estudios tienen correlación positiva, y la del Estudio A es más débil que la del Estudio B.

Mostrar retroalimentación

Solution

1. Correcto
2. Incorrecto

3.2. Rectas de regresión

Fotografía en Flickr de Cinzia A. Rizzo bajo CC

Fotografía en Flickr de ATBravo bajo CC

Si miramos las hojas o las pisadas de estas fotos como si fueran puntos de nuestras gráficas, ¿podrías decir dónde caería la próxima hoja?, ¿o dónde estaría la pisada que sigue el camino? En el primer caso sería mucha casualidad que acertases, pues la correlación es nula. Sin embargo, en el segundo caso la correlación es muy fuerte y seguro que podrías dar una respuesta bastante ajustada.

Eso vamos a hacer en este apartado, predecir resultados a partir de los datos que tenemos. Ten en cuenta que esas predicciones serán más fiables cuanto mayor sea el coeficiente de correlación.

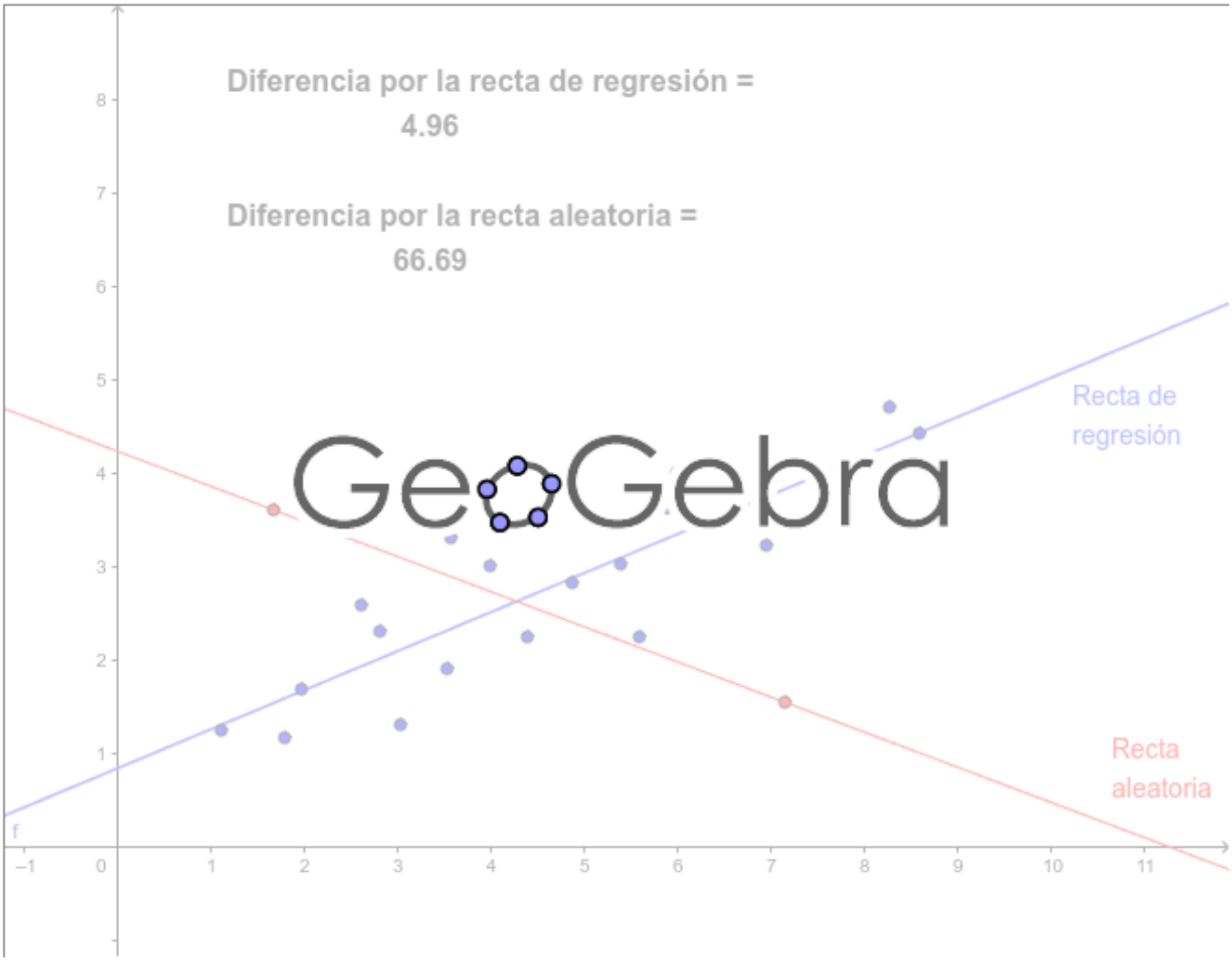
A lo largo de este tema y del anterior, has visto que en algunas gráficas aparecía una recta que se ajustaba a la nube de puntos. Esa línea se llama **recta de regresión** y es la **recta que mejor se ajusta a la nube de puntos**.

Rectas de regresión

En la siguiente escena tienes un ejemplo de una nube de puntos y su **recta de regresión** (de color azul), así como otra recta (de color rojo) que hemos llamado aleatoria. Esta última recta la puedes cambiar moviendo los puntos rojos. En la escena también se expresa la aproximación que se consigue con cada recta de la nube de puntos.

a) Mueve los puntos rojos e intenta hallar una mejor aproximación a la nube de puntos que la expresada por la recta de regresión. ¿Es posible encontrarla?

b) Si mueves los puntos azules, estarás modificando la nube de puntos y por tanto la recta de regresión. Modifica los puntos azules que quieras y repite el intento de mejorar la aproximación que viene dada por la recta de regresión.



Hasta ahora solo hemos hablado de la recta de regresión desde un punto de vista visual, a continuación veremos cuál es la ecuación de la recta que mejor "ajusta" una nube de puntos.

Importante

Para una variable estadística bidimensional, la **recta de regresión de Y sobre X** viene dada por la ecuación:

$$y = \overline{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \overline{x})$$

Una vez que tenemos la ecuación de la recta, podemos "predecir" valores de la variable Y que no conozcamos sustituyendo el valor de x. Veámoslo en el siguiente ejemplo.

Ejercicio resuelto

En el apartado anterior calculaste los parámetros para una Variable en la que se relacionaba el peso y la tensión arterial de los pacientes de un Centro de Salud.

Los valores que obtuviste fueron los siguientes:

$$\overline{x} = 96,4 \quad \overline{y} = 147,2 \quad \sigma_x = 19,52 \quad \sigma_{xy} = 440,42$$

a) Calcula la recta de regresión de Y sobre X.

Mostrar retroalimentación

Basta con que sustituyas cada uno de los valores en la expresión de la recta de regresión.

$$y = \overline{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \overline{x}) = 147,2 + \frac{440,42}{19,52^2}(x - 96,4)$$

Operando, obtendrás la ecuación de la recta en forma estándar y = m·x +n. En nuestro caso, quedará:

$$y = 1,16x + 35,77$$

b) Usando la recta de regresión anterior, haz una estimación sobre la tensión que tendrá una persona que pese x=100 kg.

Mostrar retroalimentación

Si sustituimos x=100 en la ecuación de la recta anterior, nos quedaría:

$$y = 1,16 \cdot 100 + 35,77 = 151,77$$

Este valor, 151,77 será la estimación de presión arterial para una persona que pese 100 kg.

Como hemos visto, la recta de regresión de Y sobre X es la que mejor aproxima los valores de la variable Y a una recta, pero también podemos aproximar los valores de X. Para ello usamos otra recta, la **recta de regresión de X sobre Y**, que tiene la siguiente ecuación:

$$x = \overline{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y - \overline{y})$$

En la siguiente escena descubrirás la representación gráfica de las dos rectas. Cuanto más próximo sea el valor del coeficiente de correlación de Pearson a 1 o -1, mayor coincidencia habrá entre ambas rectas. Puedes modificar los puntos para ver cómo varían ambas rectas.

El punto verde es el **Centro de Gravedad** de la distribución, y sus coordenadas son $(\overline{x}, \overline{y})$.

Comprueba lo aprendido

Utilizando los datos del Estudio B del apartado anterior, calcula las ecuaciones de las rectas de regresión de Y sobre X y de X sobre Y.

$\bar{x} = 4{,}83$ $\bar{y} = 3{,}40$	$\sigma_x = 2{,}96$ $\sigma_y = 1{,}54$	$\sigma_{xy} = 4{,}32$

- [ ] Recta de regresión de Y sobre X: y = 0,49x+1,02
- [ ] Recta de regresión de Y sobre X: y = 0,49x+2,01
- [ ] Recta de regresión de X sobre Y: x = 1,82y-3,16
- [ ] Recta de regresión de X sobre Y: x = 1,82y-1,36

Mostrar retroalimentación

Solution

1. Correcto
2. Incorrecto
3. Incorrecto
4. Correcto

A continuación un par de vídeos con dos ejercicios completos de regresión lineal, en ambos tienes que calcular los coeficientes de correlación, alguna recta de regresión y por último, hacer una estimación basándote en los resultados obtenidos:

Vídeos en youtube de juanmemol

Ejercicio resuelto

Curso 2009/2010

Una cooperativa aceitera quiere realizar un estudio sobre la influencia de las campañas publicitarias en sus cifras de ventas. Para ello dispone del gasto estimado en publicidad y del volumen de ventas de los últimos 5 años (ambos en miles de euros):

X: Gasto en publicidad	2.5	2.8	2.9	3.1	3.5
Y: Ventas	200	221	230	239	248

a) Obtenga la recta de regresión de Y sobre X. ¿Cuál será el volumen de ventas si la inversión en publicidad ascendiera a 3.8 millones de euros?

b) Calcule el coeficiente de correlación lineal e interprete su valor.

Mostrar retroalimentación

Nos piden calcular la recta de regresión de Y sobre X, para lo que utilizamos la fórmula

$$y = \overline{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \overline{x})$$

Usando los datos que nos dan en el problema construimos la siguiente tabla, sabiendo que N=5:

	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
	2,5	200	6,25	40000	500
	2,8	221	7,84	48841	618,8
	2,9	230	8,41	52900	667
	3,1	239	9,61	57121	740,9
	3,5	248	12,25	61504	868
TOTAL	14,8	1138	44,36	260366	3394,7

Calculamos la media y las desviaciones típicas marginales:

14,8

4. Apéndice

Aunque es más frecuente encontrarnos con preguntas de Estadística Unidimensional y de Cálculo de Probabilidades, el cálculo de la covarianza, del coeficiente de correlación, de las rectas de regresión y la estimación de datos a partir de los parámetros anteriores, también han aparecido ocasionalmente.

Como en unidades anteriores en este apartado te ofrecemos, unos recursos que te pueden ayudar a agilizar el cálculo. Además, puedes ampliar conocimientos con las "Curiosidades" y "Para saber más".

Importante

Con el siguiente gráfico puedes realizar un rápido repaso por todo lo visto en el tema:

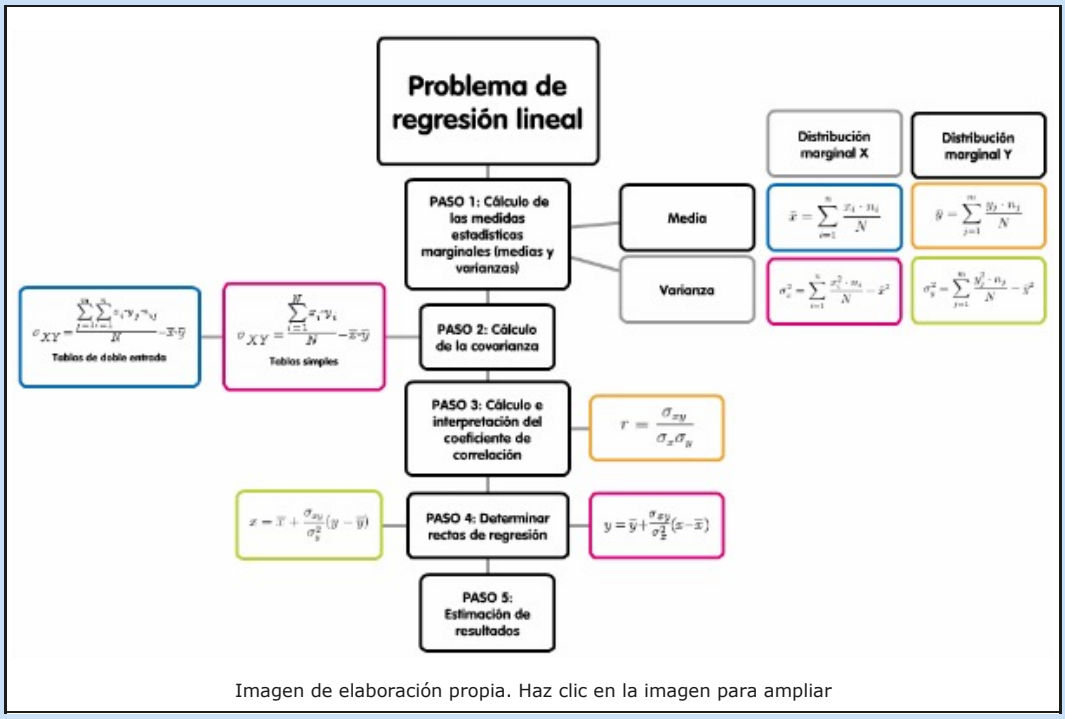


Imagen de elaboración propia. Haz clic en la imagen para ampliar

4.1. Recursos

Como ya sabes no dispondremos de alguna de estas herramientas en la prueba, pero si en algún momento decides practicar con ejercicios de los que no tengas la solución, esto te ayudará a corregirlos.

Calculadora

A continuación, un vídeo de cómo se trabaja con variables bidimensionales con una calculadora Casio. Si te animas puedes intentar rehacer algunas de las actividades anteriormente propuestas:

Hoja de cálculo

Como te habrás dado cuenta ya, calcular un coeficiente de correlación es fácil pero laborioso e incluso tedioso. Con una hoja de cálculo todas las operaciones necesarias para calcular el coeficiente de correlación se hacen de manera automática.

En este vídeo te mostramos cómo se puede calcular con la hoja de cálculo EXCEL.

En un vídeo anterior, vimos cómo se calculaba el coeficiente de correlación con la hoja de cálculo EXCEL.

En este vídeo, te presentamos cómo puede hacerse el cálculo de la recta de regresión, la estimación de un valor y el cálculo del coeficiente de correlación con la hoja de cálculo CALC de la suite ofimática OpenOffice, suite de libre distribución y totalmente gratuita. Esta hoja de cálculo es muy similar a la EXCEL que vimos antes.

Curiosidad

Algo de Historia

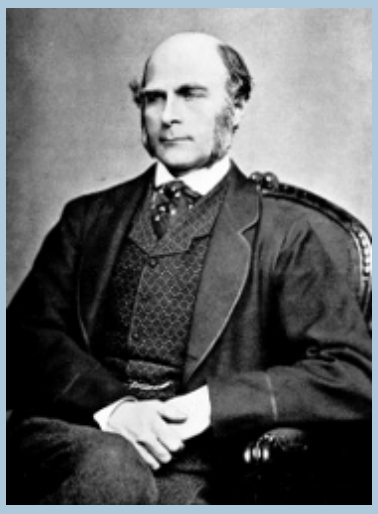


Imagen en Wikimedia Commons de Fastfission bajo Dominio Público

Sir Francis Galton (16 de febrero de 1822 – 17 de enero de 1911). Antropólogo, geógrafo, explorador, inventor, meteorólogo, estadístico y psicólogo británico.

No tuvo cátedras universitarias y realizó la mayoría de sus investigaciones por su cuenta. Sus múltiples contribuciones recibieron reconocimiento formal cuando, a la edad de 87 años, se le concedió el título de Sir o caballero del Reino.

De intereses muy variados, Galton contribuyó a diferentes áreas de la ciencia como la psicología, la biología, la tecnología, la geografía, la estadística o meteorología. A menudo, sus investigaciones fueron continuadas dando lugar a nuevas disciplinas.

Primo segundo de Charles Darwin, aplicó sus principios a numerosos campos, principalmente al estudio del ser humano y de las diferencias individuales.

En el campo de la estadística, la principal contribución de Galton fue el concepto de correlación entre pares de atributos, aplicándolo a problemas sobre herencia y genética.

Karl Pearson (1857-1936), matemático y filósofo de las ciencias británico, se le conoce por haber desarrollado algunas de las técnicas centrales de la moderna estadística, y por aplicar estas técnicas a los problemas de la herencia biológica.

A principios de 1900, Pearson se interesó por el trabajo de Francis Galton, que intentaba encontrar relaciones estadísticas para explicar cómo las características biológicas iban pasando a través de sucesivas generaciones. La investigación de Pearson colocó en gran medida las bases de la estadística del siglo XX, definiendo los significados de correlación, análisis de la regresión y desviación típica. En 1911 Pearson alcanzó el cargo de profesor de eugenesia en el University College, examinando la recopilación y análisis de la información en el sentido que las características como inteligencia, criminalidad, pobreza y creatividad se trasmiten a través de generaciones. Pearson confiaba en aplicar estas intuiciones con el fin de mejorar la raza humana.

Si quieres saber más sobre el trabajo de K. Pearson, sigue el enlace.

Curiosidad

¿Leyendas urbanas?

Aunque hemos visto que la correlación indica que una variable esté relacionada con otra, esto no quiere decir que exista una relación de causa y efecto entre una y otra. Mira los siguientes ejemplos:

- Según la DGT, el 20 % de los

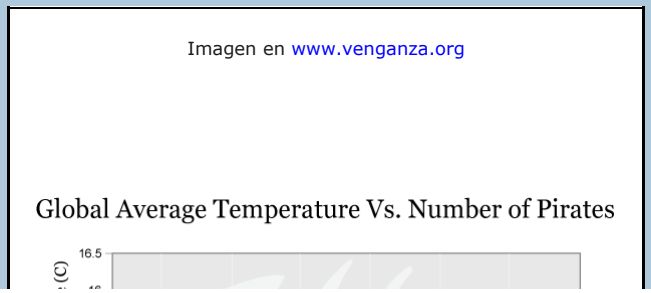
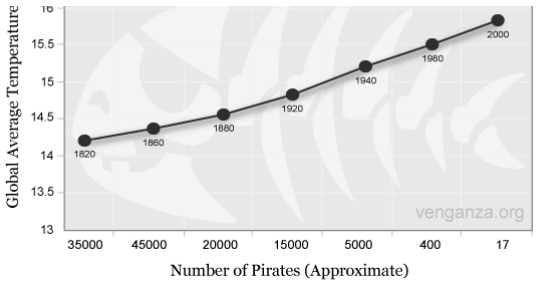


Imagen en www.venganza.org

motoristas fallecidos en accidentes de tráfico no llevaban casco.

- La mayoría de los accidentes de tráfico se producen entre vehículos que ruedan a una velocidad moderada.
- Los días de luna llena se produce un aumento en el número de nacimientos.
- El cambio climático provoca el aumento de tornados en el hemisferio norte.



En el primer y segundo caso, no se puede establecer que sea mejor no llevar casco o circular a una velocidad excesiva, pues lo que no se dice es que la mayoría de los motoristas llevan casco, y la mayoría de los coches circulan a una velocidad moderada, de ahí que sea también mayor el número de accidentes.

El tercer ejemplo es el típico caso en el que, a pesar de haber sido refutadas con estudios estadísticos, siguen formando parte de las leyendas urbanas.

El cuarto, es un ejemplo de mal uso de las estadísticas para crear buenos titulares. Los estudios indican que, aunque haya aumentado el número de tornados en el hemisferio norte, no ha sido así a nivel mundial y no se puede establecer que el aumento de tornados sea causa directa del cambio climático.

El gráfico que tienes arriba, de la web www.venganza.org, representa el aumento de las temperaturas globales y la disminución en el número de piratas en los últimos siglos. Es un ejemplo muy curioso de cómo se pueden relacionar dos variables que no tienen nada que ver, y obtener una correlación muy fuerte ¿Han desaparecido los piratas debido al cambio climático?

Resumiendo, como nos explican en Microsiervos, hay que aprender a interpretar los estudios estadísticos y a ser crítico con las noticias que nos llegan a diario.

Curiosidad

Representación gráfica para tablas de doble entrada

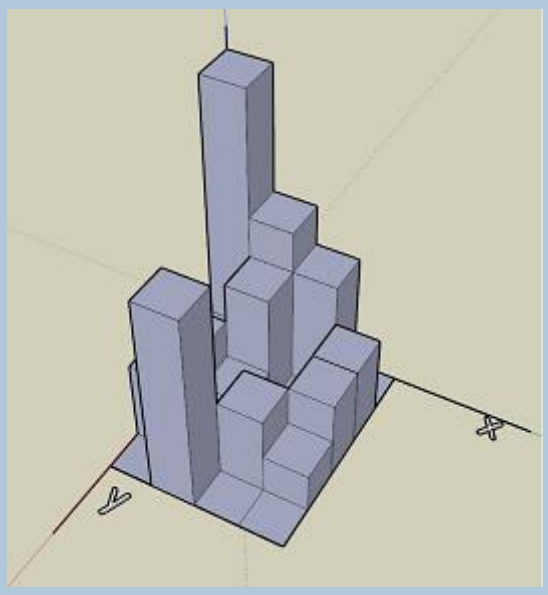
Si volvemos a la tabla de doble entrada que vimos en el ejercicio resuelto en el que comparábamos el número de días mensuales en los que se superaba la concentración máxima de NO_2 y de Ozono en el aire:

	Y	

		$y_1=0$	$y_2=1$	$y_3=2$	$y_4=3$	$y_5=4$	n_i
X	$x_1=0$	7	1	2	2	0	12
	$x_2=1$	4	3	1	1	5	14
	$x_3=2$	3	0	0	2	0	5
	$x_4=3$	0	2	2	1	0	5
	n_j	14	6	5	6	5	36

El par (0,0) se podría representar como un punto en una gráfica habitual de ejes cartesianos, pero en este caso tenemos que hacer ver de algún modo que la frecuencia de ese par es 7. A continuación, verás algunos ejemplos:

a) **Histograma tridimensional**:



Para representar la información partimos de tres ejes cartesianos.

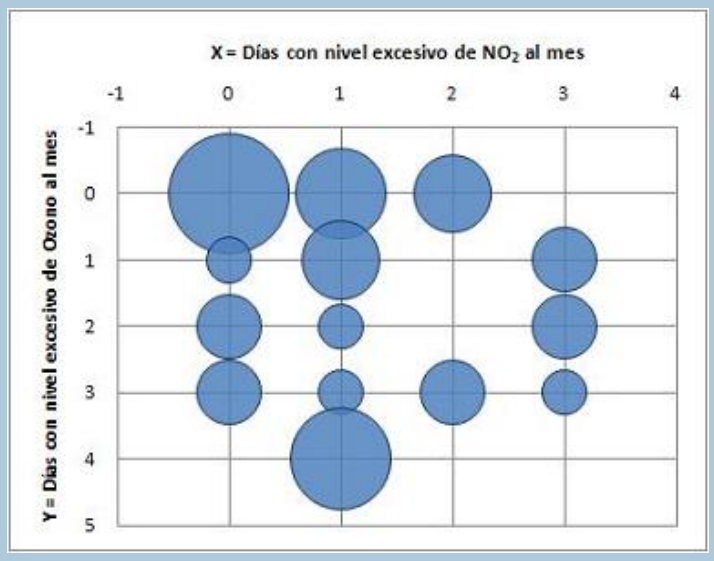
En los ejes X e Y marcamos los posibles valores de cada variable (en nuestro caso 0, 1, 2 y 3 para X, y 0, 1, 2, 3 y 4 para Y). Cada cuadrado representa un par de valores.

La altura de cada cuadrado será la correspondiente frecuencia de ese par de valores.

Fíjate cómo en nuestro caso el par con mayor frecuencia es el (0,0), que se repite 7 veces, y por tanto es el prisma de mayor altura.

El siguiente sería el (1,4) que tiene frecuencia 5.

b) **Diagrama de dispersión o de Burbujas**:



En este caso partimos de un par de ejes cartesianos X e Y en los que representamos los valores de ambos parámetros.

En lugar de puntos, representamos circunferencias en las que su **superficie es proporcional a la frecuencia**. Ojo, no son proporcionales los radios sino las superficies.

Los pares de datos que tienen frecuencia 0 no se representan.

Para saber más

Un repaso a la Estadística Bidimensional

En la siguiente página del proyecto Descartes, puedes repasar todos los contenidos vistos en el tema, además de trabajar con otros ejemplos.

Para saber más

Otra correlación: Parabólica

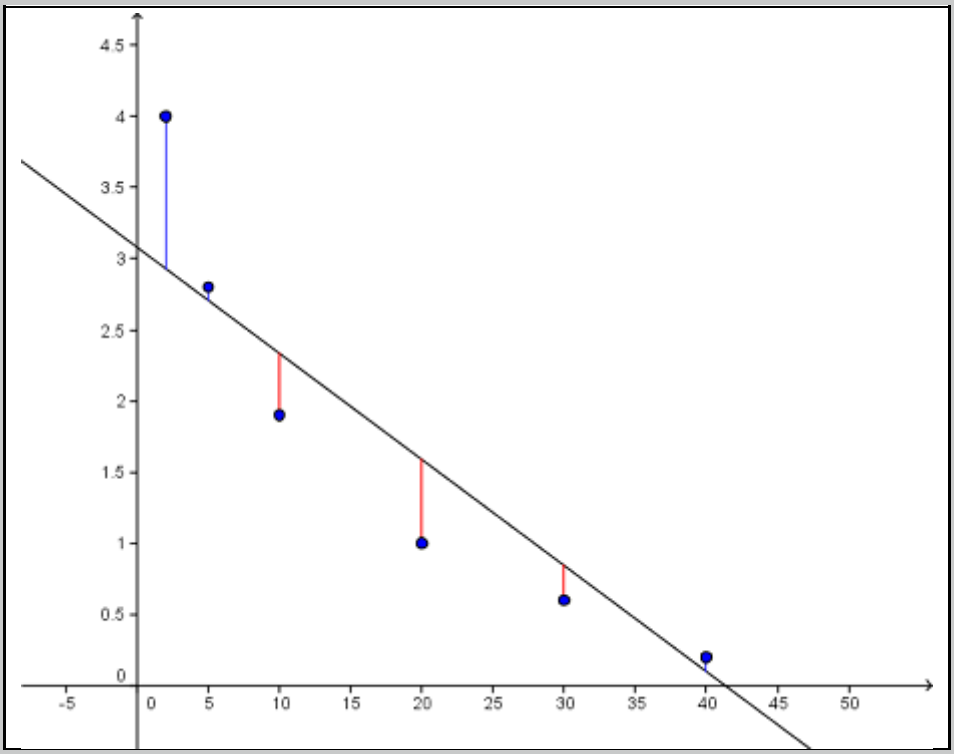
Hemos estado hablando siempre de correlación lineal, de ajustar la nube de puntos a una recta. Pero como comprenderás, hay numerosas situaciones en las que las gráficas de otras funciones se ajustan mejor a la distribución de nuestros puntos.

Así, no solo existe correlación lineal, sino que también existe correlación parabólica, correlación exponencial, correlación logarítmica, etc. En todos los casos la idea es la misma, buscar la función que mejor se ajusta a nuestra nube de puntos.

En el siguiente vídeo se analiza la influencia de los gastos en publicidad (X) y las ventas obtenidas (Y) en los distintos establecimientos de Telepizza. Además de repasar diversos conceptos importantes, al final vemos que para este caso, la correlación parabólica es mejor que la lineal.

¿De dónde sale la recta de regresión? Método de los mínimos cuadrados

Quizás te hayas preguntado cómo se consigue la recta que **mejor se ajusta**. Fíjate en la siguiente gráfica:



Si tomamos una recta cualquiera, medimos la distancia en vertical desde cada punto hasta la recta. Si el punto está por encima de la recta, el resultado es positivo (segmento azul), y si queda por debajo, el resultado es negativo (segmento rojo).

Lo que queremos es que la suma de esas distancias sea lo menor posible para que el ajuste sea mejor, pero si sumamos números positivos con negativos unos anularán a los otros.

Un método para conseguir que todas las distancias sean positivas es elevarlas al cuadrado. Ahora sí podemos sumar todos los resultados y buscar la recta que hace que esa suma sea lo menor posible.

Este método, que se llama **Método de los mínimos cuadrados**, es el que nos da la recta de regresión que hemos estudiado en este tema.