1º de Bachillerato

Matemáticas Aplicadas a las Ciencias Sociales I

Contenidos

Estadística: Distribuciones bidimensionales

1. Concepto de distribución 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.

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:

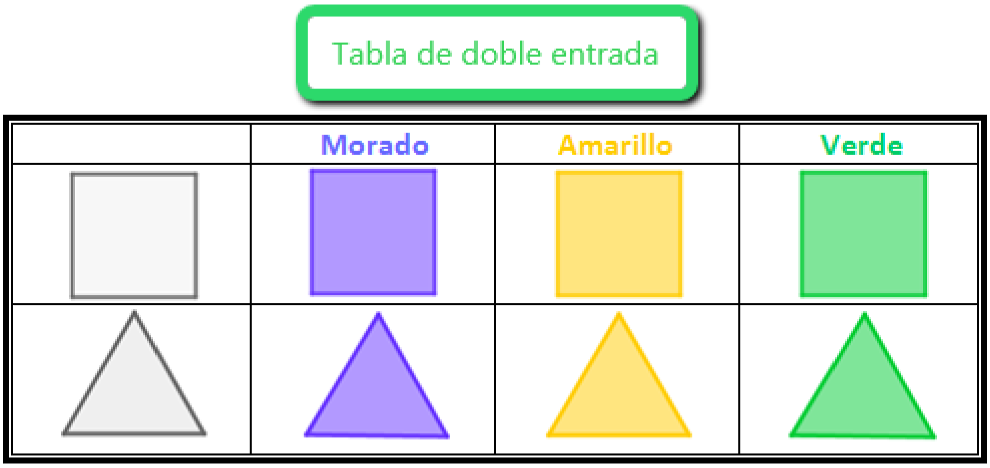


Imagen de elaboración propia

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.

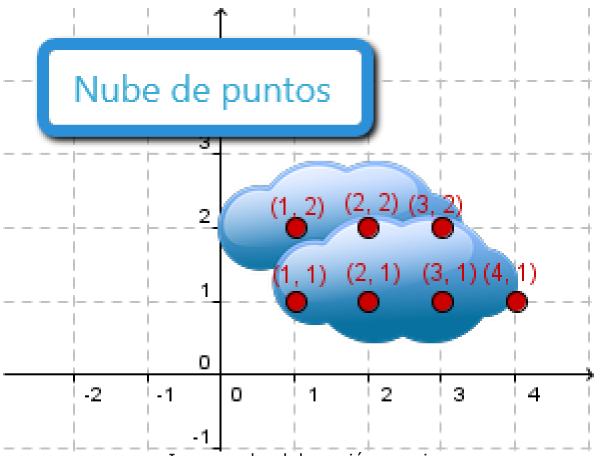


Imagen de elaboración propi

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.

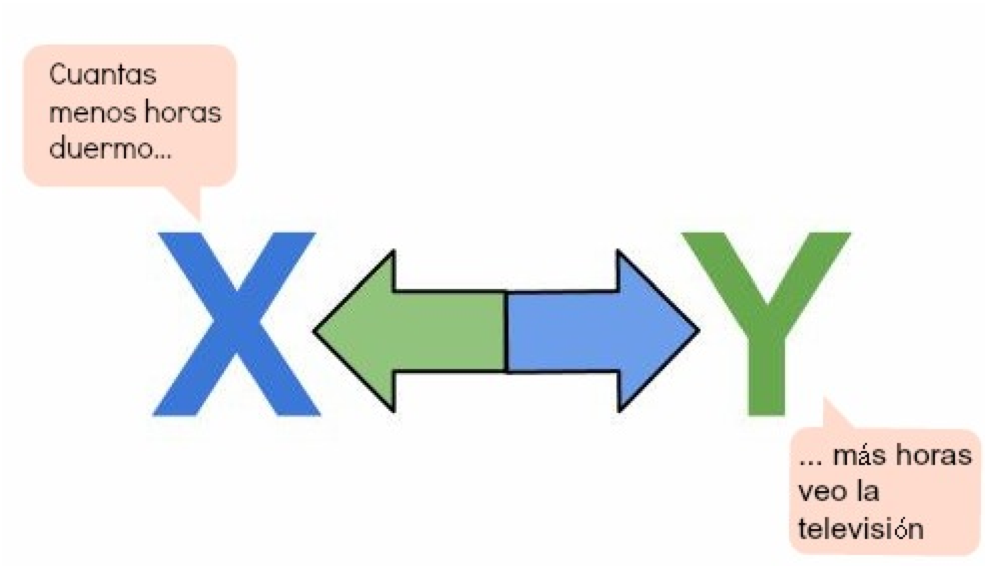


Imagen de elaboración propia

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 las notas 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

Ejemplo de tabla simple

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

En el siguiente vídeo puedes ver una explicación detallada sobre cómo construir una tabla de doble entrada:

Vídeo de estudiia alojado en **Youtube**

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.

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 =

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**.

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

La **distribución condicional** de X dado Y=yj: expresa cómo se distribuye X en la subpoblación que cumple la condición de presentar el valor Y=yj.

La **distribución condicional** de Y dado X=xi: expresa cómo se distribuye Y en la subpoblación que cumple la condición de presentar el valor X=xi.

En el siguiente vídeo puedes ver en un ejemplo práctico qué son las distribuciones marginales y condicionales:

Vídeo de KhanAcademyEspanol alojado en **Youtube**

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:

	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

Veamos un ejemplo de cómo se calcula la media y la varianza de una distribución condicional.

Ejercicio resuelto

Volvamos al ejemplo anterior. Supongamos que queremos estudiar el número de días por mes en los que se supera el límite permitido de concentración de NO_2, sabiendo que el número de días por mes en los que se supera el límite permitido de concentración de ozono es 2.

		$y_1=0$	$y_2=1$	$y_3=2$	$y_4=3$	$y_5=4$	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

Calcula la media y la desviación típica de X/Y=2.

Mostrar retroalimentación

Para calcular la media y la varianza de la condicional operamos como si se tratase de una distribución unidimensional:

$$\bar{x}|_{Y=3} = \frac{0\cdot 2 + 1\cdot 1 + 2\cdot 0 + 3\cdot 2}{5} = \frac{7}{5} = 1,4$$

$$\sigma^2_{|Y=3} = \frac{0^2\cdot 2 + 1^2\cdot 1 + 2^2\cdot 0 + 3^2\cdot 2}{5} - (1,4)^2 = 1,84$$

Importante

Uno de los objetivos del análisis de distribuciones bidimensionales es estudiar si existe asociación o relación entre las variables X e Y.

Las variables X e Y se dicen que son estadísticamente independientes si los valores de una de ellas no afecta a la distribución de la otra. Esto equivale a decir que todas las distribuciones condicionadas sean iguales.

De modo equivalente se dice que las variables X e Y son estadísticamente independientes si se cumple que la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales.

Representación de datos

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 en una 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. Veamos cómo representar datos recogidos en una **tabla simple**.

Importante

Dados los pares (x_i, y_j) de una variable estadística bidimensional (X, Y), a la representación cartesiana de estos puntos se le denomina diagrama de dispersión.

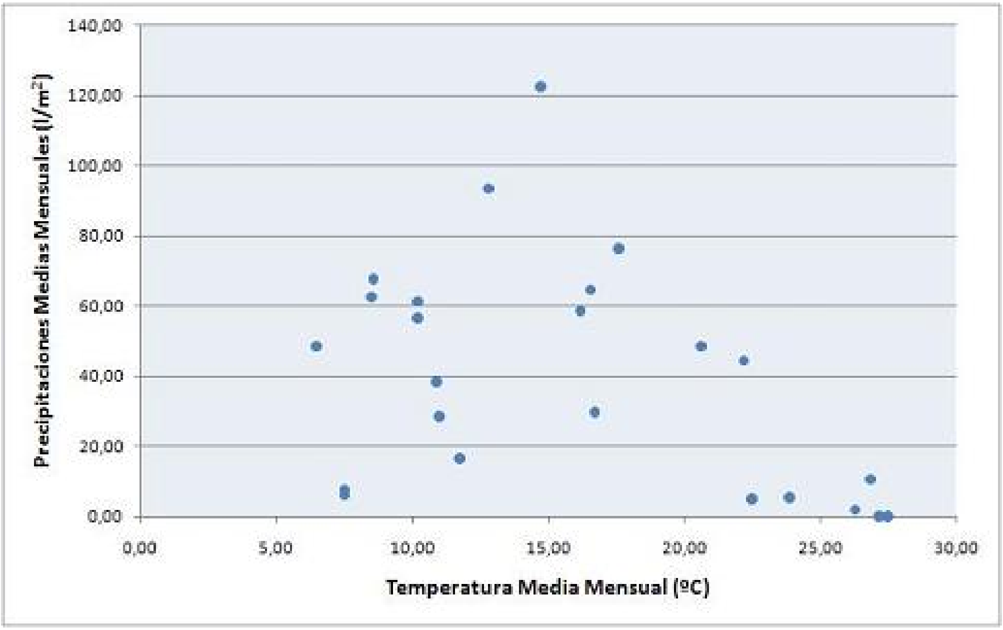
El punto $(\overline{X},\overline{Y})$ cuyas coordenadas son las medias aritméticas se le llama centro de gravedad o **centro de masas**

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:

La nube de puntos se representa sobre un par de ejes cartesianos. En este caso, cada punto representa un par de datos de la Variable Estadística Bidimensional.



En el siguiente vídeo puedes ver otro ejemplo con una explicación muy detallada:

Vídeo de estudiia alojado en **Youtube**

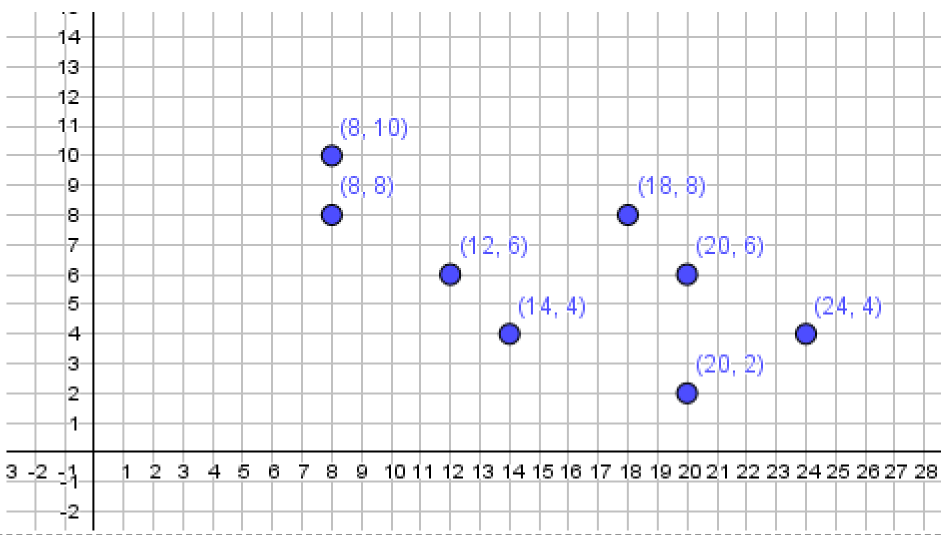
Ejercicio resuelto

Dada la tabla del apartado 2

Horas TV (X)	8	8	12	12	14	18	20	20
Horas deporte (Y)	8	10	6	10	4	8	2	6

Representa la nube de puntos.

Mostrar retroalimentación



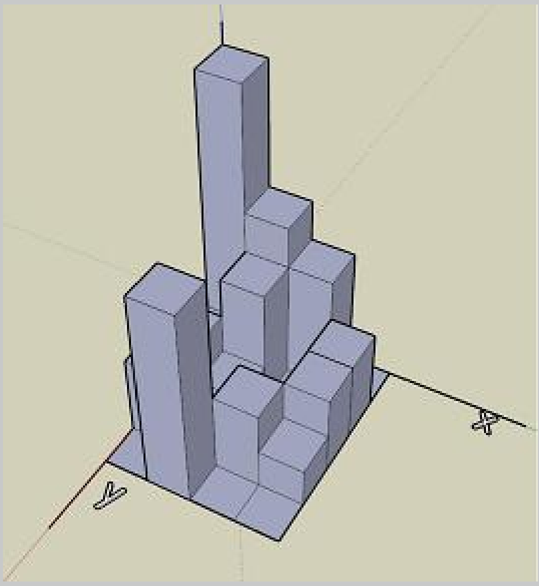
Para saber más

Representación gráfica para datos recogidos en 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
	$x_1=0$	7	1	2	2	0	12
	$x_2=1$	4	3	1	1	5	14
X	$x_3=2$	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:



Para representar la información partimos de tres ejes cartesianos.a) **Histograma tridimensional**:

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.

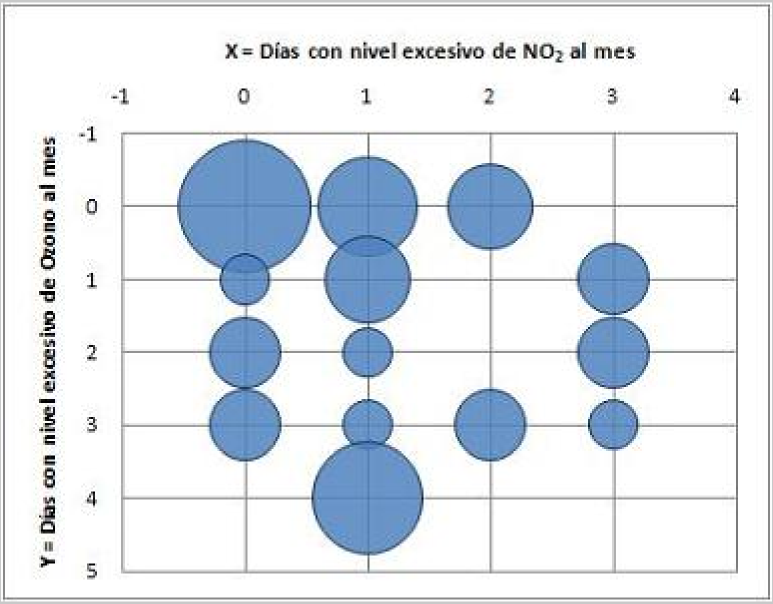
El siguiente sería el (1,4) que tiene frecuencia 5.

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.

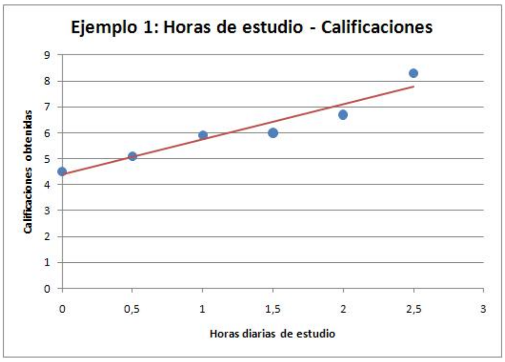


Dependencia

El objetivo de un estudio bidimensional es estudiar la relación entre dos variables, es decir, el grado de dependencia entre ambas. Las Nubes de Puntos nos ayudan a ver la dependencia entre las variables.

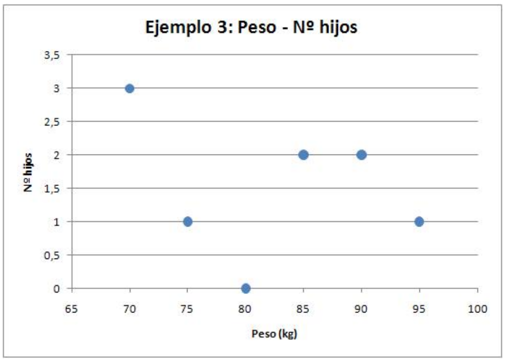
- **Dependencia positiva**: Al aumentar la variable X, también aumenta la Y.
- **Dependencia negativa**: Al aumentar la variable X, disminuye la Y.
- **Sin dependencia**: No se observa ninguna relación entre las dos variables.
- **Dependencia funcional**: Podemos encontrar una relación exacta entre ambas 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:

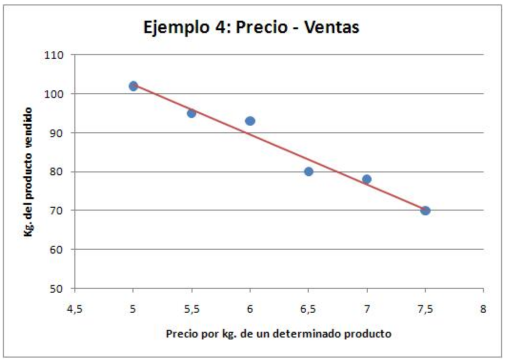


Dependencia positiva aleatoria

Dependencia positiva funcional



Sin dependencia



Dependencia negativa aleatoria

En la siguiente tema podrás ver un ejemplo muy completo sobre los tipos de dependencia y cómo estudiarlos con más profundidad.

Comprueba lo aprendido

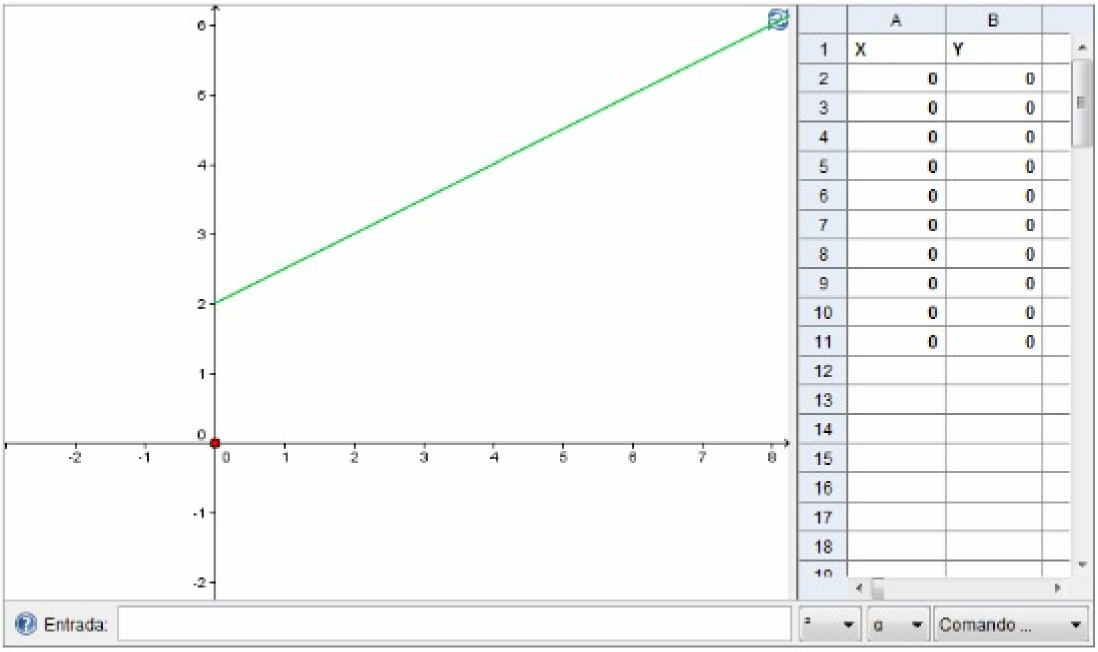
Una compañía química está estudiando el uso de un fertilizante líquido que han inventado en una determinada planta. Para ello miden dos variables: X = "cantidad diaria de fertilizante que se aporta a la planta (en ml)" e Y = "crecimiento de la planta al cabo de 10 días (en cm)".

Los resultados son los siguientes pares de datos:

X	1	2	3	4	5	6	7	8	9	10
Y	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

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

Importante

Recuerda que 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.

En el siguiente vídeo, puedes repasar este concepto con varios ejemplos:

Vídeo de estudiia alojado en **Youtube**

Los datos de las variables estadísticas se pueden organizar y representar en:

- Tablas simples y nubes de puntos
- Tablas de doble entrada o diagramas de burbujas.

Importante

El objetivo de un estudio bidimensional es estudiar la relación entre dos variables, es decir, el grado de dependencia entre ambas. Las Nubes de Puntos nos ayudan a ver la dependencia entre las variables.

- **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, 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.