

NUBES DE PUNTOS. CORRELACIÓN

Éstas son las notas de 12 estudiantes en Matemáticas y en Física:

Alumno	a	b	c	d	e	f	g	h	i	j	k	l

Es una **distribución bidimensional** porque a cada individuo le corresponden los valores de dos variables. Si tomamos esos dos valores como las coordenadas de un punto, la distribución puede ser representada mediante 12 puntos: **nube de puntos**.

Se aprecia una relación entre las dos variables: a mejor nota en *Matemáticas* mejor nota en *Física*, pero solo a grandes rasgos, *grosso modo*. Se dice que existe correlación entre esas dos variables.

Relacionemos ahora las notas de *Matemáticas* de los mismos alumnos con las de otra asignatura, *Filosofía*.

Tanto si nos fijamos en la tabla de datos como en la nube de puntos, apreciamos que también hay **correlación** entre estas dos variables, pero es **más débil** que la anterior.

Una jugadora de baloncesto lanza a canasta, desde distintas distancias, 10 balones cada vez. Lógicamente, encesta más cuanto más cerca está.

En este caso hay **correlación fuerte y negativa**, pues al aumentar una variable tiende a disminuir la otra.

La tendencia a variar conjuntamente las dos variables en una distribución bidimensional se marca mediante la **recta de regresión**. Cuanto más próximos estén los puntos a la recta, más fuerte es la correlación.

MEDIDA DE LA CORRELACIÓN

Hemos visto que la correlación entre dos variables (más o menos fuerte, positiva o negativa) se aprecia mediante el grado de apertura de los puntos de la nube. Vamos a confeccionar una fórmula que sirva para obtener su valor de forma numérica e inequívoca.

Centro de gravedad de una distribución bidimensional

MEDIA DE LA VARIABLE

MEDIA DE LA VARIABLE

El punto se llama **centro de gravedad** de la distribución.

Covarianza

Correlación

El **coeficiente de correlación**, r , tiene las siguientes propiedades:

No tiene dimensiones. Es decir, no depende de las unidades en las que se expresan los valores de las dos variables. Por tanto, si se realiza un cambio de unidades, el valor de r no varía.

El valor de r está comprendido entre -1 y 1 .

- Si la correlación es perfecta (puntos de la nube alineados), entonces $|r| = 1$, es decir, $r = 1$ ó $r = -1$.
- Si la correlación es fuerte, $|r|$ es próximo a 1.
- Si la correlación es débil, $|r|$ es próximo a 0.

EJERCICIOS RESUELTOS

2	1	4	1	2
3	3	9	9	9
4	2	16	4	8
4	4	16	16	16
5	4	25	16	20
6	4	36	16	24
6	6	36	36	36
7	4	49	16	28
7	6	49	36	42
8	7	64	49	56
10	9	100	81	90
10	10	100	100	100
72	60	504	380	431

Utilizando la fórmula anterior, calcular la correlación entre las variables nota en Matemáticas, x , nota en Física, y . Para ello, calcular previamente

El centro de gravedad es el punto (6,5). Observamos que este punto no tiene por que ser de la distribución.

Por tanto,

Es una correlación muy alta.

Método de los mínimos cuadrados

Partimos de la nube de puntos .

Hemos de encontrar la recta que mejor se ajuste a la nube.

¿Qué criterio seguimos para ese mejor ajuste?

Consideramos todas las posibles rectas $y = A + Bx$ y nos quedaremos con aquella para la cual los cuadrados de las distancias, $\sum d^2$, sumen lo menos posible: mínimo.

De ese modo se llega (utilizando métodos matemáticos superiores a este curso) a lo siguiente:

- La recta buscada pasa por el centro de gravedad de la distribución.
- Su pendiente es

El signo del coeficiente de correlación y el del coeficiente de regresión coinciden, pero aquí termina la

coincidencia: puede ser que la recta de regresión tenga pendiente alta y, sin embargo, el coeficiente de correlación sea bajo. O al contrario.

EJERCICIOS RESUELTOS

1. En la distribución Notas de Matemáticas–Notas de Física, cuyos parámetros hemos calculado en la página anterior, obtener la recta de regresión de Y sobre X.

- Pendiente:
- Ecuación:

La recta de regresión para hacer estimaciones

La recta de regresión se amolda a la nube de puntos y describe, a *grosso modo*, su tendencia. Por eso, a partir de la recta de regresión obtenemos, de forma aproximada, el valor esperado de y para un cierto valor de x, o viceversa. A estos valores se les llama *estimaciones*.

es el **valor estimado de y** correspondiente a sobre la recta de regresión.

es el **valor estimado de x** correspondiente a sobre la recta de regresión.

- Las estimaciones siempre se realizan aproximadamente y en términos de probabilidad [es probable que si entonces y valga, aproximadamente,].
- La aproximación es tanto mejor cuanto mayor sea $|r|$, pues para valores de r próximos a 1 o a -1 , los puntos están muy próximos a la recta.
- Las estimaciones sólo deben hacerse dentro del intervalo de valores utilizados o muy cerca de ellos.

EJERCICIOS RESUELTOS

1. La tabla adjunta da los alargamientos de una barra metálica por efecto de cambios en la temperatura. Calcular la recta de regresión y hacer algunas estimaciones.

TEMP. (°C)	ALARG. (mm)
0	0
8	1
16	2
25	3
40	5
50	6
60	7
75	9

La recta de regresión es .

Estimamos y para

Estimamos x para

Las estimaciones son buenas porque la correlación es muy grande. Además, está entre los valores manejados (entre 0°C y 75°C) y lo mismo le ocurre a .

No sería buena la estimación para

HAY DOS RECTAS DE REGRESIÓN

Como ya hemos dicho, la recta de regresión obtenida, es la **recta de regresión de Y sobre X**.

Si el criterio que siguiéramos para ajustar la recta a la nube de puntos fuera hacer mínima la suma de los cuadrados de las diferencias de abscisas del punto y de la recta (mínimo) obtendríamos otra recta llamada **recta de regresión de X sobre Y**.

Posiciones de las dos rectas de regresión

Cuando la correlación es casi nula, las dos rectas forman un ángulo muy grande (próximo a 90°):

Si la correlación es fuerte, el ángulo que forman las dos rectas es pequeño:

Si $|r|$ es próximo a 1, las rectas son casi coincidentes:

TABLAS DE DOBLE ENTRADA

Recordemos que las distribuciones de una variable, cuando el número de observaciones es pequeño, se dan, simplemente, enumerando los datos de forma ordenada. Pero cuando el número de datos es grande, se recurre a la tabla de frecuencias.

Del mismo modo, en las distribuciones bidimensionales, cuando hay pocos pares de valores se procede como hemos hecho hasta ahora: enumerándolos. Si algún par está repetido, se pone dos veces.

Pero cuando el número de datos es grande, se recurre a las tablas de **doble entrada**:

En cada casilla se pone la frecuencia correspondiente al par de valores que definen esa casilla. Por ejemplo, hay 11 individuos para los cuales $x=0$, $y=1$. Es decir, el par (0,1) está 11 veces.

La representación gráfica de estas distribuciones se hace:

Hinchando los puntos proporcionalmente a su frecuencia.

y_i	0	1	2	3	4
0	24	6	1	0	0
1	11	19	2	3	0
2	7	8	6	2	0
3	2	3	3	7	1

4	1	0	2	4	5
---	---	---	---	---	---

O bien levantando barras de altura proporcional a la frecuencia de cada casilla.

CUESTIONES TEÓRICAS

20. ¿Qué punto tienen en común las dos rectas de regresión? El centro de gravedad.

21. ¿Qué condición debe cumplir r para que las estimaciones hechas con la recta de regresión sean fiables? Que el valor este comprendido entre 1 y -1 .

22. Prueba que el producto de los coeficientes de regresión y es igual al cuadrado del coeficiente de correlación.

23. De una distribución bidimensional (x, y) conocemos los siguientes resultados:

- Recta de regresión de Y sobre X :

$$y = 8,7 - 0,76x$$

- Recta de regresión de X sobre Y :

$$y = 11,36 - 1,3x$$

- Calcula el centro de gravedad de la distribución.
- Halla el coeficiente de correlación.

8

Se llama covarianza al parámetro:

Ambas expresiones, como es lógico, coinciden. La segunda de ellas es más cómoda para obtener numéricamente la covarianza.

El valor de la correlación entre las dos variables de una distribución bidimensional viene dado por la expresión:

es la covarianza

son las desviaciones típicas de cada variable

La recta que hace mínima la suma, tiene por ecuación:

Se llama **recta de regresión de Y sobre X** .

A la pendiente, b , se la llama **coeficiente de regresión**.

La ecuación de la **recta de regresión de X sobre Y** es:

El número b se llama **coeficiente de regresión de X sobre Y** .

No es la pendiente de la recta, sino su inversa.

y_i