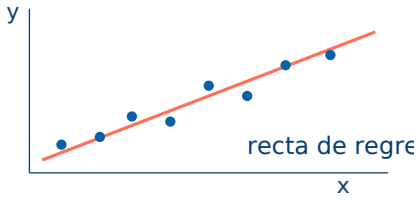
Apuntes - Estadística bidimensional

Curso: 1 Bachillerato Ccss

Con frecuencia interesa saber si dos variables están relacionadas: la publicidad y las ventas, las horas de estudio y la nota, el precio y la demanda. La estadística bidimensional estudia pares de datos (x_i, y_i) y permite medir esa relación y hacer predicciones.

Actividad de introducción: ¿más horas, mejor nota?

Pregunta a diez compañeros cuántas horas estudian a la semana y qué nota sacaron en el último examen, y escribe cada pareja como un punto en un papel cuadriculado. Es probable que los puntos formen una nube que sube de izquierda a derecha, aunque no todos estén sobre una recta: algún alumno estudia poco y saca buena nota, y otro estudia mucho y aprueba justo. La estadística bidimensional estudia precisamente este tipo de datos: dos variables observadas a la vez en los mismos individuos. Nos permite medir el grado de relación con un número, el coeficiente de correlación, dibujar la recta que mejor resume la nube y utilizarla para hacer predicciones razonables. También nos enseña a desconfiar de conclusiones precipitadas: que dos variables estén relacionadas no significa que una sea la causa de la otra.



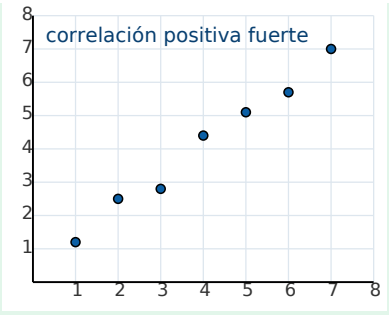
Distribuciones bidimensionales y nube de puntos

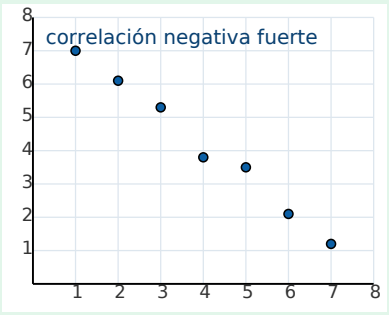
Definición: nube de puntos

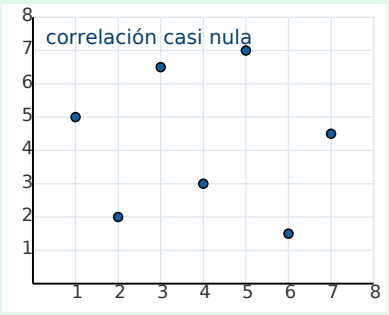
Una **variable estadística bidimensional** (X, Y) recoge dos datos por individuo. Se organiza en una tabla de pares o en una **tabla de doble entrada** (frecuencias de cada par, con distribuciones marginales en los márgenes). Su representación gráfica es la **nube de puntos** o diagrama de dispersión.

Propiedad: tipos de correlación

Si los puntos suben de izquierda a derecha, la **correlación es positiva** (directa); si bajan, **negativa** (inversa); si no siguen ninguna pauta, la correlación es **nula**. Cuanto más alineados estén los puntos, más **fuerte** es la correlación.







Covarianza y coeficiente de correlación

Definición: covarianza

La **covarianza** es

$$\sigma_{xy} = \frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y}.$$

Su signo indica el tipo de relación: positivo, correlación directa; negativo, inversa. Depende de las unidades.

Definición: coeficiente de correlación lineal

$$r = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y}, \qquad -1 \leq r \leq 1.$$

No depende de las unidades. $|r|$ cerca de 1: relación lineal fuerte; cerca de 0: no hay relación lineal. $r = \pm 1$: todos los puntos en una recta.

Método: calcular r con una tabla

1. Calcula $\bar{x}$, $\bar{y}$, $\sum x_i y_i$, $\sum x_i^2$ y $\sum y_i^2$.
2. Halla $\sigma_x^2 = \frac{\sum x_i^2}{n} - \bar{x}^2$, σ_y^2 y σ_{xy}.
3. Calcula $r = \sigma_{xy}/(\sigma_x \sigma_y)$.
4. Interpreta: signo y fuerza.

Ejemplo 1

Gasto en publicidad x (miles de €) y ventas y (decenas de miles de €): $(1,2), (2,4), (3,5), (4,4), (5,6)$. Calcula r.

$\bar{x} = 3$, $\bar{y} = 4{,}2$. $\sum xy = 71$, $\sum x^2 = 55$, $\sum y^2 = 97$. $\sigma_x^2 = 2{,}00$, $\sigma_y^2 = 1{,}76$, $\sigma_{xy} = 1{,}60$.

$r = \dfrac{1{,}60}{\sqrt{2{,}00 \cdot 1{,}76}} = 0{,}853$: correlación positiva fuerte.

Recta de regresión

Regla: recta de regresión de Y sobre X

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

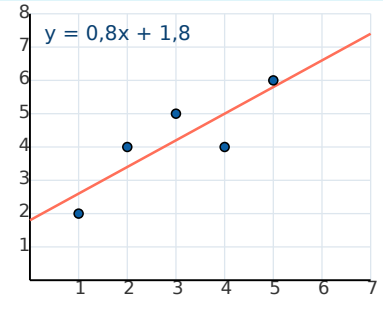
Pasa por el punto medio $(\bar{x}, \bar{y})$, y su pendiente $\dfrac{\sigma_{xy}}{\sigma_x^2}$ es el **coeficiente de regresión**.

Sirve para **predecir** y a partir de x. La de X sobre Y es $x - \bar{x} = \dfrac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$.

Ejemplo 2

Recta de regresión para los datos del ejemplo 1 y predicción para $x = 6$.

Pendiente $\dfrac{1{,}60}{2{,}00} = 0{,}80$. $y - 4{,}2 = 0{,}80(x - 3) \Rightarrow y = 0{,}80x + 1{,}80$. Para $x = 6$: $y = 6{,}6$ decenas de miles de euros.



Propiedad: fiabilidad de la predicción

Una predicción es tanto más fiable cuanto mayor es $|r|$ (el **coeficiente de determinación** r^2 indica la proporción de variación de Y explicada por X) y cuanto mayor es el número de datos. Predecir dentro del rango de datos (**interpolación**) es más fiable que hacerlo fuera (**extrapolación**).

¡Atención!

- Correlación no significa causalidad: dos variables pueden crecer a la vez por una causa común (por ejemplo, las ventas de helados y de gafas de sol).
- Si r es próximo a 0, la recta no sirve para predecir.
- No se extrapola lejos del rango de los datos.
- Las dos rectas de regresión solo coinciden si $r = \pm 1$.

Ejemplo 3: horas de estudio y nota

Horas de estudio x: $1, 2, 3, 4, 5, 6$; notas y: $3, 4, 5, 5, 7, 8$. Calcula r y la recta.

$\bar{x} = 3{,}5$, $\bar{y} = 5{,}333$, $\sigma_{xy} = 2{,}833$, $\sigma_x^2 = 2{,}917$, $\sigma_y^2 = 2{,}889$. $r = 0{,}976$ (muy fuerte). Recta: $y = 0{,}971x + 1{,}933$. Con 3,5 horas se predice nota $5{,}33$.

Observación

La covarianza cambia si cambiamos las unidades (de euros a miles de euros, por ejemplo), pero el coeficiente r no. Si r es positivo, ambas rectas de regresión tienen pendiente positiva; si es negativo, ambas negativa. Los datos económicos reales casi nunca tienen $r = \pm 1$: se buscan relaciones aproximadas y se calcula su fiabilidad antes de usar la recta para decidir.

Ampliación: tablas de doble entrada

Cuando las dos variables son cualitativas o están agrupadas en pocos valores, los datos se presentan en una tabla de doble entrada, que muestra la frecuencia de cada pareja de valores. Sumando por filas y por columnas se obtienen las **distribuciones marginales**, que describen cada variable por separado. Fijando un valor de una variable y calculando las proporciones de la otra se obtienen las **distribuciones condicionadas**. Si las proporciones condicionadas son iguales a las marginales, las variables son independientes; si cambian, hay relación entre ellas.

Ejemplo 4

En una encuesta a 200 clientes de un centro comercial, 120 son mujeres. 54 mujeres y 30 hombres compran ropa deportiva. ¿Hay relación entre el sexo y la compra de ropa deportiva?

Entre las mujeres compran ropa deportiva $\frac{54}{120} = 45\,\%$; entre los hombres, $\frac{30}{80} = 37{,}5\,\%$. En el conjunto, $\frac{84}{200} = 42\,\%$. Las proporciones condicionadas (45 % y 37,5 %) son distintas, pero próximas: hay una ligera relación, que con una muestra de 200 personas podría ser casual.

Ampliación: el procedimiento paso a paso con una tabla de cálculo

Método: calcular r y la recta de regresión

1. Construye una tabla con las columnas x, y, x^2, y^2 y xy y suma cada una.
2. Calcula las medias $\bar{x} = \sum x/n$ e $\bar{y} = \sum y/n$.
3. Calcula las varianzas $\sigma_x^2 = \sum x^2/n - \bar{x}^2$ y $\sigma_y^2 = \sum y^2/n - \bar{y}^2$, y la covarianza $\sigma_{xy} = \sum xy/n - \bar{x}\bar{y}$.
4. Calcula $r = \sigma_{xy}/(\sigma_x \sigma_y)$ y decide el tipo y la fuerza de la correlación.
5. Escribe la recta $y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$ y úsala para estimar.

Ejemplo 5

Con los datos $(1,2)$, $(2,3)$, $(3,5)$, $(4,6)$ calcula r y la recta de y sobre x.

$n = 4$: $\sum x = 10$, $\sum y = 16$, $\sum x^2 = 30$, $\sum y^2 = 74$, $\sum xy = 47$. Medias: $\bar{x} = 2{,}5$, $\bar{y} = 4$. Varianzas: $\sigma_x^2 = 7{,}5 - 6{,}25 = 1{,}25$ y $\sigma_y^2 = 18{,}5 - 16 = 2{,}5$. Covarianza: $\sigma_{xy} = 11{,}75 - 10 = 1{,}75$. Entonces $r = \frac{1{,}75}{\sqrt{1{,}25 \cdot 2{,}5}} = 0{,}990$ y la pendiente es $\frac{1{,}75}{1{,}25} = 1{,}4$. La recta es $y - 4 = 1{,}4(x - 2{,}5)$, es decir, $y = 1{,}4x + 0{,}5$.

Propiedad: cómo valorar el coeficiente de correlación

- $-1 \le r \le 1$, y r no tiene unidades.
- Si $|r|$ está cerca de 1, la relación lineal es fuerte; si está cerca de 0, es débil o inexistente.
- $r > 0$ indica relación directa y $r < 0$ relación inversa.
- El coeficiente de determinación r^2 indica la proporción de variación de y explicada por la recta.
- Las dos rectas de regresión se cortan en $(\bar{x}, \bar{y})$; cuanto mayor es $|r|$, más próximas están.

Ampliación: predecir con prudencia

Una recta de regresión es una herramienta de predicción, pero tiene límites. Dentro del intervalo de valores observados se habla de **interpolación**, y la estimación es tanto más fiable cuanto más cerca de 1 esté $|r|$. Fuera de ese intervalo se habla de **extrapolación**, que puede ser absurda: un modelo lineal del precio de un coche usado puede dar un precio negativo para un coche muy antiguo. Además, hay que conservar el sentido de las unidades: la pendiente se mide en unidades de y por unidad de x (por ejemplo, euros por metro cuadrado o clientes por anuncio) y siempre debe interpretarse en el contexto del problema. Por último, la estimación de x a partir de y debe hacerse con la recta de regresión de x sobre y, no despejando de la de y sobre x.

Ejemplo 6

Una cafetería de Salamanca ajusta la recta $y = 2{,}5x + 40$, donde x son los grados de temperatura exterior e y los cafés fríos vendidos al día, con datos entre 15 y 35 grados. Interpreta la pendiente y estima las ventas con 30 grados.

La pendiente 2,5 significa que, por cada grado más, se venden unos 2,5 cafés fríos más. Con 30 grados: $y = 2{,}5 \cdot 30 + 40 = 115$ cafés. Al estar 30 dentro del intervalo 15-35, es una interpolación y la estimación es razonable. En cambio, con 2 grados el modelo daría 45 cafés fríos, una extrapolación dudosa. Además, el valor 40 de la ordenada no tiene por qué tener sentido real: representa lo que predice la recta para 0 grados, fuera del rango estudiado.

Errores muy comunes en estadística bidimensional

- Pensar que una correlación alta demuestra una relación de causa y efecto: puede haber una variable oculta que explique ambas.
- Olvidar que r solo mide relaciones lineales: unos datos en forma de U pueden tener $r \approx 0$ y estar muy relacionados.
- Extrapolar muy lejos de los datos observados.
- Confundir las dos rectas de regresión o despejar x de la recta de y sobre x.
- No interpretar la pendiente con sus unidades.
- Dejar que un dato atípico distorsione el análisis sin revisarlo.

¿Sabías que?

La palabra «regresión» procede del científico británico Francis Galton, que a finales del siglo XIX observó que los hijos de padres muy altos solían ser algo más bajos que sus padres, y los de padres muy bajos algo más altos: sus alturas «regresaban» hacia la media. Su discípulo Karl Pearson definió el coeficiente de correlación que usamos hoy. Los datos que manejó Galton, alturas de padres e hijos, son todavía un ejemplo clásico en los libros de estadística.

El consejo de Otto

Dibuja siempre la nube de puntos antes de calcular nada: si no se ve una pauta lineal, la recta de regresión no tiene sentido aunque la calculadora te dé un número.