Apuntes - Estadística bidimensional

Curso: 1 Bachillerato Ciencias

En este tema aprenderás: distribución bidimensional, parámetros: medias, varianzas y covarianza, correlación lineal, rectas de regresión y ajustes no lineales. Cada idea va en un cuadro de color, con ejemplos resueltos paso a paso.

Actividad de introducción: ¿estudiar más sube la nota?

Un profesor de un instituto de Salamanca sospecha que los alumnos que dedican más horas a estudiar obtienen mejores notas. Pregunta a ocho estudiantes y anota dos datos por cada uno: las horas semanales de estudio y la nota final. Con una variable sola se calcula una media; con dos variables a la vez la pregunta es otra: **¿están relacionadas?** Si lo están, ¿cómo de fuerte es la relación? Y, lo más útil, ¿podemos predecir la nota de un alumno que estudia seis horas? La estadística bidimensional responde con tres herramientas: la nube de puntos, que permite ver la relación; el coeficiente de correlación, que la mide con un número entre -1 y 1; y la recta de regresión, que sirve para predecir.

Distribución bidimensional

Propiedad

La estadística bidimensional estudia dos variables X e Y medidas sobre los mismos individuos. Los datos son pares (x_i, y_i) y se representan en una **nube de puntos** (diagrama de dispersión). Si hay muchos datos repetidos se organizan en una **tabla de doble entrada**, con frecuencias conjuntas f_{ij} y **distribuciones marginales** (la suma de cada fila o columna). La **distribución condicionada** estudia una variable fijando un valor de la otra: $f(y \mid x) = \frac{f_{xy}}{f_x}$.

Ejemplo 1

De 40 alumnos, 18 aprueban Matemáticas y Física, 4 aprueban Matemáticas y suspenden Física, 3 suspenden Matemáticas y aprueban Física, y 15 suspenden ambas.

Aprueban Matemáticas $18 + 4 = 22$; aprueban Física $18 + 3 = 21$. La frecuencia relativa conjunta (aprueba las dos) es $\frac{18}{40} = 0{,}45$, y la proporción que aprueba Física entre quienes aprueban Matemáticas es $\frac{18}{22} \approx 0{,}82$.

Parámetros: medias, varianzas y covarianza

Propiedad

Se calculan las medias $\bar{x}$ y $\bar{y}$ (el **centro de gravedad** de la nube), las varianzas $\sigma_x^2 = \frac{\sum x_i^2}{n} - \bar{x}^2$, $\sigma_y^2 = \frac{\sum y_i^2}{n} - \bar{y}^2$ y la **covarianza**

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

La covarianza es positiva si al crecer x tiende a crecer y, y negativa en caso contrario.

Ejemplo 2

Con los datos $(2,3), (4,5), (6,8), (8,9), (10,12)$: $\bar{x} = 6$, $\bar{y} = 7{,}4$; $\sum x^2 = 220$, $\sigma_x^2 = 44 - 36 = 8$; $\sum y^2 = 323$, $\sigma_y^2 = 64{,}6 - 54{,}76 = 9{,}84$; $\sum xy = 266$, $\sigma_{xy} = 53{,}2 - 44{,}4 = 8{,}8$.

Correlación lineal

Definición: coeficiente de correlación lineal

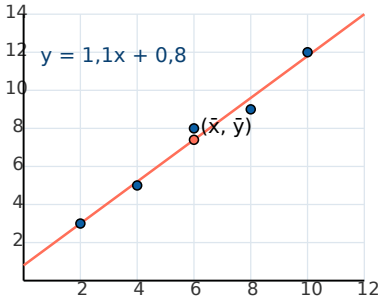
El **coeficiente de correlación lineal** es $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, con $-1 \le r \le 1$. Cuanto más cerca de ± 1, más se ajusta la nube a una recta; cerca de 0, no hay relación lineal. El signo indica si la correlación es directa ($r > 0$) o inversa (r). Aquí $r = \frac{8{,}8}{\sqrt{8 \cdot 9{,}84}} \approx 0{,}992$: correlación positiva muy fuerte. El **coeficiente de determinación** r^2 es la proporción de variación de Y explicada por el modelo lineal ($r^2 \approx 0{,}984$, un 98,4 %).

Rectas de regresión

Definición: recta de regresión de Y sobre X

La **recta de regresión de Y sobre X** es $y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$; minimiza los cuadrados de las distancias verticales y sirve para **predecir** y a partir de x. La de X sobre Y es $x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$. Ambas pasan por $(\bar{x}, \bar{y})$, y el producto de sus pendientes es r^2.

Para el ejemplo: $m = \frac{8{,}8}{8} = 1{,}1$, así que $y - 7{,}4 = 1{,}1(x - 6)$, es decir, $y = 1{,}1x + 0{,}8$.

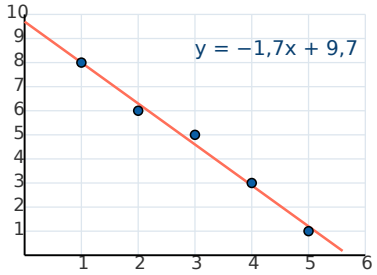


Propiedad

Para $x = 5$ predice $y = 6{,}3$, una **interpolación** fiable; para $x = 12$, $y = 14$, una **extrapolación** menos fiable porque está fuera del rango de datos. La predicción es tanto más fiable cuanto mayor es $|r|$ y cuantos más datos haya.

Ejemplo 3: correlación negativa

Con $(1,8), (2,6), (3,5), (4,3), (5,1)$: $\bar{x} = 3$, $\bar{y} = 4{,}6$, $\sigma_{xy} = -3{,}4$, $\sigma_x^2 = 2$, $\sigma_y^2 = 5{,}84$, $r \approx -0{,}995$ y la recta es $y = -1{,}7x + 9{,}7$.



Cómo se hace: obtener una recta de regresión y usarla

1. Calcula $\bar{x}$, $\bar{y}$, σ_x^2, σ_y^2 y σ_{xy}.
2. La recta de Y sobre X es $y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$.
3. Calcula $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$: si $|r|$ es cercano a 1, la predicción es fiable.
4. Predice solo dentro del rango de los datos o muy cerca de él.

Ejemplo 4: coeficiente de determinación

Si $r = -0{,}8$, ¿qué porcentaje de la variación de y explica la recta?

El coeficiente de determinación es $r^2 = 0{,}64$: la recta explica el 64 % de la variación de y. La correlación es negativa y bastante fuerte.

Ajustes no lineales

Propiedad

Si la nube sigue una curva, un modelo lineal no es adecuado aunque r sea alto. Los puntos $(1,1), (2,4), (3,9), (4,16)$ siguen $y = x^2$ (**regresión cuadrática**); la recta $y = 5x - 5$ da $r \approx 0{,}984$, pero en $x = 5$ predice 20 cuando el valor real es 25. Conviene siempre mirar la nube antes de calcular r.

¡Atención! Errores frecuentes

- Interpretar correlación como causalidad: que dos variables se muevan juntas no significa que una cause la otra.
- Usar la recta de Y sobre X para predecir x.
- Extrapolar muy lejos del rango de datos.
- Olvidar que r mide solo relación *lineal*.

Tablas de doble entrada: distribuciones marginales y condicionadas

Definición: distribución marginal y condicionada

Cuando los datos se agrupan en una **tabla de doble entrada**, cada casilla contiene la frecuencia f_{ij} de los alumnos que tienen a la vez el valor x_i y el valor y_j. La **distribución marginal** de X se obtiene sumando cada fila (ignorando Y); la de Y se obtiene sumando cada columna. Una **distribución condicionada** estudia una variable solo entre los individuos que cumplen una condición sobre la otra, por ejemplo, Y entre los que tienen $X = 2$.

Ejemplo 5

En una tabla de 40 alumnos hay 10 que tienen $X = 1$, y de ellos 4 tienen $Y = 2$. Calcula la frecuencia relativa de $Y = 2$ condicionada a $X = 1$.

Hay que dividir entre el total de la condición, no entre 40: la frecuencia condicionada es $\frac{4}{10} = 0{,}4$, es decir, el 40 %. Si dividiéramos entre 40 se obtendría la frecuencia conjunta $\frac{4}{40} = 0{,}1$, que responde a otra pregunta distinta.

Cómo interpretar el coeficiente de correlación

Propiedad: valores de r

El coeficiente $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ siempre cumple $-1 \leq r \leq 1$. El signo es el de la covarianza: positivo si las variables crecen juntas, negativo si una crece mientras la otra decrece. Su valor absoluto mide la fuerza: si $|r|$ es mayor que $0{,}9$ la correlación es muy fuerte; entre $0{,}7$ y $0{,}9$ es fuerte; entre $0{,}4$ y $0{,}7$ es moderada; y por debajo de $0{,}4$ es débil. Si $r = \pm 1$ todos los puntos están sobre una recta y si $r = 0$ no hay relación *lineal*.

Propiedad: coeficiente de determinación

El cuadrado r^2 se llama **coeficiente de determinación** y es el porcentaje de la variación de Y que explica la recta de regresión. Con $r = 0{,}9$ tenemos $r^2 = 0{,}81$: el $81\,\%$ de la variación de Y se explica por X y el $19\,\%$ restante depende de otros

factores. Además, $r^2 = b_{yx} \cdot b_{xy}$, el producto de las pendientes de las dos rectas.

¡Atención! correlación no es causalidad

Que dos variables estén muy correlacionadas no significa que una cause la otra. Las ventas de helados y los rescates en las playas crecen a la vez porque ambas dependen del calor y de la cantidad de gente, no porque los helados provoquen accidentes. Antes de afirmar una causa hace falta un argumento fuera de las matemáticas. Tampoco olvides que $r = 0$ solo indica ausencia de relación lineal: unos datos pueden seguir una parábola perfecta y tener r casi nulo.

Fiabilidad de una predicción

Cómo se hace: predecir con la recta de regresión

1. Dibuja la nube de puntos y comprueba que la forma es aproximadamente lineal.
2. Calcula $\bar{x}$, $\bar{y}$, las varianzas y la covarianza, y después r.
3. Elige la recta adecuada: la de Y sobre X si quieres estimar y a partir de x, y la de X sobre Y en el caso contrario.
4. Sustituye el valor conocido y calcula la estimación.
5. Valora la fiabilidad: ¿$|r|$ es grande? ¿El valor está dentro del rango de los datos (**interpolación**) o fuera (**extrapolación**)? ¿El número de datos es suficiente?

Ejemplo 6

Una recta de regresión de la nota (Y) sobre las horas de estudio (X) es $y = 0{,}8x + 2$, con datos entre 1 y 8 horas y $r = 0{,}95$. Estima la nota para 5 horas y para 20 horas.

Con 5 horas: $y = 0{,}8 \cdot 5 + 2 = 6$, fiable porque es una interpolación y la correlación es alta. Con 20 horas saldría $y = 18$, que no es posible en una escala sobre 10: es una extrapolación absurda. Una buena predicción exige que el valor esté dentro del rango observado.

Cambios de unidad y ajustes no lineales

Propiedad: invariancia de r

Si cambiamos las unidades con una transformación lineal $x' = ax + b$, $y' = cy + d$ con a y c del mismo signo, el coeficiente de correlación no cambia; si tienen signo distinto, solo cambia el signo de r. En cambio, las pendientes de las rectas sí dependen de las unidades: pasar la altura de centímetros a metros multiplica por 100 la pendiente de la recta del peso sobre la altura.

Ejemplo 7

Una muestra de pares (semana, seguidores) crece como $3,\ 6,\ 12,\ 24,\ \ldots$ ¿Puede ajustarse con una recta?

No bien: los datos se duplican cada semana, de modo que siguen $y = 3 \cdot 2^{x}$. Si tomamos logaritmos, $\ln y = \ln 3 + x \ln 2$, una recta de pendiente $\ln 2 \approx 0{,}693$. Basta calcular la regresión lineal de $\ln y$ sobre x y deshacer el logaritmo con la exponencial. Lo mismo ocurre con los modelos potenciales $y = a x^{b}$, que se linealizan tomando logaritmos de las dos variables.

¿Sabías que? La regresión hacia la media

La palabra «regresión» nació con Francis Galton, un científico inglés del siglo XIX. Midió la estatura de padres e hijos y observó que los hijos de padres muy altos eran, de media, altos pero algo menos que sus padres, y los de padres muy bajos eran algo más altos que los suyos. Los valores extremos «regresaban» hacia la media. Su discípulo Karl Pearson dio la fórmula del coeficiente de correlación que hoy llamamos *r de Pearson*.

El consejo de Otto

Dibuja siempre la nube de puntos antes de calcular nada: el ojo detecta enseguida si una recta es un buen modelo.