# Tema 8

# Análisis de dos variables: dependencia estadística y regresión

## Contenido



## 8.1. Introducción

Las **distribuciones bidimensionales** recogen la información de dos características o variables medidas sobre los mismos individuos.

Hay dos formas básicas de representar la información de las distribuciones bidimensionales: las tablas de datos apareados y las tablas de doble entrada o tablas de contingencia.

Las **tablas de datos apareados** se utilizan cuando los distintos pares de modalidades se repiten pocas veces y representan el listado de datos de todos los individuos de la muestra.

Las **tablas de doble entrada** o **tablas de contingencia** muestran las modalidades de una de las variables en la primera fila, las de la otra en la primera columna y en el cruce de cada par de modalidades, muestra la frecuencia con la que aparecen a la vez esos dos valores.

A veces es necesario estudiar cada una de las características por separado, a pesar de disponer de datos bidimensionales. Cuando se tienen datos apareados, esto se puede hacer trivialmente considerando la fila (o columna) correspondiente a cada variable por separado. Cuando se tienen tablas de doble entrada, para conseguir la frecuencia de cada valor se debe sumar la frecuencia de cada fila o columna. Estas frecuencias se suelen anotar en el margen de la tabla, por lo que se llaman **distribuciones marginales**.

Una vez que se tienen las distribuciones marginales, se pueden realizar los mismos análisis que se planteaban en temas anteriores con cada una de las variables por separado.

En ocasiones, interesa trabajar sólo con una parte de los datos que se tienen. Las distribuciones de frecuencias de este tipo reciben el nombre de **distribuciones condicionadas**, porque se seleccionan los datos que verifican una condición.

**Problemas propuestos:** Problemas 8.1 y 8.2.

## 8.2. Dependencia/independencia estadística

Se dice que dos **variables** son **estadísticamente independientes** cuando conocer el valor que toma una de ellas no aportaría ninguna información acerca de la distribución de la otra variable.

En general se puede comprobar si dos variables son estadísticamente independientes verificando si las distribuciones relativas de una variable condicionada a cualquier valor de las otras son las mismas. Matemáticamente se puede comprobar que dos variables son independientes si la *frecuencia relativa de cada casilla es igual al producto de las marginales relativas correspondientes*. Una forma muy común de comprobar la independencia es observar si se verifica esa condición para *todas* las casillas, aunque la mayor parte de los programas estadísticos ayudan a verificar la independencia sin necesidad de realizar operaciones.

**Problema propuesto:** Problema 8.3.

## 8.3. Representación gráfica: diagrama de dispersión

Para detectar si existe algún tipo de relación o dependencia entre dos variables cardinales es muy útil dibujarlas para visualizar cómo es esa relación. Para ello se suele utilizar un gráfico denominado **nube de puntos** o **diagrama de dispersión** consistente en representar sobre un eje de coordenadas todos los pares de modalidades que aparezcan en la muestra.

Habitualmente se representa en el eje de las equis lo que se llama la **variable independiente**, que se suele denotar por $X$, y en el eje de las ies la **variable dependiente**, que se suele denotar por $Y$. En problemas en los que hay algún tipo de causa-efecto lógico, la variable dependiente $Y$ es la que se cree que varía en función de la otra (es decir, en función de la independiente $X$). En otro caso $Y$ será la que se pretende aproximar o predecir una vez que se conoce el valor de $X$.

**Problema propuesto:** Apartado a) del Problema 8.4.

## 8.4. Regresión

Al realizar un diagrama de dispersión entre dos variables $X$ e $Y$ pueden surgir algunas de las siguientes situaciones representadas en la Figura 8.1.

En la Figura 8.1 (a) se observa una relación matemática exacta entre $X$ e $Y$, es decir, dado un valor de $X$ podemos calcular el valor de $Y$ mediante una fórmula (dependencia matemática).

En la Figura 8.1 (b) no se observa ninguna relación entre las variables, es decir, conocer $X$ no sirve en absoluto para calcular $Y$ (independencia estadística).

En las Figuras 8.1 (c) y (d) aunque no hay una dependencia matemática exacta, sí que se observa una relación aproximada (dependencia estadística).

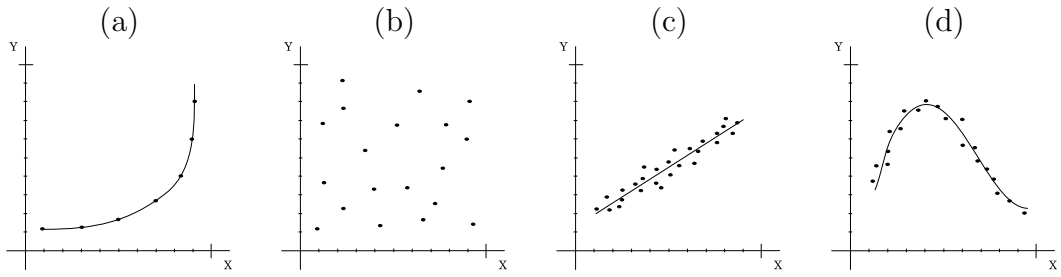


Figura 8.1: Nubes de puntos para distintas relaciones entre $X$ e $Y$.

En concreto en la Figura 8.1 (c) la nube de puntos se "parece" a una recta. Si se sabe calcular la ecuación de esa recta, se podría "aproximar" el valor de $Y$ una vez conocido el valor de $X$. El problema de encontrar la ecuación de la recta que más se parezca (o que mejor se ajuste) a la nube de puntos se conoce como **regresión lineal** y es lo que se analizará en la próxima sección.

La Figura 8.1 (d) es similar a la (c), aunque en este caso la nube de puntos se "parece" a una curva y no a una recta. Encontrar la ecuación de esa curva sería un problema de **regresión no lineal** y eso se tratará un poco más adelante.

Aunque para la regresión lineal se mostrarán las fórmulas, se recomienda el uso de programas estadísticos para realizar los cálculos. Las fórmulas de las regresiones no lineales son muchas veces aproximadas y se obtienen realizando transformaciones que no son el objetivo de este curso.

### 8.4.1. Regresión lineal

La regresión lineal es la recta que mejor aproxima la variable $Y$ para cada punto fijado de la variable $X$ en media. Como la fórmula de cualquier recta es $y(x) = b_0 + b_1 x$, para determinarla, basta calcular los valores concretos $\widehat{b}_0$ y $\widehat{b}_1$ a partir de los datos de la muestra $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ que hacen que $\widehat{y}(x) = \widehat{b}_0 + \widehat{b}_1 x$ sea la que más se aproxima a la nube de puntos. Se puede comprobar que

$$\widehat{b}_1 = \frac{S_{xy}}{S_x^2} \quad \text{y} \quad \widehat{b}_0 = \overline{y} - \widehat{b}_1 \overline{x},$$

donde $S_{xy} = \overline{xy} - \overline{xy}$ es la covarianza de $X$ e $Y$. Para calcular la covarianza hay que calcular primero la media del producto, que involucra el producto de todos los datos y su frecuencia. Cuando se tienen $n$ datos apareados es simplemente

$$\overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n}.$$

La recta de regresión se puede utilizar para explicar la relación aproximada entre dos variables. El valor de $\widehat{b}_1$ dice cuánto cambia $y$ por cada unidad en la que se incremente $x$ (aprox.).

La recta de regresión también se puede utilizar para realizar predicciones si se conoce un valor de la variable independiente **que se encuentre entre el mínimo y el máximo de la muestra** (interpolación). No se puede utilizar, sin embargo, si el valor de la variable independiente está fuera de ese rango (extrapolación) porque

las condiciones fuera de lo recogido por la muestra podrían cambiar y por tanto la recta hallada podría no ser válida.

**Problema propuesto:** Apartado b) del Problema 8.4.

### 8.4.2. Correlación lineal

En la sección anterior se buscaba la forma de encontrar la fórmula de la recta que mejor se aproximase a la nube de puntos para poder hacer predicciones a partir de ella. Sin embargo, para poder confiar en esas predicciones hay que comprobar que esa aproximación es buena. Los estudios de **correlación** tratan de medir cómo de buena es la recta (o, en general, más adelante será la curva) de regresión para realizar predicciones.

La recta de regresión será una buena aproximación cuando los puntos de la muestra están próximos a ella (ver Figura 8.2 (a)) y será mala cuando estén alejados (ver Figura 8.2 (b)).

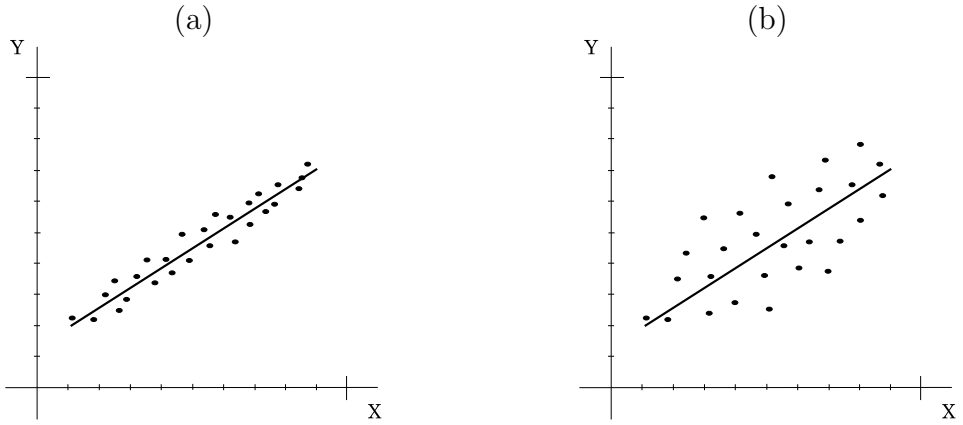


Figura 8.2: Nubes de puntos para distintas correlaciones entre $X$ e $Y$.

La fiabilidad se puede cuantificar numéricamente mediante el **coeficiente de determinación**, que se denota habitualmente por $R^2$ y que indica la proporción de variación de la variable $Y$ que se explica por su relación lineal con $X$ (es decir, "la parte de $Y$" que queda determinada por la recta).

El coeficiente de determinación es el cuadrado del **coeficiente de correlación de Pearson**, $R$, también muy utilizado, aunque con una interpretación menos intuitiva. Sus fórmulas son:

$$R = \frac{S_{xy}}{S_x S_y} \quad \text{y} \quad R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}.$$

Como $R^2$ es una proporción, siempre toma valores entre 0 y 1. Si $R^2 = 0$, significa que la recta no explica nada de la variación de $Y$, por lo que se diría que no hay dependencia lineal (la recta no serviría en absoluto para hacer predicciones).

Si $R^2 = 1$ significa que el 100 % de la variación de $Y$ queda determinada por la recta, es decir, todos los puntos de la nube estarían justo encima de la recta y las predicciones serían completamente fiables. En general, cuanto más se aproxime $R^2$ a 1 mejor será la aproximación y cuanto más se acerque a 0, peor.

**Problemas propuestos:** Apartado c) del Problema 8.4 y Problema 8.5.

### 8.4.3. Regresión y correlación curvilínea

En los apartados anteriores se consideraron únicamente modelos lineales para simplificar, sin embargo, en la práctica aparecen otros modelos que pueden funcionar mejor que las rectas de regresión para realizar predicciones.

Como el coeficiente de determinación indica lo bueno que es un modelo, se pueden calcular distintos **modelos curvilíneos** y elegir el mejor de ellos para hacer la aproximación.

Las regresiones más habituales son la lineal, la cuadrática, la cúbica, la logarítmica, la inversa, la potencial y la exponencial.

La mayor parte de los modelos curvilíneos habituales dependen de dos parámetros $\widehat{b}_0$ y $\widehat{b}_1$, pero algunos, como el cuadrático o el cúbico, dependen de más. Es mejor elegir modelos con pocos parámetros, así que si los $R^2$ son similares, es mejor elegir el modelo más simple.

La decisión entre un modelo y otro puede depender también del conocimiento que tengamos sobre el tema, ya que a menudo tiene más lógica un modelo que otro. La referencia visual también puede ayudar a determinar qué tipo de relación es la más conveniente en cada caso.

Al igual que la regresión lineal, cualquier regresión curvilínea se puede utilizar para realizar predicciones si conocemos un valor de la variable independiente **que se encuentre entre el mínimo y el máximo de la muestra** (interpolación). No se puede utilizar si el valor de la variable independiente está fuera de ese rango (extrapolación) porque las condiciones fuera de lo recogido por la muestra podrían cambiar y por tanto la fórmula hallada podría no ser válida.

**Problema propuesto:** Problema 8.6.