Análisis exploratorio de variables bidimensionales

PRAUZ-5294

1. Introducción

El alumno al finalizar esta práctica ha de ser capaz de:

- Identificar situaciones en las que la relación entre dos variables de interés sea la solución a un problema.
- Aplicar e interpretar herramientas gráficas para identificar patrones entre dos variables.
- Aplicar e interpretar medidas numéricas para cuantificar la relación entre dos variables.
- Obtener e interpretar la recta de regresión. Usar la recta de regresión para predecir el valor de una variable conocido el que toma la otra.

2. Conceptos clave

En ciertas ocasiones nos interesa determinar si entre dos variables de interés, denotadas como X e Y, existe algún tipo de relación. Este sería el caso, por ejemplo, del peso y la altura de una persona. X se denomina **variable explicativa**, **covariable** o **predictor**, mientras que Y se conoce como **variable respuesta**. Los papeles entre ambas no son intercambiables. La estatura determina el peso de una persona pero no al revés. Si nos dicen que una persona mide 1.85 podríamos establecer alguna conjetura de su peso y no esperaríamos que este fuera de 60 kilos. Sin embargo, si nos dicen que una persona pesa 60 kilos, el rango de alturas posibles es muy amplio. Por consiguiente, X es la estatura e Y, el peso.

En estos casos, la información está dada en forma de datos bivariados (x_i, y_i) donde x_i e y_i son, respectivamente, el valor de las variables X e Y en el individuo i de la muestra de tamaño n. Si se observa algún tipo de relación entre las variables, se podrá utilizar esa relación para predecir el valor de Y conocido el valor de X para un nuevo individuo que no esté en la muestra inicial que conduce a establecer la relación entre X e Y. El tipo de análisis dependerá de la naturaleza de las dos variables: cualitativa-cualitativa, cualitativa-numérica ó numérica-numérica.

2.1. Descripción gráfica y numérica de la relación entre dos variables

2.1.1. X categórica e Y categórica

Las herramientas más útiles para representar las distintas categorías de una variable frente a las de otra son:

- Gráficos de barras.
- Gráficos circulares.

Con ellos se puede advertir si alguna determinada categoría de la Y tiende a ocurrir más cuando la X está en ciertas categorías concretas.

El efecto anterior se puede cuantificar mediante una **tabulación cruzada**. Este tipo de tablas presenta un número de celdas igual al producto del número de categorías de la variable X por al número de categorías de Y. El número de veces que se observa cada combinación de categorías aparece en el interior de la celda correspondiente.

2.1.2. X categórica e Y numérica

La dependencia de X respecto a Y se analiza mediante gráficos de caja de los valores que toma Y en los distintas categorías de X. A partir de su comparación podremos determinar:

1. Si los valores que toma Y tienden a ser mayores en alguna categoría de la X.
2. Si la dispersión o variabilidad de la Y es similar en todas las categorías de la X o, por el contrario, depende del valor de X:

Para analizar las posibles diferencias se pueden obtener las medidas numéricas descriptivas (media, mediana, desviacion típica) que toma la variable Y, separadamente, para cada categoría de X. Cuanto más elevadas sean tales difencias, mayor sera dependencia de Y con X.

2.1.3. X numérica e Y numérica

La representacción gráfica a aplicar es un **diagrama de dispersión** en el que los datos (x_1, y_i), $i = 1, 2 \ldots, n$, son una "nube" de puntos en el plano. Las tendencias o patrones que se advertan en esa nube pueden indicar una relación entre X e Y. Para ayudar en la detección de patrones se puede realizar un **suavizado** que actúa a modo de "filtro", resaltando la relación entre X e Y frente a las observaciones más discrepantes con ella. Podría entenderse como una operación de aislar una **señal** que está perturbada por un **ruido**.

Las relaciones entre las variables pueden ser de cualquier tipo: lineal, cuadrática, logarítmca, exponencial, etc. Sin embargo, el patrón lineal es, entre todos, el más sencillo de identificar y de interpretar. Así, un medida numérica de interés es el **coeficiente de correlación** que mide, exclusivamente, el grado de relación lineal entre dos variables. Su módulo y signo son determinantes para establecer si existe y cómo es la relación lineal entre X e Y.

Cuando el estudio de gráfico y numérico anterior indique que entre X e Y existe una relación lineal, podemos expresarla mediante la **recta de regresión,** $Y = aX + b$. La interpretación de la pendiente de la recta, a, resume la relación entre X e Y. A diferencia de cualquier otra recta, la consideración de b, en la recta de regresión, como el valor que toma Y cuando $X = 0$, está condicionada a que el valor $X = 0$ sea posible en la práctica. Este sería el caso, por ejemplo, en la recta de regresión del peso frente a la estatura puesto que no existen personas cuya altura sea igual a 0. Por consiguiente, si bien la recta de regresión tiene la expresión formal de todas las rectas, puede diferir en su interpretación de b. Otro elemento diferenciador es que la recta de regresión no se puede extender fuera del rango medido en las variable X, es decir, pierde el carácter "infinito".

2.2. Consejos en el análisis exploratorio de variables bidimensionales

1. Empezar el estudio identificando la naturaleza de las dos variables: cualitativa-cualitativa, cualitativa-numérica ó numérica-numérica.
2. Utilizar el gráfico y medida numérica que toque en cada caso.
3. Si no se usa el gráfico o la medida numérica adecuada, el software utilizado puede dar un mensaje de error o llevar a un resultado aberrante.
4. Dada una colección de observaciones (x_1, y_i), $i = 1, 2 \ldots, n$, siempre es posible obtener el coeficiente de correlación y la recta de regresión, por tanto, a partir de ellos no se puede concluir que la verdadera relación entre X e Y sea lineal.

3. Material

3.1. Guiones de prácticas con un software específico

3.1.1. Grado en Enfermería de la Facultad de Ciencias de la Salud (R-Commander)

La carpeta **Enfermería-Estadística Aplicada Ciencias Salud (R-Commander)** contiene tres prácticas de Estadística descriptiva bivariante del curso Estadística, en el Grado en Enfermería que se imparte en la Facultad de Ciencias de la Salud. El software utilizado es R Commander.

La primera carpeta **EACS-EstadísticaDescriptivaBivariante-2 variables cualitativas** se centra en el análisis de dos variables cualitativas. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- EACS-Estadística descriptiva bivariante Dos variables cualitativas.pdf es el guión de la práctica.
- 8 ficheros de datos .RData.
- 1 fichero de datos .xlsx.

La segunda carpeta **EACS-EstadísticaDescriptivaBivariante-2 variables cuantitativas** se centra en el análisis de dos variables numéricas. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- EACS-Estadística descriptiva bivariante Dos variables cuantitativas.pdf es el guión de la práctica.
- 7 ficheros de datos .RData.
- 1 fichero de datos .xlsx.

La tercera carpeta **EACS-EstadísticaDescriptivaBivariante-Comparación de distribuciones** se centra en la comparación de la distribución de una variable numérica según los valores definidos por una variable cualitativa. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- EACS-Estadístca descriptiva bivariante Comparación de distribuciones.pdf es el guión de la práctica.
- 4 ficheros de datos .RData.

3.1.2. Grado en Ingeniería de Tecnologías Industriales (Minitab)

La carpeta **IngenieriaTecnologiasIndustriales-Estadistica (Minitab)** contiene la práctica de Estadística descriptiva bivariante de la asignatura Estadística del Grado en Ingeniería de Tecnologías Industriales que se imparte en la Escuela de Ingeniería y Arquitectura. Este material ha sido elaborado por Jesús Asín, Lola Berrade y Carmen Galé. El software utilizado es Minitab. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- Laboratorio-04-DescriptivaBidimensional.pdf es el guión de la práctica.
- 12 ficheros de datos .mtw.

3.1.3. Grado en Óptica y Optometría de la Facultad de Ciencias (R-Commander)

La carpeta **Óptica-Métodos Estadísticos para óptica y optometría (R-Commander)** contiene dos prácticas de Estadística descriptiva bivariante de la asignatura Métodos Estadísticos para óptica y optometría del Grado en Óptica y Optometría que se imparte en la Facultad de Ciencias. El software utilizado es R-Commander.

La primera carpeta **Práctica 3. Correlación y regresión** se centra en el análisis de dos variables numéricas. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- Practica3.pdf es el guión de la práctica.
- 2 ficheros de datos .rda.
- 1 fichero de datos .txt.

La segunda carpeta **Práctica 4. Tablas de contingencia** se centra en el análisis de dos variables cualitativas. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- Practica4.pdf es el guión de la práctica.
- 2 ficheros de datos .rda.
- 1 fichero de datos .xls.

3.1.4. Grado en Relaciones Laborales y Recursos Humanos de la Facultad de Ciencias Sociales y del Trabajo (R-Commander)

La carpeta **Relaciones Laborales y Recursos Humanos-Estadística (R-Commander)** contiene cinco prácticas de Estadística Descriptiva bivariante de la asignatura Estadística del

grado en Relaciones Laborales y Recursos Humanos de la Facultad de Ciencias Sociales y del Trabajo. El software utilizado es R-Commander.

La primera carpeta **Práctica 5 - Tablas de contingencia** se centra en el análisis descriptivo numérico de dos variables cualitativas. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- Guion.pdf es el guión de la práctica.
- 1 fichero de datos .xlsx.

La segunda carpeta **Práctica 6 - Representación gráfica bivariante de variables cualitativas** se centra en el análisis descriptivo gráfico de dos variables cualitativas. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- Guion.pdf es el guión de la práctica.
- 1 fichero de datos .xlsx.
- 1 fichero de datos .RData.

La tercera carpeta **Práctica 7 - Diagrama de dispersión y coeficiente de correlación** se centra en el análisis descriptivo gráfico de dos variables numéricas. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- Guion.pdf es el guión de la práctica.
- 1 fichero de datos .xlsx.
- 2 fichero de datos .RData.

La cuarta carpeta **Práctica 8 - Recta de regresión** se centra en el ajuste de la recta de regresión entre dos variables numéricas. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- Guion.pdf es el guión de la práctica.
- 2 fichero de datos .RData.

La quinta carpeta **Práctica 9 - Una variable cualitativa y una cuantitativa** se centra en el análisis de una variable numérica según los valores de una variable cualitativa. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- Guion.pdf es el guión de la práctica.
- 1 fichero de datos .RData.
- 1 fichero de datos .xlsx.

3.2. Guiones de prácticas sobre colecciones de datos

3.2.1. Análisis bivariante: Accidentes de tránsito en Guatemala

La carpeta **Accidentes Guatemala Bivariante** contiene el guión y el fichero de datos de una práctica de análisis bivariante en la que se analiza un conjunto de datos real que contiene información de víctimas de accidentes de tránsito recopilada por la Policía Nacional Civil de Gautemala entre enero y junio de 2023. Este material ha sido elaborado por Miguel Lafuente.

3.2.2. Análisis bivariante: VIH en Malawi

La carpeta **Malawi VIH bivariante** contiene el guión y el fichero de datos de una práctica de análisis bivariante en la que se analiza un conjunto de datos procedente de un experimento de campo realizado en zonas rurales del sur de Malawi durante 2004-2006. Este material ha sido elaborado por Miguel Lafuente.

3.2.3. Médicos vs IA: ¿Quién quieres que te atienda?

La carpeta **Médicos vs IA** contiene un conjunto de datos reales procedente de un estudio publicado en 2025 en el que se analiza el impacto del uso de inteligencia artificial (IA) en la toma de decisiones clínicas por parte de médicos. Este material ha sido elaborado por Miguel Lafuente.

4. Referencias

- Henderson, R.G. (2011). Six Sigma. Quality Improvements with Mininitab, 2nd ed. Wiley.
- Myatt, G.J., Johnson, W.P. (2014). Making Sense of Data I. A Practical Guide to Exploratory Data Analysis and Data Mining, 2nd ed. Wiley.