Análisis exploratorio de una variable

PRAUZ-5294

1. Introducción

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

- Extraer la información contenida en un conjunto de datos.
- Distinguir y aplicar las herramientas exploratorias de variables cualitativas y cuantitativas.
- Interpretar los gráficos y las medidas numéricas descriptivas.

2. Conceptos clave

Dada una colección de datos la primera tarea a realizar es organizar y resumir la información contenida en los mismos para conocer su distribución. Se trata de explorar las características más importantes de la variable representada en los datos. Además, las técnicas básicas para la descripción de datos resultan de gran utilidad en la depuración de muestras o identificación de valores anómalos y errores. El análisis exploratorio de cualquier variable incluye calcular las medidas numéricas adecuadas y realizar los gráficos necesarios.

Supongamos que se dispone de datos relativos a una **variable** de interés como la edad o el sexo de un colectivo de personas. Los datos pueden ser de naturaleza numérica o cualitativa. La consideración de una variable como cualitativa, numérica discreta o continua condiciona su tratamiento descriptivo. Son **cualitativos**, denominados también **categóricos**, cuando clasifican a los individuos en diferentes categorías que se distinguen por alguna característica no numérica, como el sexo. Aunque estos datos estén codificados con valores numéricos, por ejemplo mujer=1 y hombre =0, su tratamiento se realiza como variable cualitativa ya que los valores 1 y 0 son sólo etiquetas sin valor numérico. Son **cuantitativos** o cuando proceden de una medición numérica, como el peso, la altura o el número de hermanos de una persona. En este caso se distingue entre **numéricas discretas** como el número de hermanos y **numéricas continuas**, como el peso y la altura. Las variables numéricas discretas están asociadas al conteo de un determinado evento y pueden tomar un número pequeño de valores, como el número de hermanos, o un número muy grande, como la población de la ciudad de nacimiento de la persona. Las variables continuas se denominan así porque, dados dos valores siempre existe un número entre ambos. Así, entre las alturas 1.82 y 1.825 se encuentra, por ejemplo, la altura 1.823. Las variables continuas corresponden a medidas que involucran a números decimales. El primer paso en el análisis estadístico de una variable es identificar de qué tipo es.

2.1. Análisis descriptivo de variables cualitativas

La distribución de una variable cualitativa, o de una variable discreta con pocos valores, se describe mediante su tabla de frecuencias y frecuencias relativas y su representación gráfica habitual es el diagrama de sectores o el diagrama de barras. Para expresar la mayor o menor

frecuencia de las categorías de datos cualitativos se construye un gráfico Pareto. En este gráfico, las barras se organizan según la frecuencia.

2.2. Análisis descriptivo de variables numéricas

En primer lugar, para obtener una idea rápida sobre la distribución de los datos se considera la realización de gráficos. Cuando el número de datos a representar no es muy elevado se puede utilizar un **gráfico de puntos** que conserva los valores exactos de cada una de las observaciones. Si se realiza un **histograma**, se agrupan los datos en clases y se pierde el valor exacto de los mismos, a cambio suaviza la forma de su distribución. En general, el software empleado determina por defecto el número de clases definidas. Es habitual que en el histograma se superponga la curva que describe el patrón de un modelo teórico, por ejemplo la campana de Gauss del modelo normal. La observación de cualquier gráfico de una variable contribuye a identificar la presencia de **datos atípicos (outliers)** que hay que comprobar con el tomador de datos. Notemos que, el histograma proporciona una primera idea sobre la densidad de probabilidad de la variables representada: los valores más probables, la simetría o no de la distribución y la concentración o no de los valores en torno a algún valor. Estos gráficos resultan adecuados para analizar variables numéricas continuas y discretas con muchos valores.

La interpretación de un gráfico tiene cierta componente subjetiva. Por este motivo, cualquier análisis exploratorio se acompaña del cálculo de un conjunto de medidas estadísticas. Las medidas que aportan información sobre la posición donde se sitúan los datos son medidas de localización, y las que aportan información sobre la variabilidad de los datos son medidas de dispersión. Además hay otras medidas que informan sobre la forma de la distribución.

Dada una colección con n datos de una variable numérica, las medidas de localización más habituales son:

Mínimo El valor mínimo de todos los datos.

Máximo El valor máximo de todos los datos.

Media La media aritmética de los n datos:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Mediana Valor de la variable que supera el 50 % de los datos.

Media truncada al 5 % Media de los datos después de eliminar el 5 % de los valores más bajos y el 5 % de los valores más altos.

Primer cuartil El 25 % de los datos es inferior a este valor.

Tercer cuartil El 75 % de los datos es inferior a este valor.

Para valorar la variabilidad entre los datos de la muestra son de gran interés las medidas de dispersión:

Varianza Es un medida de la dispersión de los datos. La cuasivarianza muestral se calcula:

$$\hat{S}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}$$

La desviación típica se define como la raíz cuadrada de la varianza.

Coeficiente de variación Una medida adimensional que permite comparar la dispersión de variables medidas en diferente magnitud:

$$CV = \frac{\hat{S}}{\overline{X}}$$

Rango Diferencia entre el máximo y el mínimo.

Rango intercuartílico Diferencia entre el tercer y el primer cuartil.

Finalmente, las medidas de forma:

Coeficiente de asimetría expresa el grado de simetría en que se disponen los datos de la muestra alrededor de su media. Un valor cercano a 0 indica una distribución de los datos simétrica.

Coeficiente de curtosis Cuando se estudia una distribución de carácter simétrico se compara cómo es de apuntada en comparación con la campana de Gauss (distribución normal). Dado que esta distribución se toma como referencia, se le asigna un valor de curtosis igual a cero. Si el perfil de la distribución de datos es más plano que la distribución normal, la curtosis correspondiente es negativa y si es más puntiagudo, la curtosis es positiva. La curtosis resulta muy útil para identificar distribuciones de datos con forma de campana pero que su comportamiento no se ajusta por una distribución normal.

La **media truncada** es una medida de localización o tendencia central menos sensible a valores extremos que la media. Si toman valores muy diferentes indica que existen valores extremos. Cuando existen valores extremos por debajo y por arriba, ambas medias toman valores próximos. El hecho de que la **media** y la **mediana** no sean próximas es un indicio de que la distribución no es simétrica.

Otro gráfico interesante es el diagrama de caja. Los lados de la caja del gráfico boxplot pasan por el primer y tercer cuartil y la línea central se sitúa a nivel de la mediana. Por tanto, la amplitud de la caja coincide con el rango intercuartílico. Los segmentos van desde el extremo de la caja hasta los datos más extremos que se encuentran a una distancia menor de 1.5 veces el rango intercuartílico desde el borde de la caja. Los valores más alejados de esos límites se representan con *.

El diagrama de caja además de informar sobre la forma de la distribución, identifica el valor de los cuartiles, la mediana y el rango intercuártilico y los valores atípicos. Estos valores requieren un análisis especial. Si se reconocen como errores en la captura de datos se eliminan y se realiza de nuevo el análisis, ya que tanto los gráficos como las medidas numéricas están muy influidos por la presencia de datos atípicos o “outliers”. En otro caso, son motivo de un análisis más exhaustivo.

2.3. Consejos en el análisis exploratorio de una variable

1. El primer paso en el análisis consiste en identificar qué variables son categóricas y cuáles son numéricas. En el caso de ser numérica, si se trata de una variable discreta o de una variable continua. De acuerdo con el tipo de variable se realiza el análisis con medidas numéricas y gráficos adecuados. El análisis es diferente para cada tipo de variable.

2. En el análisis de una variable categórica debe incluir un diagrama de sectores circulares, sólo cuando el número de categorías no es elevado, o un diagrama de barras que represente la frecuencia relativa, o el porcentaje, de casos de cada categoría.

3. En variables numéricas debe analizarse la posición central (media, media recortada, mediana) y la variabilidad (desviación típica, rango intercuartílico, coeficiente de variación). Además de las medidas de localización hay que establecer conclusiones también sobre la variabilidad.

4. En variables categóricas o cualitativas no tiene sentido elaborar una descriptiva con medias, medianas, percentiles, desviación típicas, medidas de forma o representar un histograma.

5. Los gráficos para variables numéricas más adecuados son el histograma, el dotplot y el boxplot. El gráfico de barras no es adecuado, porque tiene un eje X en el que no existe una escala numérica y hay que utilizarlo para variables categóricas o numéricas discretas.

6. En la redacción de conclusiones hay que tener especial cuidado con el uso de la palabra "significativo", en Estadística tiene una connotación especial.

7. Un valor atípico no debe eliminarse de la muestra sistemáticamente, sino sólo si hay evidencia de que es un dato falso o que no corresponde a la población que se muestrea.

3. Material

3.1. Guiones de prácticas con un software específico

3.1.1. Grado en Ciencias Ambientales de la Escuela Politécnica de Huesca (R-Commande)

La carpeta **CienciasAmbientales-Estadística(R-Commander)** contiene la práctica de Estadística descriptiva univariante del curso Estadística, en el Grado en Ciencias Ambientales que se imparte en la Escuela Politécnica de Huesca. Además incluye el tratamiento descriptivo de dos variables cualitativas. El software utilizado es R Commander. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- Guión Práctica 2 - Estadística descriptiva (una y dos variables).pdf es el guión de la práctica.
- 5 ficheros de datos .RData.

3.1.2. Grado en Enfermería de la Facultad de Ciencias de la Salud (R-Commander y Excel)

La carpeta **Enfermeria-EstadísticaAplicadaCienciasSalud (R-Commander)** contiene la práctica de Estadística descriptiva univariante del curso Estadística Aplicada a Ciencias de la Salud, en el Grado en Enfermería que se imparte en la Facultad de Ciencias de la Salud de la Universidad de Zaragoza. Este material ha sido elaborado por Ana Pérez, Fernando Plo y Javier Tejel. El software utilizado es R Commander, aunque también se utiliza Excel como herramienta auxiliar para gestionar las bases de datos. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- EACS-Estadísticadescriptivaunivariante.pdf es el guión de la práctica.
- 8 ficheros de datos .RData.
- Scripts de R.

3.1.3. Grado en Ingeniería Informática de la Escuela de Ingeniería y Arquitectura (R-Commander)

La carpeta **Ingeniería Informática-Estadística (R-Commander)** contiene la práctica de Estadística descriptiva univariante del curso Estadística, en el Grado en Ingeniería Informática que se imparte en la Escuela de Ingeniería y Arquitectura de la Universidad de Zaragoza. Además incluye el tratamiento descriptivo de una variable numérica según los valores de una variable cualitativa. El software utilizado es R Commander. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- Estadistica descriptiva (Practicas 1 y 2).pdf es el guión de la práctica.
- 7 ficheros de datos .RData.

3.1.4. Grado en Ingeniería de Tecnologías Industriales (Minitab)

La carpeta **Ingenieria Tecnologias Industriales-Estadistica (Minitab)** contiene la práctica de Estadística descriptiva univariante 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-01-DescriptivaUnidimensional.pdf es el guión de la práctica.
- 9 ficheros de datos .mtw.

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

La carpeta **Óptica-Métodos Estadísticos para óptica y optometría (R-Commander)** contiene la práctica de introducción a R-Commander y Estadística descriptiva univariante 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. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

- Practica1.pdf es el guión de la práctica.
- 1 fichero de datos .RData.
- 1 fichero de datos .xls.

3.1.6. 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 guiones correspondientes a tres prácticas de Estadística Descriptiva 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. En esta carpeta se incluye la información suficiente para poderlo utilizar de forma autónoma:

La primera práctica incluye

- Guion 2.pdf es el guión de la primera prácticas con la descriptiva básica univariante de variables cualitativas y cuantitativas.
- Guion 3.pdf es el guión de la segunda práctica con las medidas de síntesis para resumir distintos aspectos de un conjunto de datos.
- Guion 4.pdf es el guión de la tercera práctica sobre cómo realizar transformaciones lineales de las variables.
- 2 fichero de datos .RData.
- 2 fichero de datos .xlsx.
- 1 fichero de datos .txt.

3.2. Guiones de prácticas sobre colecciones de datos

3.2.1. Análisis univariante: Accidentes de tránsito en Guatemala

La carpeta **Accidentes Guatemala Univariante** contiene el guión y el fichero de datos de una práctica de análisis univariante 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 univariante: VIH en Malawi

La carpeta **Malawi VIH Univariante** contiene el guión y el fichero de datos de una práctica de análisis univariante 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.

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.
- Grima Cintas, P., Marco Almagro, L., Tort-Martorell Llabrés, X. (2022) Estadística con MINITAB. Aplicaciones para el control y la mejora de la calidad. Ed. Garceta.