

Práctica Estadística Descriptiva y Análisis Exploratorio de Datos

Introducción al análisis de datos con R
A2B2C

Octubre 2020

Análisis exploratorio de datos

Esta práctica tiene como objetivo realizar una primera aproximación a la comprensión de nuestros datos, utilizando para ello las técnicas del análisis exploratorio de datos vistas. Para esta práctica trabajaremos con un dataset de COVID-19 de la OMS disponible en <https://data.humdata.org/dataset/> y con el dataset de antropometria.

1. Medidas de resumen

- a) Cargar el dataset antropometria
- b) Analizar la estructura del dataset. Ayuda: `?class`, `?str`, `?colname`, `?nrow`, `?ncol`, `?head`, `?tail`
- c) ¿Cuántos hombres y cuántas mujeres tiene el dataset? Ayuda: `?table`
- d) Calcular máximo, mínimo, media, mediana, varianza y desviación estándar de las alturas separadas por sexo. ¿Cuál tiene mayor media? ¿Y mayor varianza? Ayuda: `?max`, `?min`, `?mean`, `?median`, `?var`, `?sd`, `?summary`
- e) Cargar el dataset `cases-covid-19-by-country.csv` de una forma apropiada.
- f) ¿Cuántos datos u observaciones contiene el dataset? ¿Cuántos atributos tiene cada dato?
- g) ¿De qué tipo son los atributos? ¿Cómo se llaman?
- h) Mostrar los primeros 5 datos y los últimos 5 datos.
- i) Realizar una descripción general del dataset a partir de este análisis con tus palabras.
- j) Calcular la media, la mediana y la moda para las variables numéricas.
¿Son similares o muy diferentes?
- k) Calcular el rango, el desvío estandar y el IQR.
A partir de esto indicar si los valores están muy concentrados o si por el contrario, están muy dispersos.
- l) En base a lo visto anteriormente, ¿Qué medidas usarías para resumir los valores?

2. Visualización

- a) Cargar el dataset antropometria
- b) ¿Existe relación entre la altura y la edad? Graficar en un scatter plot altura vs. edad y decidir. ¿Hubiera cambiado la conclusión de haber realizado el mismo gráfico separado por sexo? Ayuda: `?plot`
- c) Agregar una recta vertical en rojo en donde comienza el plateau. Ayuda: `?abline`
- d) Para remover el efecto de la edad, utilizar el gráfico del punto anterior para decidir a qué edad la misma deja de tener efecto en la altura y subsetear el dataset original en un vector nuevo llamado “adultos”.
- e) Realizar un boxplot de altura para adultos separados por hombre y por mujer ¿Varía la altura entre sexos? ¿Y para el peso? Agregar una leyenda al gráfico. Ayuda: `?boxplot` `?legend`
- f) ¿Existe correlación entre la altura y el peso? Graficar en un scatter plot altura vs. peso para hombre y mujer por separado. Para cuantificar este efecto, calcular la correlación. Ayuda: `?cor`
- g) Realizar un histograma de la altura para varones adultos y para mujeres adultas en el mismo gráfico. ¿Qué pinta tienen? ¿Qué pasa si se cambia la cantidad de bins de los mismos? Agregar una leyenda al gráfico. Ayuda: `?hist` y parámetro `add = TRUE` de `hist`
- h) Cargar el dataset `cases-covid-19-by-country.csv` de una forma apropiada.
 - i) Realizar un histograma para cada variable y decidir si es más conveniente separar los datos en dos grupos para su análisis. En ese caso, ¿qué criterio podríamos usar?
 - j) Graficar en un mismo gráfico un histograma y un boxplot para cada variable. ¿Siguen apareciendo valores extremos en los boxplot? ¿Por qué?
 - k) Realizar un gráfico de scatterplot entre las variables numéricas. ¿Es necesario separar los casos más extremos para su análisis? ¿Existe relación entre la cantidad de infectados y la cantidad de fallecidos? ¿Si existe, es en todas las escalas? ¿De qué tipo?