Practice

Marcus Nunes

17 November, 2019

Contents

1 Instructions

Save a version of this file and call it `practice_answers.Rmd`. Remove this first section and practice what we learned today, writing a small report analyzing the `iris` dataset.

After your work is done, change the code `output: bookdown::html_document2` on line 5 to `output: bookdown::pdf_document2` in order to have a pdf. It is preferable to work with the HTML version of the report first to save time and create the pdf only after the analysis is complete.

2 Introduction

The Iris flower dataset is a well known multivariate dataset introduced in Fisher (1936). It contains 150 observations of three species of plants called *Iris setosa*, *Iris versicolor*, and *Iris virginica*. In this report we will analyze this dataset and show why it is interesting.

3 Descriptive Statistics

We will describe the most important features about the Iris flower dataset in this section.

3.1 Tables

1. Create a table with the means for `iris` dataset. Describe your findings.

3.2 Plots

2. Make scatter plots showing the relationships between all the numeric variables in this dataset. What you can see?

3. Plot the scatter plot between the two variables with the highest correlation, coloring the points according to the flower species. Is there anything special about this plot?

4 Modeling

4. Fit a linear regression model between the two variables you found on question 3. Use the variable with the highest mean as the predictor variable.

5. Plot the regression line on the scatter plot.

6. Make a boxplot comparing the observations of the `Sepal.Width` variable between the three plant species. Do you think there is a group whose mean is different from the others?

7. Test the hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_0 : \text{at least one pair } \mu_i \neq \mu_j, \text{ if } i \neq j$$

where μ_i is the mean for the variable `Sepal.Width` for the groups

- $i = 1$ (setosa)
- $i = 2$ (versicolor)
- $i = 3$ (virginica)

What is your conclusion?

5 Conclusion

As we could see, this dataset is great to practice statistical concepts.

References

Fisher, R. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7 (7): 179–88.