Master BeNeFri in Computer Science

Course: Statistical Learning Methods
Spring 2016

Exercise #2: Plot() and *t*-test (for the mean)

Download from the ILIAS website the dataset "Education dataset (filename: Education.txt) and read the description of the data given in the pdf file "Education description". Ignore the outliers.

1. Using the plot() function, show *graphically* the possible relationship between the independent variable `Eduation` and the dependent variable `Wage`. Can you change the main title of the plot as well as the labels appearing in the two axes?

2. As the variable `Gender` is a factor (categorical / binary data), select the wage and education values corresponding to each of the two possible gender values. Show graphically the possible relationship between the variable independent `Eduation` and the dependent variable `Wage` but only when considering the observations corresponding to *men* then with *women*. Can you also plot on the same graphics the information for *men* and *women* (with two different color)?

Download from the ILIAS website the dataset Mean20 dataset (filename: Mean20.txt). This dataset is composed by a single variable (`time`), the time delay in minutes between two calls in an info-center.

3. Compute the mean, the median, the standard deviation, the minimum and maximum value of the variable `time`. Do you need to preprocess this list of values in some way?

4. We suppose that the mean delay between two calls is 7.05 minutes. Can you test this hypothesis with the data available? What is your conclusion? Do you see a difference when considering the original values and the preprocessed values?

5. For Mary, the delay cannot be smaller than 7.05. Thus the only credible alternative hypothesis must take account of this (well-known) fact. How can you test Mary's hypothesis?