Dr Isabella Gollini

Probability and Statistics

Examples 1

1. The data to be analysed are measurements of the speed of light in air, from an experiment carried out by Albert Abraham Michelson between 5th June and 2nd July, 1879. The data arise from five experiments, each consisting of 20 consecutive runs. The recorded response variable is the speed of light in km/s, less 299000, which we shall refer to as `speed`. The data set, `morley` has 3 columns and 100 rows. The first column contains the value of `Speed`, the second column contains the number of the run, `Run`, which takes the values 1 ... 20, and the third column has the the number of the experiment, `Expt`, which takes the values 1 ... 5.

(i) Load the dataset `morley` that is available in `R`. Check the dataset. Remember to write all the commands into the Editor, and save the `R` script as, say, "morley.R".

(ii) For the `speed` data in the first column, obtain the descriptive statistics that we saw in Lecture 1.

What, in particular, was the average recorded speed of light to the nearest km/s?

What was the interquartile range?

(iii) Obtain a boxplot of the `speed` data, including the title "speed of light in km/s, less 299000".

What are the values of the three "outliers" in the boxplot?

(iv) Obtain a histogram of the `speed` data, including the title "speed of light in km/s, less 299000", and using class boundaries from 600 to 1100 in steps of 50.

2. A simple experiment was carried out where a coin was tossed ten times and the number of heads out of the ten tosses was recorded. This experiment was repeated 200 times. The resulting data set is contained in the data file "cointossing.dat", which has a single column of data of length 200.

(i) Download the data file "cointossing.dat" from Moodle or from the course materials web page.

Import the data from the file (Hint: To load the data use the command: `cointossing <- read.table(file = "cointossing.dat")` if you are in the working directory, or `cointossing <- read.table(file = file.choose())` and find the file).

Rename the first column `heads`. (Hint: `colnames(cointossing) <- "heads"`)

Remember to write all the commands into the Editor, and to save the `R` script as, say, "cointossing.R".

(ii) Obtain the descriptive statistics.

What, in particular, was the average number of heads observed in ten tosses of the coin?

What was the smallest number of heads observed in ten tosses, and what was the largest number of heads observed?

(iii) Using `R` obtain the observed frequency distribution.

(iv) Obtain a bar chart to illustrate the frequency distribution of the data, including the title "Number of heads in ten tosses of a coin". (Hint: Use the material from Lecture 1, in particular see Section 1.5)

3. The dataset `faithful` contains data on the duration time of eruptions and the waiting times between successive eruptions, both measured in minutes, for the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA. The data set has 272 rows and two columns.

(i) Load the dataset `faithful` that is available in `R`

(ii) Produce the standard descriptive statistics for both variables.

(iii) Produce the boxplots for both variables.

(iv) Produce the default histograms for both variables, labelling them with the corresponding class frequencies. For the variable `waiting`, produce different histograms using class boundaries of your own choice, for example, using class intervals 2, 5 and 10 respectively, to cover the range of values from 40 to 100 that are taken by the variable.

(v) Which of the histograms for the variable `waiting`, in your judgement, best illustrates the data?

For both variables, what feature of the data, which is not obvious from examination of the descriptive statistics or the boxplot, is brought out by the histogram?