

Assignment #2 contains problems about exploratory data analysis. Problems which require the use of R have the symbol $\langle R \rangle$.

Please include **your name** and **the discussion section (day/time) that you attend** on your homework.

Instructions to install R on your own computer are on the course web page. R is also available in many campus computing labs including CALS, Memorial Union, and Union South.

This assignment is worth 50 points in total, and points for each problem are indicated.

If you feel challenged by these problems, I encourage you to do additional problems on your own. Many problems have answers in the back of the textbook.

Your assignment must be turned in during lecture or to your TA's mailbox by 5pm on the due date. We will not grade late homework. If there are special circumstances, please speak to Professor Larget, preferably in advance, for consideration.

Historically, most students who have made an honest effort to do all of their homework on time have received the maximum credit for the semester. Missing some assignments can have a large negative effect on your course grade. Even if an assignment is incomplete, it is best to turn in the problems you have completed on time.

1. *[20 points]* Summarize the data in Exercise 2.10 (page 25).
 - (a) Display the data using a stem-and-leaf display.
 - (b) Construct a frequency table for the data.
 - (c) Display the data in a histogram.
 - (d) Find the median, lower, and upper quartiles of the distribution.
 - (e) Display the data using a modified boxplot (using fences that are 1.5 IQR from the box).
 - (f) Compute the mean and standard deviation of the data.
 - (g) What proportion of the data is within 1 standard deviation of the mean? What proportion of the data is within 2 standard deviations of the mean?
 - (h) Is the data skewed? If so, in what direction is the skew? Is this consistent with the relative sizes of the mean and median of the data?
2. *[5 points]* Without calculation, use your intuitive understanding of the definitions of the mean and median to estimate the mean and median of the data displayed in Exercise 2.28 (page 32). Which estimate is larger? In which direction is the skew?
3. *[5 points]* By hand or using R, construct parallel boxplots of the data in Exercise 2.33 (page 39). These commands will do the trick in R.

```
> male = c(6,0,2,1,2,4.5,8,3,17,4.5,4,5)
> female = c(5,13,3,2,6,14,3,1,1.5,1.5,3,8,4)
> boxplot(list(male=male,female=female))
```

4. (R) [20 points] Download the file `trypanos.txt` from the course web page. A link to the data is on the course schedule near where you found this homework assignment. Read the data into R using the `read.table` command, for example, by doing the following.

```
> prob1.2 = read.table(file.choose(),header=T)
> str(prob1.2)
> attach(prob1.2)
```

The first command opens a dialog box that allows you to browse for the file. The argument `header=T` tells R that the first row of the file is a header row that contains the variable names. This command creates a data frame (data set) called `prob1.2`.

The second command shows the structure of the data frame `prob1.2`. There should be a single quantitative variable named `length`.

The third command attaches variable names in `prob1.2` to the list of special names that R knows about. This allows you to refer to `length` in subsequent commands rather than using the more cumbersome `prob1.2$length`.

- Display the data using a stem-and-leaf display. Use the command `stem(length)`.
- Display the data in a histogram. Use the command `hist(length)`.
- What feature of the histogram suggests that the sample is a mixture of two distinct types?
- Display the data in a histogram using only five classes. Use the command `hist(length,nclass=5)`. Explain why this histogram masks important features of the data.
- Find a five number summary of the data using the command `fivenum(length)`.
- Display the data using a modified boxplot using the command `length`.
- Compute the mean and standard deviation of the data using the commands `mean` and `sd`.
- What proportion of the data is within 1 standard deviation of the mean? What proportion of the data is within 2 standard deviations of the mean? Use the following commands.

```
> m = mean(length)
> s = sd(length)
> w1 = sum( (length > m-s) & (length < m+s) )
> w2 = sum( (length > m-2*s) & (length < m+2*s) )
> n = length(length)
> w1/n
> w2/n
```

Compare these proportions to the empirical rule.