

## Stat 359 Assignment 1

1. Reading: Chapters 1, 2, 3, 4
2. Suppose the following data comes from a study on plant growth (mm) where 2 plants are in each pot, 3 pots are within each plot and 2 plots are given one of two fertilizer treatments.

| Plot 1 | Treatment 1 |       |       | Treatment 2 |       |       |
|--------|-------------|-------|-------|-------------|-------|-------|
|        | Pot 1       | Pot 2 | Pot 3 | Pot 1       | Pot 2 | Pot 3 |
|        | 14.6        | 13.2  | 16.4  | 7.1         | 6.8   | 10.0  |
|        | 15.2        | 12.9  | 12.2  | 7.7         | 6.0   | 8.3   |
| Plot 2 | Pot 1       | Pot 2 | Pot 3 | Pot 1       | Pot 2 | Pot 3 |
|        | 18.5        | 22.2  | 24.7  | 9.7         | 6.8   | 10.4  |
|        | 16.7        | 18.8  | 20.3  | 8.8         | 9.0   | 11.3  |

- (a) Arrange the data into a dataframe so that it can be analysed. Print out this dataframe.
  - (b) Sort the data by plant growth.
  - (c) Calculate the mean, and standard deviation of the data.
  - (d) Plot the data using a histogram (R function `hist()`). Clearly label the axis, title and use bin sizes of 2 mm.
3. Write a function that uses the short cut formula to calculate the sample variance of a data vector. Use the vector  $y=(11,11,10,8,11,3,15,11,7,6)$  as your test vector.
  4. On the course webpage you will find a dataset with filename 'tv.txt'. The data arise from a study examining the time teenagers spend watching tv. A random sample of  $n = 100$  eighth grade American high school students was obtained, and the number of minutes spent watching TV during the first week of October was recorded. A similar sample of  $m = 90$  Canadian students was also obtained. In this study it is of interest to compare the TV watching habits of the teenagers from the two different countries, specifically to determine if Canadian students watch less TV than their American counterparts.
    - (a) Compare the two samples using appropriate descriptive statistics, including side-by-side boxplots.
    - (b) Write an R function `z.test(y1,y2,H1)` to compute the p-value for a large sample z-test (discussed in lecture) for testing equality of two population means ( $H_0 : \mu_1 = \mu_2$ ). The arguments to this function are:  $y1$ , a vector containing the sample measurements from the first population;  $y2$ , a vector containing the sample measurements from the second population; and  $H1$ , a string variable, which takes one of three possible values: 'two.sided', 'less' or 'greater' specifying the alternative hypothesis. To complete this question you will need to use the R function `pnorm()` which computes standard normal probabilities, see the help file `?pnorm`.
    - (c) Apply your function to the TV data, computing the p-values for each of the three possible alternative hypotheses.

- (d) Which of the three alternative hypotheses is relevant for the particular question being asked in this study? Comment on the results.