

Stat 401, Section F Homework 1

Due Date: Wednesday, August 29

Please do not include any computer output besides graphs. All other information should be copied from R into your written answer.

1. Do Japanese automakers produce better cars? To answer this question, a group of investigators gathered the following data, consisting of the number of assembly defects per 100 cars which were built in 1989. Data was collected at 11 Japanese-owned plants, and 16 non-Japanese plants.

Non-Japanese	Japanese
29	39
38	38
68	42
67	48
69	50
69	55
70	56
68	57
69	56
79	61
84	89
87	
98	
100	
140	
170	

Use R to answer the following questions.

- (a) Compute the five number summaries for each group (copy these answers from R into the written answer to this question, do not include any computer output).
- (b) Construct side-by-side boxplots that allow you to visually compare the distribution of assembly defects for the Japanese produced cars, versus the non-Japanese ones. To plot side-by-side boxplots in R you can use the same command as for one data set, except you will now use as input both data sets separated by a comma. To produce PDF files which are saved in your working directory, you can include the following two lines of code before and after the `boxplot` command (assume my vectors are called `j1` and `j2`):

```
pdf("NameOfMyGraph.pdf")
boxplot(j1, j2)
dev.off()
```

2. Some short problems on summary statistics (feel free to use R if you need to):

- (a) Provide a set of numbers for which exactly 3 of the 5 numbers in the five-number summary are equal to one.
- (b) Provide a set of 5 numbers that has an average of 9 and a standard deviation 0.

(c) For which data set is the standard deviation larger?

2, 2, 2, 2, 9, 9, 9, 9 or 2, 3, 4, 5, 6, 7, 8, 9

3. Find the approximate distance (“as the crow flies”) from your place of birth to Ames, Iowa. (See <http://www.indo.com/distance/> to get an approximate distance.) Write down this distance in miles. Consider the distribution of such distances obtained by collecting this information from all your colleagues in Stat401 Section F students. Which would be larger, the mean or the median distance? Why? (I am not looking for a numerical answer here, a guess followed by an explanation would be just great for this question!)
4. A team of researchers is investigating the effect of two drugs designed to help people quit smoking. They found that 39 people out of 90 who decided to use Drug A at the beginning of 1998 were no longer smoking at the end of 1998. In contrast, only 17 people out of 115 who chose to use Drug B at the beginning of 1998 had quit smoking by the end of 1998. The researchers concluded that Drug A is superior to Drug B when it comes to helping people quit smoking.
 - (a) Is this an experiment or an observational study? Explain.
 - (b) Comment on the validity of researcher’s conclusion. Can you think of any other reasons why more individuals in the first group might have quit smoking?
5. Download the data set named `4set_data.txt` from the course web site (in the “Computing” section). Using the `read.table` command read this data set into R. Then use R and report the following:
 - (a) the 5 number summary and standard deviation for each of the 4 columns of the data set.
 - (b) a histogram for each of the 4 columns
 - (c) 4 side-by-side boxplots for the four columns of the data set
 - (d) the 5 number summary and variance for the entire data set
 - (e) a histogram, a boxplot, a stem-and-leaf graph, and a dot-plot for the entire data set