Chapter 2 - Summarizing Data

Chunjie Nan

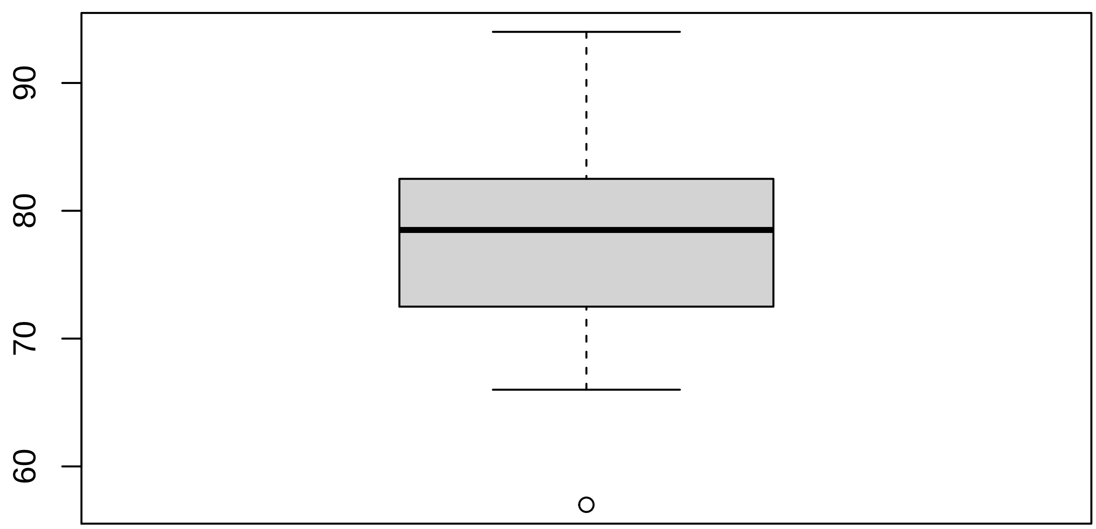
Stats scores. (2.33, p. 78) Below are the final exam scores of twenty introductory statistics students.

57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

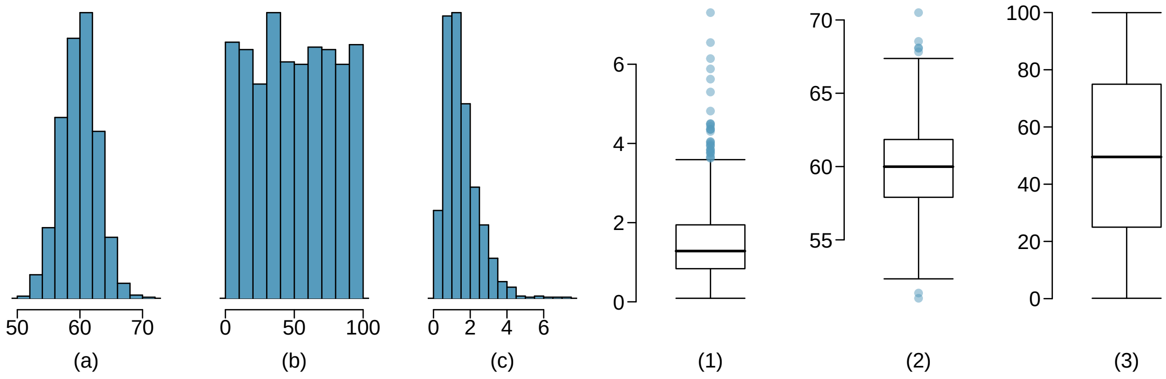
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

```
data_frame<-data.frame("scores"=scores)
boxplot(data_frame$scores)
```



Mix-and-match. (2.10, p. 57) Describe the distribution in the histograms below and match them to the box plots.



Distributions and appropriate statistics, Part II. (2.16, p. 59) For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Housing prices in a country where 25% of the houses cost below \$350,000, 50% of the houses cost below \$450,000, 75% of the houses cost below \$1,000,000 and there are a meaningful number of houses that cost more than \$6,000,000.

The distribution is right skewed since a lot of the data is concentrated on the left. In this case, median is better to show the variability.

(b) Housing prices in a country where 25% of the houses cost below \$300,000, 50% of the houses cost below \$600,000, 75% of the houses cost below \$900,000 and very few houses that cost more than \$1,200,000.

the distribution is very balanced, so this is a symmetric normal distribution. Mean and standard deviation both are good to be represented.

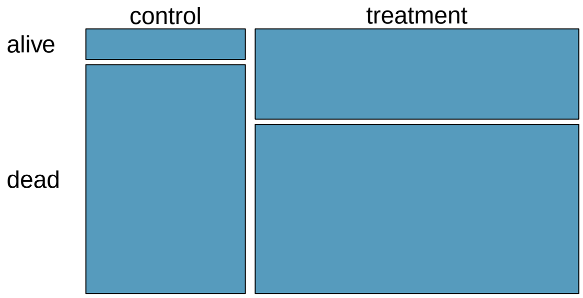
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

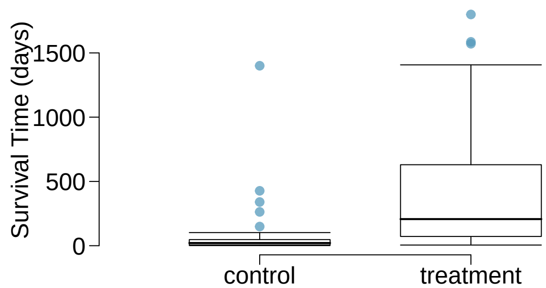
Assume that students under 21 years old are not alcoholic drinker, it shows the right skewed distribution. In this case, median is a good variability.

(d) Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than the all other employees.

Only a few high salaries and the overwhelming average salaries contribute a symmetric distribution. Therefore, mean and standard deviation good to present the variability.

Heart transplants. (2.26, p. 76) The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable *transplant* indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Of the 34 patients in the control group, 30 died. Of the 69 people in the treatment group, 45 died. Another variable called *survived* was used to indicate whether or not the patient was alive at the end of the study.





(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

From the mosaic plot, we can conclude that the suvrvival depend on for patients who got transplant because it increase the survival days compare who didn't get the transplant.

(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.

The tratment group has survived more days than the controlled group.

(c) What proportion of patients in the treatment group and what proportion of patients in the control group died?

45/69 30/34

In the treatment group, 65% patients died, and 88% died in the control group.

(d) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

i. What are the claims being tested? The claim tested hypothesis and null hypothesis.

hypothesis: survivality is depend on the trasplant. null hypothesis: survivality is not depend on the getting transplants or not.

ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

We write *alive* on **28** cards representing patients who were alive at the end of the study, and *dead* on **75** cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size **69** representing treatment, and another group of size **34**representing control. We calculate the difference between the proportion of *dead* cards in the treatment and control groups (treatment - control) and record this value. We repeat this

100 times to build a distribution centered at **0**. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are **24/69 - 4/34 = 0.23**. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.

iii. What do the simulation results shown below suggest about the effectiveness of the transplant program?

The graph shows that very little fractions on 0.23 point, therefore we can reject the null hypothesis. Also, we can conclude that the getting transplant can increase the survivality.