Bryan Persaud

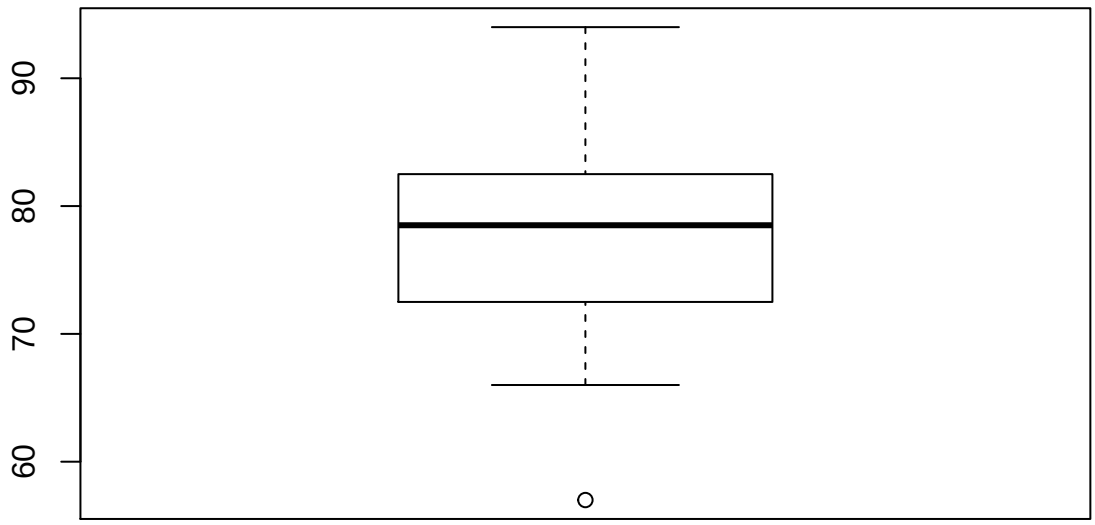
title: “Chapter 2 - Summarizing Data” author: "" output: pdf_document: extra_dependencies: [“geometry”, “multicol”, “multirow”] editor_options: chunk_output_type: console —

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

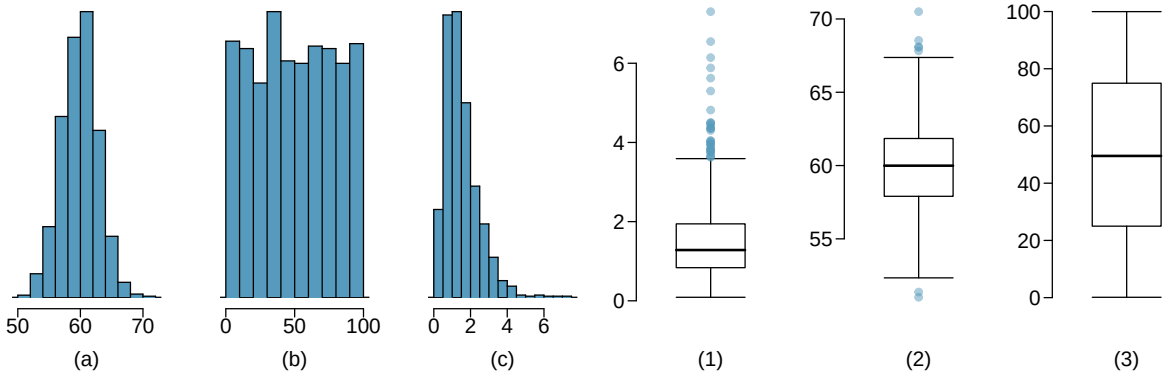
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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



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



Histogram (a) is unimodel and looks to almost be symmetrical and could be normally distributed. It matches with box plot (2).

Histogram (b) could be symmetric but is not normally distributed and almost has a rectangular shape. It matches with box plot (3).

Histogram (c) is unimodel and is right skewed. It matches with box plot (1).

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.
(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.
(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.
(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.

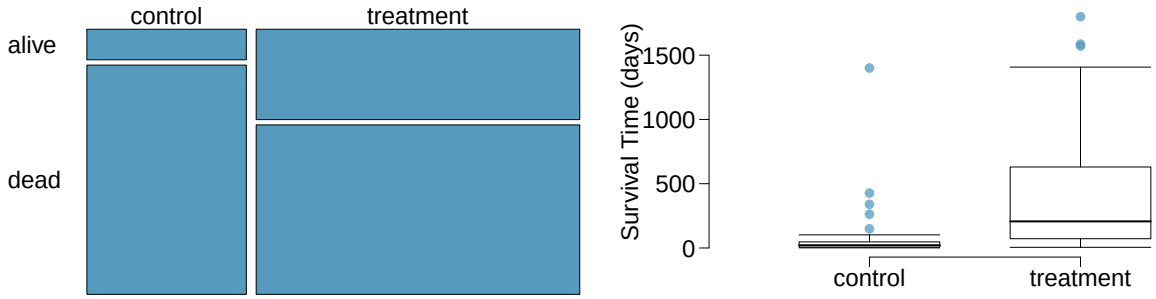
(a) The distribution would be right skewed. You would use the median to best represent a typical observation. IQR would best be used to represent the variability of observations. This is because there are a good number of expensive houses and this would increase the mean and standard deviation.

(b) The distribution would be symmetric. The mean would best represent a typical observation. The standard deviation would best be used to represent the variability of observations. This is because there are not a lot of expensive houses and the ones that are expensive do not cost that much more compared to the data set.

(c) The distribution would be right skewed. You would use the median to best represent a typical observation. IQR would best be used to represent the variability of observations. This is because there are not a big number of excessive drinkers.

(d) The distribution would be right skewed. You would use the median to best represent a typical observation. IQR would best be used to represent the variability of observations. This is because there are only a few that have a much higher salary than most of the other employees.

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.

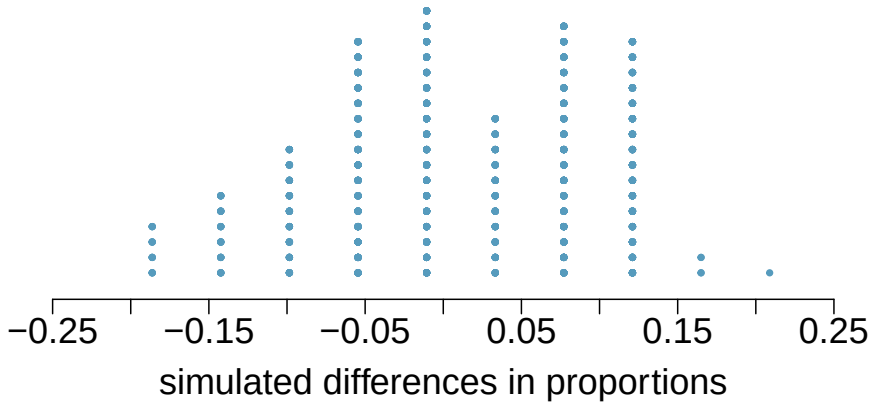


(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.
(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.
(c) What proportion of patients in the treatment group and what proportion of patients in the control group died?
(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?
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 _____45/69 - 30/34_***. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are*** -0.230179_____. 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?



(a) Based on the mosaic plot, survival is dependent on whether or not someone gets a transplant since the ones who a transplant had a higher chance of living.

(b) The box plots below suggest that the heart transplant treatment increases the survival time for the patients.

(c)

```
library(data.table)
heartTr <- as.data.table(heartTr)
x <- heartTr[, .(count = .N), by = .(transplant, survived)]
treatment_dead_ratio <- x[transplant == "treatment" & survived == "dead"]$count / sum(x[transplant == "
control_dead_ratio <- x[transplant == "control" & survived == "dead"]$count / sum(x[transplant == "cont
print(paste("The proportion of patients that died in the treatment group is ", treatment_dead_ratio))
```

```
## [1] "The proportion of patients that died in the treatment group is  0.652173913043478"
```

```
print(paste("The proportion of patients that died in the control group is ", control_dead_ratio))
```

```
## [1] "The proportion of patients that died in the control group is  0.882352941176471"
```

(d) i. The claims being tested are whether an experimental heart transplant program increases lifespan.

iii. The simulaton results shown below suggest that the transplant program was effective in showing that a patient who gets a heart transplant is more likely to live longer than if the patient does not get a heart transplant.