# Chapter 2 - Summarizing Data

Donald Butler

09/12/2021

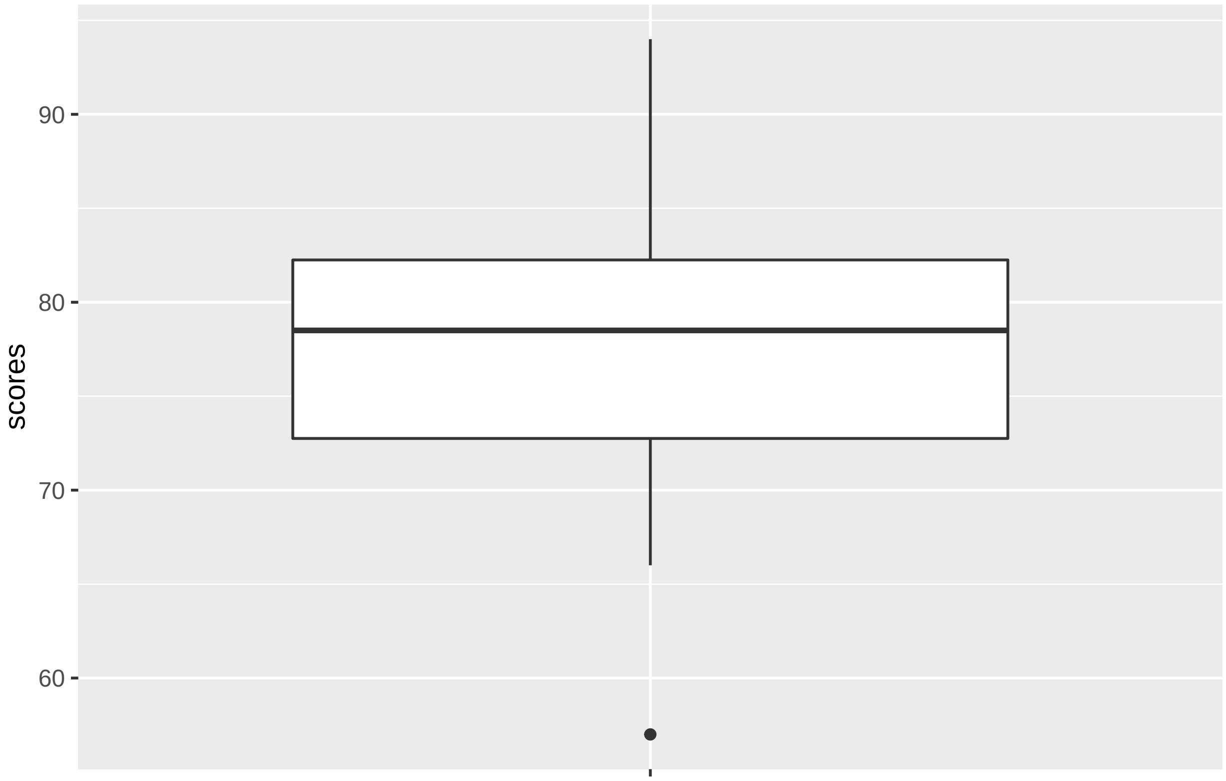
```
library(tidyverse)
```

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

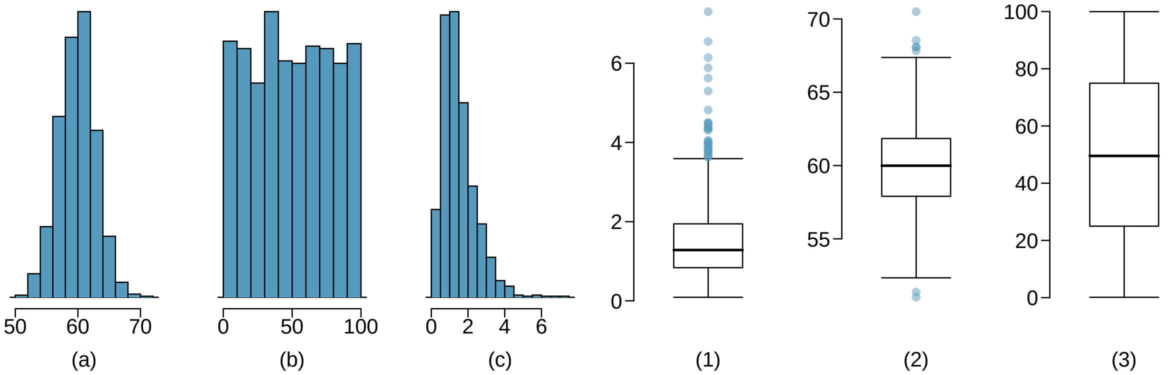
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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 a symmetric distribution and matches up with boxplot (2).
Histogram (b) is a uniform distribution and matches up with boxplot (3).
Histogram (c) is a right skewed distribution and matches up with boxplot (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.

The distribution of housing prices will be right skewed. With the significant number of outliers, the median would be a typical observation and using IQR will be the best test of variability in the market.

(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 of housing prices are symmetrical. With few outliers in the market price, the mean is a good representation of a typical observation and the standard deviation can be used to express variability in the market.

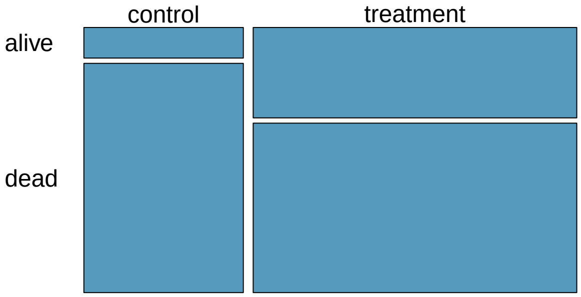
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

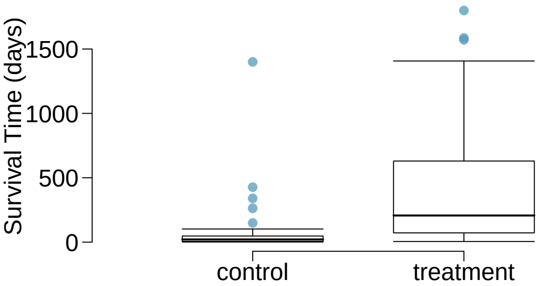
The distribution is likely to be bimodal and right skewed. Due to the significant number of 0 drinks, and a few outliers, it would be best to use the median and IQR.

(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.

The distribution will be right skewed. While there are only a few executives, their salaries could be significantly higher then non-executive employees. The median and IQR would be better used to describe the variability of the distribution.

**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.

Based on the mosaic plot, it appears that survival rates are better in the treatment group, so survival does not appear to be an independent variable with respect to receiving a heart transplant.

(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?

In the control group, 30 of 34 (88%) of the patients died during the study. In the treatment group, 45 of 69 (65%) of the patients died during the study.

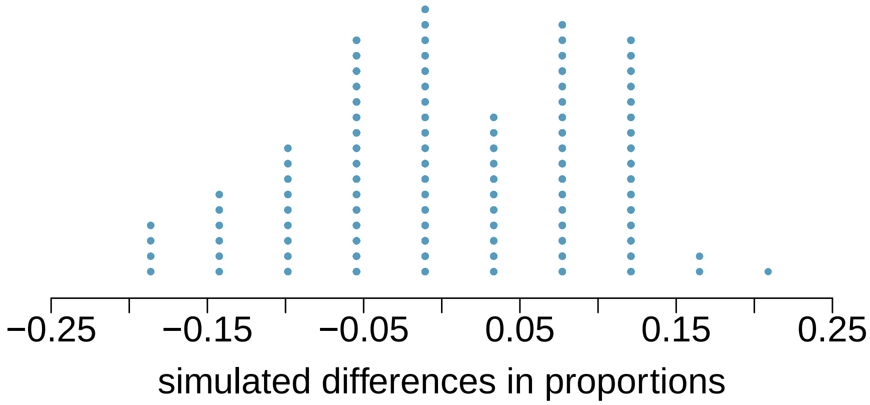
(d) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

i. What are the claims being tested?
Does a heart transplant increase the lifespan of gravely ill patients that have been designated as a transplant candidate?

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 **similar to the observed difference**. 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?



None of the simulated outcomes produced a difference between treatment and control groups of -23%. If the heart transplant treatment played no role in the increased survival rate, then the simulations would have produced the observed difference in the study, so we conclude that the heart transplant did improve survival rates.