Chapter 2 - Summarizing Data

```
# Load needed libraries
library(devtools)
library(tidyverse)
library(RCurl)
library(plyr)
library(knitr)
```

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

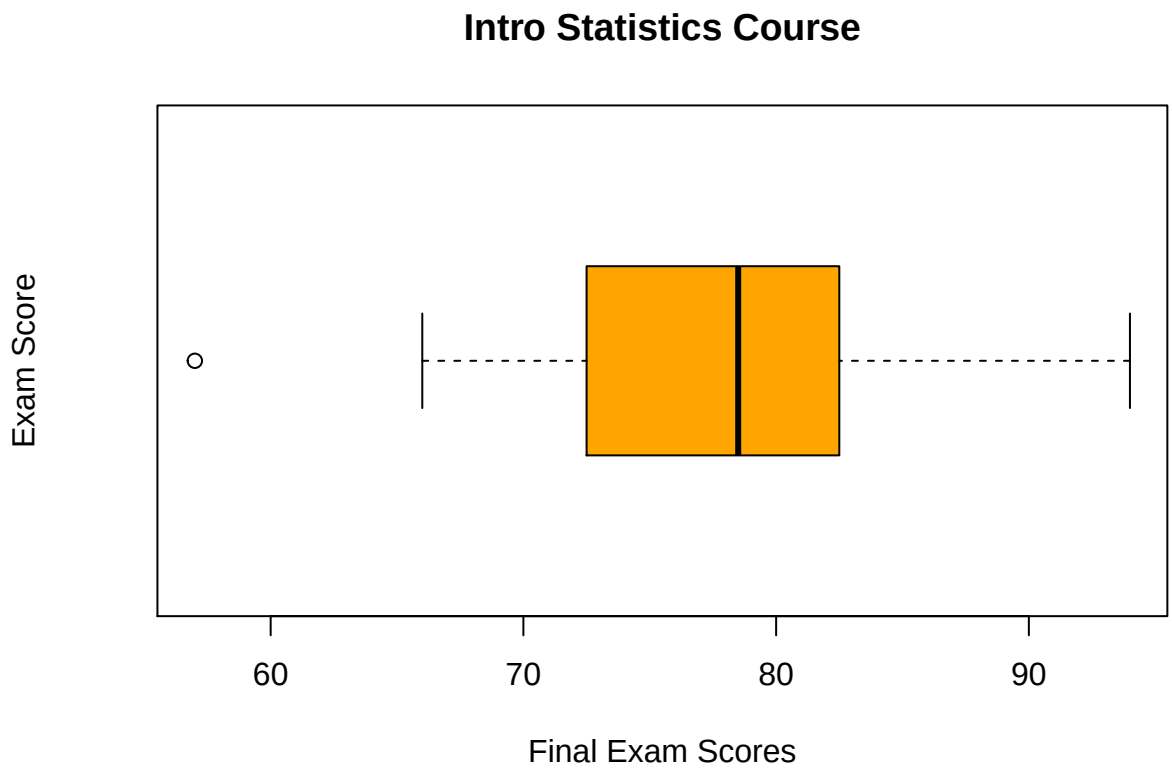
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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

```
##  [1] 57 66 69 71 72 73 74 77 78 78 79 79 81 81 82 83 83 88 89 94
```

```
bxp <- boxplot(scores, main="Intro Statistics Course", xlab="Final Exam Scores",
               ylab="Exam Score", col="Orange", border="black",
               horizontal=TRUE, notch=FALSE)
```




```
sprintf("Mean: %s", mean(scores))
```

```
## [1] "Mean: 77.7"
```

```
sprintf("Min: %s", scores[1])
```

```
## [1] "Min: 57"
```

```
sprintf("Qtr1: %s", bxp$stats[2])
```

```
## [1] "Qtr1: 72.5"
```

```
sprintf("Median(Qtr2): %s", bxp$stats[3])
```

```
## [1] "Median(Qtr2): 78.5"
```

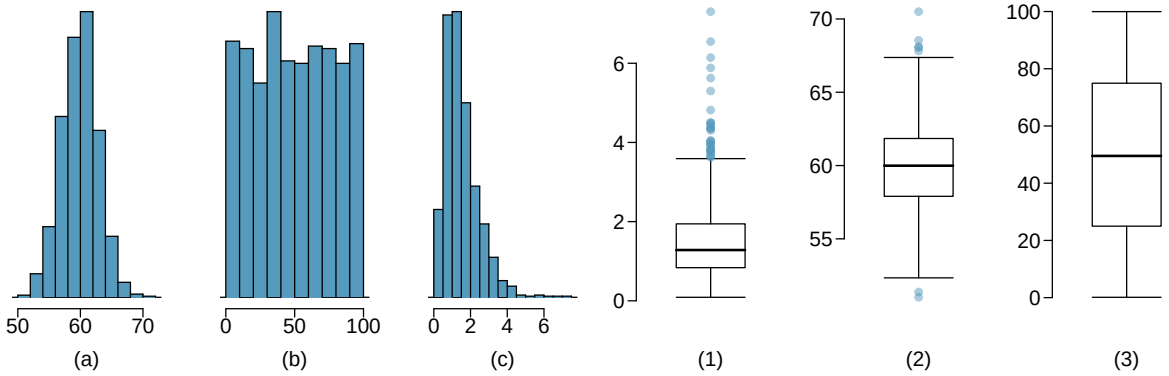
```
sprintf("Qtr3: %s", bxp$stats[4])
```

```
## [1] "Qtr3: 82.5"
```

```r
sprintf("Max: %s", bxp$stats[5])
```

```
## [1] "Max: 94"
```

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



Answers:

(a) is bimodal since it has one prominent peak, mode is approximately 67. It is symmetrical because the left and right tails appear to be equal and so do the other bins on either side of the peak. Boxplot (2) is the match. Note the observations just on the other side of the upper and lower whiskers.

(b) is multimodal as there are more than two prominent peaks. In fact they are all prominent which means that the density of the data is equally distributed. Boxplot 3 is the match for histogram (3). There are no outliers.

(c) is a bimodal because it has two prominent peaks, however, the histogram is right skewed. This means that most of the data bins lie between 0 and 4 and there are some outliers off to the right. Boxplot (1) is its match and from that you can see the number of values above the upper whisker. The lower whisker is much closer to QTR1 and the IQR box is relatively narrow. The upper whisker is further from the QTR3. This matches closely to what the histogram shows.

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 would be skewed right due to the meaningful number of houses that cost more than $6 million. The median would best represent the typical observation. The homes greater than $6 million would skew the mean.

(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 would be symmetrical compared to (a) since there is not so much of a gap between QTR3 and the upper whisker of $1.2 million. The mean would best represent the distribution since it is more symmetrical.

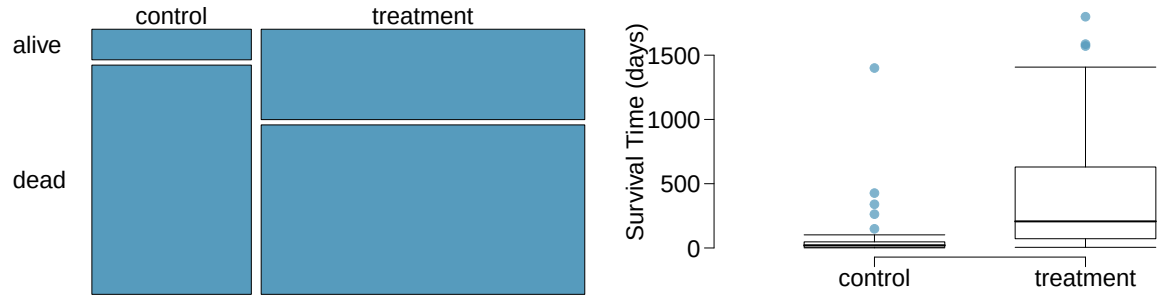
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

Since drinks would be very close to 0 for students younger than 21 years old, the distribution would be left skewed and the number of drinks would increase sharply greater than 21 years old. As a result, the median would better represent the distribution because the much lower number of drinks for those below 21 years of age would significantly lower the mean.

(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 would be symmetrical since there are only a few high level executive salaries and so there would not be a large gap between QTR3 and the upper whisker. The mean would best represent the distribution since it is more symmetrical.

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 looking at the mosaic plot, one could think that survival is independent of the of whether a patient got a transplant or not because the mortality rate in both scenarios is very high. In both cases, the mortality rate is over 50%.

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

The box plots show a different story. Patients who received transplants tend to live much longer than those that did not receive one. It is apparent how much longer transplant patients survived by the size of the box while the size of the box for those that did not recieve a transplant is almost non-existent.

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

```
transplant_prop_img <- "/Users/Audiorunner13/CUNY MSDS Course Work/DATA606 Spring 2021/Week 2/Homework2/
# attr(transplant_prop_img, "info")
include_graphics(transplant_prop_img)
```

	Outcome		
Treatment	Dead	Alive	Total
Transplant	45	24	69
No Transplant	30	4	34
Total	75	28	103

	Outcome		
Treatment	Dead	Alive	Total
Transplant	0.65	0.35	1.00
No Transplant	0.88	0.12	1.00

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

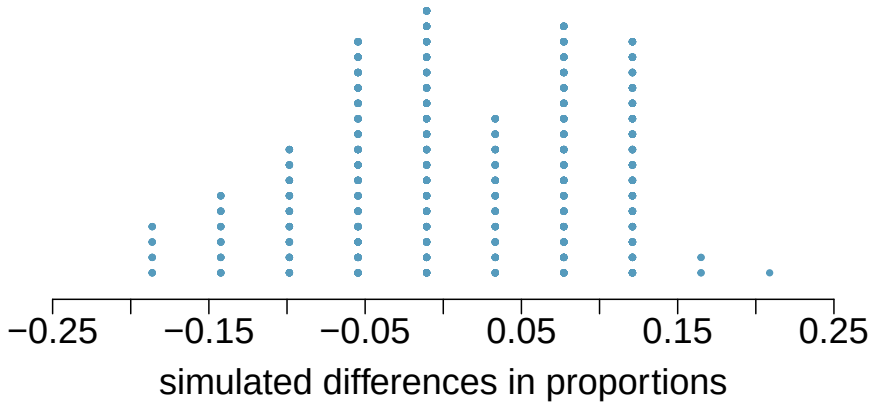
i. What are the claims being tested?

The independence model is being tested - that the treatment and outcome have not relationship. That the observed difference between the proportion who died in teh two groups 23% was due to choice.

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 **index** cards representing patients who were alive at the end of the study, and *dead* on **index** 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 *near zero*. 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?



By the stacked plot it appears that the difference of at least 23% in survival rate was due to chance alone would occur only 1% of the time, this indicates a rare event. When formal studies are conducted, the notion that a rare event was observed is rejected. The independence model reject in favor of the alternative and conclude that the 23% difference in survival rate was due to the heart transplant.