Shana Green

DATA 606 - Homework 2

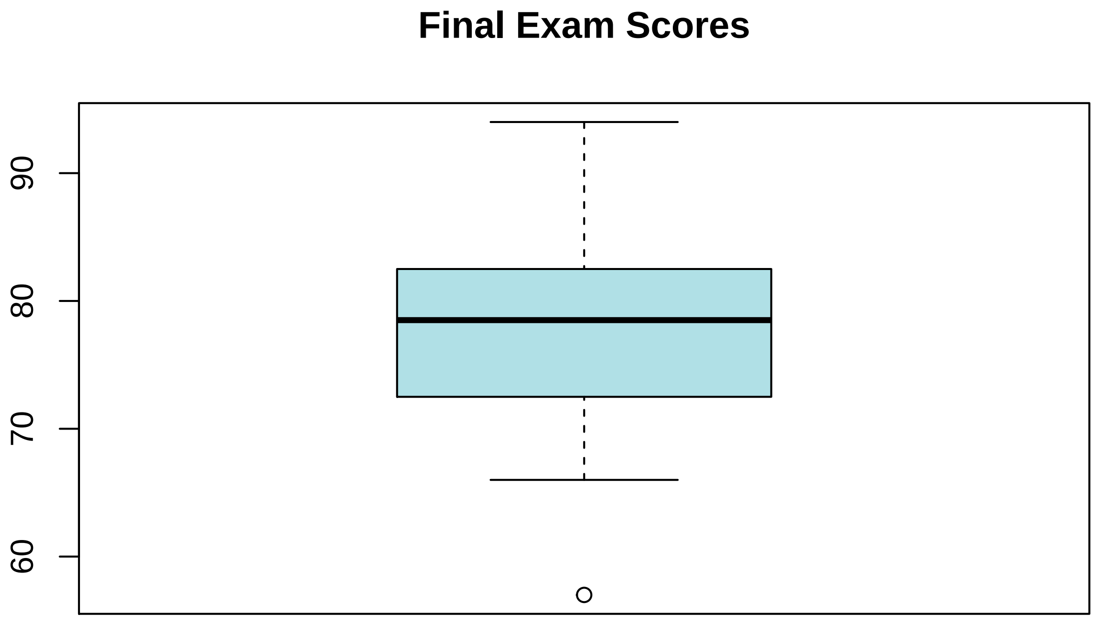
Due Date: 9/06/2020

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.

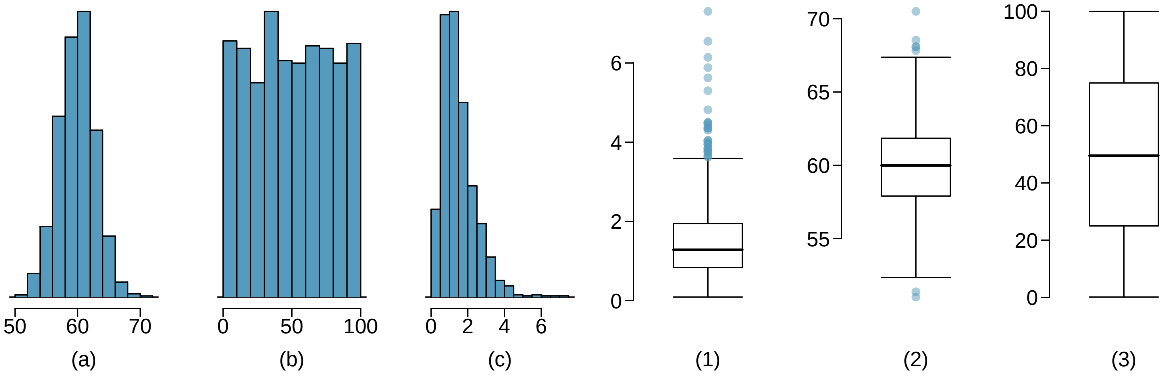
Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94



```r
summary(scores)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   57.00   72.75   78.50   77.70   82.25   94.00
```

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



Answer: 1) Histogram (a) matches with and box plot (2). This histogram is bell shaped and the distribution is symmetrical with a peak at the center.

2) Histogram (b) matches with box plot (3). This histogram has a rectangular shape and the distribution is symmetrical.

3) Histogram (c) matches with box plot (1). This histogram has a long tail at the right side and the distribution is right skewed.

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.

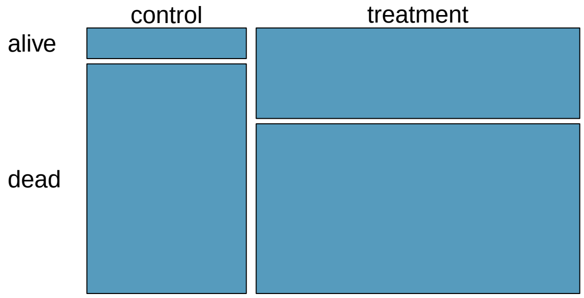
Answer: (a) In this scenario, the distribution is believed to be right skewed. If we look at the range between Q1 and Q2 ($450,000 - $350,000 = $100,000) vs. Q2 and Q3 ($1,000,000 - $450,000 = $550,000), the values are significantly less than the other. Using the median would best represent a typical observation in the data since the data is skewed. The variability of observation in the data would be best represented by IQR, since quartiles were provided in this scenario and the distribution is skewed.

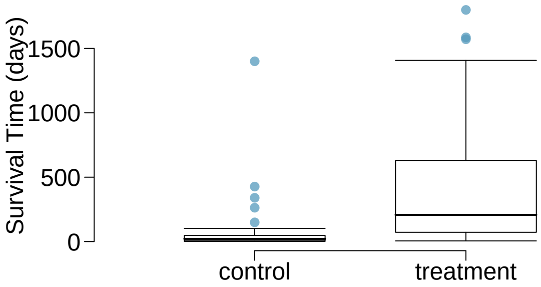
(b) In this scenario, the distribution is believed to be symmetrical because there is no difference between the quartile lengths. If we look at the range between Q1 and Q2 ($600,000 - $300,000 = $300,000) vs. Q2 and Q3($900,000 - $600,000 = $300,000), both values are symmetrical. The measurement that would be best to represent a typical observation in the data is either the mean or median. Since the data is symmetric, both mean and median can be used for best representation of a typical observation of the data. The variability of observation in the data would be best represented by standard deviation or IQR, since the distribution is symmetrical.

(c) In this scenario, the distribution is believed to be right skewed because the data represents that most of the students don't drink. Only few students drink excessively, hence the distribution is right skewed. Since the data is right skewed, the measurement that would be best to represent a typical observation in the data is the median. Outliers are present in the data, this is why the median is best used to represent the typical observation in the data. The variability of observation in the data would be best represented by IQR, since quartiles were provided in this scenario and the distribution is right skewed.

(d) In this scenario, the distribution is believed to be right skewed because very few employees earned more than other employees. The measurement that would best represent a typical observation is the median. The variability of observation in the data would be best represented by IQR.

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.

Answer: Survival is not independent of transplant because the number of people who survived is not the same across both groups.

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

Answer: The heart transplant increases the survival rate for a longer period of time.

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

Answer: From the article we can find as follows:

Control Group alive = 4 dead = 30

total control = alive + dead = 4 + 30 = 34

Treatment Group alive = 24 dead = 45

total treatment = alive + dead = 24 + 45 = 69

Control Group dead population: 30/34 Treatment Group dead population: 45/69

(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?

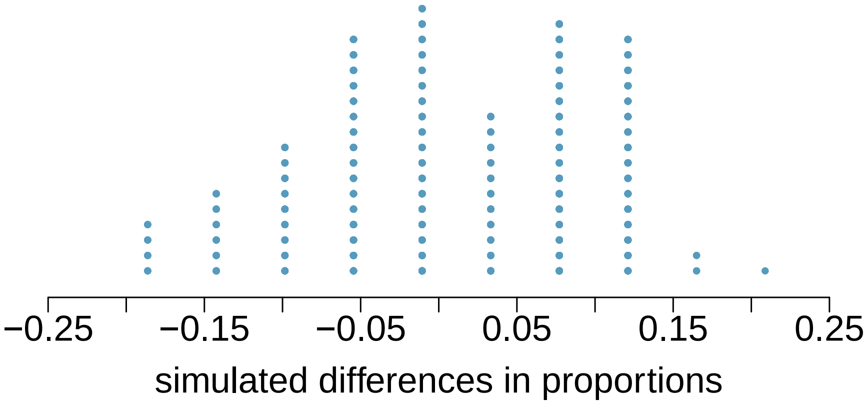
Answer: We want to test whether heart transplant treatments increase lifespan. Null hypothesis: heart transplant treatment has no effect on increasing lifespan.

Alternative hypothesis: heart transplant treatment has an effect on increasing lifespan.

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 **23.02%**_. 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?



Answer: The simulation graph shows simulated differences in proportions less than .25. This number confirms the actual simulated differences in proportion of 0.23.
We can conclude that having heart transplants increases lifespan.