Chapter 2 - Summarizing Data

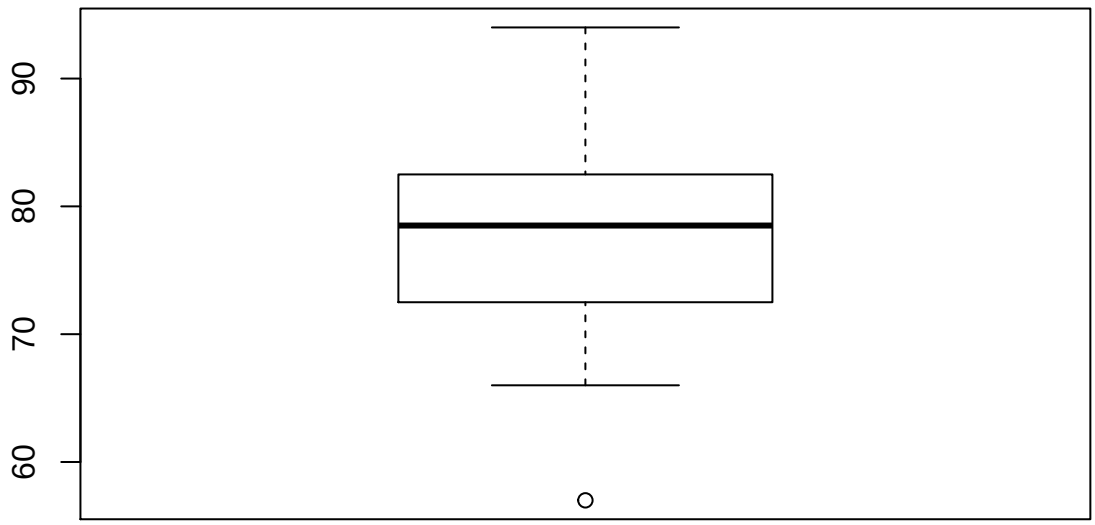
Stats scores. (2.33, p. 78) Below are the final exam scores of twenty introductory statistics students.

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

```
boxplot(scores)
```

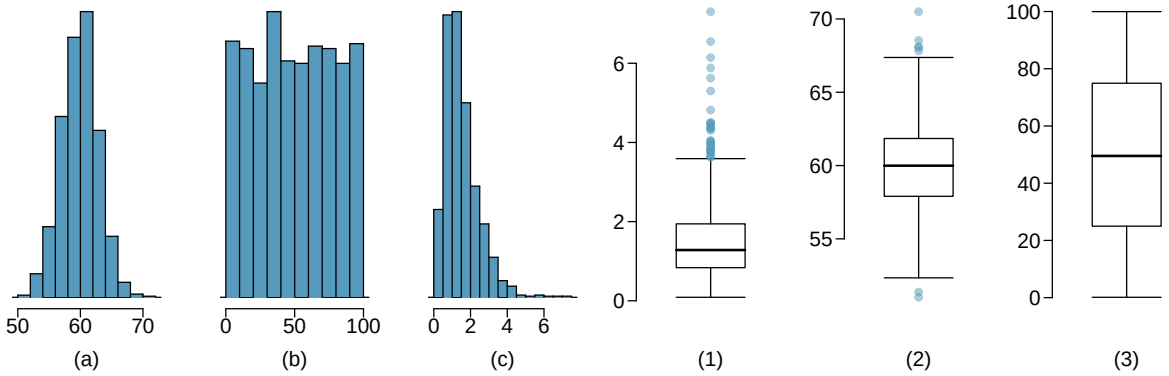


See above for my boxplot of scores. Below, is the corresponding summary of the scores.

```
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.



After examining the distributions and the box plots, it looks like the matches are:

(a) matches with (2)

(b) matches with (3)

(c) matches with (1)

(a) is a unimodal distribution that is symmetric, since there is roughly equal trailing off in both directions.
(b) is a multimodal distribution, with more than two prominent peaks.
(c) is a unimodal distribution that is right skewed, since there is trailing off towards the right.

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.

I expect the distribution to be right skewed, with a long right tail. Robust statistics would be better to represent this data than non-robust statistics. The median would be best to represent a typical observation in this data, and the variability of observations would be best represented using IQR. I believe this is a right skewed distribution because the distance between the median and Q3 ($550,000) is much larger than the distance between the median and Q1 ($100,000), suggesting that a large subset of housing prices fall within a range of $450,000 and below. Additionally, with a meaningful number of houses costing over $6,000,000, which would be considered outliers in this dataset, the IQR, which is more sensitive to numbers near Q1, the median, and Q3, and is thus much more representative than the standard deviation, which would be more sensitive to these outliers.

(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.

I expect the distribution to be pretty symmetrical. Since there isn't much of an indication of outliers, the mean and standard deviation may be a better representation of the typical observation in the data. I believe this is a symmetrical distribution because the distance between the median and Q3 ($300,000) is the same distance as the distance between the median and Q1 ($300,000), suggesting that there is a pretty equal distribution of observations that fall within the IQ. Additionally, with very few houses that cost more than $1,200,000 (which is still considered within the range of the upper whisker in a box plot), there is an indication that outliers wouldn't really affect the mean calculation all that much.

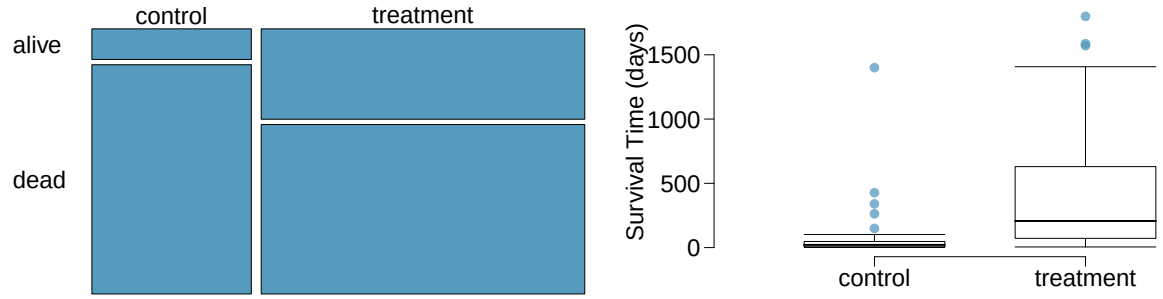
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

I expect the distribution to be right skewed. Robust statistics would be better to represent this data than non-robust statistics. The median would be best to represent a typical observation in this data, and the variability of observations would be best represented using IQR. I believe this is a right skewed distribution because modality here would indicate that most students would consume zero drinks in a given week. Therefore, there will be a large cluster of students around zero, and a median number of drinks close to zero. Since a few students drink excessively, there may be a few outliers (to the far right) in this dataset, so with the IQR being more sensitive to numbers near Q1, the median, and Q3, this statistic would be more representative than the standard deviation, which would be more sensitive to these outliers.

(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.

I expect the distribution to be pretty symmetrical. Since there isn't much of an indication of outliers (only a few high level executives earn much higher salaries than all other employees), the mean and standard deviation may be a better representation of the typical observation in the data. I believe this is a symmetrical distribution because there will likely be a lot of similar salaries will fall within the IQ (not much spread), and there is an indication that outliers wouldn't really affect the mean and standard deviation calculations too much.

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 box plot, survival does not seem to be independent of whether or not the patient got a transplant. I think this is the case because there appears to be a larger proportion of patients who received the treatment that stayed alive than the proportion of patients that did not receive the treatment who stayed alive. However, we need to be cautious about making this assumption, especially given that there appears to be unequal patients represented in the control vs. treatment group (width of the plots are different), which suggests that further testing should occur to confirm that this wasn't merely due to chance.

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

The box plots below suggest that the treatment may have had an effect in extending the survival time of patients (in number of days), from those that did not receive the treatment. With a much larger IQR that we can visually see in the box plot of the treatment group than the control group, this shows that there was a wider spread of survival time in days in the treatment group relative to the control group.

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

First, we create a table to generate the frequencies of these two columns in the dataset:

```
table(heartTr$survived, heartTr$transplant)
```

```
##
##          control treatment
##   alive       4        24
##   dead       30        45
```

Then, we can calculate the proportions:

```r
# total number of patients in the treatment group
total_treatment <- 24 + 45
# total number of patients in the treatment group that died
dead_treatment <- 45
# total number of patients in the control group
total_control <- 4 + 30
# total number of patients in the control group that died
dead_control <- 30
```

Total number of patients in the control group that died:

```r
dead_control / total_control
```

```
## [1] 0.8823529
```

Roughly 88% of patients in the control group died.

Total number of patients in the treatment group that died:

```r
dead_treatment / total_treatment
```

```
## [1] 0.6521739
```

Roughly 65% of patients in the treatment 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?

Independence model: The variables of transplant and survived are independent. They have no relationship, and the observed difference between the proportion of patients who died in the two groups, ~23%, was due to chance.

Alternative model: The variables of transplant and survived are not independent. The difference in death rates of ~23% was not due to chance, and the treatment affected the death rate.

```r
table(heartTr$survived)
```

```
##
## alive  dead
##    28    75
```

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 **greater than or equal to 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 simulation results shown below suggest that a small fraction of simulated differences in proportions were at least 23%. Therefore, this suggests that our initial observation with proportion differences of 23% may have been a rare event. However, when conducting a formal study such as this, we usually reject the notion that such a rare event could have occurred. Thus, we can conclude the evidence is sufficiently strong to reject our null hypothesis and assert that the treatment was effective - we reject the notion that we observed a rare event in our first observation.