Chapter 2 - Summarizing Data

Jeff Shasmp - DATA 606 HW 2

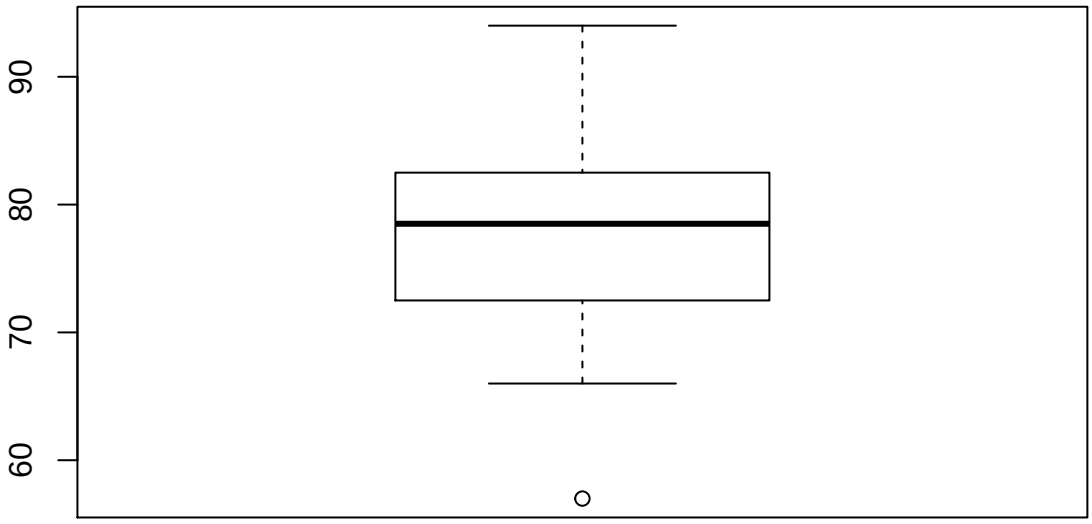
Stats scores. (2.33, p. 78) Below are the final exam scores of twenty introductory statistics students.

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.

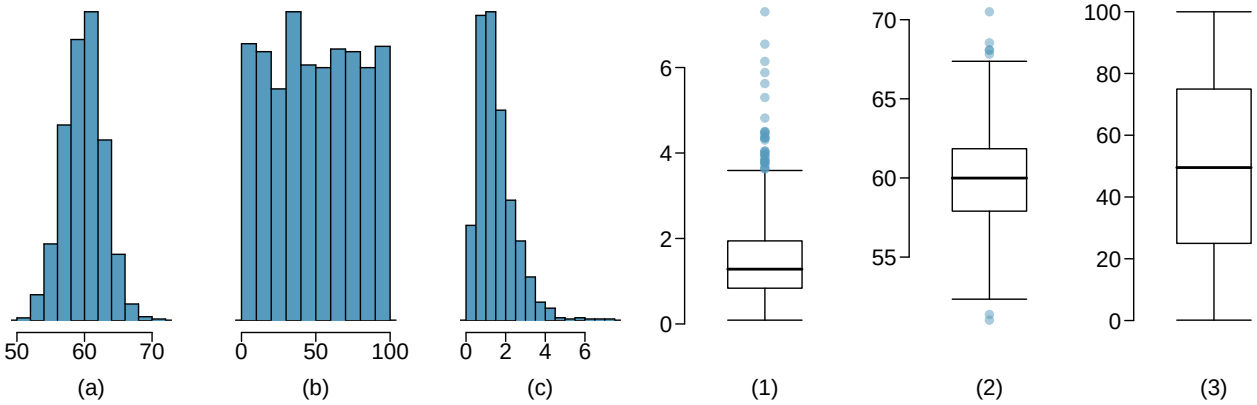
Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

```
boxplot(scores)
```



These results appear to be typical exam scores in my experience.

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



The pairs are as follows: a & 2, b & 3, c & 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.

This would be right skewed since the second quartile is small compared to the third and the third is small compared to the fourth. Since so many values are beyond the third quartile, median and IQR seem like more reasonable descriptors of typical observations and variability. This is due to the robustness of median and IQR. The more very large values (around 6mm) the more likely average and std are going to be adversely affected.

(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.

This data appears to be more normalized and as such average and standard deviation will likely be fine descriptors of typcial observations and variability.

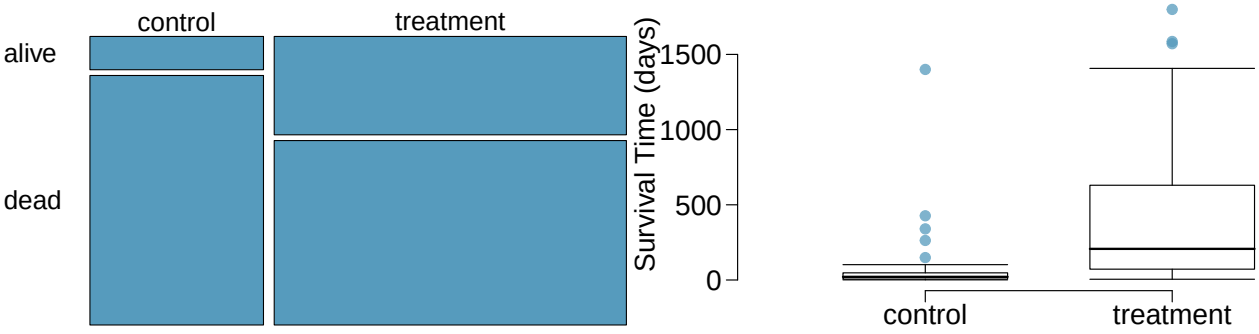
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

College students not drinking as a result of being under 21 years old is a bold assumption, but given what is written this would likely be fairly close to normalized data. As such average and standard deviation should be good descriptors.

(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.

This would represent a right skew where many of the salaries are on the low end ($30-45K) and a long shrinking tail extends into the millions. United Technologies would be a good example of this type of company. The average salary at UTC is $60K (with minor equity grants) and the CEO makes 17.5mm in total compensation.

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 solely on the mosaic plot, we could make the conjecture that survival is not independent of transplant. That is, the 23% change in survival for control and treatment groups are most likely not the result of chance. This would not be a conclusive analysis, and more work would need to be done beyond a mosaic plot, but it does indicate a interesting difference between treatment and control.

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

It appears as thouugh the survival time of the individuals is increased. That is, a typical patient would likely see an increase in life span as the result of a transplanted heart. Here the outliers for the control group (except one) are within the IQR of the treatment group and the median of the treatment group is in outlier range for the control group.

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

```
morality_rate_control<- 30/34
morality_rate_control
```

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

```
morality_rate_treat<- 45/69
morality_rate_treat
```

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

(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?

That receiving this heart transplant increases the lifespan of patients in gravely ill heartcase situations. That recieving this treatment produces a better than random chance increase in lifespan of affected patients.

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 **zero**_. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are ***similar to the experimental result***. 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 observed survival rate of ~23% occurs in 1% of the random simulations. This is a rare event.