Chapter 2 - Summarizing Data

Zachary Safir

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

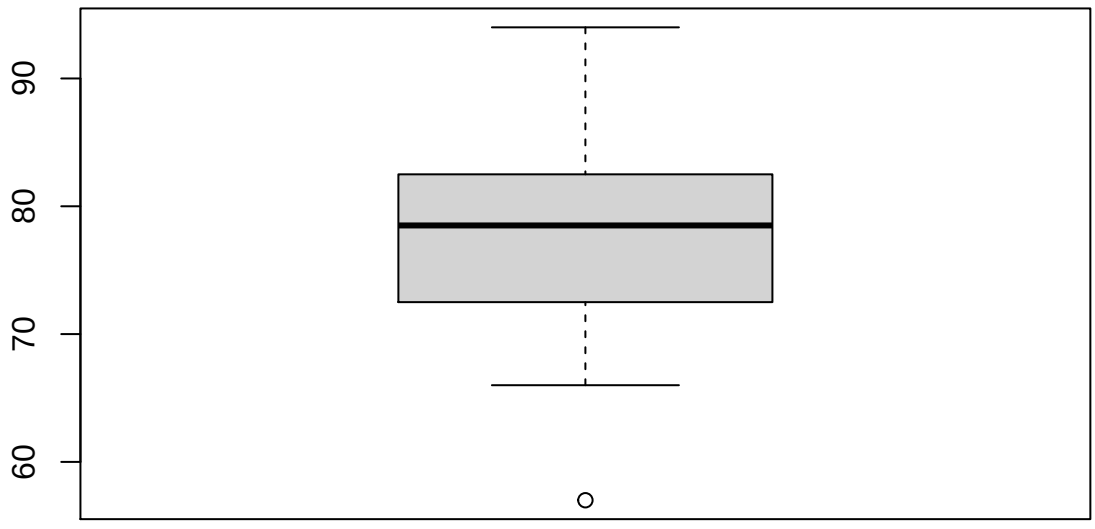
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

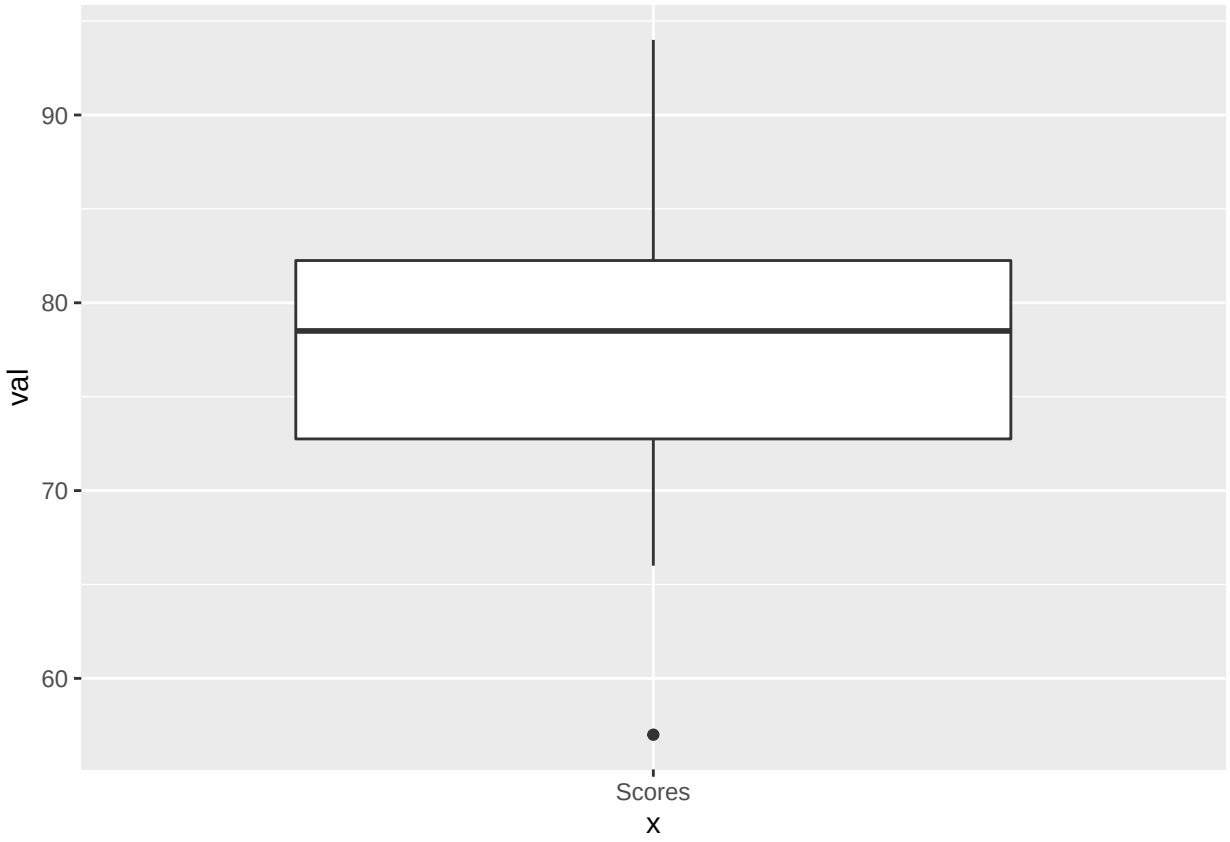
Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

Two methods for generating the box plot.

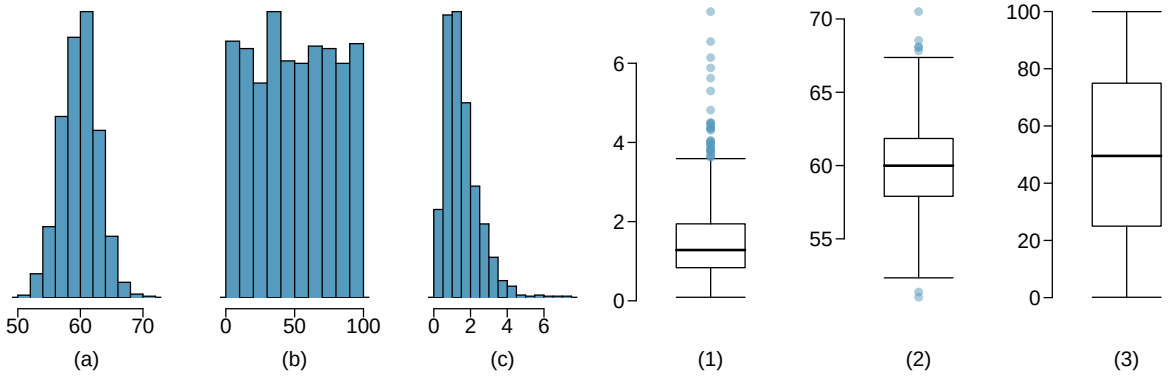
```
 library(ggplot2)
library(dplyr)


boxplot(scores)
```



```
data_frame(val=scores) %>%
ggplot(aes(x='Scores',y=val)) +
  geom_boxplot()
```



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



A is a unimodal, symmetric, nonskewed histogram. It matches with boxplot 2. B is seemingly uniform in shape. It matches with boxplot 3. C is skewed right and corresponds to 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.

Q1 are houses below $350,000, median $450,000, Q3 houses below $1,000,000 and the IQR is 650,000. The problem states that here are a meaningful number of houses that cost more than $6,000,000. This data is right skewed and we should use the median to best represent the data. Should use the IQR due to the extreme 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.

Q1 is $300,000, median is $600,000, Q3 is $900,000, and the IQR is 600,000. This distribution of this data is symmetrical. We should use the mean and the standard deviation.

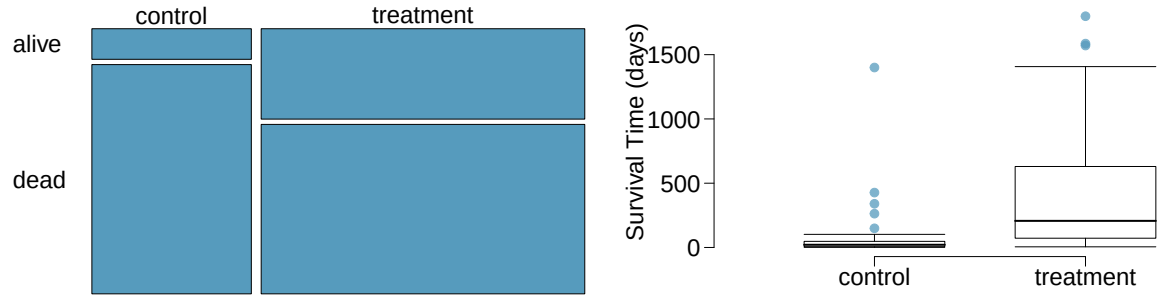
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

Right skewed distribution as the majority of values will be on the lower end with excessive drinkers as outliers that will skew the data.

(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.

It depends on the size of the company and the distribution of the non high level employees. In a large company, the few high level executives would likely not skew the data, and it would have a normal distribution. In a smaller company, with non high level employees that are paid significantly less, we might see a skewed 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.

```
control_percent_s <- (4/34) * 100
treat_percent_s <- (24/69) * 100

control_percent_s
```

```
## [1] 11.76471
```

```
treat_percent_s
```

```
## [1] 34.78261
```

As we can see from the percentage of survivors, and from the box plot above, the difference between each group is significant enough that we can say that survival is not independent of getting a transplant

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

The median survival rate and overall distribution of the treatment group is significantly grater than the control group. It is apparent that the ones in the treatment group live for a much longer time.

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

```
control_percent_d <- (30/34) * 100
treat_percent_d <- (45/69) * 100

control_percent_d
```

```
## [1] 88.23529
```

```
treat_percent_d
```

```
## [1] 65.21739
```

(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?

To determine whether an experimental heart transplant program increased 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 24 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 (45/69) - (30/34). Lastly, we calculate the fraction of simulations where the simulated differences in proportions are -.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?

It shows what we found above, that the treatmeant does increase the lifespan of the patients in the study.