Chapter 2 - Summarizing Data

Sam Reeves

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

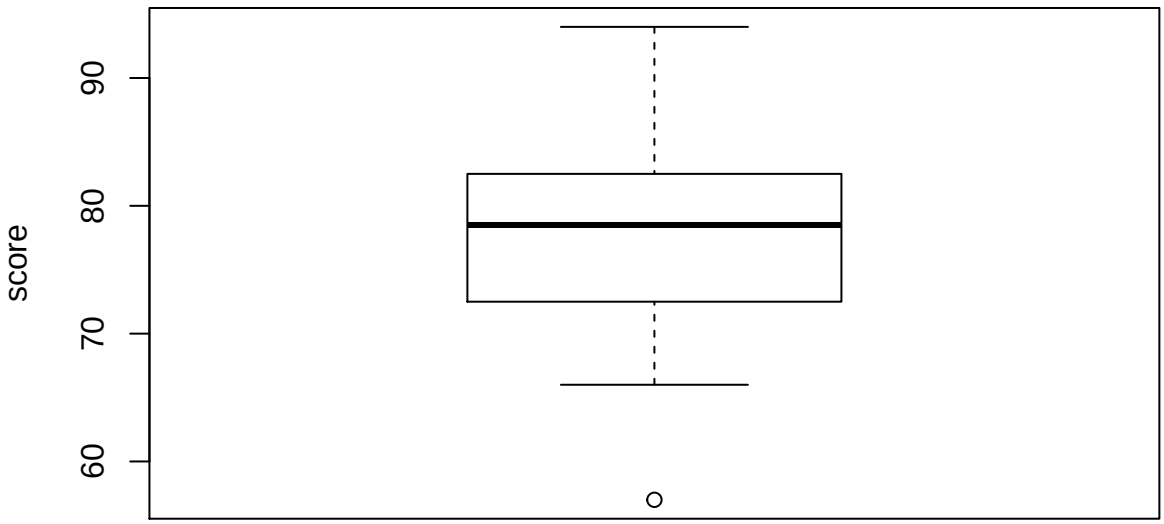
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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

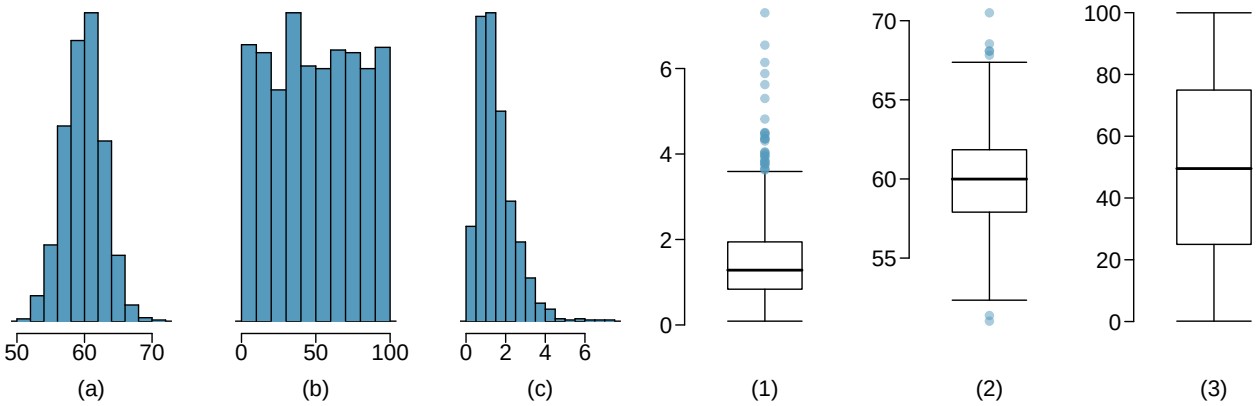
Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

```
boxplot(scores,
        ylab = 'score')
```



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



(a)

(3) normally distributed, symmetrical, monomodal

(b)

(2) polymodal

(c)

(1) monomodal, skewed 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.

–

Median is best, because the significant number of observations with a higher-than-normal magnitude will offset the mean. This would be right skewed. Standard deviation would give a better picture of varying prices.

(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.

–

Still, a median is best, because there could be a few extreme outliers that would affect the average. This would be slightly right skewed. IQR might give a better understanding of the regularly distributed data.

(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.

–

The mean score would be more useful in this case because there are potentially many observations which are zero. You would have a more accurate budget for your next party. IQR would help account for all the normies.

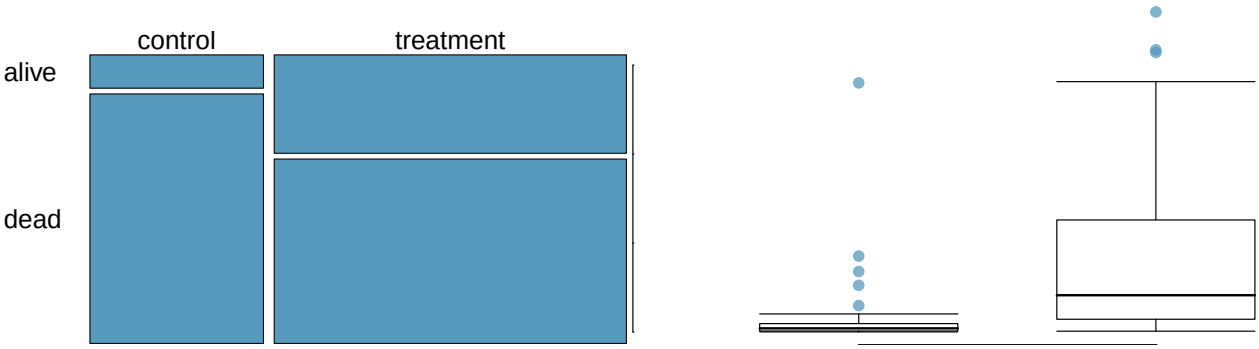
(d)

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.

–

In this case, median score would tell you a real value instead of a simulated one based on the entire populaiton. The data would be skewed hard right, and a standard deviation would be the most telling feature for scale.

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.

Yes, it looks like a significant number of people survived longer with treatment. Maybe there isn't enough information to be sure.

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

They suggest that a robust portion of the candidates survived longer with treatment.

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

```
t_death <- nrow(filter(heart_transplant,
                  transplant == 'treatment',
                  survived == 'dead'))
c_death <- nrow(filter(heart_transplant,
                  transplant == 'control',
                  survived == 'dead'))

cat("Treatment group deathtoll: ",
    t_death / sum(heart_transplant$transplant == 'treatment'))
```

```
## Treatment group deathtoll:  0.6521739
```

```
cat("Control group deathtoll: ",
    c_death / sum(heart_transplant$transplant == 'control'))
```

```
## Control group deathtoll:  0.8823529
```

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

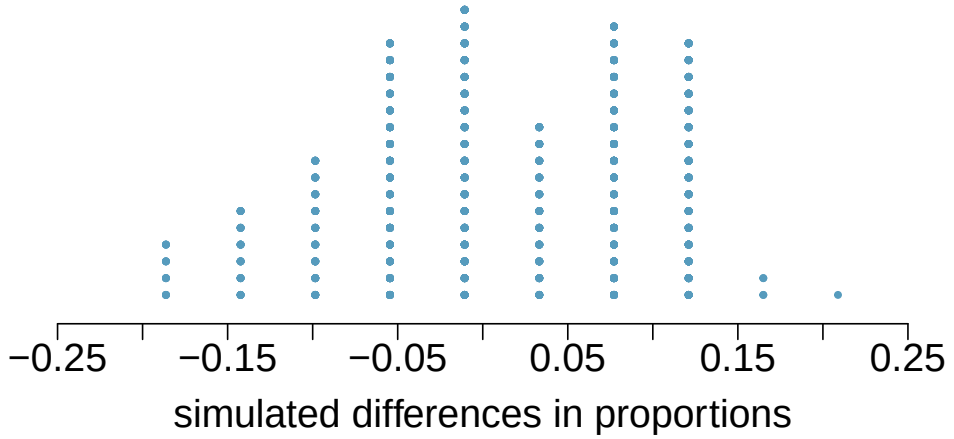
i. What are the claims being tested?

Heart transplant positively affects survival rate in this group of patients who were already in extremely poor health.

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 **THE MEDIAN**. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are **0**. 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 is entirely possible that these results occurred by chance. This is not a study which can be considered representative of the population.