Chapter 2 - Summarizing Data

David Simbandumwe

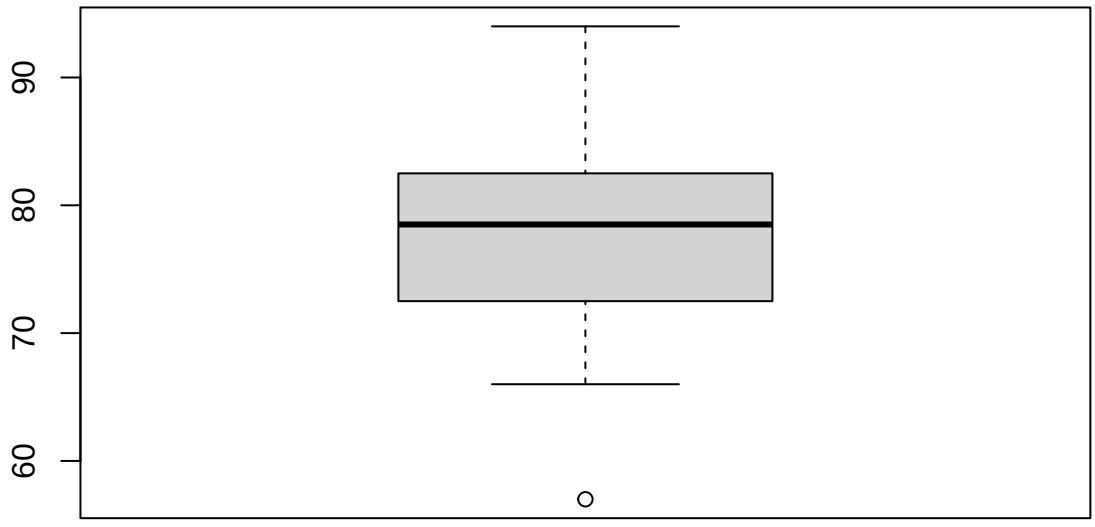
Stats scores. (2.33, p. 78) Below are the final exam scores of twenty introductory statistics students.

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

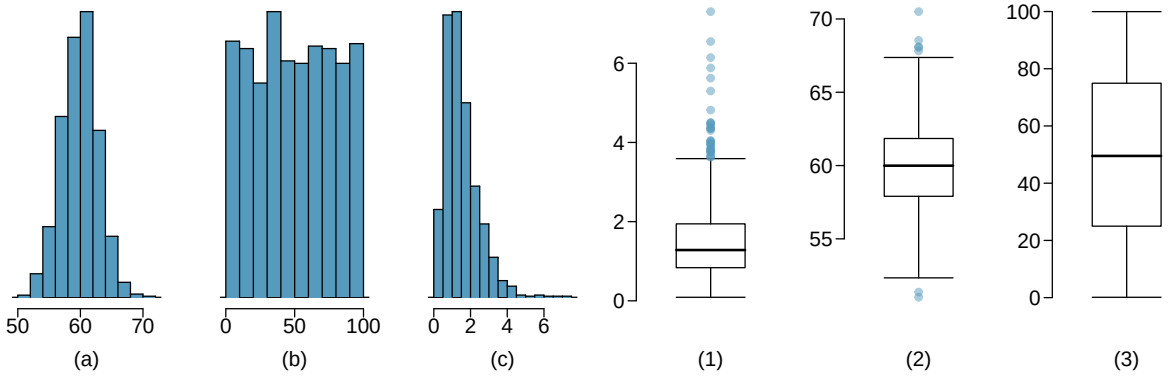
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

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

Questions 1



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



Question 2

```
*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.

- distribution: right skewed
- typical observation: median
- variability: IQR

(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.

- distribution: symmetric
- typical observation: mean
- variability: standard deviation

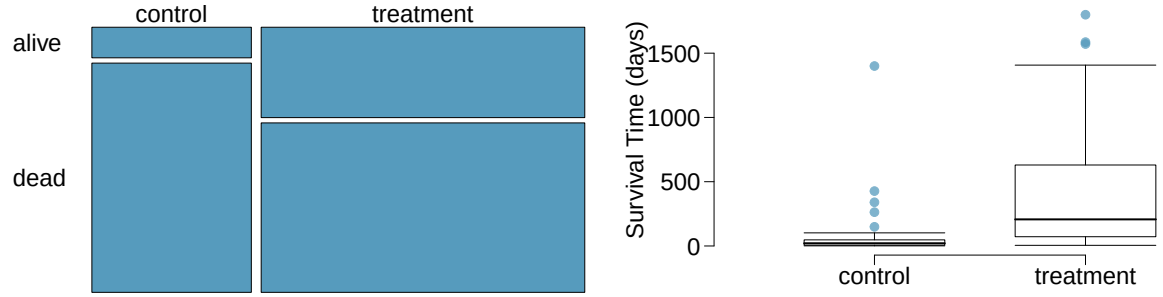
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

- distribution: left skewed
- typical observation: median
- variability: IQR

(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.

- distribution: right skewed
- typical observation: median
- variability: 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.

```
##        id            acceptyear        age          survived     survtime
##  Min.   :  3.00   Min.   :68.0   Min.   :19.00   alive:24   Min.   :   5.0
##  1st Qu.: 32.00   1st Qu.:69.0   1st Qu.:42.00   dead :45   1st Qu.:  72.0
##  Median : 56.00   Median :71.0   Median :47.00              Median : 207.0
##  Mean   : 54.32   Mean   :70.9   Mean   :45.48              Mean   : 415.4
##  3rd Qu.: 79.00   3rd Qu.:72.0   3rd Qu.:52.00              3rd Qu.: 630.0
##  Max.   :100.00   Max.   :74.0   Max.   :64.00              Max.   :1799.0
##  prior        transplant      wait
##  no :57   control  : 0   Min.   :  1.00
##  yes:12   treatment:69   1st Qu.: 10.00
##                          Median : 26.00
##                          Mean   : 38.42
##                          3rd Qu.: 46.00
##                          Max.   :310.00
```

```
##        id            acceptyear         age          survived     survtime
##  Min.   :  1.00   Min.   :67.00   Min.   : 8.00   alive: 4   Min.   :   1.00
##  1st Qu.: 17.50   1st Qu.:68.00   1st Qu.:39.25   dead :30   1st Qu.:   6.00
##  Median : 39.50   Median :70.00   Median :46.00              Median :  21.00
##  Mean   : 45.47   Mean   :70.06   Mean   :42.94              Mean   :  96.62
##  3rd Qu.: 72.75   3rd Qu.:71.00   3rd Qu.:51.75              3rd Qu.:  47.50
##  Max.   :103.00   Max.   :74.00   Max.   :59.00              Max.   :1400.00
##
##  prior        transplant      wait
##  no :34   control  :34   Min.   : NA
##  yes: 0   treatment: 0   1st Qu.: NA
##                          Median : NA
##                          Mean   :NaN
##                          3rd Qu.: NA
##                          Max.   : NA
##                          NA's   :34
```

** From the data it appears that survival is dependent on if a patient receives a transplant. The median and mean of the treatment group are higher than *treatment mean 415.4 vs control mean 96.62 and treatment median 207.0 vs control median 21.0* also the first and the 3rd quartile are higher. with the 3rd quartile being substantially higher in the treatment group meaning that 75% of the patients live substantially longer than control group patients.

**

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

+The plots indicate that having a heart transplat substantially increase the median survival time and the overall survival time for 75% of the treatment group.

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

```
heart_transplant %>%
  group_by(transplant) %>%
  summarise(survivalRate = sum(survived == 'dead') / n())
```

```
## # A tibble: 2 x 2
##   transplant survivalRate
##   <fct>             <dbl>
## 1 control           0.882
## 2 treatment         0.652
```

- control: 88.2% died
- treatment: 65.2% 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?

- Does a heart transplant increase the likelihood of survival or increase the patients 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.

```
h_trans <- heart_transplant %>%
  group_by(survived,transplant) %>%
  summarize(
    count = n()
  )
```

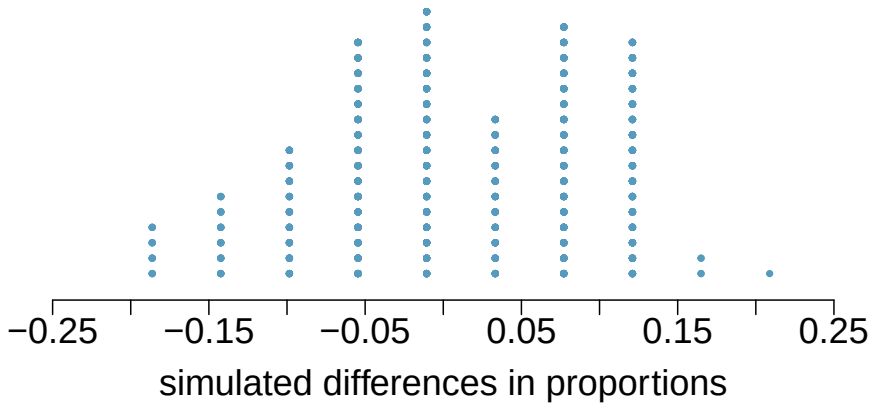
```
## `summarise()` has grouped output by 'survived'. You can override using the `.groups` argument.
```

```
h_trans
```

```
## # A tibble: 4 x 3
## # Groups:   survived [2]
##   survived transplant count
##   <fct>    <fct>      <int>
## 1 alive    control        4
## 2 alive    treatment     24
## 3 dead     control       30
## 4 dead     treatment     45
```

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 _____-0.2302_____. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are _____less than_____. 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 results of the simulation suggest that the treatment increases the likelihood of survival for the patient.

This document is available at [**RPubs**] and on [**Github**]