# Chapter 2 - Summarizing Data

*Joe Rovalino*

**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

```
y <-as.integer(c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94))
y
```

```
##  [1] 57 66 69 71 72 73 74 77 78 78 79 79 81 81 82 83 83 88 89 94
```

```
typeof(y)
```

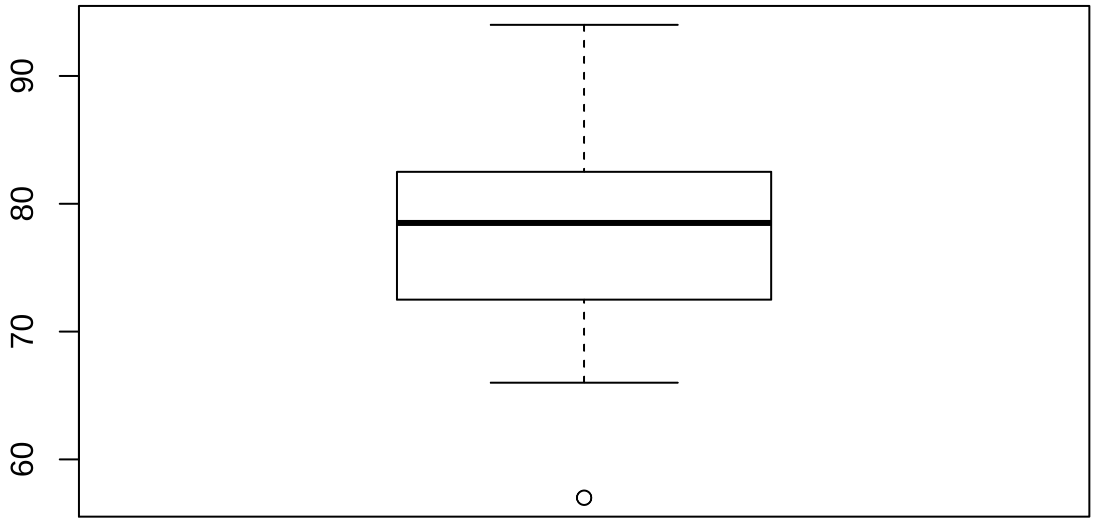
```
## [1] "integer"
```

```
ss <- data.frame(x = c(1:20), y)
ss
```

```
##     x  y
## 1   1 57
## 2   2 66
## 3   3 69
## 4   4 71
## 5   5 72
## 6   6 73
## 7   7 74
## 8   8 77
## 9   9 78
## 10 10 78
## 11 11 79
## 12 12 79
## 13 13 81
## 14 14 81
## 15 15 82
## 16 16 83
## 17 17 83
## 18 18 88
## 19 19 89
## 20 20 94
```

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

```
boxplot(ss$y)
```

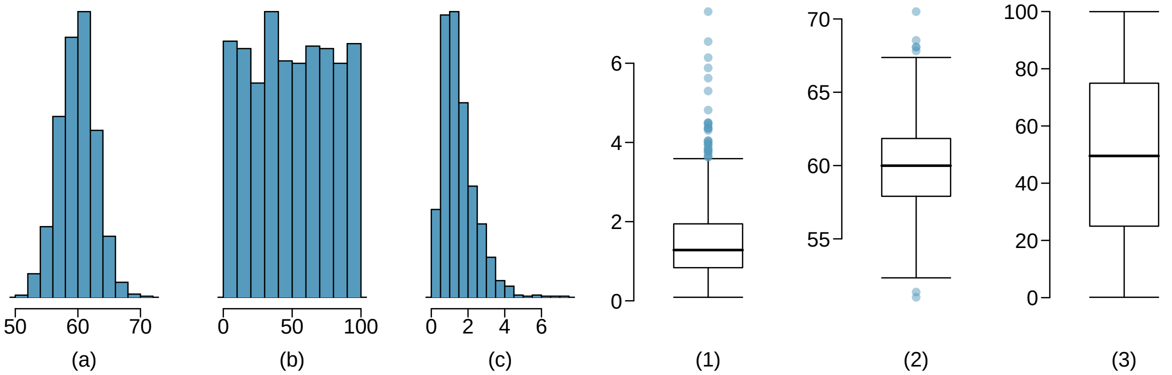


```r
summary(ss$y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   57.00   72.75   78.50   77.70   82.25   94.00
```

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

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



JR Answer:

histplot a matches with boxplot 2 histplot b matches with boxplot 3 histplot c matches with 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.

JR Answer: (a) is right skewed. Median would be best representative as it is "more robust statistic because extreme observations have little effect" pg 51 OSv4.pdf. IQR would best represent the variability as the robust statistic.

(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.

JR Answer: (b) is symmetric. Either the mean or median would represent a typical observation. IQR or SD would represent the variability since symmetric.

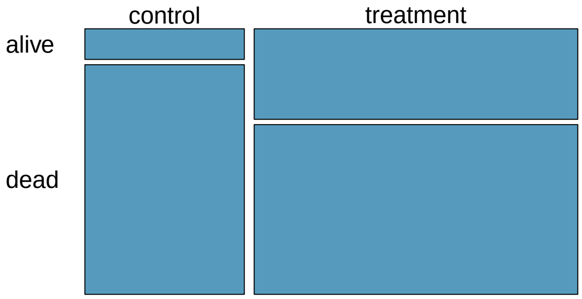
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

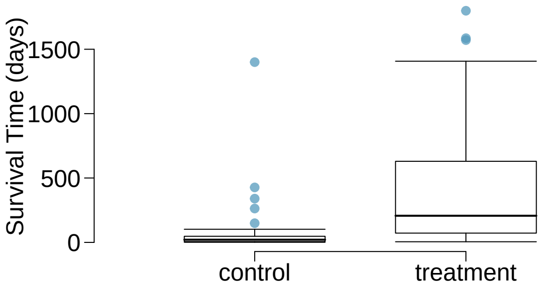
JR Answer: (c) is right skewed. Median would be best to have least impact from outliers. IQR qould better handle the variability from the outliers.

(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.

JR Answer: right skewed. Median would be best to have least impact from outliers. IQR qould better handle the variability from the outliers.

**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.

JR Answer: Looking at the Mosiac plot the treatment group had a little more then 12% better alive metrics (30/34 = .88 died vs 45/69 = .65 died). If survival was independent of a transplant - would epxect to see a closer number to .88.

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

JR Answer: The median is higher for survival in the transplant category. Also the Q3 (75 quartile is longer in the treatment category so survimal is clearly longer. (IQR is larger number in the treatment group then in control group)

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

JR Answer: (30/34 = .88 died vs 45/69 = .65 died)

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

JR Answer: 4 Techniques: controlling, randomization, replication and blocking.

i. What are the claims being tested?

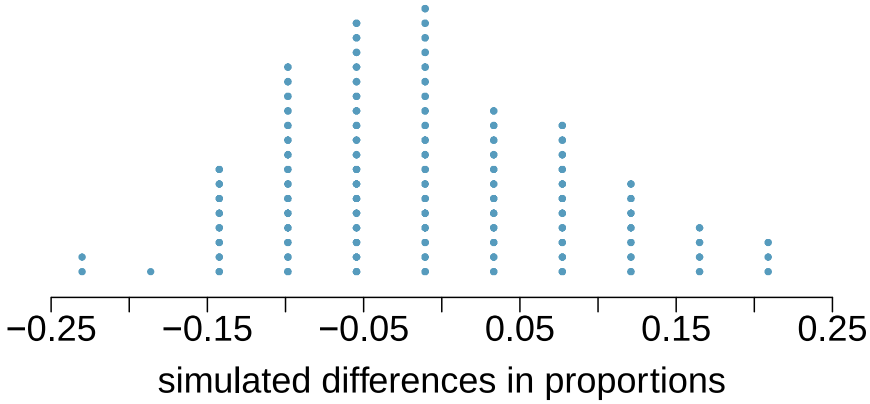
JR answer: whether an experimental heart transplant program increased lifespan.

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 ***Note***______ cards representing patients who were alive at the end of the study, and *dead* on ***note***_____ 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 _______normal_***. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are*** _*normal*_______. 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?

JR Answer: The effetiveness is not significant since highest proportions are around -1 to 0.


simulated differences in proportions