Chapter 2 - Summarizing Data

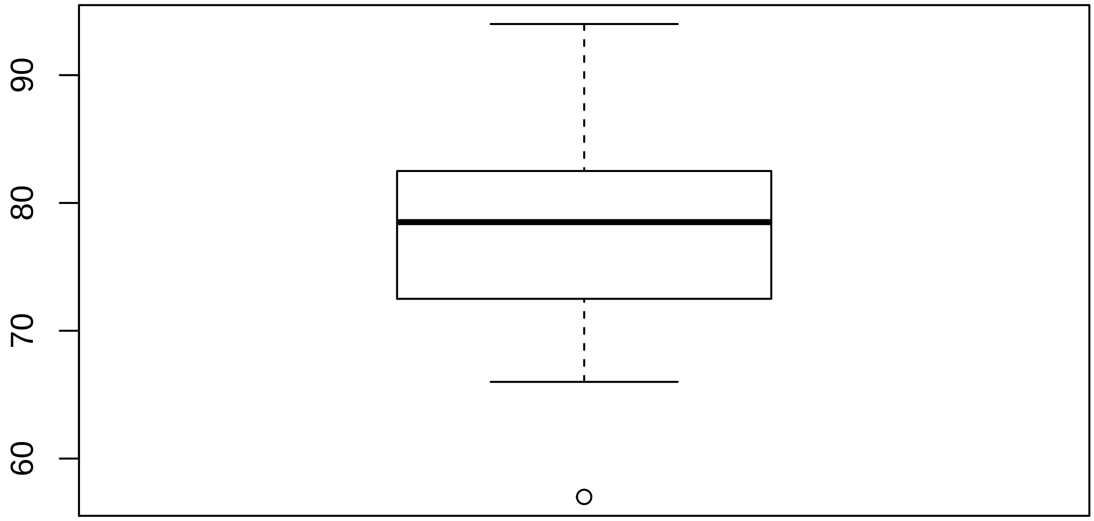
Jai Jeffryes

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

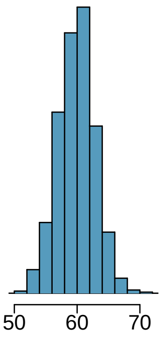
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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

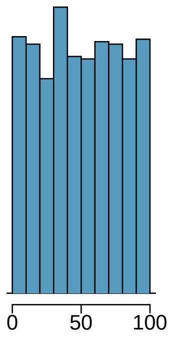
Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94



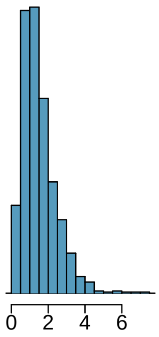
Mix-and-match. (2.10, p. 57) Describe the distribution in the histograms below and match them to the box plots.



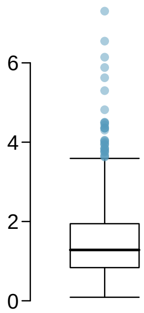
(a)



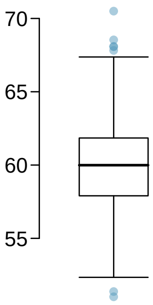
(b)



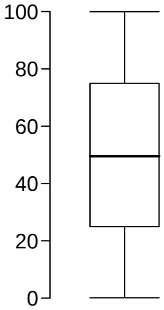
(c)



(1)



(2)



(3)

- Histogram (a) matches boxplot (2). Its distribution is symmetric and unimodal. The median value is about 60. The spread is not very wide, though there are a few outliers at the high and low ends.
- Histogram (b) matches boxplot (3). Its distribution is multimodal. There is no clearly prominent peak, with frequencies roughly uniform across the range. The median is about 50 and the half the values (2nd and 3rd quartiles) fall within about 15 units. The range extends from 0 to 100.
- Histogram (c) matches boxplot (1). The distribution is unimodal and skewed to the right. The median is a little greater than 1. There is a strong central tendency with a narrow spread. There are many outliers on the high end.

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.

ANSWER

Right skewed. The median is $450,000 and with many cases above $1M, the mean of this dataset is higher than the median, skewing to the right. Since the extreme values are said to be meaningful, I would select the measures of central tendency and variability which are less robust, namely median and standard deviation, so that they would reflect those extreme values.

(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.

ANSWER

This distribution is symmetric and will appear pretty flat, in other words widely spread. The mean and median will be pretty close and either statistic will represent well a typical observation. I would use the IQR to understand the variability since it is more robust than SD and will be a better measure of variability, representing the greater part of the data and affected less by the extreme values.

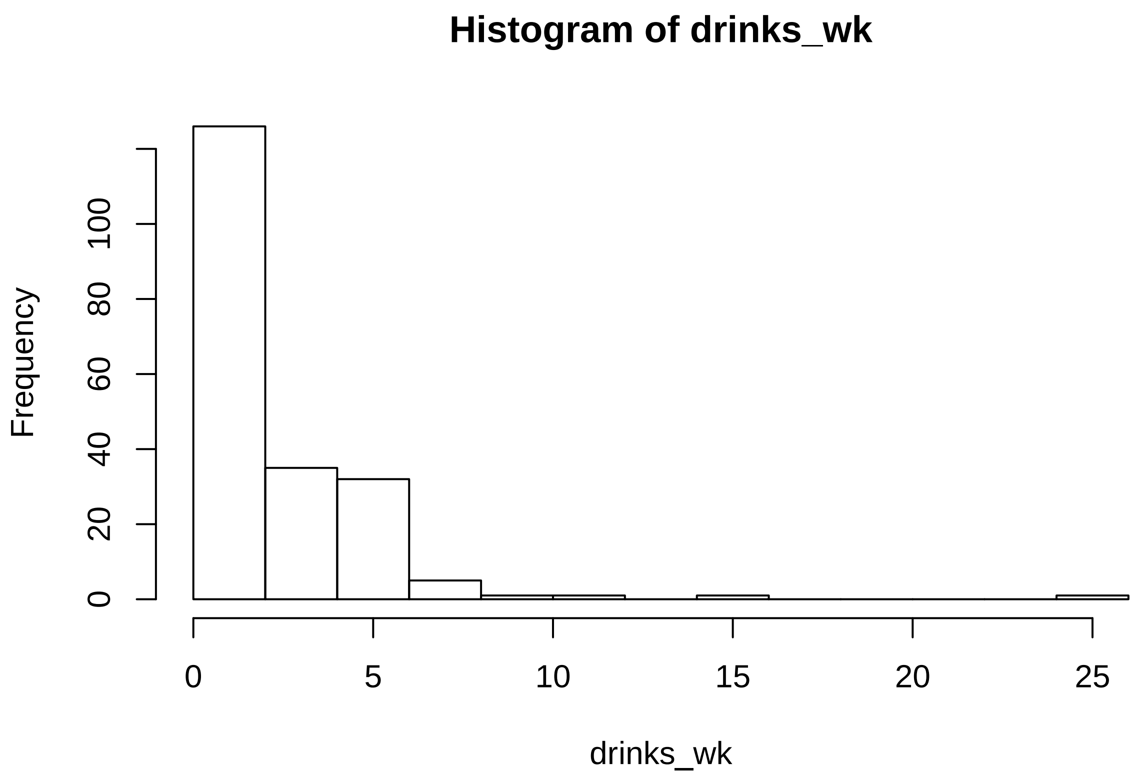
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don’t drink since they are under 21 years old, and only a few drink excessively.

ANSWER

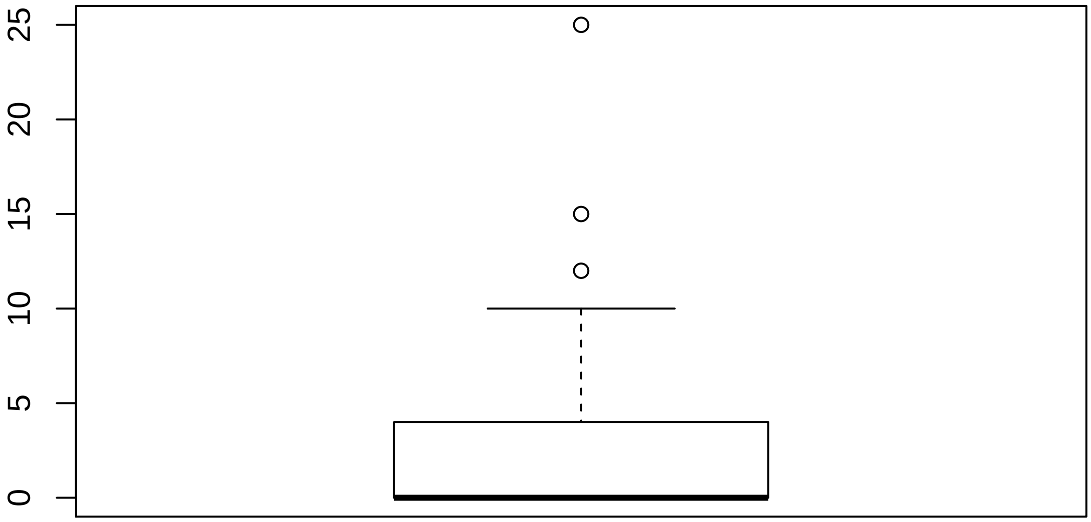
I think this distribution will be bimodal. There will be a peak at 0 because of the underage students, and then a symmetric, normal distribution of students who are able to drink. I would rely on median as a measure of the typical case since it lies closer to the category of underage, non-drinking (presumably!) students. I would rely on IQR for variability because of the high number of students who do not fall in the normal distribution range of legal age students.

I was curious what this would look like, so I manufactured a toy dataset with my assumptions and plotted it. I would be curious to learn a better way than mine. I wanted to create a random distribution within a range. `rnorm()` can extend into negative values, so I just threw those out of my toy dataset. Then I rounded down to get whole numbers.

```r
# Toy data set.
set.seed(23)
drinks_wk <- c(rep(0, 100), rnorm(100, mean = 4, sd =2))
drinks_wk <- drinks_wk[drinks_wk >= 0]
drinks_wk <- floor(drinks_wk)
drinks_wk <- c(drinks_wk, 10, 12, 15, 25)
hist(drinks_wk)
```



```
boxplot(drinks_wk)
```



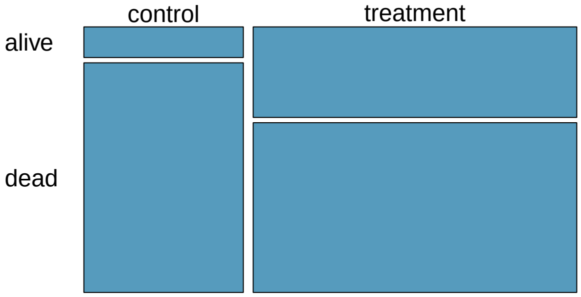
My fake data came out with a clear right skew. Even with my assumptions, it didn't appear bimodal. I don't know if that's because I was just wrong about that, or my approach at manufacturing the data didn't implement my assumptions.

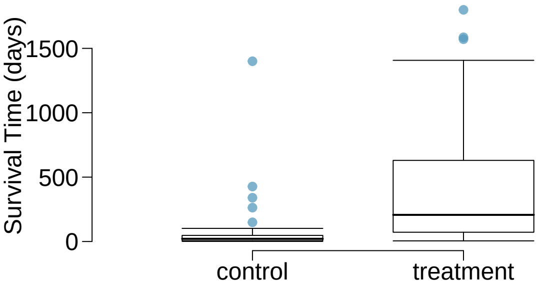
(d) Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than the all other employees.

ANSWER

Though there are a few high outliers, I expect the distribution to be symmetric. However, the executive salaries will affect the mean and the standard deviation. Therefore, for typical values and variability I would rely on median and 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.

ANSWER

No, survival is dependent on a transplant, or in other words, survival and transplants are associated. The mosaic plot presents a differing vertical alignment between the two horizontal groups, control and treatment, and since there is a difference in outcome, the results suggest an association.

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

ANSWER

The IQR of the treatment group shows that survival is extended a year and a half beyond the narrow IQR for the control group. The median survival days and the maximum are also higher for the treatment group. Heart transplants appear to be effective treatment.

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

ANSWER

```r
results <- data.frame(cbind(c(45, 30), c((69 - 45), (34 - 30))))
colnames(results) <- c("dead", "alive")
rownames(results) <- c("treatment", "control")
results
```

```
##           dead alive
## treatment   45    24
## control     30     4
```

```r
# Proportion of deaths
## Treatment group
dead_prop_treat <- results["treatment", "dead"] / sum(results["treatment", ])
dead_prop_treat
```

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

```r
## Control group
dead_prop_control <- results["control", "dead"] / sum(results["control", ])
dead_prop_control
```

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

```r
prop_diff <- dead_prop_control - dead_prop_treat
```

(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?

ANSWER

We're testing two hypotheses.

1. Null hypothesis: The difference in proportion of deaths for the treatment and control groups observed in the experiment is due to chance.
2. Alternative hypothesis. The difference resulted from heart transplants.

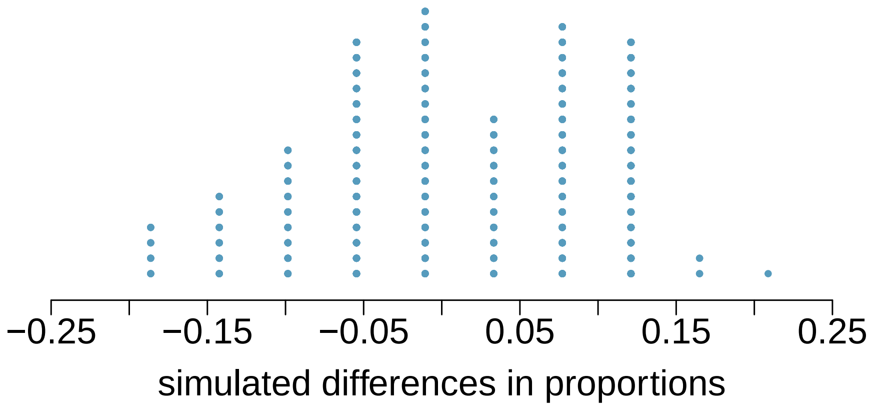
ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

ANSWERS

I liked the idea of putting the results in a little contingency table and then using inline `R` statements to fill in the answers.

> 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 **at least as great as the experimental difference in proportions, namely 0.230179**. 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?



ANSWER

A difference of at least the experimental difference of 0.230179 would happen due to chance with a frequency of about 1%. That low probability justifies rejection of the null hypothesis and concluding that heart transplants reduce the death rate for patients who are gravely ill.