Chapter 2 - Summarizing Data

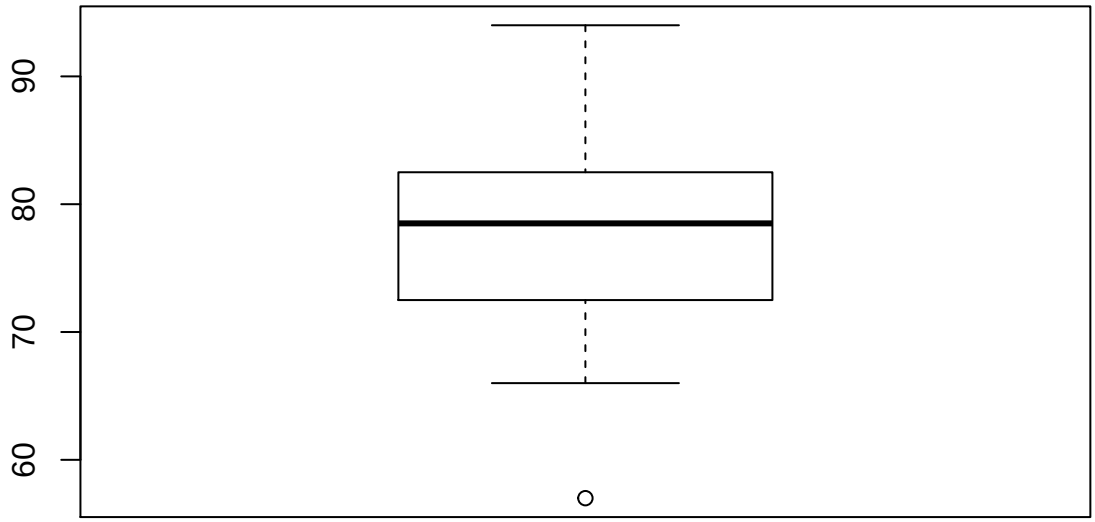
Joshua Registe

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

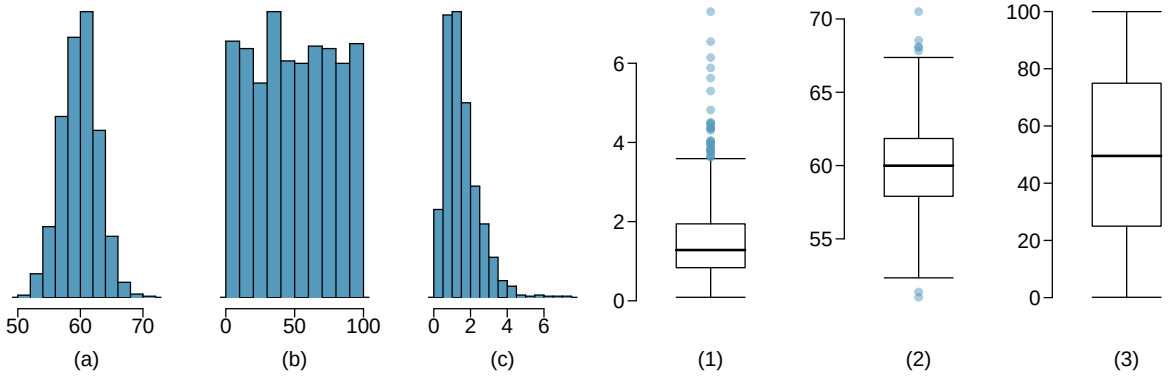
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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.



The distribution for histogram (a) is a normal (gaussian) distribution with the median ~ 60. This distribution matches Box plot (2), The box plots shosws the IQR approximately between 58 and 62 with a few outliers outside of the 1.5xIQR whiskers.

The distribution for histogram (b) is a uniform distribution where there seems to be nearly equal probabilities of the values falling anywhere between 0 and 100. This histogram matches box plot (3) where the distribution of values lie between 0 and 100 on the y axis.

Finally, the third histogram (c) is a rightly skewed distribution where the many of the data points lie on the left of the x axis. This distribution matches boxplot (1) where Majority of points lie between 0 and 2 and there are several outliers outside the upper whisker.

(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.
(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.
(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.
(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.

A.

```
IQR<-1000000-350000
IQR
```

```
## [1] 650000
```

```
1.5*IQR
```

```
## [1] 975000
```

```
IQR+1.5*IQR
```

```
## [1] 1625000
```

This distribution is rightly skewed. This is because majority of homes are less than 450,000 and there are a significant number of homes greater than 6 million. the IQR is approximately 650,000 (1,000,000-350,000). 1.5x the IQR is 975,000 and because many homes go above 6 million, those outliers skew the data right. This data is best represented with the median since the skewed >6million dollar homes can skew the mean. The variability observed is best explained using the IQR since it is not affected by outliers as much as mean and standard deviation.

```
n<-sample(0:300000,2500,replace = TRUE)
n<-append(n, sample(300000:600000,2500, replace = TRUE))
n<-append(n, sample(600000:900000,2500, replace = TRUE))
n<-append(n, sample(1200000:2000000,3, replace = TRUE))
mean(n)
```

```
## [1] 450214.9
```

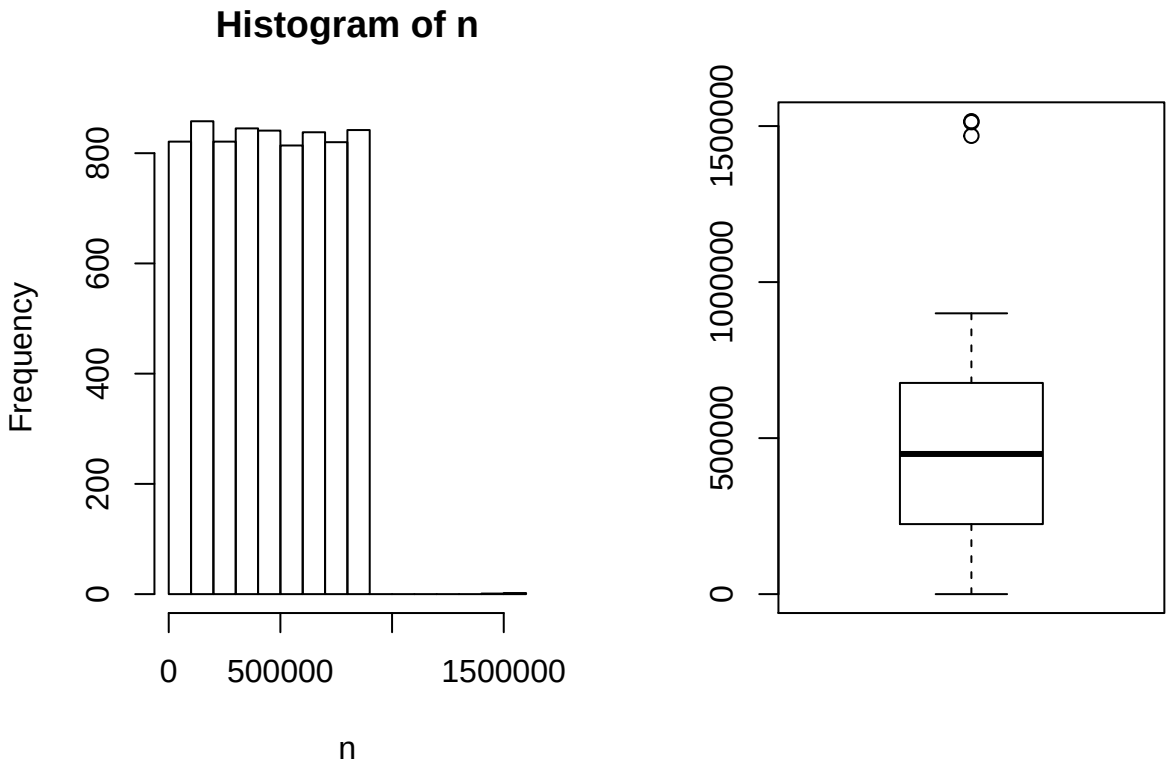
```
median(n)
```

```
## [1] 449211
```

```r
sd(n)
```

```
## [1] 260962
```

```r
par(mfrow=c(1,2))
hist(n)
boxplot(n)
```

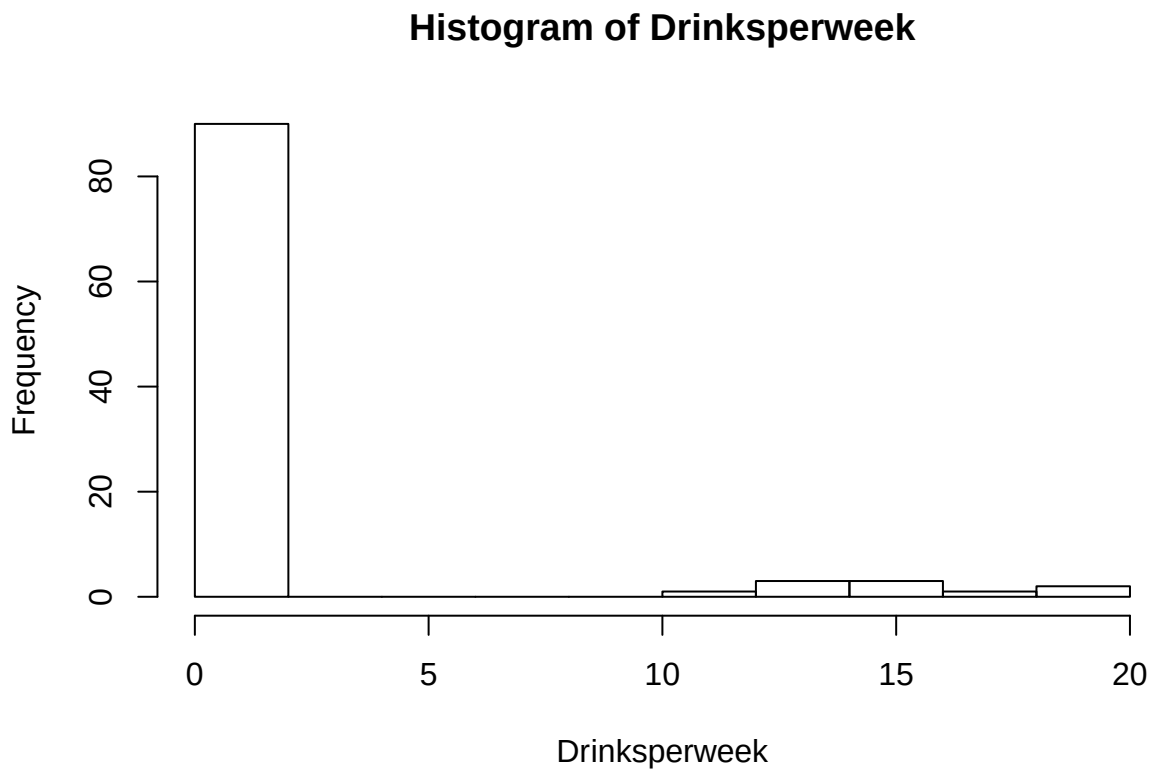


B. This distribution is a uniform distribution as shown by the simulated sampling distribution above. The probabilities that these houses are in any bin around the distribution are close. there is however still some skewness with the few outliers >1.2 million dollars. The mean or the median works well in the case of this distribution because of the uniformity. I would still give preference to the median for data like this. Interquartile range provides a robust statistic of variance.

C. For this distribution it is rightly skewed with most students not drinking, and few students who drink excessively as exemplified below with a simulation assuming 10% of students are heavy drinkers and all others do not drink. Skewness of this type merits the use of medians, it would be innappropriate to say on average students drink X alcoholic beverages per week where this is primarily driven by the few outliers, while the median is 0. the standard deviation gives a good idea of how large the variance is within a dataset of this type.

```r
#assuming 10-20 drinks
studentswhodrink<-sample(10:20,10, replace = TRUE)
studentswhodontdrink<-sample(0:0,90, replace = TRUE)
```

```
Drinksperweek<-append(studentswhodrink, studentswhodontdrink)
par(mfrow=c(1,1))
hist(Drinksperweek)
```




```
mean(Drinksperweek)
```

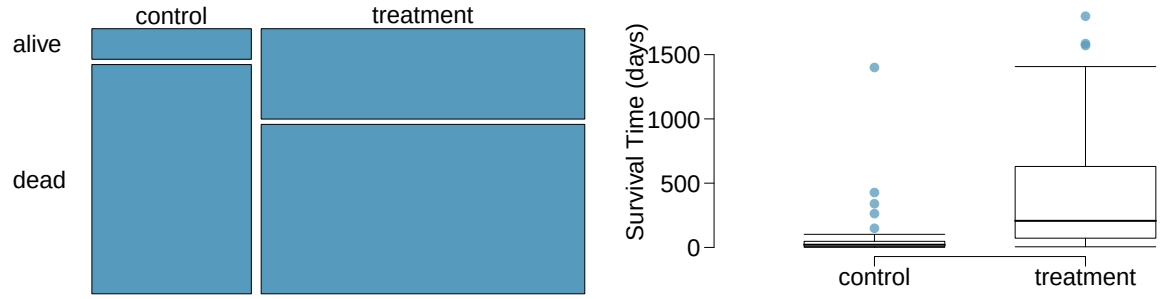
```
## [1] 1.54
```

```
median(Drinksperweek)
```

```
## [1] 0
```

D. This problem is very similar to the student drinking question preceding it and the answer is the same. The data will be rightly skewed due to the executive salaries and the median would be a better representation of the population because of the tendency of outliers to skew the mean. Standard deviation and IQRs can both provide information on the variance.

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.

Based on the mosaic plot, survival does not seem independent of whether or not the patient got a transplant as there are more patients who survived the treatment group. However, simulations should be done in order to test whether or not this was an artifact of randomness or if the null hypothesis (survival is independent of treatment) holds true.

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

The box plot suggest that individuals in the treatment group had much greater survival times than those from the control group. This hold true for majority of the distribution, however there are still some outliers showing a few control group individuals who survived longer than the treatment group. The overall median for the treatment group shows better survival times.

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

```
prop.table(table(heartTr$transplant, heartTr$survived))
```

```
##
##                   alive       dead
##   control   0.03883495 0.29126214
##   treatment 0.23300971 0.43689320
```

based on the contingency table, approximately 29% of all patients died in the control group. This is also 88% of the patients in the control group only. Approximately 43% of all patients died from within the treatment group, this is about 65% of patients within the treatment group.

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

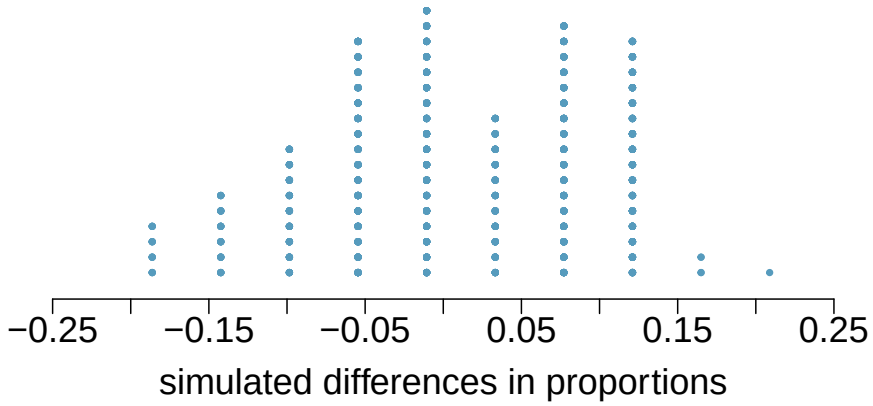
i. What are the claims being tested?

The claims being tested are the following hypotheses: Null hypothesis: The heart transplant is independent of whether an individual will survive and live longer. Alternate hypothesis: The heart transplant has an effect on survivability and longetivity positively.

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 30 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. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are treatment-control. 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 figure shows that the difference in proportion of 0.23 is rarely observed when simulating the proportions. because of this, we reject the null hypothesis in favor of the alternate hypothesis where the transplant indeed has an effect on survivability and longevity and the result of the study was likey not due to chance.