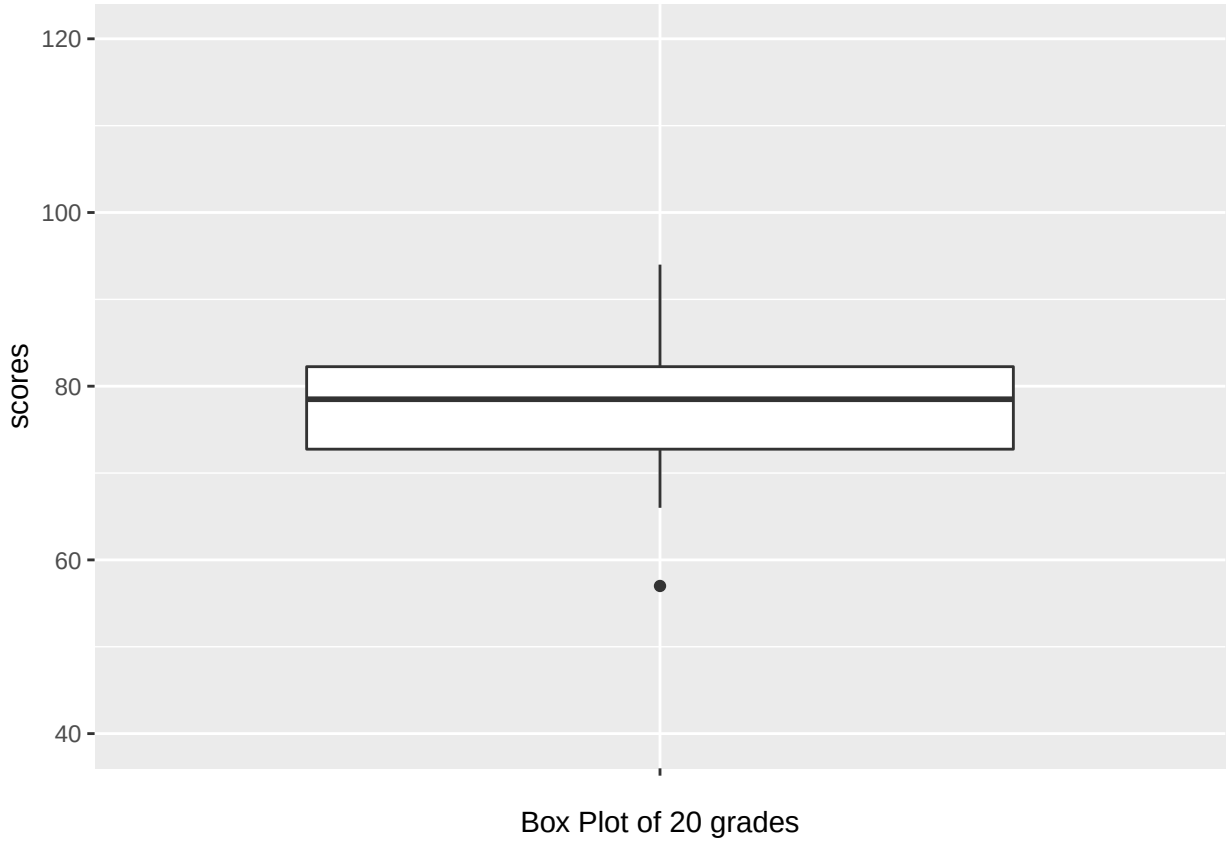
Chapter 2 - Summarizing Data

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

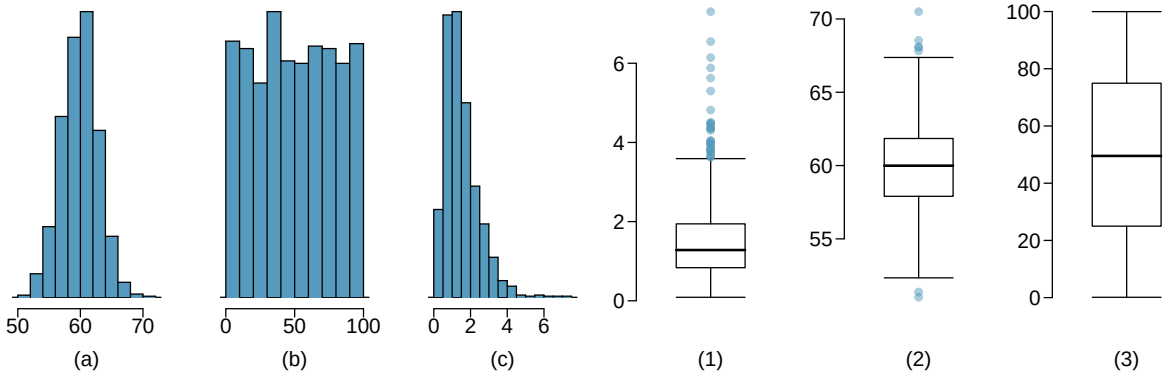
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

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



Box Plot of 20 grades

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



Histogram(a) and Box Plot(2) represent a Normal Distribtuion *produced by the rnorm() function. The mean is 60 and the standard deviation is 3. So a normal distribution predicts 68.2% of all observations will fall between 57 and 63.*

Histogram(b) and Box Plot(3) represent a Uniform Distribution *of 1000 numbers from 0 to 100. Unlike the sample() function which returns discrete random numbers, runif() returns continous random numbers.*

Histogram(c) and Box Plot(1) reflect a Gamma Distribution *which is defined as a Continuous Probability Distribution that models variables that are always positive and have skewed distributions. An example might be the wait time while standing on a line.*

The shape and scale parameter, often referred to as alpha and theta, effect the curvature of the curve, as opposed to shifting the curve the way that standard deviation or mean would do.

Note: *All "r" functions produce a random set of numbers that theoretically fit the parameters of the given distribution. However the seed function fixes the intitial "random" element so the above code will always return the same results when run again and again.*

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) Right skewed. The median is around 450,000 but the mean will be greater. I suppose the median is more reflective of the average home-owner.

The mean isnt useful by itself because it cant distinguish between 10 homes worth 6MM or 1 home worth 60MM. The IQR is useful if you specifically know what Q1 and Q3 are. Youve already told me that Q1=350K and Q3=1MM and that tells me a lot. in terms of how I visualize the economy of that country. Standard deviation is only useful if you also know the mean. You also want to know if the curve is skewed since it is weighted equally by values below and above the mean.

(b) Normalized. I already know the mean and the IQR. It sounds like very few outliers but the mean would be useful to know.

If the mean greately exceeded the median, it would tell me that there were some very expensive homes.

(c) Right skewed. In fact it sounds like Q1, Q2 and beyond are 0. Given that, I would like to know the IQR to understand Q3.

The mean would be useful to get a sense of how much the drinkers are drinking.

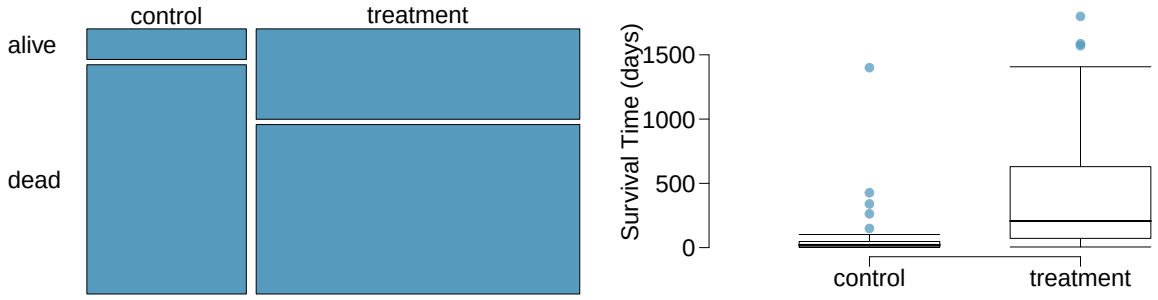
(d) Right skewed. In this case, I dont know much except a few outliers exist on the right tail (upper tail).

I would like to know the median and that is because I feel I would understand the salaries of most employees better The median will not be altered by the outliers whereas the outliers will weigh up the mean. Similarly, the IQR will not be effected by the few executives. If I know the median, and the IQR range, I will understand the salaries of most employees.

Note: *The statistic that is most helpful depends upon the context of what you already know and what you are trying to find out.*

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.

```
## [1] 30
```



(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

I would say tentatively there is a dependency. The survival rate increased from 65% to 88% with treatement. However, the sample size is not that great, and medical treatments are inheritently unpredictable.

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

The box plots reveal that the treatment also improves survival time in those patients who eventually died. Its not clear what it means when survived="alive" and yet survtime has a number.

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

```
treated_died<-nrow(subset(heart_transplant,survived=="dead" & transplant=="treatment"))
treated_total<-nrow(subset(heart_transplant, transplant=="treatment"))

untreated_died<-nrow(subset(heart_transplant,survived=="dead" & transplant=="control"))
untreated_total<-nrow(subset(heart_transplant, transplant=="control"))

sprintf("The percentage of treated patients who died is %.2f", treated_died/treated_total )
```

```
[1] "The percentage of treated patients who died is 0.65"
```

```
sprintf("The percentage of untreated patients who died is %.2f", untreated_died/untreated_total )
```

```
[1] "The percentage of untreated patients who died is 0.88"
```

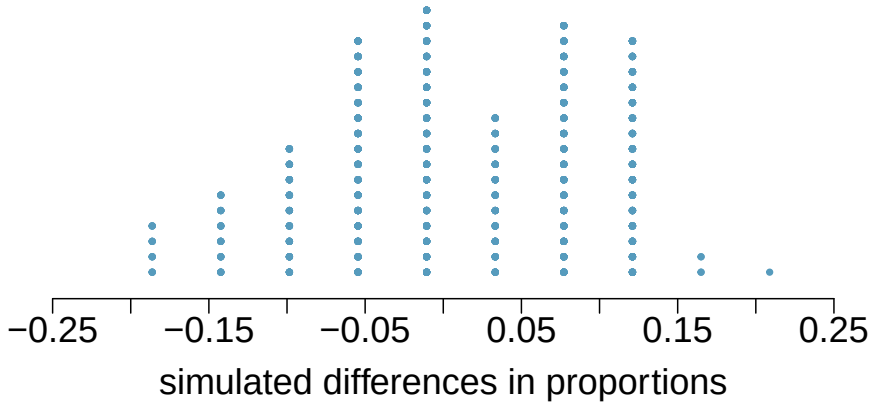
(d) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

i. What are the claims being tested?

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 ___ 0 ***. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are*** 0.23 _____. 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 actuatl trial revealed a difference of .23, while the simulation did not reveal one difference that was as large as .23

Note: *Im not clear how this simulation incorporates a standard deviation that is appropriate to this medical treatment.*