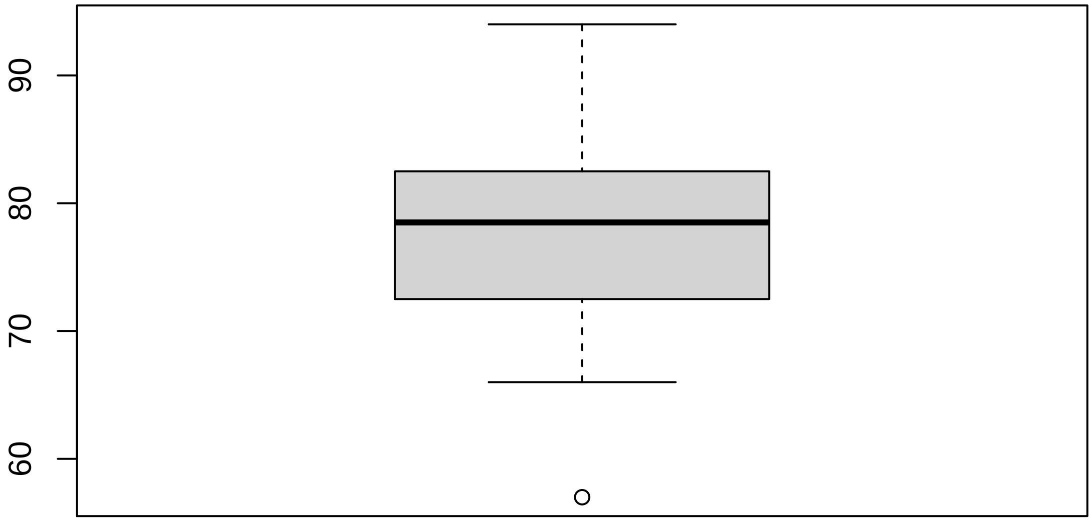
Chapter 2 - Summarizing Data

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

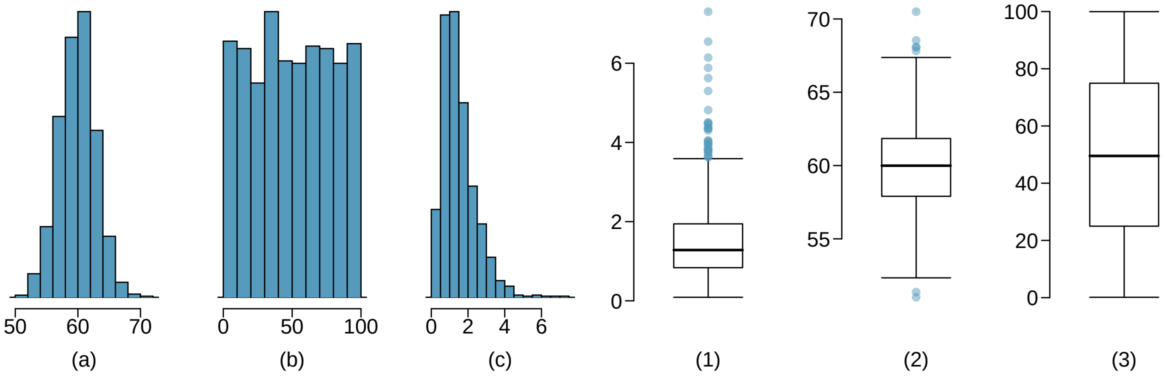
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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



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



a. Symmetric Unimodal Distribution - Boxplot Figure 2
b. Multimodal Distribution - Boxplot Figure 3
c. Right Skewed Unimodal Distribution - Boxplot Figure 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.

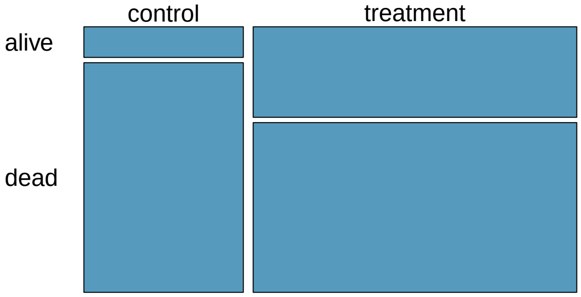
a. Because most of the houses are less than $1,000,000 this would be Right Skewed Distribution. The median and IQR would be the best representation as the super expensive houses would skew the average.

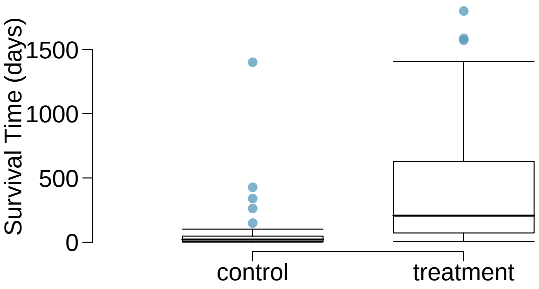
b. Symmetric Distribution because all the house prices are distributed evenly throughout all the pricing brackets. The mean and using the standard deviation would be the best representation as this is a normal distribution therefore it is better to use the mean and standard deviation

c. If the distribution is of the ages groups that drinks then it would be a Left Skewed as all of the students under 21 do not drink (unlikely). The median and IQR would be the best representation as using the average and standard deviation would be inaccurate since we have a lot of students that are under 21 that do not drink

d. This would be a right skew as most of the employees will be earning low income while only a handful of executive will be earning huge salaries. The median and IQR would be recommended as the executives would greatly effect the average salaries

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.

No survival is dependent on the patient getting the transplant because based on the box plot patients who got the transplant were able to survive longer compared to patients who did not get the transplant.

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

The boxplot shows that patients who got the transplant were able to survive longer compared to patients who did not get the transplant. Even though there were some patients who did not get the transplant survived as long as patients who did get the transplant, in general patients who got the transplant survived longer than patients who did not. We can also see that the median for patients that got the transplant overtook the control groups Maximum which shows how effective the transplant is.

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

treatment group - 69 people, 45 died = 65%

```
45/69
```

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

control group - 34 people, 30 died = 88%

```
30/34
```

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

(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?

Null Hypothesis - There is no survival rate difference if a patient got the heart transplant or not

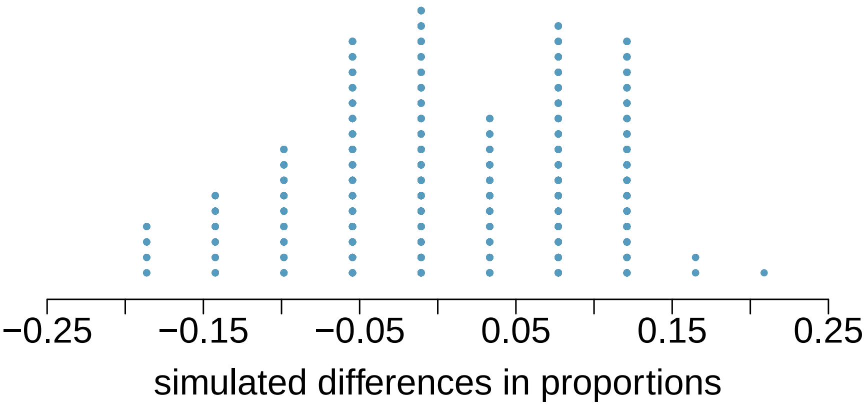
Alternative Hypothesis - Patients who got the heart transplant lived longer

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 (24/69)-(4/34)=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 simulation shows that the results are moving away from 0 therefore we should accept the alternative hypothesis instead of the null hypothesis. This suggest that the heart transplant program is therefore effective.


simulated differences in proportions