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

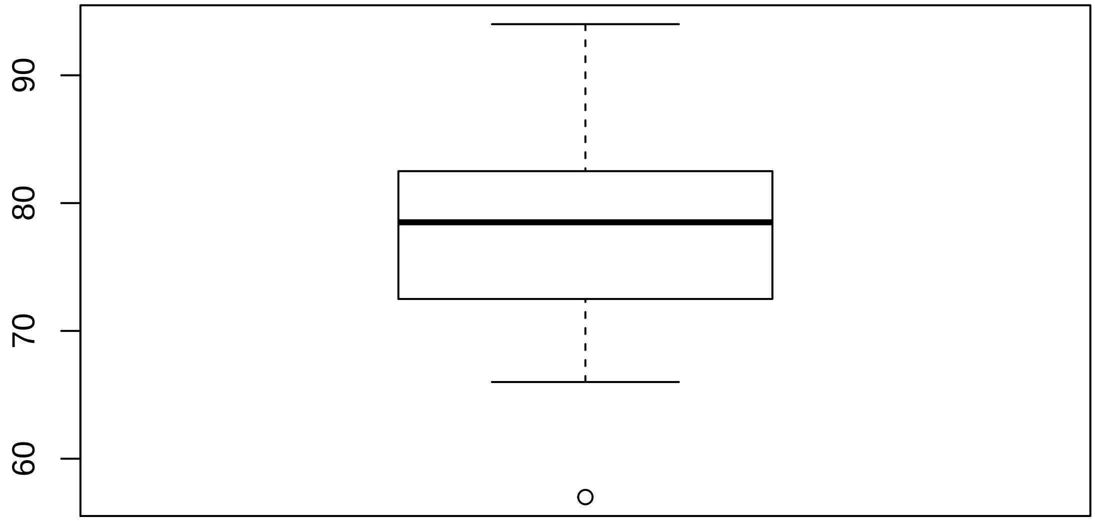
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

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

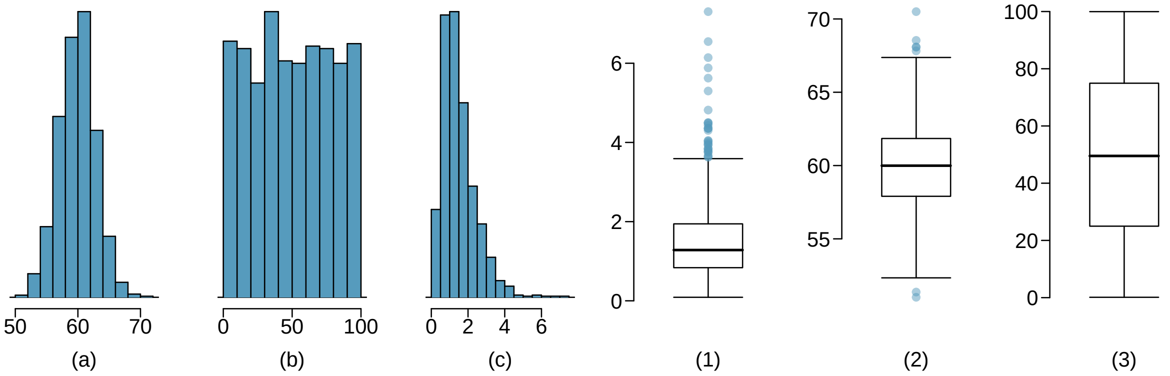
Answer Stats Scores:

```r
#creating box plot using boxplot function

boxplot(scores)
```



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



Answers Mix and Match:

(a): This histogram is unimodal and symmetric with potential unusual observation at 70. 60 is the mean and it matches number (2) box plot.

(b): This histogram is almost uniform. 50 is the median and it matches the number (3) box plot.

(c): This histogram is unimodal and right skewed, with potential unusual observation at 6. The median is 1 and it matches number (1) box plot.

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.

Answers - Distributions and appropriate statistics.

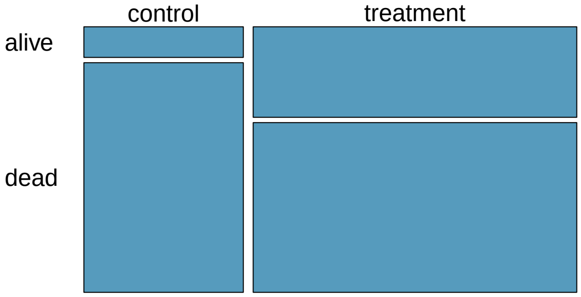
(a)
- Even though the distribution of the houses that cost below 350K, below 450K and below 1 million (we can see that 25% of the houses are below 350K, 25% of the houses are between 350-450 and 25% of the houses between 450K-million) seems to be uniformed there are meaningful number of houses that cost more than \$6,000,000. Hence the distribution will be right skewed. Median and IQR would be the best to use and describe the center and spread.

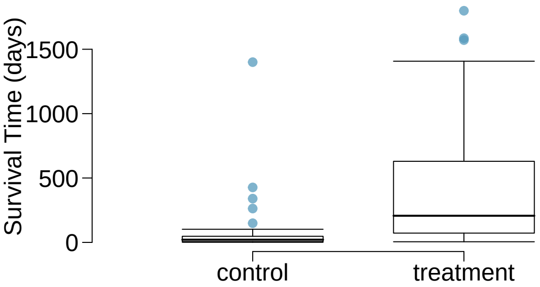
(b) The distribution of the houses are symmetric. When we look at the house pricing we see that 25% of them are 300K , 25 % of them are between 300-600K, 25% of them is between 600-900K. There are few houses cost more \$1,200,000 however would not cause to make the distribution skewed. The Mean would be around 600K. Mean and Standard deviation would be the best to use and describe the center and the spread.

(c) The distribution of the number of students that consumes alcoholic drinks are right skewed. Considering they are under 21 years old and as explained only few students will drink (so there are still students that will drink) the number of students that drink will go lower and skewed to the right side of the long tail. Median and IQR would be the best to use and describe center and spread.

(d) The distribution of the salaries are symmetric as only few of the high level executives earn much higher (outliers). Mean and standard deviation would be best to describe the center and the spread.

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.

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

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

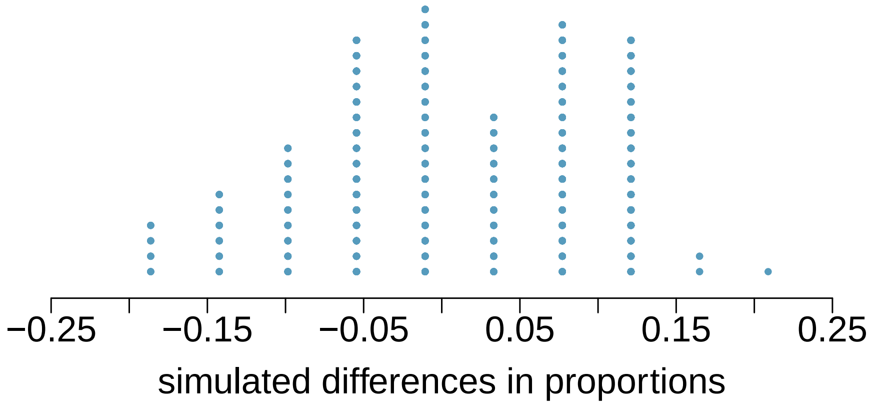
(d) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

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 ______________ cards representing patients who were alive at the end of the study, and *dead* on ______________ cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size ______________ representing treatment, and another group of size ______________ 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 ______________. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are ______________. 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?



Answers - Heart Transplants

(a) The survival is dependent on if the patient was in treatment or control group. Around 65% of the patients died in treatment group and around 88% of the patients died in control group. The patients that were in the treatment group had more chance of survival.

(b)When we look at the box plot , we see that there is a bigger survival time if the patient was in treatment group. Median survival time is much higher in the treatment group compare to the control group. Q1 and Q3 survival time is much higher for treatment group compare to the control group. The box plot suggest that the treatment is effective for survival time.

(c) 30/34 = 88% of the patients in control group died. 45/69=65% of the patients died in treatment group.

(d)

i) Claims being tested are; treatment does have relationship with survival and survival time (days) of the patients. Treatment does not have relationship with survival and survival time of the patients. (It does not matter if the patient receives a treatment or not , their survival time and chance would not change)

ii) We write *alive* on **24+4=28** cards representing patients who were alive at the end of the study, and *dead* on **30+45=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 **30/34 - 45/69 = 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) The result shows that the survival is dependent on group (treatment/control). Simulated differences in proportions are 0.23 and low. It is not by chance patients survive. With treatment, survival of patients will improve.