Chapter 2 - Summarizing Data

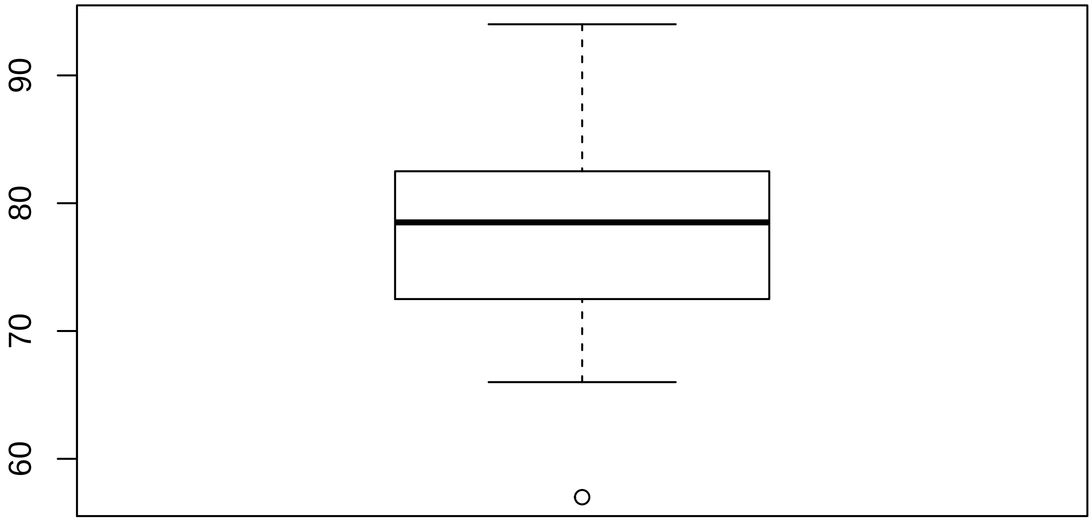
Adam Gersowitz

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

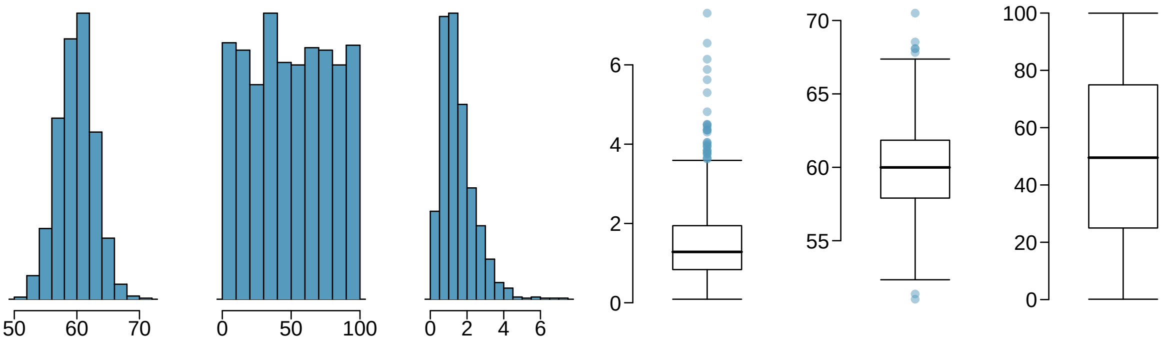
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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 is approximately a normal distribution and it matches with boxplot 2.

Histogram b is a uniform distribution and it matches with boxplot 3.

Histogram c is a skewed right distribution and it matches with boxplot 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.

The data is Right-skewed because 75% of houses are below 1 million but there are a significant number of houses above 6 million. Median is a better representation of a typical observation because of the large number of outliers above 6 million. I would use IQR for variability for the same reasons.

(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.

The data is uniformly distributed. The mean is the best representation of a typical obersvation becasue there aren't any extreme outliers. I would use standard deviaiton because of the lack of outliers.

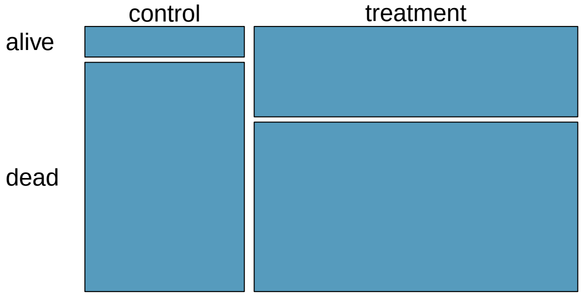
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

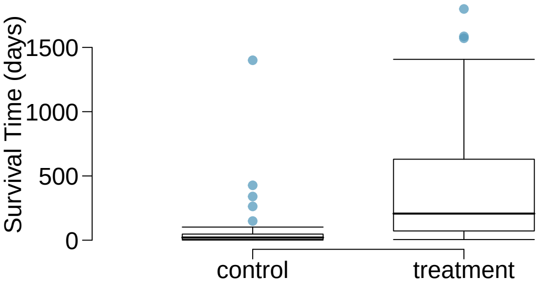
This data will be right-skewed because most students don't consume drinks. Due to the large number of students who don't drink the center is best described by the median and the variability is best described by the IQR.

(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.

This data will likely be right skewed. There will likely be a large number of entry-level positions that pay a lesser amount and the higher you go up the salary scale there will be less of those positions. Due to the large number of lower-paid people the center is best described by the median and the variability is best described by the 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.

No, survival is not independent of whether someone got a transplant. It seems like nearly three times the ratio of patients were alive in the treatement group.

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

They suggest the treatment is very effective.

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

Treatment group: $45/69 = 0.652$ Control group: $30/34 = 0.882$

(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 observed difference in survival rate represents the efficacy of the treatment.

The observed difference in survival rate is due to random chance.

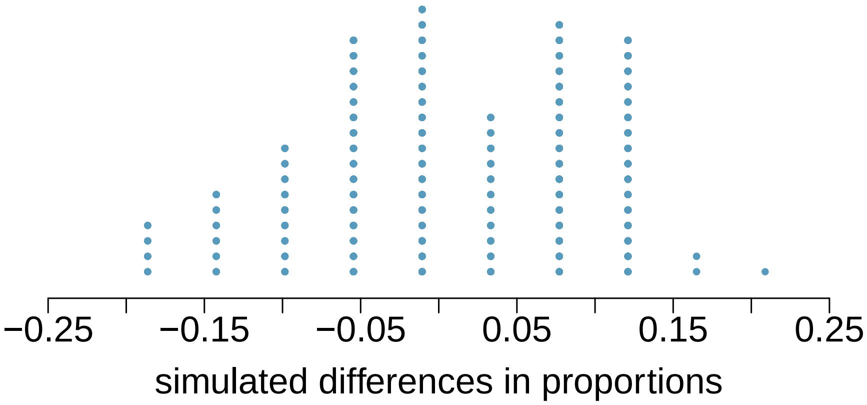
ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

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 *from chace alone*___________. 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 results show the program is effective. The difference between the groups is 0.882-.0652 = 0.23. In the simulations only one simulation showed a differenc ein proporitons that was as high as 0.23. We will reject the independence model for the alternative.


simulated differences in proportions