Summarizing Data

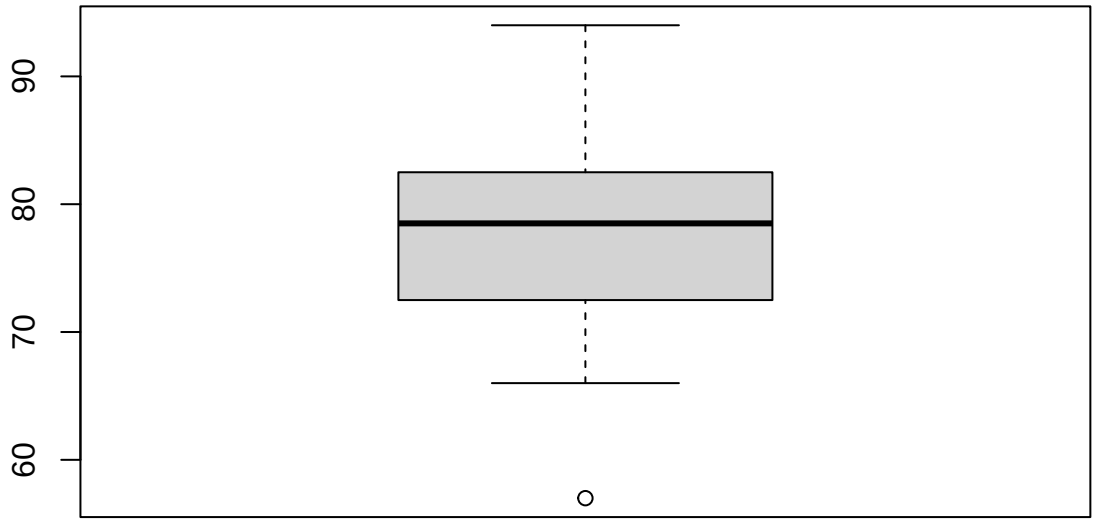
Nnaemeka Newman Okereafor

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

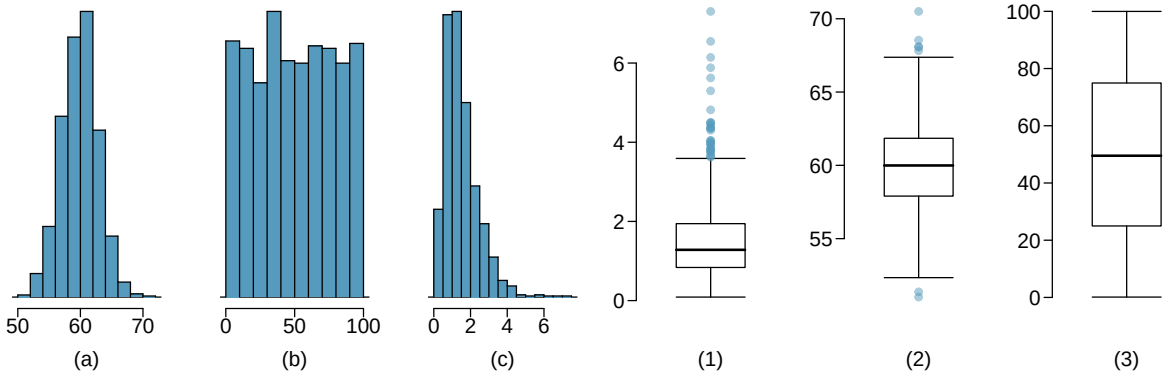
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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 boxplots.



Histogram (a)

Histogram (a) matches boxplot (2). The distribution of Histogram (a) is roughly bell shaped with a center of 60, a range of about 22 (50 to 72), and outliers are apparent on both the higher end and lower end.

Histogram (b)

Histogram (b) matches the boxplot (3). The distribution of the values are uniform with a center of approximately 50, a range of about 100 (0 to 100), and no apparent outliers.

Histogram (c)

Histogram © is synonymous to boxplot (1). The histogram is positively skewed. In other words, it is skewed to the right, centered at about 1 with most of the data concentrated between 0.4 and 0.6, a range of approximately 8 (0 to 8), and outliers estimated to exist at point 3.8 and above.

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.

Answer

Skewness: Left skewed

Measure of center: Median

Measure of variability: IQR

Reason: Houses with \$6,000,000 prices are significant outliers that could alter the center of the data and its variability when mean and standard deviation are used respectively.

(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.

Answer

Skewness: symmetrical

Measure of center: Mean

Measure of variability: standard deviation

Reason: Few houses that cost $1,200,000 prices are not extreme outliers that could alter the center of the data and its variability significantly.

(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.

Answer

Skewness: Right skewed

Measure of center: Median

Measure of variability: IQR

Reason: Few students drinking excessively will impact greatly on the center of the data and its variability significantly.

(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.

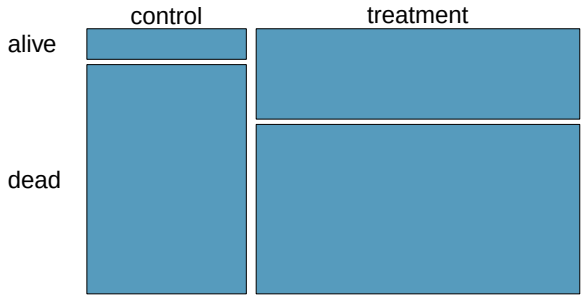
Skewness: Right skewed

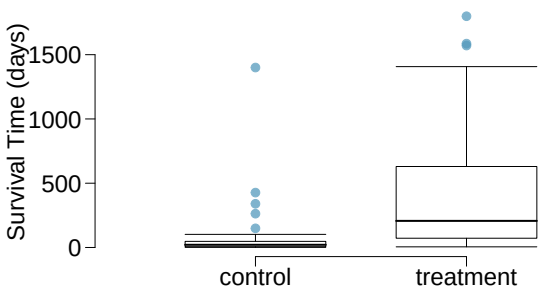
Measure of center: Median

Measure of variability: IQR

Reason: Few high level executive salaries will impact greatly on the center of the data and its variability significantly.

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.

I cannot say because the given information is not enough to determine the relationship between survival and whether a patient got a transplant or not. Having patients in both the control and treatment group with survival time of 500 days and 1500 days indicates that further analysis and hypothesis testing is needed to ascertain if survival is independent of whether the patient has a transplant or not

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

The boxplot shows that most patients in the treatment group had a survival time between 0 and 1500 and a few outliers.

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

30/34 (i.e 88.2 percent) of patient in control group died. While 45/69 (i.e 65.2 percent) of patients in treatment 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?

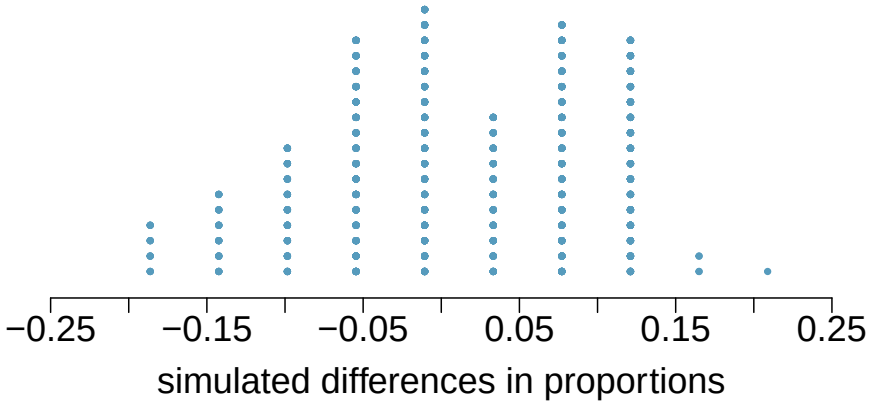
Transplants increase survival rates and time

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 *75* on *dead* 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 the *0*. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are *equal to or greater than about -25%*. 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?

\begin{center}



The simulation results show that the null hypothesis should be rejected because treatment and heart transplants have a significant effect on patient survival