Chapter 2 - Summarizing Data

John Mazon

install.packages("devtools") install.packages("backports") install.packages("glue")

data(package = 'openintro')

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

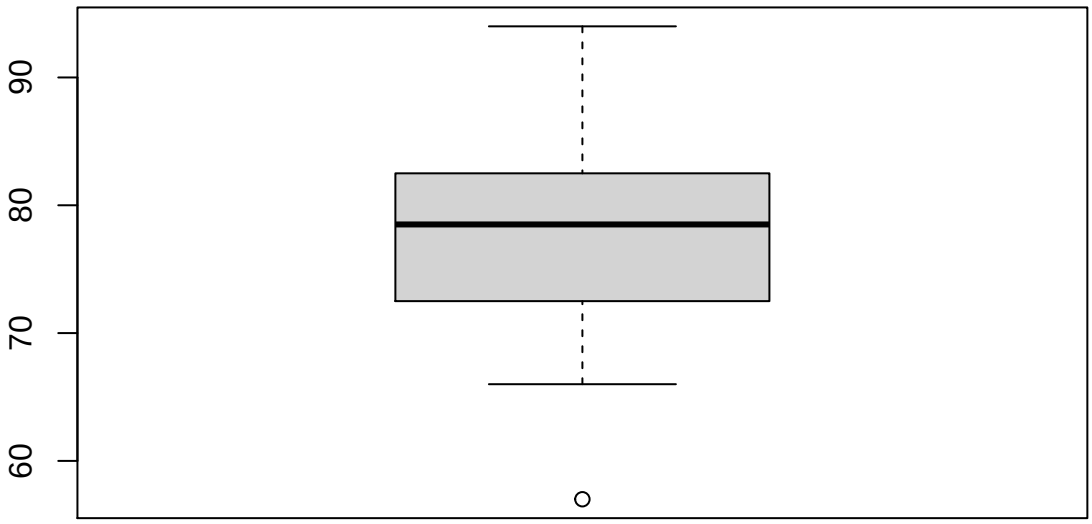
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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

```
scores <- c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94)
summary(scores)
```

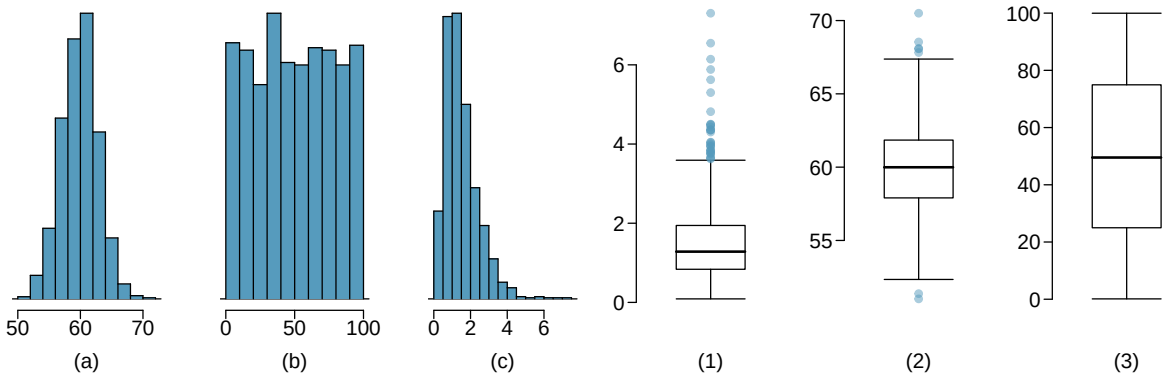
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   57.00   72.75   78.50   77.70   82.25   94.00
```

```
scores <- (c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79,
    81, 81, 82, 83, 83, 88, 89, 94))
boxplot(scores)
```



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 Number 2 box plot to match with Symmetrical/Unimodal Distribution B Number 3 box plot to match with Multimodal Distribution due to many peaks C Number 1 box plot to match with Right skew Distribution



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. Because of the meaningful number of houses above $6M it would appear to be **right skewed**. Majority of houses by price would be far left of the expensive house prices. Similar to the example in the lab with the delayed flights, due to the "Way above" **median** price of certain homes, it'd be best used median perhaps.**iqr** would best be used in my opinion since median might be heavily influenced by the most expensive homes.iqr gives a clearer representation of the "majority" range. variability of observation would be best represented this way.

#interquartile range (IQR), also called the midspread, middle 50%, or H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles

(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. In this example we witness that the home prices are not ridicously far apart in price range. One would assume the distribution to be **symmetric** as there are "very few" houses that cost more than $1.2 M. This wouldn't cause any extreme skew as in the above question. Typical observation in these cases would best be depicted by **mean**. Lastly I would use **standard deviation** since it is well known that it is affected by skewed data, however it can be more reliable when the data is normalized and not skewed.

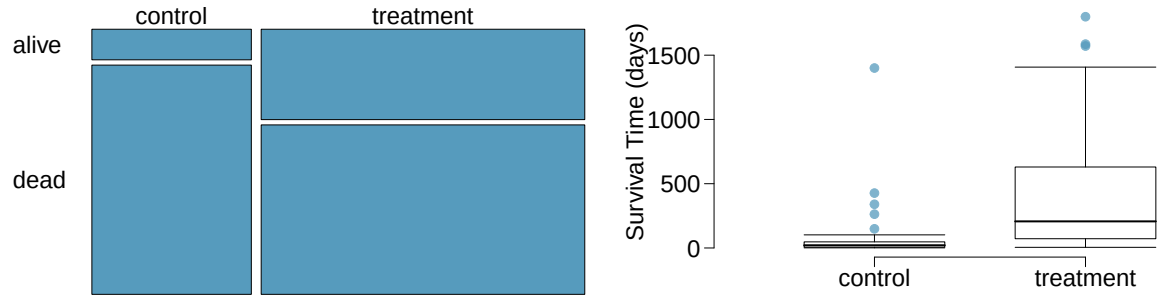
#The "mean" is the "average" you're used to, where you add up all the numbers and then divide by the number of numbers.

(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. Assuming alcoholic drinks have a starting point of Zero[0] meaning they haven't consumed at all, this would be **right skewed**. We can come to this conclusion since the question gives us data stating not many students have drank in the past and few are excessive drinkers. Using this reasoning, again we can state since the majority of the college students will show results of very few drinks consumed in a week, **median** will represent a better overall picture. Lastly **iqr** seems to be the better option for variability of observations due to the right skew.

#The "median" is the "middle" value in the list of numbers.

(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. Due to the majority of employees salaries being on the "lower end' we can conclude this will be **right skewed**. We came to this conclusion since the question is specifying only a few high level execs make the big bucks. To take in account the majority of"low level" salary employees compared to the big buck makers, we would assume **median** is best to present us a more accurate picture. Any data with extreme data endpoints will follow this path. Mean would be "too" highly influence by those making extremely large salaries. Variability of observation would be best represented by **iqr** being that few high level execs would throw off the average with their high salaries.The interquartile range is preferred when the data are skewed or have outliers.

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 indepedent of whether or not the patient got a transplant. We clearly see that those who got the transplant survived in higher numbers than those who didn't

(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment. Box plots suggest great efficacy of the heart transplant treatment based on the survival rates. We witness outliers in the control group with long survival rates however the patients lucky enough to receive the transplant showed longer survival times. Patients in the control group survived all around the same amount of time.

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

```
#control group
prop_control_group <- round((30/34)*100,2)
prop_control_group
```

```
## [1] 88.24
```

```
t=45/69
prop_treatment_group <- round((45/69) * 100,2)
prop_treatment_group
```

```
## [1] 65.22
```

(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? In a nutshell, the claims being tested are whether the heart transplants increase the survival time[measured in days] in the patients. Perhaps some may want to find out if the heart transplant has zero effect on life duration.

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** _-0.230179_____. 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?
Upon viewing the simulation results shows, closer toward the 0.15 and 0.23 shows small probability. Simulated differences closer to 0.23 is unlikely to happen.