Chapter 2 - Summarizing Data

Subhalaxmi Rout

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

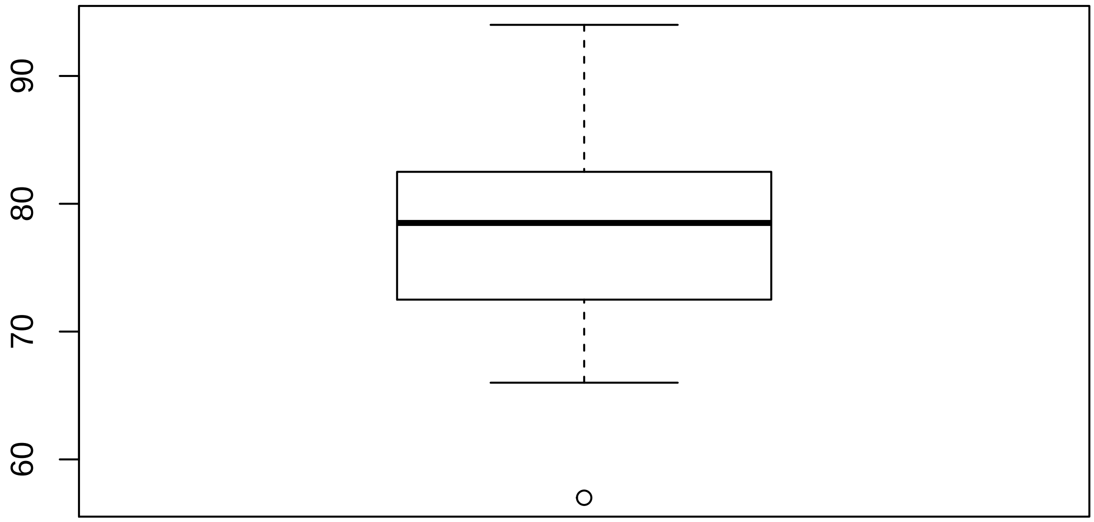
Create a box plot of the distribution of these scores. The five number summary provided below may be useful.

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

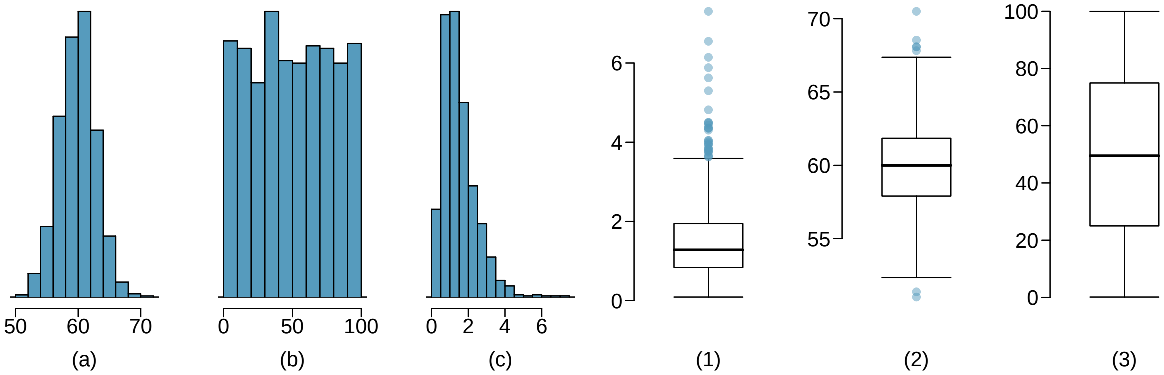
```
df<-  data.frame( "scores" = c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 
summary(df)
```

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

```
boxplot(df$scores)
```


90
80
70
60

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



(2.10)

- a: is a symmetric unimodal distribution histogram, this matches with boxplot in figure (2)
- b: is a multi-modal distribution(more than one peak point), this matches with boxplot in figure (3)
- c: is a right skewed unimodal distribution, this matches with boxplot in 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.

(2.16)

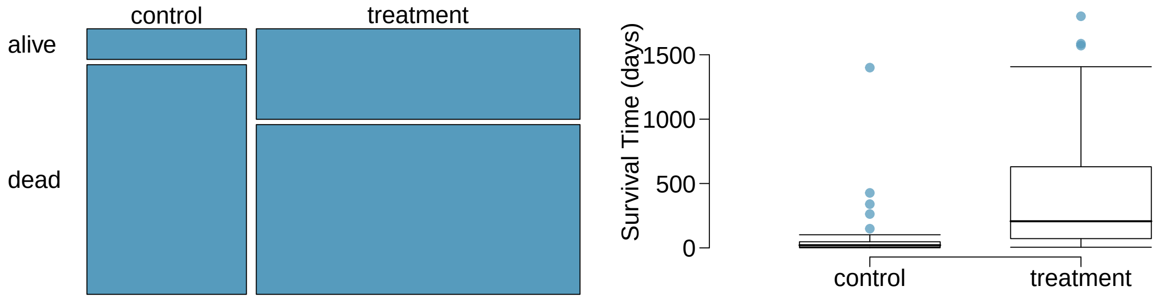
(a) Most of the data is on left, so the distribution is Right skewed distribution. So rather than Mean, Median and IQ will be the better variability.

(b) Data seems evenly distributed, so this is a symmetric distribution. This is best represnted by Mean and standard deviation.

(c) Though most of student do not drink the distribution will be Right skewed. So Median and IQ will be the better variability.

(d) Only few high executives have higher salaries than all other employees, so this will be a symmetric distribution. Mean and SD(standard deviation) will be the best represent the variability.

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.

Answer

(a) Based on the above plot, survival is dependent on the patient got a transplant because the patient who got transplant are more surival days than the patients who did not get transplant.
(b) The box plot shows that the control group patients has less survival days than treatment group. Except number of outliers, the control group most patatients survived around the same amount of time.
(c) Treatment Group Total Patients: 69 Alive : 24 Died: 45

```
percent_treat_gp <- round((45/69) * 100,2)
percent_treat_gp
```

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

Control Group Total Patients: 34 Alive: 4 Died: 30

```
percent_control_gp <- round((30/34)*100,2)
percent_control_gp
```

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

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?

Answer

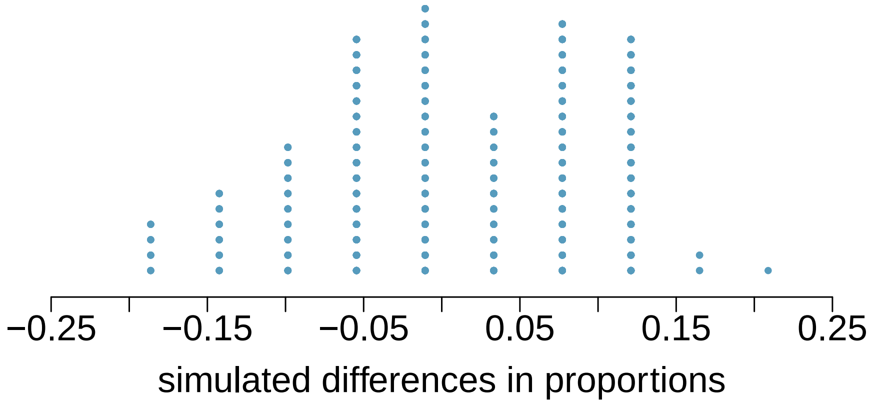
(d)

i. The claims can be tested on 2 ways i.e null hypothesis and actual hypothesis.

- Null hypothesis: Survivility is not dependent on if the patient had a transplant or not.
- Actual hypothesis: Survivality is dependent on the transplant.

ii. 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. The simulation graph shows that there are only less instances where the fraction is 0.23. Which is very rare event so we can reject null hypothesis and conclude, having heart transplant affect on the survivality.


simulated differences in proportions