Chapter 2 - Summarizing Data

Md Forhad Akbar

Stats scores

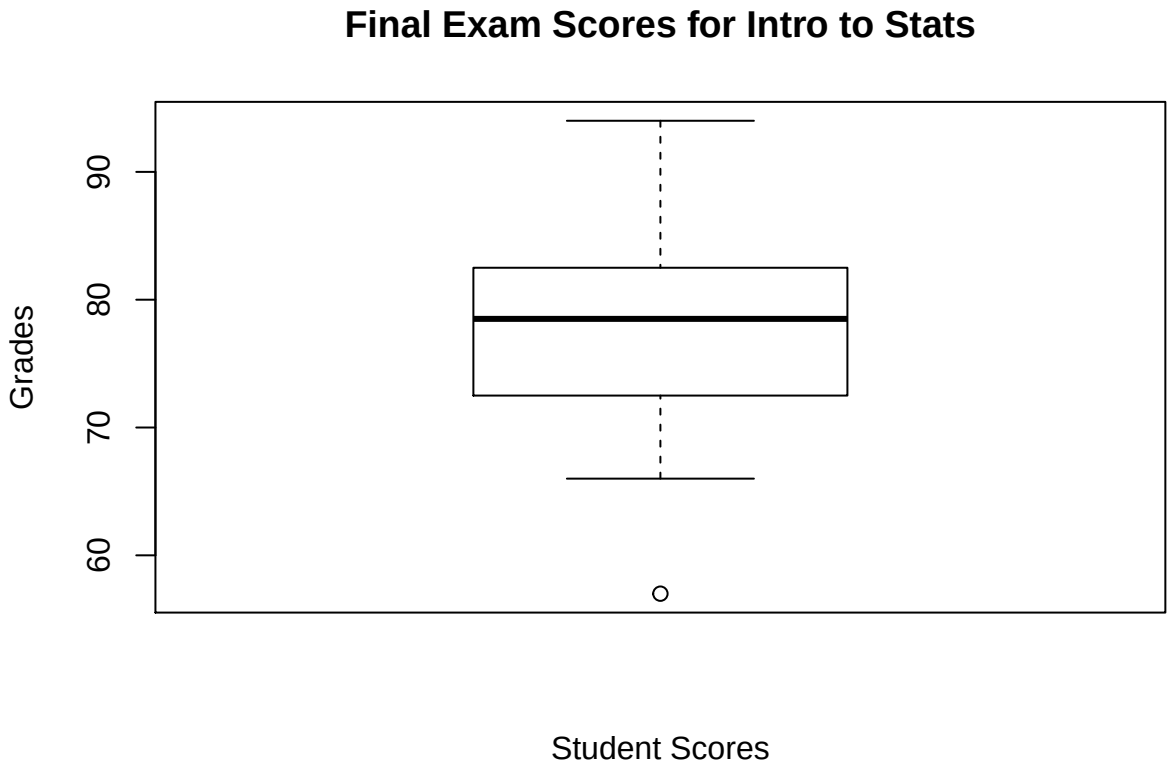
Stats scores. (2.33, p. 78) Below are the final exam scores of twenty introductory statistics students.

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.

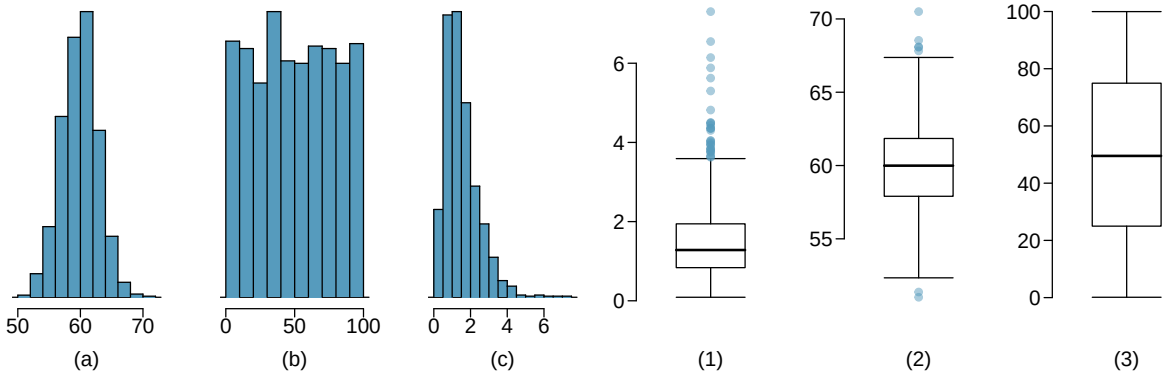
Min	Q1	Q2 (Median)	Q3	Max
57	72.5	78.5	82.5	94

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



Mix-and-match

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



a) Symmetrical and and Unimodal distribution; match will be box plot #2.

b) Symmetrical and Multimodal distribution; match will be the box plot #3.

c) Right Skewed and Unimodal distribution; match will be the box plot #1.

Distributions and appropriate statistics

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: It will be right-skewed since 3rd quartile is less densely distributed than the first 2 quartiles and since there are a meaningful number of houses worth multiple times the value of the other houses. The median would best represent the typical observation since it will mitigate the effect of the extreme values. The variability would be best represented by the IQR because the SD would be sensitive to the extreme values.

(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: It will be a mostly symetrical distribution since the quartile ranges are very similar. The median would best represent the typical observation since it will mitigate the effect of the extreme values. The variability would be best represented by the IQR because the SD would be sensitive to the extreme values.

(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.

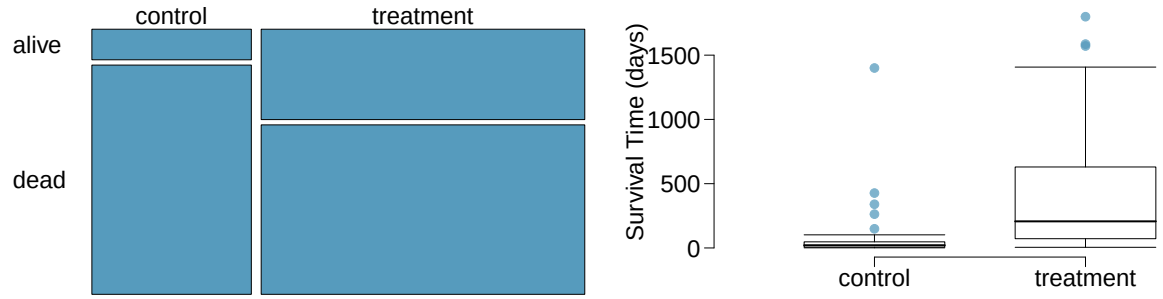
Answer: It will be left-skewed distribution since most of the students will be at the minimum value of zero and since very few drink excessively. The median would best represent the typical observation since it will mitigate the effects of the all the non-drinkers and the excessive drinkers.

(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.

Answer: It will be a mostly symetrical distribution. The median would best represent the typical observation since it will mitigate the effect of the extreme values of the high-level executives. The variability would be best represented by the IQR because the SD would be sensitive to the extreme values of the high-level executives.

Heart transplants

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.

Answer: If we see the mosaic plot, we can conclude that the survival is not independent. A significantly higher proportion of those who received the transplant survived than those who in the placebo/control group.

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

Answer: The transplant is at least marginally to moderately effective. While the transplant didn't save the majority of the patients, it did save a much greater proportion than the placebo/control group.

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

Answer: From the article we can find as follows:

proportion of patients in the control group died

```
#Control Group
controlalive<- 4
controldead<- 30
totalcontrol<-controlalive + controldead
ControlGroupdiedproportion<- controldead / totalcontrol
ControlGroupdiedproportion
```

```
## [1] 0.8823529
```

proportion of patients in the treatment group died

```r
#Treatment Group
Treatmentalive<- 24
Treatmentdead<- 45
treatmenttotal<- Treatmentalive + Treatmentdead
TreatmentGroupdiedproportion<- Treatmentdead/treatmenttotal
TreatmentGroupdiedproportion
```

```
## [1] 0.6521739
```

(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?

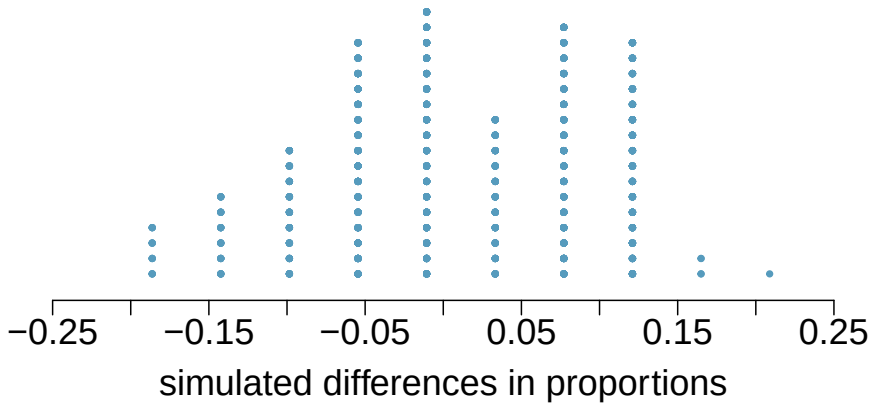
Answer: H0 the null hypothesis: The transplant does not alter the survivability of these patients.

HA the alternative hypothesis: The claim was that transplanted patients were more likely to survive than non-transplanted patient.

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 ***.2302***_____. 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: Simulation results in the assignment show that a large different such as .2302 is unlikely to occur by chance and that null hypothesis should be rejected. That is, the variable do not appear to be independent, and we reject the independence model in favor of the alternative. The study's results provide convincing evidence that heart transplant is effective.