Data606 Homework Chapter 2 - Summarizing Data

Completed by Chad Bailey

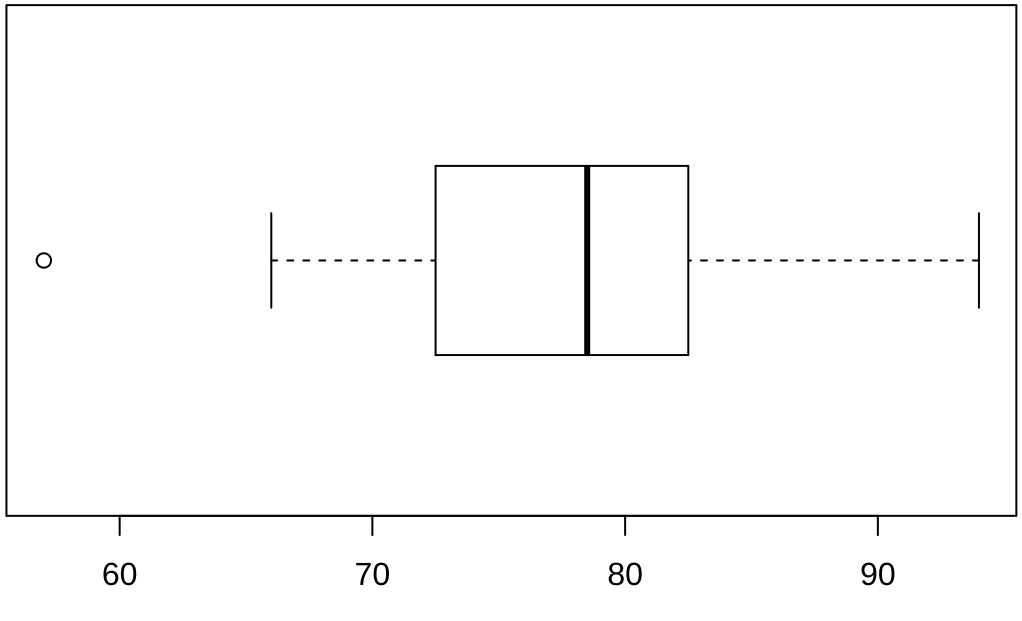
Stats scores. (2.33, p. 78) Below are the final exam scores of twenty introductory statistics students.

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

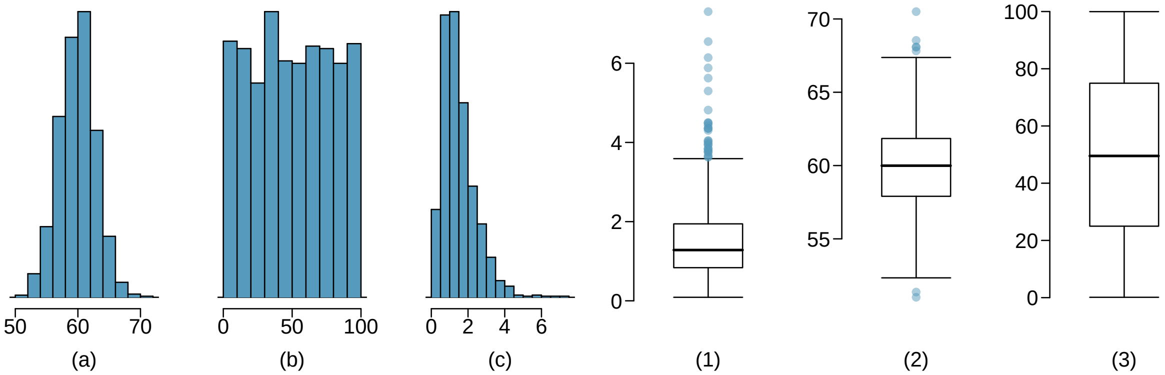
Create a box plot of the distribution of these scores. The summary provided below may be useful.

```
##    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, horizontal = TRUE)
```



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



```
## (1) and (a) are describing the same distribution which is sysmetrical
## (2) and (c) are describing the same distribution which is right skewed
## (3) and (b) are describing the same distribution which is roughly rectangular
```

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.

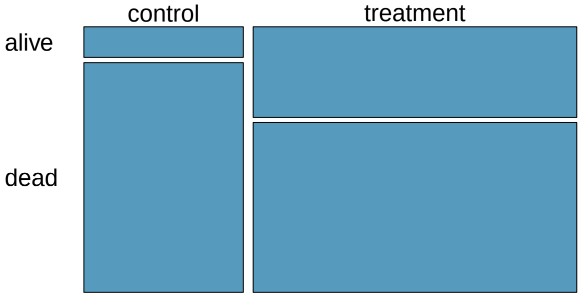
(b) Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.

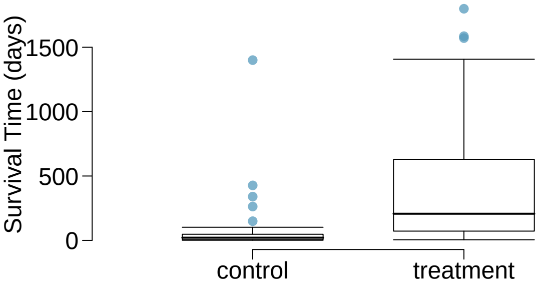
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

(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.

```
##  (a).1 The data are right skewed, seen as the distance between (Q2-Q1) and (Q3-Q2)
##        are not close to equal
##  (a).2 Median would better represent a typical observation as median is less affected
##        by skewed data
##  (a).3 IQR would better represent variability as IQR is less affected by skewed data

##  (b).1 The data are symmetrical; the distance between (Q2-Q1) and (Q3-Q2) are equal
##  (b).2 Mean would better represent a typical observation as it is more generally
##        understood and would allow for other stastistical operations
##  (b).3 Standard deviation would better represent variability as IQR is less affected
##        by skewed data as it is more generally understood and would allow for other
##        stastistical operations

##  (c).1 The data are right skewed given the assumption that most students don't drink
##        (i.e., number of drinks will be 0 for most students) since they are under 21
##        and only a few drink excessively (i.e., the right tail will be long)
##  (c).2 Median would better represent a typical observation as median is less affected
##        by skewed data
##  (c).3 IQR would better represent variability as IQR is less affected by skewed data

##  (d).1 The data are right skewed as most salaries will be close together with executive
##        being far higher (causing a long right tail).
##  (d).2 Median would better represent a typical observation as median is less affected
##        by skewed data
##  (d).3 IQR would better represent variability as IQR is less affected by skewed data
```

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.

```
## It appears survival is NOT independent of the patient receiving the transplant.
## The mosaic plot shows the control had a much higher proportion of patients that
## died by the end of the study.
```

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

```
## The box plots show that there appears to be almost a 2-quartile shift between
## the control and treatment group in the number of survival days. That is the
## treatment group's Q1 appears to be higher than the Q3 of the control group.
## This would be a relatively strong indication of an effect not likely to be
## observed by chance, and that transplant has high efficacy.
```

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

```
## get control proportion that died
control_ntotal <- nrow(subset(heartTr, transplant == 'control'))
control_ndied <- nrow(subset(heartTr, transplant == 'control' & survived == 'dead'))
control_pctdied <- round(control_ndied / control_ntotal *100, 2)

## get treatment proportion that died
treatment_ntotal <- nrow(subset(heartTr, transplant == 'treatment'))
treatment_ndied <- nrow(subset(heartTr, transplant == 'treatment' & survived == 'dead'))
treatment_pctdied <- round(treatment_ndied / treatment_ntotal *100, 2)
```

The proportion of the control group that had died by the end of the study was 88.24%.
The proportion of the treatment group that had died by the end of the study was 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?

```
## The Null Hypothesis is that the treatment had no effect on survival rate, and
## the differences observed would only occur rarely if the experiment was repeated.

## The Alternative Hypothesis is that the treatment had an effect on survival rate.
```

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.

```
table(heartTr$survived)
```

```
##
## alive  dead
##    28    75
```

```
table(heartTr$transplant)
```

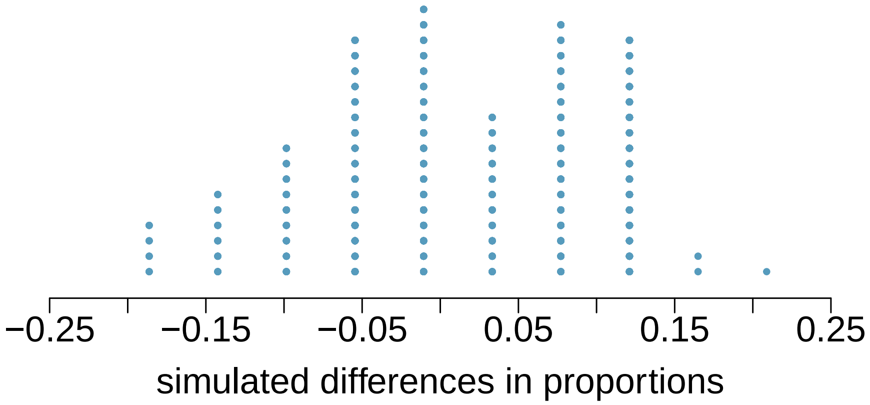
```
##
##   control treatment
##        34        69
```

```
treatment_pctdied - control_pctdied
```

```
## [1] -23.02
```

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 ______at least 23.02________. 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 shows that the occurance of a difference of at least 23.02 would
## be a VERY rare event. Given these results, the Null Hypothesis is rejected and
## the Alternative Hypothesis is accepted.
```