Chapter 2 - Summarizing Data

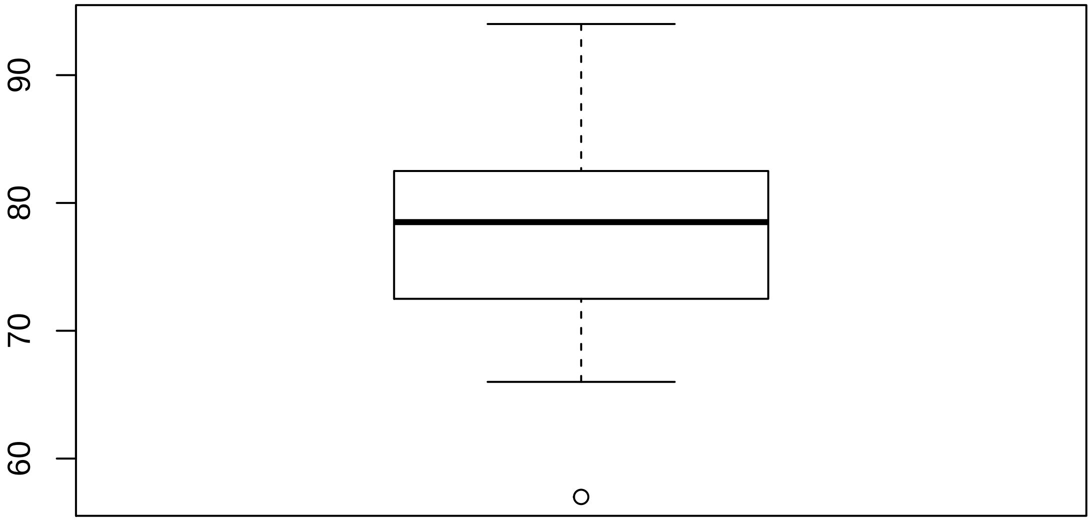
Samuel I Kigamba

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

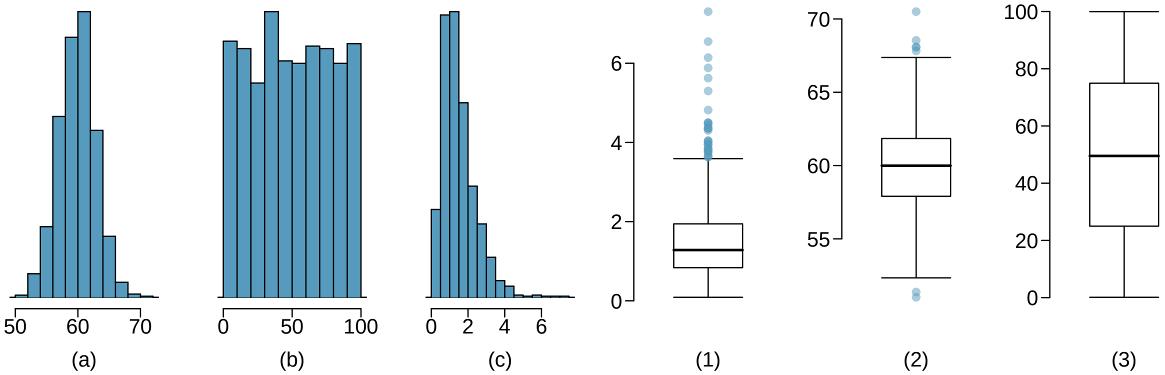
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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 box plots.



Histogram 1. The box plot that matches this histogram is (2). The distribution of this data is unimodal, symmetric, and appears it could be normally distributed.

Histogram 2. The box plot that matches this histogram is (3). The distribution of this data might be symmetric but it does not look normally distributed. It is possible it is uniform but a smaller bin width would have to be selected to see.

Histogram 3. The box plot that matches this histogram is (1). The distribution of this data is unimodal and has a right skew.

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.

This distribution would be right skewed. The median would be the best to represent a typical obser

(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.

This distribution would be symmetric since the few expensive houses wouldnt skew the data much, and

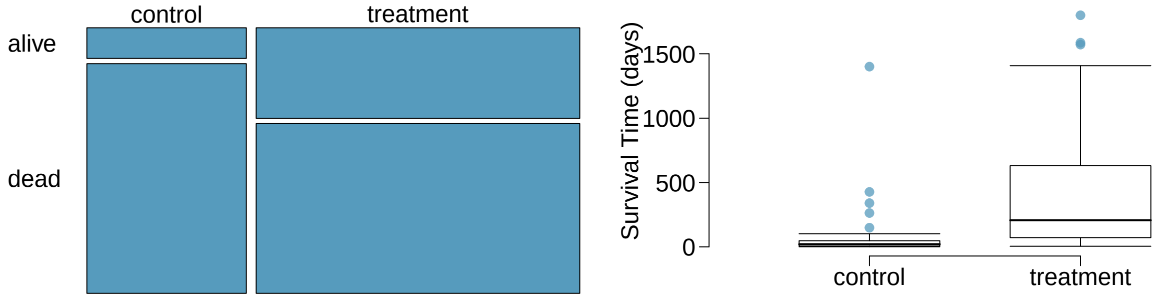
(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students dont drink since they are under 21 years old, and only a few drink excessively.

This distribution is right skewed. Since there isnt a significant number of excessive drinkers, med

(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.

As with the housing example (a), there appear to be a few highly outling salaries so the median and

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, it appears survival is dependent on whether the patient recieved a transplant. The control grou
```

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

```
The heart transplant treatment significantly increases the survival time (days) for patients. it ap
```

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

```
Treatment group = 65.22%
Control group = 88.25%
Note: I used the heartTr data set in the OpenIntro package.
```

```r
library('openintro')
data('heartTr')
summary('heartTr')
```

```
##    Length     Class      Mode 
##         1 character character
```

```r
control_all <- nrow(subset(heartTr, heartTr$transplant == 'control'))
control_dead <- nrow(subset(heartTr, heartTr$transplant == 'control' & heartTr$survived == 'dead'))
proportion_control_dead <- control_dead/control_all

proportion_control_dead
```

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

```r
treatment_all <- nrow(subset(heartTr, heartTr$transplant == 'treatment'))
treatment_dead <- nrow(subset(heartTr, heartTr$transplant == 'treatment' & heartTr$survived == 'dead'))
proportion_treatment_dead <- treatment_dead/treatment_all

proportion_treatment_dead
```

```
## [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?

```
We are testing whether or not heart transplants increase the chances of survival.
```

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.

```r
library('openintro')
data('heartTr')
summary('heartTr')
```

```
##    Length     Class      Mode
##         1 character character
```

```r
all_alive <- nrow(subset(heartTr, heartTr$survived == 'alive'))
all_alive
```

```
## [1] 28
```

```r
all_dead <- nrow(subset(heartTr, heartTr$survived == 'dead'))
all_dead
```

```
## [1] 75
```

```r
treatment_all <- nrow(subset(heartTr, heartTr$transplant == 'treatment'))
treatment_all
```

```
## [1] 69
```

```r
control_all <- nrow(subset(heartTr, heartTr$transplant == 'control'))
control_all
```

```
## [1] 34
```

```r
proportion_treatment_dead - proportion_control_dead
```

```
## [1] -0.230179
```

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 [approximately zero]. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are *highly divergent from our calculated/expected value of* -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.

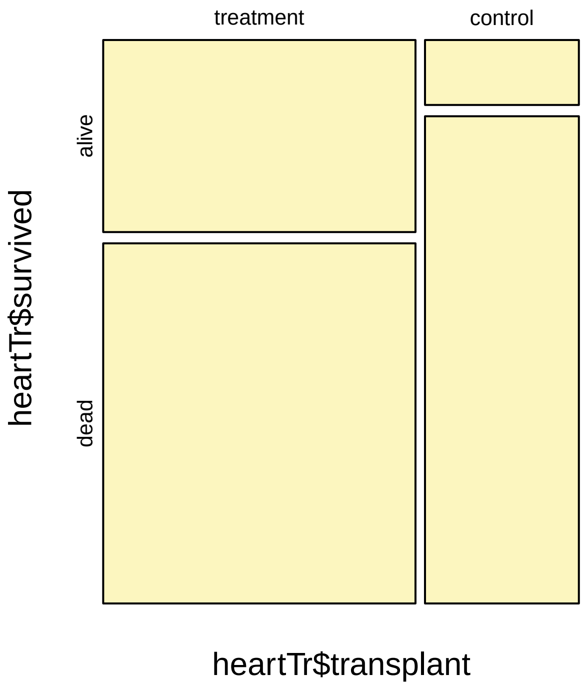
iii. What do the simulation results shown below suggest about the effectiveness of the transplant program? from the figure its apparent that only 2% of the time would a difference of atleast 23% be achieved. This is thus a rare event once that cannot be achieved easily by chance or luck.

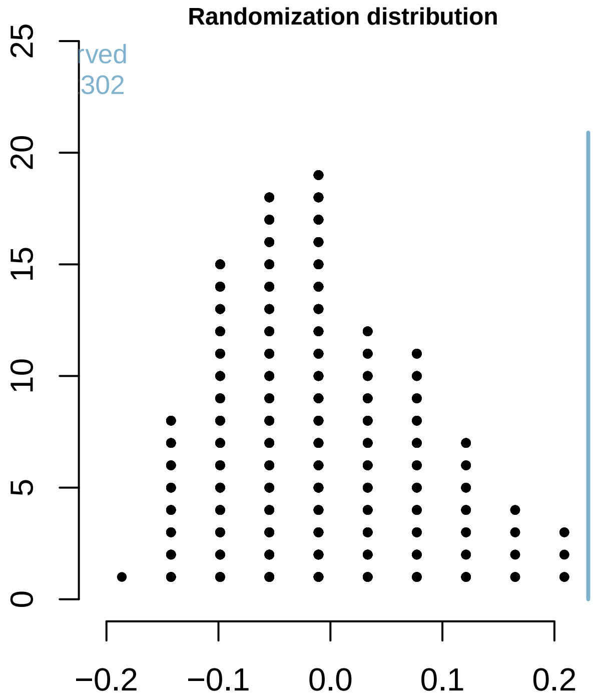
```r
# randomization -----------------------------------------------------
diffs <- DATA606::inference(heartTr$survived, heartTr$transplant,
                  success = "dead", order = c("treatment","control"),
                  est = "proportion", type = "ht", method = "simulation",
                  nsim = 100, null = 0, alternative = "twosided", simdist = TRUE,
                  seed = 95632)
```

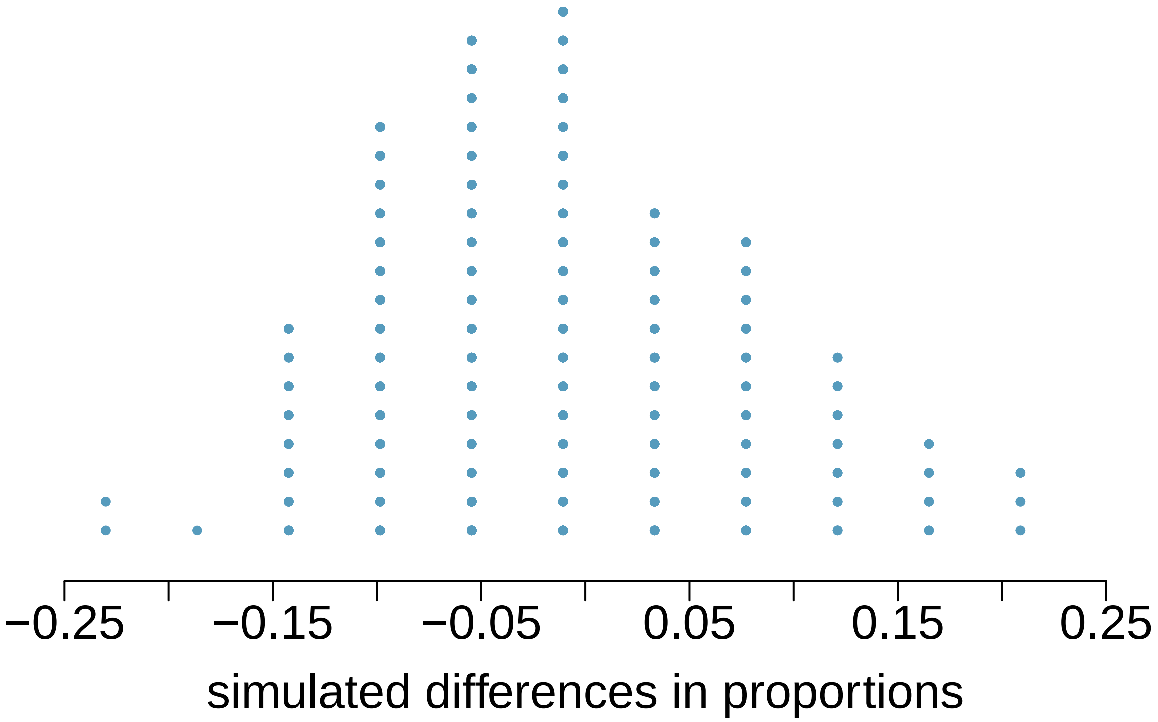
```
## Response variable: categorical, Explanatory variable: categorical
## Two categorical variables
## Difference between two proportions -- success: dead
## Summary statistics:
##         x
## y       treatment control Sum
##   alive        24       4  28
##   dead         45      30  75
##   Sum          69      34 103
```

```
## Observed difference between proportions (treatment-control) = -0.2302
##
## H0: p_treatment - p_control = 0
## HA: p_treatment - p_control != 0
```

```
## p-value =  0.04
```


treatment
control
alive
dead
heartTr$survived
heartTr$transplant


Randomization distribution
rved
302
0
5
10
15
20
25
−0.2
−0.1
0.0
0.1
0.2


−0.25
−0.15
−0.05
0.05
0.15
0.25
simulated differences in proportions