IS5 in R: Comparing Groups (Chapter 17)

Margaret Chien and Nicholas Horton (nhorton@amherst.edu)

July 11, 2018

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fifth Edition of *Intro Stats* (2018) by De Veaux, Velleman, and Bock. More information about the book can be found at http://wps.aw.com/aw_deveaux_stats_series. This file as well as the associated R Markdown reproducible analysis source file used to create it can be found at http://nhorton.people.amherst.edu/is5.

This work leverages initiatives undertaken by Project MOSAIC (http://www.mosaic-web.org), an NSF-funded effort to improve the teaching of statistics, calculus, science and computing in the undergraduate curriculum. In particular, we utilize the `mosaic` package, which was written to simplify the use of R for introductory statistics courses. A short summary of the R needed to teach introductory statistics can be found in the mosaic package vignettes (http://cran.r-project.org/web/packages/mosaic). A paper describing the mosaic approach was published in the *R Journal*: https://journal.r-project.org/archive/2017/RJ-2017-024.

Chapter 17: Comparing Groups

```
library(mosaic)
library(readr)
library(janitor)
```

Section 17.1: A Confidence Interval for the Difference Between Two Proportions

XX NH need seatbelts data and online profiles data

Section 17.2: Assumptions and Conditions for Comparing Proportions

Section 17.3: The Two-Sample z-Test: Testing for the Difference Between Proportions

XX NH need sleep data

Section 17.4: A Confidence Interval for the Difference Between Two Means

Section 17.5: The Two-Sample t-Test: Testing for the Difference Between Two Means

Step-By-Step Example: A Two-Sample t-Test for the Difference Between the Two Means

```
# page 556
BuyingCam <- read_csv("http://nhorton.people.amherst.edu/is5/data/Buy_from_a_friend.csv")
```

```
## Parsed with column specification:
## cols(
##   Friend = col_integer(),
##   Stranger = col_integer()
## )
```

By default, `read_csv()` prints the variable names. These messages can be suppressed using the `message=FALSE` code chunk option to save space and improve readability.

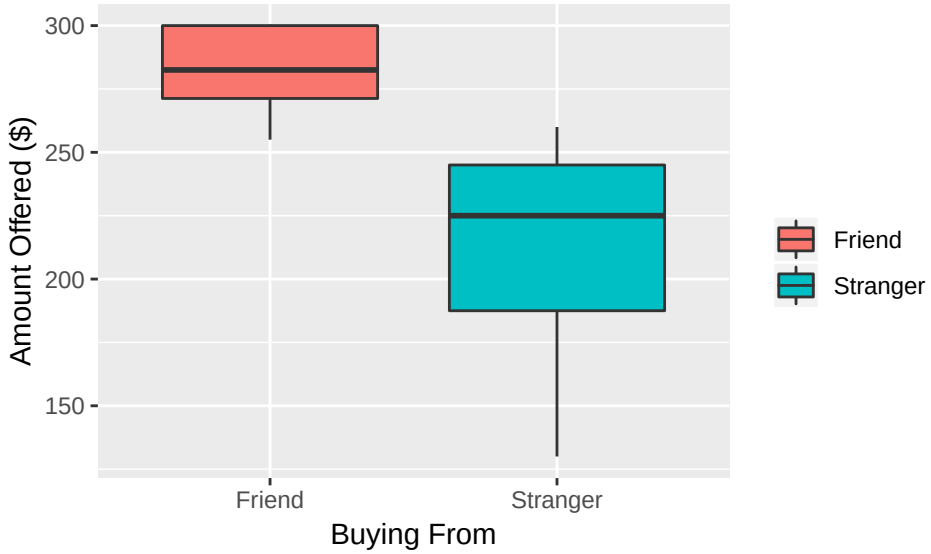
```
library(tidyr) # for gather() function
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:Matrix':
##
##     expand
```

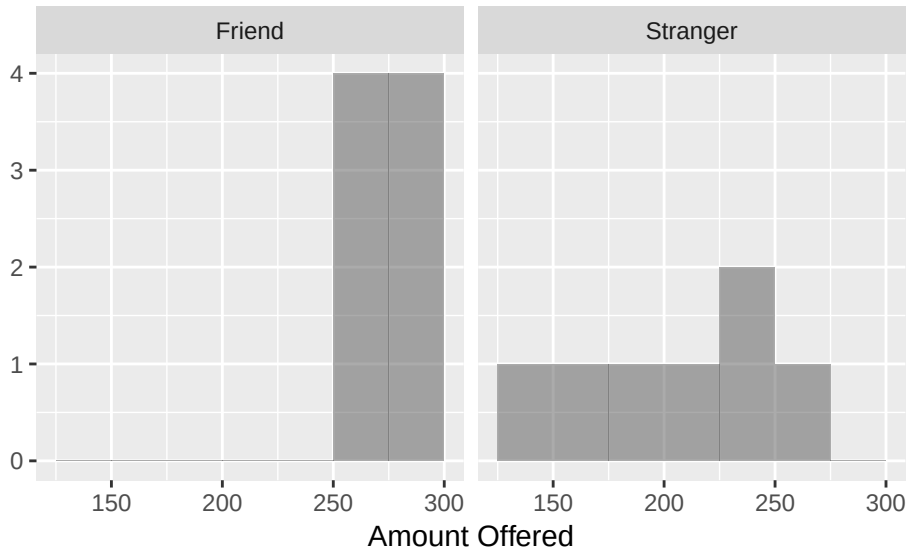
```
BuyingCam <- BuyingCam %>%
  gather(key = buying_type, value = amount_offered, Friend, Stranger)
# Model
gf_boxplot(amount_offered ~ buying_type, fill = ~ buying_type, data = BuyingCam) %>%
  gf_labs(x = "Buying From", y = "Amount Offered ($)", fill = "")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```




```
gf_histogram(~ amount_offered, binwidth = 25, center = 12.5, data = BuyingCam) %>% # doesn't exactly ma
  gf_facet_wrap(buying_type ~ .) %>%
  gf_labs(x = "Amount Offered", y = "")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```




```
# Mechanics
favstats(~ amount_offered | buying_type, data = BuyingCam)
```

```
##   buying_type min     Q1 median  Q3 max     mean       sd n missing
## 1      Friend 255 271.25  282.5 300 300 281.8750 18.31032 8       0
## 2    Stranger 130 187.50  225.0 245 260 211.4286 46.43223 7       1
```

Section 17.6: Randomization Tests and Confidence Intervals for Two Means

Section 17.7: Pooling

Section 17.8: The Standard Deviation of a Difference