Comparing Groups (Chapter 20)

Patrick Frenett, Vickie Ip, and Nicholas Horton (nhorton@amherst.edu)

June 22, 2018

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fourth Edition of *Intro Stats* (2013) 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 https://nhorton.people.amherst.edu/is4.

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.

Note that some of the figures in this document may differ slightly from those in the IS4 book due to small differences in datasets. However in all cases the analysis and techniques in R are accurate.

Chapter 20: Comparing Groups

Section 20.1: The standard deviation of a difference

We can replicate the calculations in the example on the bottom of page 543.

```
n1 <- 248
p1 <- 0.57
n2 <- 256
p2 <- 0.70
sediff <- sqrt(p1*(1 - p1)/n1 + p2*(1 - p2)/n2)
sediff
```

```
## [1] 0.04252786
```

Section 20.3: Confidence interval for a difference

We can replicate the values from the example on page 546.

```
(p2 - p1) + c(-1.96, 1.96)*sediff
```

```
## [1] 0.04664539 0.21335461
```

Section 20.4: Testing for a difference in proportions

We can replicate the values from the example on pages 550-551.

```
n1 <- 293
y1 <- 205
n2 <- 469
y2 <- 235
ppooled <- (y1 + y2)/(n1 + n2)
ppooled
```

```
## [1] 0.5774278
```

```
sepooled <- sqrt(ppooled*(1 - ppooled)/n1 + ppooled*(1 - ppooled)/n2)
sepooled
```

```
## [1] 0.0367838
```

```
z <- (y1/n1 - y2/n2)/sepooled
z
```

```
## [1] 5.398915
```

```
pval <- 2*pnorm(z, lower.tail = FALSE)
pval
```

```
## [1] 6.704501e-08
```

Section 20.6: Testing for a difference in means

```
n1 <- 8
n2 <- 7
ybar1 <- 281.88
ybar2 <- 211.43
s1 <- 18.31
s2 <- 46.43
sediff <- sqrt(s1^2/n1 + s2^2/n2)
sediff
```

```
## [1] 18.70483
```

```
t <- (ybar1 - ybar2)/sediff
t
```

```
## [1] 3.766407
```

```
pval <- 2*pt(t, df = 7.62)
pval
```

```
## [1] 1.993996
```

```
prices <- read.csv("https://nhorton.people.amherst.edu/sdm4/data/Camera_prices.csv")
prices
```

```
##   Buying.from.a.Friend Buying.from.a.Stranger
## 1                  275                    260
## 2                  300                    250
## 3                  260                    175
## 4                  300                    130
## 5                  255                    200
## 6                  275                    225
## 7                  290                    240
## 8                  300                     NA
```

```
with(prices, t.test(Buying.from.a.Friend, Buying.from.a.Stranger))
```

```
##
##  Welch Two Sample t-test
##
## data:  Buying.from.a.Friend and Buying.from.a.Stranger
## t = 3.766, df = 7.6229, p-value = 0.006003
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   26.93688 113.95597
## sample estimates:
## mean of x mean of y
##  281.8750  211.4286
```

```
ds <- with(prices,
  data.frame(price = c(Buying.from.a.Friend, Buying.from.a.Stranger),
             group = c(rep("Friend", nrow(prices)), rep("Stranger", nrow(prices)))))
ds
```

```
##    price    group
## 1    275   Friend
## 2    300   Friend
## 3    260   Friend
## 4    300   Friend
## 5    255   Friend
## 6    275   Friend
## 7    290   Friend
## 8    300   Friend
## 9    260 Stranger
## 10   250 Stranger
## 11   175 Stranger
## 12   130 Stranger
## 13   200 Stranger
## 14   225 Stranger
## 15   240 Stranger
## 16    NA Stranger
```

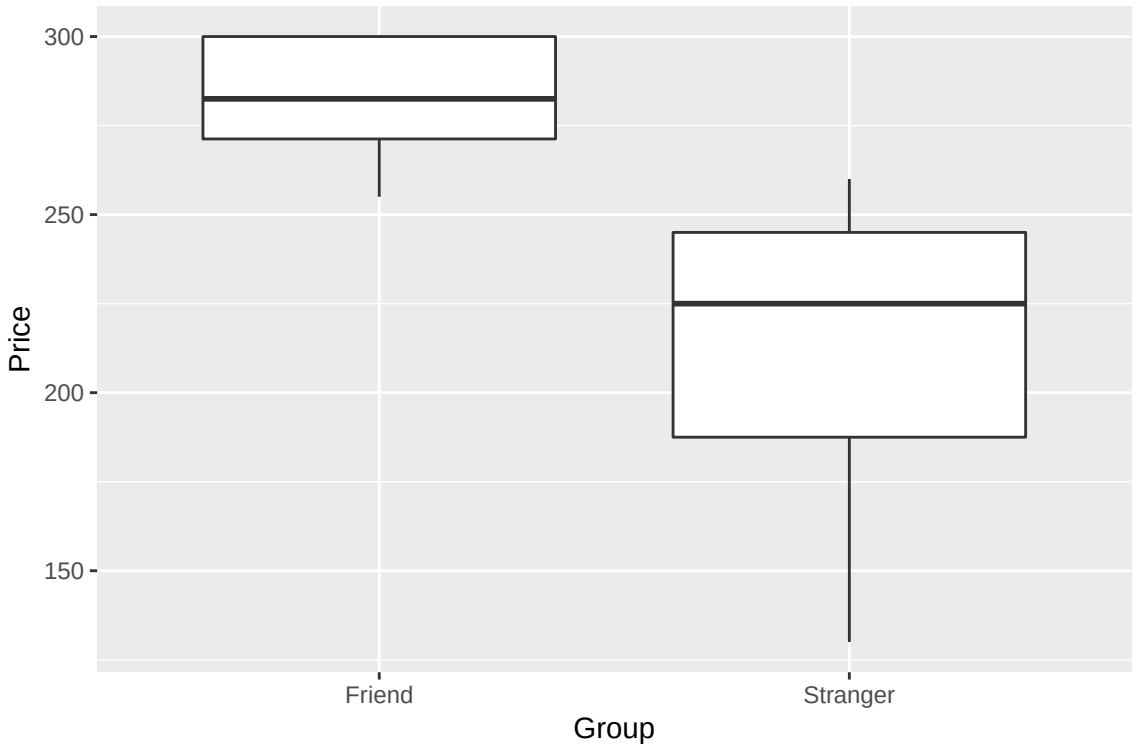
```r
t.test(price ~ group, data = ds)   # Unpooled
```

```
##
##  Welch Two Sample t-test
##
## data:  price by group
## t = 3.766, df = 7.6229, p-value = 0.006003
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   26.93688 113.95597
## sample estimates:
##   mean in group Friend mean in group Stranger
##               281.8750               211.4286
```

```r
t.test(price ~ group, var.equal = TRUE, data = ds)   # Pooled
```

```
##
##  Two Sample t-test
##
## data:  price by group
## t = 3.9699, df = 13, p-value = 0.0016
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   32.11047 108.78238
## sample estimates:
##   mean in group Friend mean in group Stranger
##               281.8750               211.4286
```

```r
gf_boxplot(price ~ group, data = ds) %>%
  gf_labs(x = "Group", y = "Price")
```


300
250
200
150
Price
Friend
Stranger
Group