SDM4 in R: Inferences about Means (Chapter 20)

Nicholas Horton (nhorton@amherst.edu)

June 13, 2018

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fourth Edition of *Stats: Data and Models* (2014) 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/sdm4.

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 20: Inferences about Means

Section 20.1: The Central Limit Theorem

Let's begin by reproducing the figure on the bottom of page 519.

```
mu <- 1309
sd <- 15.7
xpnorm(c(mu-3*sd, mu-2*sd, mu-sd, mu+sd, mu+2*sd, mu+3*sd), mean = mu, sd = sd)
```

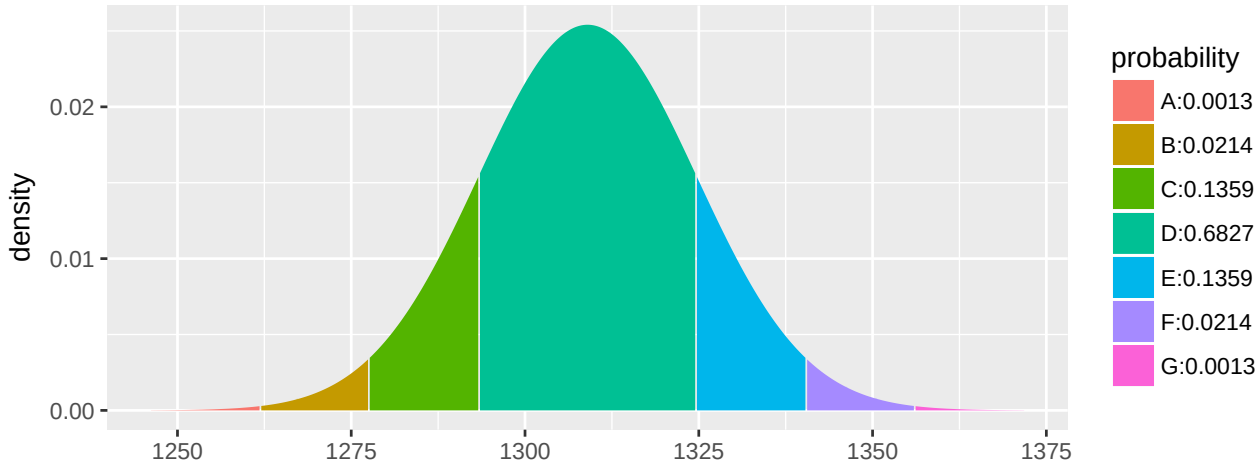
```
##

## If X ~ N(1309, 15.7), then

##  P(X <= 1262) = P(Z <= -3) = 0.00135 P(X <= 1278) = P(Z <= -2) = 0.02275 P(X <= 1293) = P(Z <= -1) =

##  P(X >  1262) = P(Z >  -3) = 0.99865 P(X >  1278) = P(Z >  -2) = 0.97725 P(X >  1293) = P(Z >  -1) =

##
```

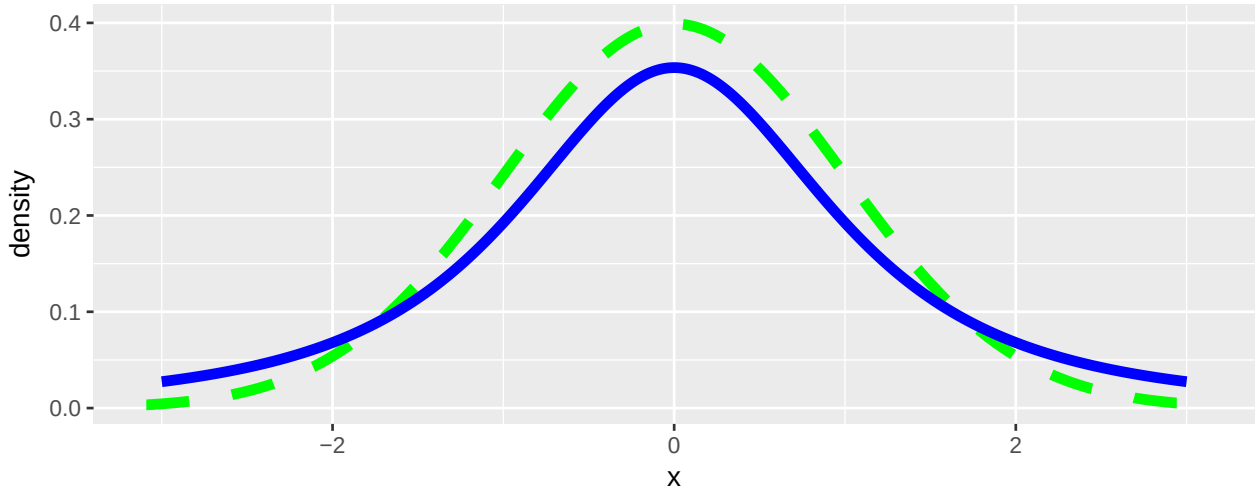


```
## [1] 0.00135 0.02275 0.15866 0.84134 0.97725 0.99865
```

Section 20.2: Gosset's t

Figure 20.1 (page 521) displays a normal curve (dashed green curve) and a t-model with 2 degrees of freedom (solid blue curve).

```
gf_dist("norm", lty = 2, col = "green", lwd = 2) %>%
  gf_dist("t", params = 2, lty = 1, lwd = 2, col = "blue", xlim = c(-3, 3))
```



We can reproduce the calculations for the Farmed salmon example (pages 523-524) using summary statistics:

```
n <- 150
ybar <- 0.0913
s = 0.0495
tstar <- qt(0.975, df = n-1)
tstar
```

```
## [1] 1.98
```

```
ybar + c(-tstar, tstar)*s/sqrt(n)
```

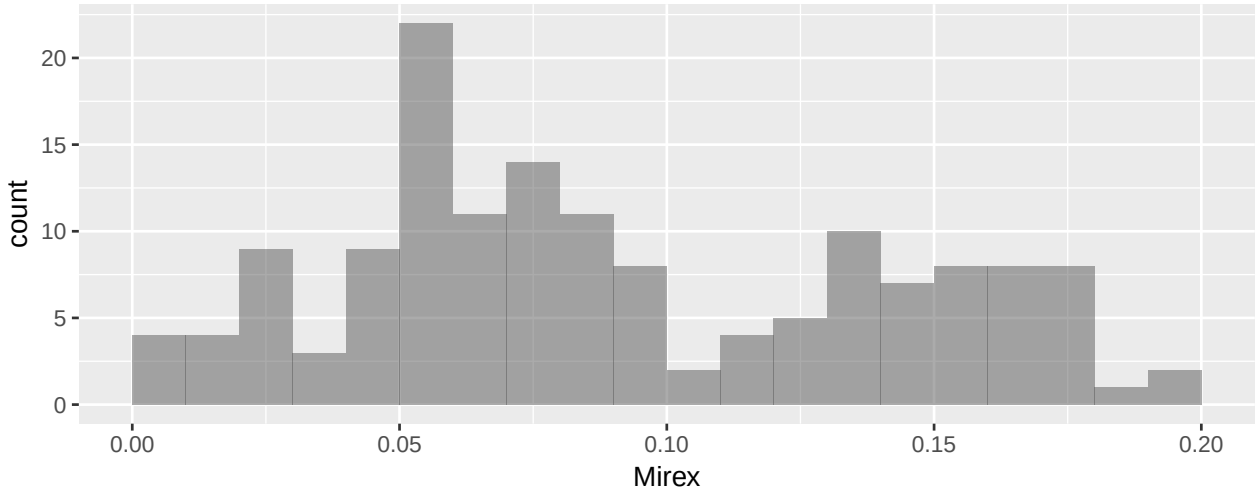
```
## [1] 0.0833 0.0993
```

or directly:

```
Salmon <- read.csv("http://nhorton.people.amherst.edu/sdm4/data/Farmed_Salmon.csv")
favstats(~ Mirex, data = Salmon)
```

```
##  min    Q1 median    Q3   max   mean     sd   n missing
##    0 0.056  0.079 0.135 0.194 0.0913 0.0495 150       3
```

```
gf_histogram(~ Mirex, binwidth = 0.01, center = 0.01/2, data = Salmon)
```




```
t.test(~ Mirex, data = Salmon)
```

```
##
##  One Sample t-test
##
## data:  Mirex
## t = 20, df = 100, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.0833 0.0993
## sample estimates:
## mean of x
##    0.0913
```

We note that the distribution of measurements is not particularly normal.

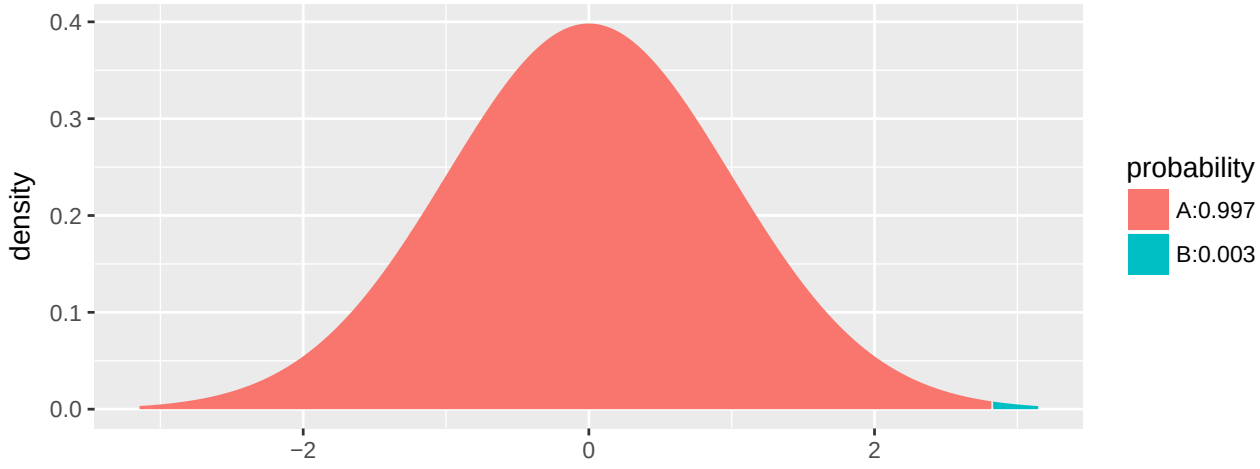
Section 20.4: A hypothesis test for the mean

We can carry out the one-sided test outlined on page 530:

```
tval <- (.0913-0.08)/0.0040
tval
```

```
## [1] 2.83
```

```
1-xpt(tval, df = 149)
```




```
## [1] 0.00269
```