

SDM4 in R: Confidence Intervals for Proportions (Chapter 18)

Nicholas Horton (nhorton@amherst.edu)

August 12, 2017

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 18: Confidence intervals for proportions

Section 18.1: A confidence interval

The Facebook survey of 156 respondents yielded 48 who updated their status every day (or more often).

```
n <- 156
binom.test(48, n)      # exact binomial

##
##
##
## data:  48 out of n
## number of successes = 48, number of trials = 156, p-value = 2e-06
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.236 0.386
## sample estimates:
## probability of success
##                0.308
```

Calculation on page 473

```
phat <- 48/n; phat
```

```
## [1] 0.308
```

```
sep <- sqrt(phat*(1-phat)/n); sep # matches value on page 473
```

```
## [1] 0.037
```

```
interval <- phat + c(-2, 2)*sep; interval
```

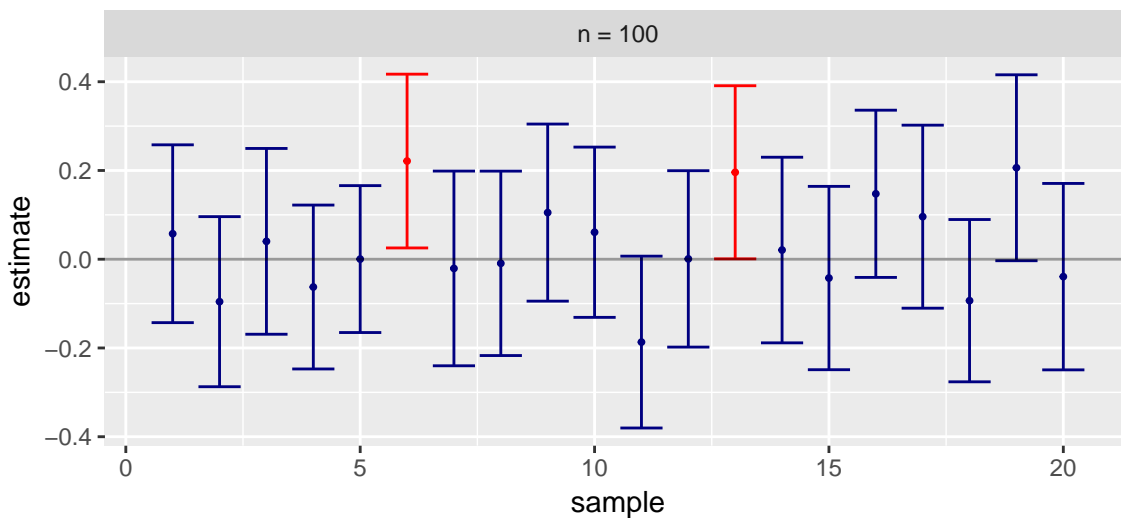
```
## [1] 0.234 0.382
```

Section 18.2: Interpreting confidence intervals

```
set.seed(1988)  
CIsim(n=100, samples=20)
```

```
## Interval coverage:
```

```
##      cover  
## n      Low Yes High  
## 100 0.0 0.9 0.1
```



We would expect 19 out of 20 of the intervals to cover the true (population) value, but here only 18 out of 20 actually covered that value (see Figure at top of page 476).

Section 18.3: Margin of error

We can replicate the calculation for the “For Example: Finding the margin of error Take 1” (page 478)

```
sdp <- sqrt(.5*.5/1010); sdp # worst case margin of error (based on p=0.5)
```

```
## [1] 0.0157
```

```
me <- 2*sdp; me
```

```
## [1] 0.0315
```

We can replicate the calculation for the “For Example: Finding the margin of error Take 1” (pages 478-479)

```
qnorm(.95, mean=0, sd=1) # z-star for 90% confidence interval
```

```
## [1] 1.64
```

```
sep <- sqrt(.4*.6/1010); sep
```

```
## [1] 0.0154
```

```
me <- 1.6445*sep; me
```

```
## [1] 0.0254
```

Section 18.4: Assumptions

We can replicate the calculation for the “For Example: choosing a sample size” (page 482)

```
zstar <- qnorm(.975, mean=0, sd=1); zstar
```

```
## [1] 1.96
```

```
me <- 0.02 # desired margin of error
```

```
p <- 0.40
```

```
n <- (zstar*sqrt(p*(1-p)))/me)^2; n
```

```
## [1] 2305
```

We will need about 2305 subjects to yield a margin of error of 2% under these assumptions.