Assignment 1

Alfonso Rojas-Alvarez

September 9, 2017

Prepare the dataset:

```
set.seed(3347)
library(ggplot2)
library(foreign)
library(Hmisc)
library(grid)
library(gridExtra)
library(easyGgplot2)
rm(list = ls())
setwd("/Users/Alfonso/Google Drive/UT/Fall 2017/RD")
```

Load Dataset:

```
students <- read.dta("STAR_students.dta")
```

Load Variable Labels:

```
var.labels <- attr(students, "var.labels")
data.key <- data.frame(var.name=names(students), var.labels)
```

Clear Mising gkschid:

```
STAR_kindergarteners <- students[!(is.na(students$gkschid)),]
```

How many students are in the population (N)? Suppose I told you that almost every Tennessee kindergartner participated in the STAR experiment. Would you believe me? Why or why not?

```
nrow(STAR_kindergarteners)

## [1] 6325
```

The population is 6325 students. In this data set, every kindergartner participated in the STAR experiment, so it would be a reasonable inference to state that almost every student in TN participated in the STAR experiment. With a sample size that big, we can be somewhat confident about our inferences, as long as we asume the data set is unbiased.

In the population, what are the mean and standard deviation of the kindergarten reading tests?

Population Parameters:

```
gktreadss <- na.omit(STAR_kindergarteners$gktreadss)
mu <- mean(gktreadss)
mu

## [1] 436.7253

sigma <- sd(gktreadss)
sigma

## [1] 31.70626
```

Samples

Now take simple random samples from the population, repeatedly. Using the sample command, take a simple random sample of n=160 students.

Set Sample Size:

```
n <- 160
```

Take Sample 1:

```
sample1 <- STAR_kindergarteners[sample(nrow(STAR_kindergarteners), n),]
```

Use your sample to estimate the population mean. Use the mean command and paste the output below.

```
gktreadss_sample1 <- na.omit(sample1$gktreadss)
xbar1 <- mean(gktreadss_sample1)
xbar1

## [1] 434.9252

sigma1 <- sd(gktreadss_sample1)
sigma1

## [1] 28.35564
```

How much does the mean of your sample differ from the mean of the population? What is this difference called?

It is called a Sampling Error, and for this case it is the following:

```
diff1 <- mu - xbar1
diff1

## [1] 1.800171
```

How large is the difference compared to the estimated standard error?

Standard Error:

```
se1 <- sigma1 / 40
se1

## [1] 0.708891
```

The standard error is much smaller.

Look at the 95% confidence interval which Stata has calculated from your sample. Does it cover the population mean?

Confidence Interval:

```
c1 <- xbar1 - qnorm(0.975) * sigma1/sqrt(n)
c2 <- xbar1 + qnorm(0.975) * sigma1/sqrt(n)

sample1_CI <- c(c1, c2)
sample1_CI

## [1] 430.5315 439.3188
```

Is the sample mean "significantly different" from 440?

Hypothesis test:

```
z1 <- (xbar1 - mu)/(sigma1/sqrt(n))
z1

## [1] -0.8030347

alpha <- 0.05
critical_values <- c(-qnorm(1-alpha/2), qnorm(1-alpha/2))
reject1 <- isTRUE(!-qnorm(1-alpha/2) <  z1 & z1 < qnorm(1-alpha/2))
reject1

## [1] FALSE
```

No, we can't reject the null hypothesis that they are different.

Reload the population. Then take a different simple random sample of 160 students.

Take Sample 2:

```
sample2 <- STAR_kindergarteners[sample(nrow(STAR_kindergarteners), n),]
```

Use your new sample to re-estimate the population mean.

```
gktreadss_sample2 <- na.omit(sample2$gktreadss)
xbar2 <- mean(gktreadss_sample2)
xbar2

## [1] 435.331

sigma2 <- sd(gktreadss_sample2)
sigma2

## [1] 30.93633
```

How much does the mean of this sample differ from the mean of the first sample? What is this difference called?

It is called a Sample Variation, and for this case it is the following:

```
sample_variation <- xbar1 - xbar2
sample_variation

## [1] -0.4058644
```

How large is the difference compared to the estimated standard error?

Standard Error:

```
se2 <- sigma2 / 40
```

It is smaller, and much closer to the SE than the Sampling Error.

In this second sample, does the 95% confidence interval cover the population mean?

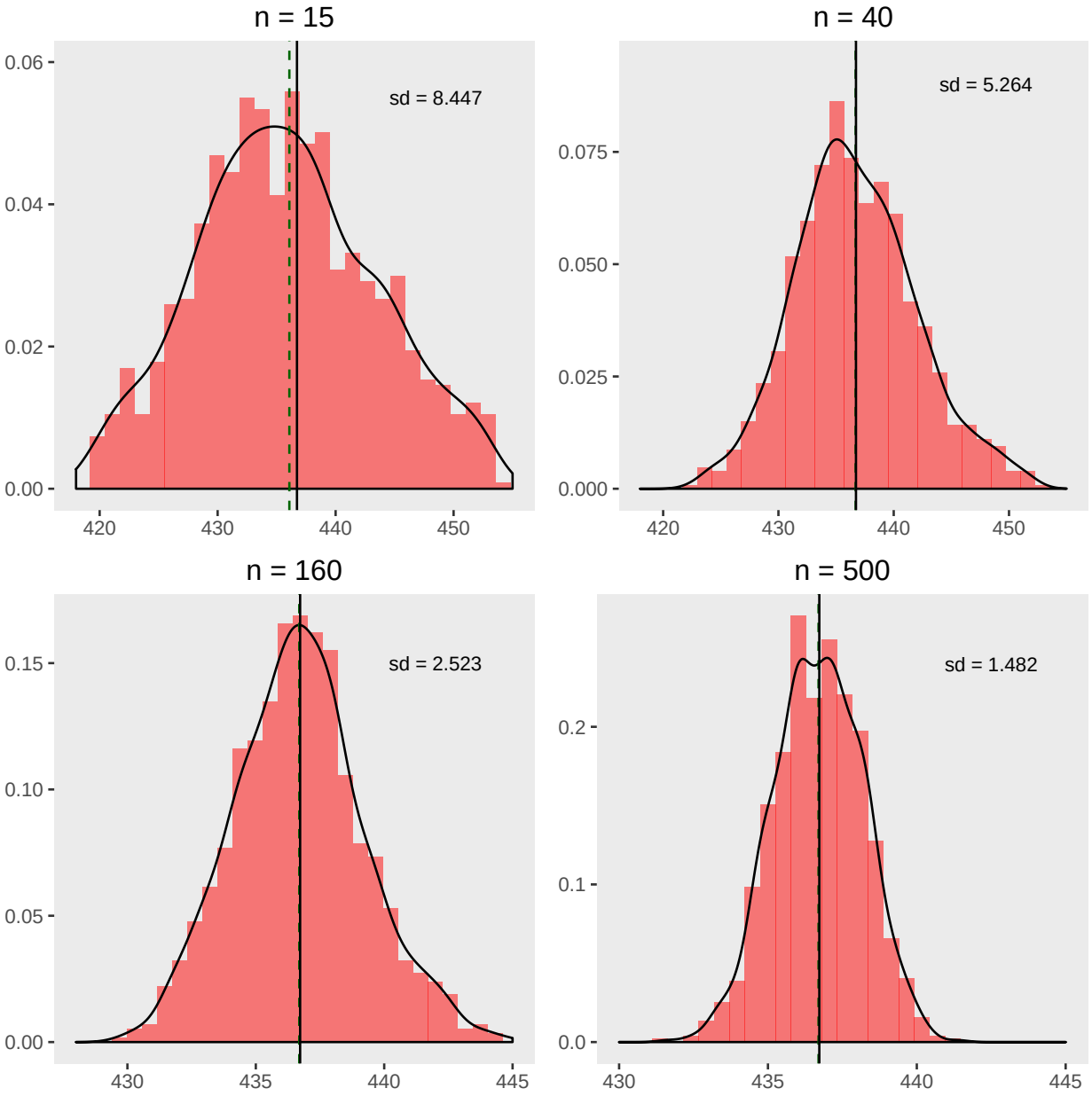
Confidence Interval:

```r
c1 <- xbar2 - qnorm(0.975) * sigma2/sqrt(n)
c2 <- xbar2 + qnorm(0.975) * sigma2/sqrt(n)

sample2_CI <- c(c1, c2)
sample2_CI

## [1] 430.5375 440.1246
```

Yes, the confidence interval we have estimated contains the population mean.

Bootstrap Method

```
ggplot2.multiplot(a,b, c, d, cols=2)
```




```
round(mu, digits=3)

## [1] 436.725
```

```
##       Mean       SD   Mu-Sigma       C1       C2          Z Reject
## 1 433.1739 29.96556  3.5514281 430.5474 435.8005 -2.6501205      1
## 2 439.6746 34.04862 -2.9492278 436.6901 442.6590  1.9368402      0
## 3 436.2775 32.62128  0.4478081 433.4182 439.1369 -0.3069559      0
## 4 434.2581 31.07733  2.4672067 431.5341 436.9821 -1.7751980      0
## 5 434.1856 28.27658  2.5397516 431.7071 436.6641 -2.0083960      1
## 6 435.4677 31.26556  1.2576352 432.7272 438.2082 -0.8994426      0
```