Week 3 HW

Zai Rutter

1/29/2022

```r
############
## Homework 3
## Zai Rutter
## Jan 26 2022
############

############
# Question 1
############

#### A
# Expected value: 76
# Variance of 35
#prob that average is between 73 and 78, in 50 obvs.

# should we divide by 50?
## bilal: we need -

pnorm(78, mean = 76, sd = sqrt(35/50)) - pnorm(73, mean = 76, sd = sqrt(35/50))
```

```
## [1] 0.9914182
```

```r
# standarderrormean
sem <- sqrt(35/50)

#### B
# Find a symmetric 99 percent confidence interval on the mean in a
# random sample of 75 observations?
semb1 <- sqrt(35/75)

qnorm(0.0045, 76, sd = semb1)
```

```
## [1] 74.21563
```

```r
qnorm(.9955, 76, sd = semb1)
```

```
## [1] 77.78437
```

```r
76 + qnorm(0.0045)*semb1
```

```
## [1] 74.21563
```

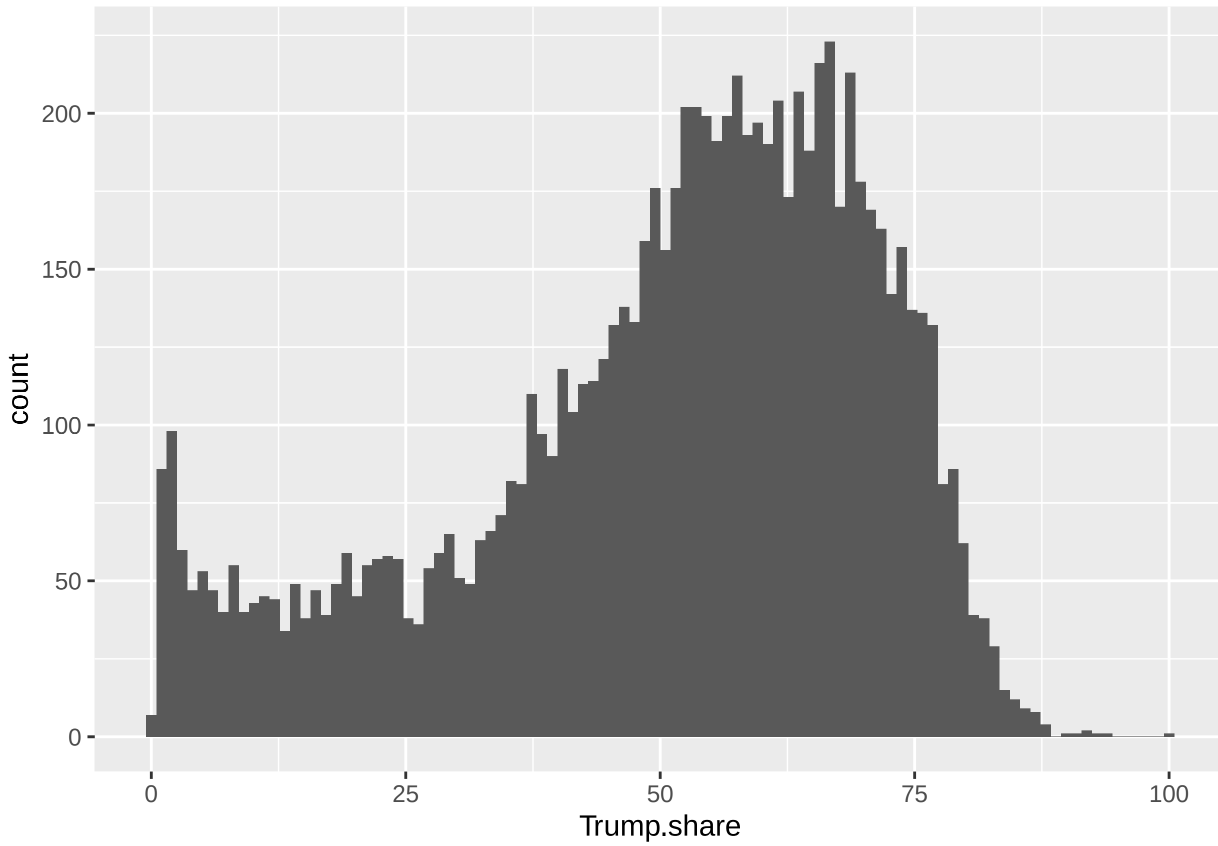
```r
76 + qnorm(.9955)*semb1
```

```
## [1] 77.78437
```

```r
#### C
# How large must my sample be in order for there to be a 95 percent probability
# that my sample average is between 75.5 and 76.5?
# (Note, you will likely have to round the answer you find to the nearest whole
# number. We can't sample a fraction of person. YET.)

# CI = Mean +/- z-score * sd of distribution
# Z-score is just the number of standard deviations away from the mean where you
# would expect this probability

# 1.96 * se = margin of error
# se
# 1.96 * se = 0.5
# se = sqrt var/n
# 1.96 * root var/n = 0.5

(qnorm(.975))^2 * 35 / ((.5)^2)
```

```
## [1] 537.8042
```

```r
############
# Question 2
############

#### A
library(readxl)
library(ggplot2)
Ohio2016 <- read_excel("~/Documents/Upenn/Data 310/Week 3/Homework/Ohio2016.xlsx")

Ohio2016$Percent.Trump<-(Ohio2016$Trump / Ohio2016$Ballots)
Ohio2016$Trump.share <- Ohio2016$Percent.Trump  * 100

Meantrump <- weighted.mean(Ohio2016$Trump.share,Ohio2016$Ballots)



ggplot(Ohio2016, mapping = aes(x = Trump.share)) +
  geom_histogram(bins=100)
```




```
# Almost, but there is too many data points at the left tail for this to be a good
# Candidate for Normal Distribution.

#### B

# Not really because the data from those who did not vote are not included

# Population Mean
Meantrump <- weighted.mean(Ohio2016$Trump.share,Ohio2016$Ballots)
Meantrump
```

```
## [1] 50.6631
```

```
TrumpVar <- var(Ohio2016$Percent.Trump)
TrumpVar
```

```
## [1] 0.04156068
```

```
#### C
# 99 percent confidence interval would be on this statistic.
## Sample size 40 , Find 99% CI using Central Limit Thereom

semC2 <- sqrt(TrumpVar/40)
semC2
```

```
## [1] 0.03223379
```

```r
Meantrump + qnorm(.995)*semC2
```

```
## [1] 50.74613
```

```r
Meantrump + qnorm(.005)*semC2
```

```
## [1] 50.58008
```

```r
qnorm(0.995, mean = Meantrump,
      sd = semC2)
```

```
## [1] 50.74613
```

```r
qnorm(0.005, mean = Meantrump,
      sd = semC2)
```

```
## [1] 50.58008
```

```r
#### D
# Calculate the bounds of this 99 percent confidence interval if 80 or 120
# precincts were sampled instead.

Meantrump + qnorm(.995)*sqrt(var(Ohio2016$Trump.share)/120)
```

```
## [1] 55.45677
```

```r
Meantrump + qnorm(.005)*sqrt(var(Ohio2016$Trump.share)/120)
```

```
## [1] 45.86944
```

```r
Meantrump + qnorm(.995)*sqrt(var(Ohio2016$Trump.share)/80)
```

```
## [1] 56.53412
```

```r
Meantrump + qnorm(.005)*sqrt(var(Ohio2016$Trump.share)/80)
```

```
## [1] 44.79209
```

```r
#### E
# To confirm that the bounds calculated by the Central Limit Theorem are correct,
# run the following simulation

# is this the right variable?

Q3E <- rep(NA,10000 )


for (i in 1:length(Q3E)) {
  Q3E[i]<-mean(sample(Ohio2016$Trump.share,120))
}

quantile(Q3E,probs = c(.005, .995))
```

```
##      0.5%    99.5%
## 45.79972 55.17172
```

```
plot(density(Q3E))
```