Problem Set 4

Political Data Science - Spring 2020

Gencer, Alper Sukru

Due May 1, 10:00 AM (Before Class)

Instructions

1. The following questions should each be answered within an R script. Be sure to provide many comments in the script to facilitate grading. Undocumented code will not be graded.
2. Work on git. Fork the repository found at https://github.com/domlockett/PDS-PS3 and add your code, committing and pushing frequently. Use meaningful commit messages – these may affect your grade.
3. You may work in teams, but each student should develop their own R script. To be clear, there should be no copy and paste. Each keystroke in the assignment should be your own.
4. If you have any questions regarding the Problem Set, contact the TAs or use their office hours.
5. For students new to programming, this may take a while. Get started.
6. You will need to install ggplot2 and dplyr to complete this dataset.

Question 1 - Sample Statistics

Load the following data: http://politicaldatascience.com/PDS/Datasets/GSS-data.csv.

The variable poleff11 asks participants to rate their level of agreement with the statement "People like me don't have any say about what the government does" (see the codebook for more information on all variables in this dataset at: http://politicaldatascience.com/PDS/Datasets/gss_codebook.csv).

1. Convert this variable into a numeric where higher values indicate higher levels of political efficacy (1- strongly agrees with the statement; 5- strongly disagrees with the statment) and all other values ('Cant choose' etc.) become NA's.
2. What is the proportion of individuals from the entire sample who feel as though they "have a say in the government?"
3. Using a sample of 25 from this dataset. What is the average proportion who feel as thought hey have a say?
4. Pull a random sample of 25 from the poleff11 data and calculate the mean for this outcome. Now repeat this process 500 times and store these values in a variable called trials 25.
5. Now create a variable called trials 100 where we do 500 trials with $n = 100$ instead of 25.
6. Draw a histogram of the sampling distribution for the two trials ($n = 25$ vs. $n = 100$) you just conducted. Give the plots meaningful titles and axis labels. Save these plots in your repository.
7. What notable difference occur when we use a larger sample size in our trials?

Answer 1 - Sample Statistics

Load the following data: http://politicaldatascience.com/PDS/Datasets/GSS-data.csv.

```
rm(list = ls())
gss.data <- read.csv("http://politicaldatascience.com/PDS/Datasets/GSS-data.csv")
gss.data <- gss.data[-c(2349, 2350), ] ## included some unrelated information
```

Convert this variable into a numeric where higher values indicate higher levels of political efficacy (1- strongly agrees with the statement; 5- strongly disagrees with the statment) and all other values ('Cant choose' etc.) become NA's:

```
levels(gss.data$poleff11)
```

```
## [1] ""                           "Agree"
## [3] "Cant choose"                "Disagree"
## [5] "Neither agree nor disagree" "No answer"
## [7] "Not applicable"             "Strongly agree"
## [9] "Strongly disagree"
```

```
gss.data$poleff11.recoded[gss.data$poleff11 == "Strongly agree"] <- 1
gss.data$poleff11.recoded[gss.data$poleff11 == "Agree"] <- 2
gss.data$poleff11.recoded[gss.data$poleff11 == "Neither agree nor disagree"] <- 3
gss.data$poleff11.recoded[gss.data$poleff11 == "Disagree"] <- 4
gss.data$poleff11.recoded[gss.data$poleff11 == "Strongly disagree"] <- 5
gss.data$poleff11.recoded[gss.data$poleff11 == "Cant choose"] <- NA
gss.data$poleff11.recoded[gss.data$poleff11 == "Not applicable"] <- NA
gss.data$poleff11.recoded[gss.data$poleff11 == "No answer"] <- NA
```

What is the proportion of individuals from the entire sample who feel as though they "have a say in the government?":

```
########  For this, I'll be looking at the proportion of people who "disagree (4) \ strongly disagree (

#---  including NA values:
prop_have.a.say.na <-sum(table(gss.data$poleff11.recoded)[4:5])/nrow((gss.data))
prop_have.a.say.na
```

```
## [1] 0.1937819
```

```
#---  excluding NA values:
prop_have.a.say <-sum(table(gss.data$poleff11.recoded)[4:5])/length(na.omit(gss.data$poleff11.recoded))
prop_have.a.say
```

```
## [1] 0.3953084
```

Using a sample of 25 from this dataset. What is the average proportion who feel as thought hey have a say?:

```r
set.seed(20200425)

#---  including NA values:
sample25_na <-sample(gss.data$poleff11.recoded, size =25)
sum(sample25_na>=4, na.rm = T)/25
```

```
## [1] 0.24
```

```r
#---  excluding NA values:
sample25_no.na <-sample(na.omit(gss.data$poleff11.recoded), size =25)
sum(sample25_no.na>=4, na.rm = T)/25
```

```
## [1] 0.36
```

\subsection{Pull a random sample of 25 from the poleff11 data and calculate the mean for this outcome. Now repeat this process 500 times and store these values in a variable called \textbf{trials_25}:}

```r
########  For convenience, I will be solving by removing NA's.

#---  Random one sample:
mean(sample(na.omit(gss.data$poleff11.recoded), size =25, replace = T))
```

```
## [1] 2.6
```

```r
#---  Random 500 samples:
trials_25 <- NULL
for(i in 1:500){
  trials_25 <- c(trials_25, mean(sample(na.omit(gss.data$poleff11.recoded), size =25, replace = T)))
}
```

\subsection{Now create a variable called trials_100 where we do 500 trials with n=100 instead of 25:}

```r
########  For convenience, I will be solving by removing NA's.

#---  Random one sample with n = 100:
mean(sample(na.omit(gss.data$poleff11.recoded), size =100, replace = T))
```

```
## [1] 2.92
```

```r
#---  Random 500 samples with n = 100:
trials_100 <- NULL
for(i in 1:500){
  trials_100 <- c(trials_100, mean(sample(na.omit(gss.data$poleff11.recoded), size =100, replace = T)))
}
```

Draw a histogram of the sampling distribution for the two trials (n=25 vs. n=100) you just conducted. Give the plots meaningful titles and axis labels. Save these plots in your repository.:

```r
library(tidyverse)
##install.packages("ggpubr")
library(ggpubr)

df.trials <- as.data.frame(cbind(trials_25, trials_100))
```

```r
p_trials_100 <- ggplot(df.trials) +
  geom_histogram(aes(trials_100), bins = 100, fill = "red") +
  xlab("Sample Means") +
  ggtitle("Sampling Dist. with Size 100") +
  ylim(0,50) +
  xlim(2,4) +
  theme_light()

p_trials_25 <- ggplot(df.trials) +
  geom_histogram(aes(trials_25), bins = 100, fill = "blue") +
  xlab("Sample Means") +
  ggtitle("Sampling Dist. with Size 25") +
  ylim(0,50) +
  xlim(2,4) +
  theme_light()

ggarrange(p_trials_100, p_trials_25,
                    ncol = 2, nrow = 1)
```

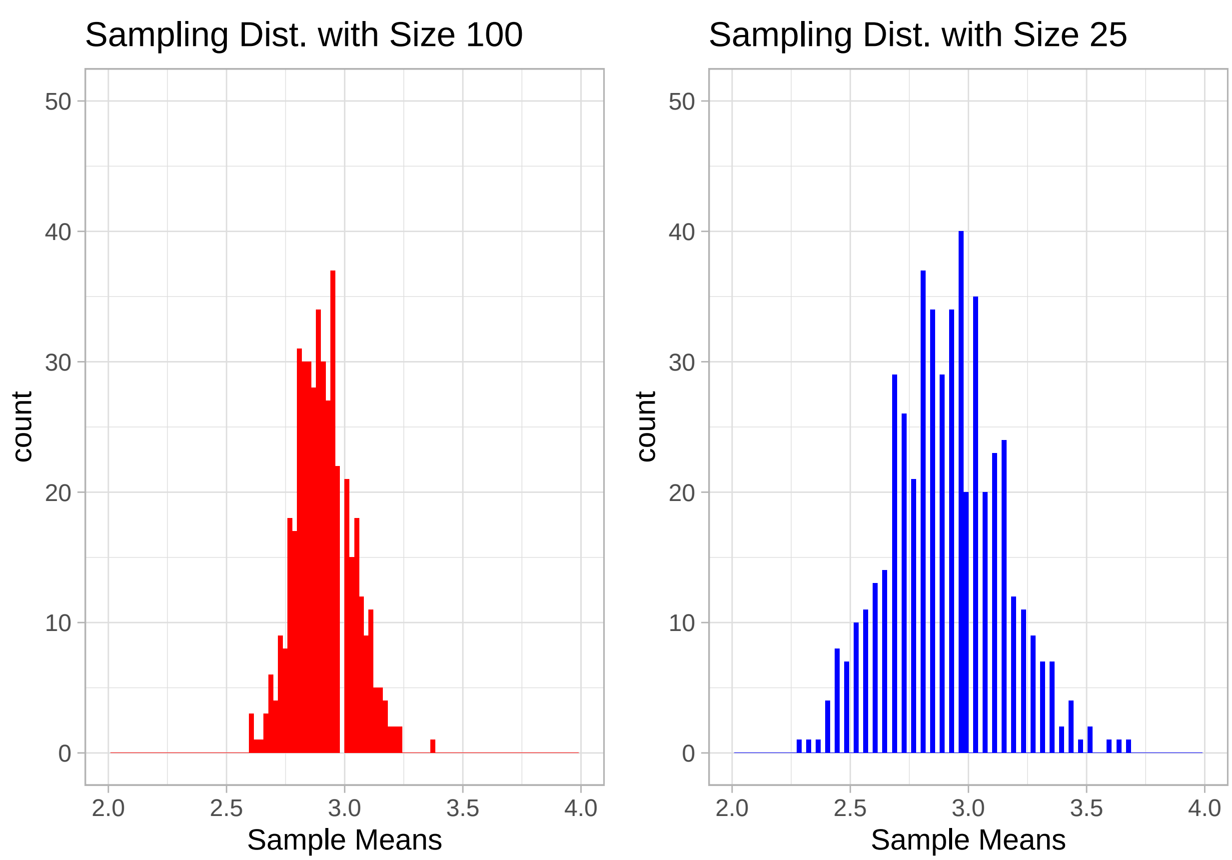

Figure 1: Sampling Distributions with Varying Sample Sizes

What notable difference occur when we use a larger sample size in our trials?:

As it can be seen in the graph, the spead of the sampling distribution with sample size 100 is narrower than the one with sample size 25.

```
library(knitr)

table <- cbind(mean(na.omit(gss.data$poleff11.recoded)), mean(trials_100),
      sd(trials_100), mean(trials_25), sd(trials_25))

colnames(table) <- c("True Mean", "Size-100 Mean", "Size-100 SD",
                  "Size-25 Mean", "Size-25 SD")

kable(table)
```

True Mean	Size-100 Mean	Size-100 SD	Size-25 Mean	Size-25 SD
2.908775	2.91692	0.1183373	2.90648	0.2374067

Also see from the table that the mean of size-100 trials is closer to the true mean and at the same time, the sampling distribution of sample mean has less variance (smaller standard error).

Q2 - tidyverse