API - 209 | Problem Set 3

Prof. Dan Levy

Due on Tuesday, September 20 2022 at 10:00 am.

INSTRUCTIONS

To successfully complete this problem set, please follow these steps:

1. **Download this RMarkdown document file into your computer.**
2. **Insert all your answers into this document.** Guidance **here** on how to insert objects such as handwritten work or screenshot images in your answers.
3. **SAVE your work frequently**.
4. To make things easier to visualize in RStudio, you can set the view mode as "Visual" instead of as "Source" in the top left of your screen (just below the Save button).
5. Once your document is complete, please save it as a PDF by clicking the **KNIT** button.
6. Please submit an electronic copy of the PDF (and any separate requested files) to the Canvas course page.

   6.a) If you want to check a PDF version of this problem set before starting to work on it, you can always knit it. In fact, you can knit the document at any point.

   6.b) If you cannot Knit and it's time to submit the problem set, submit the RMarkdown file and make an appointment with a member of the teaching team
7. Remember to consult the R resources from math camp, particularly the HKS R cheat sheet (available **here**, which contains many of the commands needed to answer the questions in this problem set.

IDENTIFICATION

1. Your information

```
Last Name:   Chaturvedi
First Name: Shreya
```

2. Group Members (please list below the classmates you worked with on this problem set):

```
Group members: Sohaib Nasim, Manisha Jha, Kusha
```

3. Compliance with Harvard Kennedy School Academic Code: Do you certify that my work in this problem set complies with the Harvard Kennedy School Academic Code[1] (mark with an X below)?

[1]We abide by the Harvard Kennedy School Academic code (available here) for all aspects of the course. In terms of problem sets, unless explicitly written otherwise, the norms are the following: You are free (and encouraged) to discuss problem sets with your classmates. However, you must hand in your own unique written work and code in all cases. Any copy/paste of another's work is plagiarism. In other words, you can work with your classmate(s), sitting side-by-side and going through the problem set question-by-question, but you must each type your own answers and your own code. For more details, please see syllabus.

[X] YES [ ] NO

QUESTION 1 – COLOMBIAN PRESIDENTIAL ELECTIONS[2]

On June 19, 2022, Colombia voted to elect their first left-wing President Gustavo Petro (a former M-19 rebel also previously serving as Mayor of Bogotá) and first Afro-Colombian female Vice President Francia Márquez. The elections leading up to this historic result were hotly contested. Ten candidates announced a run for the Presidency, six made it to the ballots on 29 May 2022, but none cleared the threshold to win by earning 50% of the vote. The electoral procedure then dictated the top two candidates to battle it out in a second round, which was famously *too close to call* even just days before the runoffs.

When Gustavo Petro of the Pacto Histórico por Colombia alliance finally triumphed over former Mayor of Bucaramanga, Rodolfo Hernández of the Liga de Gobernantes Anticorrupción, he was a whisker above the halfway mark. Petro received 50.4% of the vote while Hernández received 47.3%, and 2.2% of the votes were cast blank. Many consider Petro and Marquez's victory to be a turning point, but how easily could the elections have gone the other way?

In this problem set you will use data from the Colombian Presidential Run-off Election 2022 to help internalize some key concepts about statistical inference. The exercise will ask you to draw random samples from this population using R. There are some hints in the **Appendix** provided to help you with the questions below. My advice is that you refer to the appendix only after having tried to figure out how to do it by yourself; this will better allow you to develop your R skills.

The file "Colombia Election 2022.csv" contains the actual data of the presidential second round from all voting tables in the Colombian election. Assume our population of interest comprises all voting tables in this election. Please note that in some sub-questions we ask you to compare your results with answers in earlier parts so be sure to store calculated values for easy reference.

Import the data into R by writing the relevant commands into your R script and running it. Browse the data just to get a feel for it. Now answer the questions below.

1. Based on the population (i.e. all voting tables), calculate the population parameter (Petro's win margin) and population standard deviation of the win margin in the whole dataset. The win margin is the difference in proportion of votes between Petro and Hernández in the population. In other words, the difference between Petro's and Hernández's votes over the total votes cast[3]. This is the population parameter of interest to us.

```
## [1] 0.03032751
```

```
## [1] 0.4739024
```

Answer:

Population Win margin = Total number of votes for Petro (across all rows) - Total number of votes for Hernandez (across all rows) / (Total number of votes cast across all rows)

= 0.0303

S.D. of Table Level Win Margin = 0.4739 (We calculate table-wise win margin by doing the same calculation as above but per each row, and then get the SD for the whole column)

---

2. Now draw a random sample of 5 voting tables[4]. Look at the sample and imagine this is the sample you were going to use to decide who won the election.
    a. How many possible samples of 5 tables can be drawn from the total number of tables? Write a formula or a number.
    b. Look at your sample of 5 tables and imagine you were going to use it to declare the winner. Estimate Petro's win margin in this sample.
    c. Who won the election according to your sample of 5 tables? Is this what you would expect?

---

[2]My gratitude to Camila Valencia for helping locate the data for this problem set question and to Vaishnavi Prathap for her help writing it.

[3]To do this, calculate the proportion of votes that Petro received by dividing the total number of votes that Petro received by the total number of votes in the election. Call this number A. Then do the same for Hernández, and call this number B. The population parameter of interest is A-B.

[4]Use the command `set.seed()` to make results replicable (i.e., you will get the same random sample when you re-run your code). One way to draw the random sample is the `sample_n()` function or `slice_sample()`.

```
## [1] 0.1996602
```

Answer:

2a. How many possible samples? If there are N = 103364 voting tables, we can have NC5 voting tables where NC5 = N*(N-1)*(N-2)*(N-3)*(N-4)/5*4*3*2*1 and N = 103364. We can also calculate this using choose(103364,5) = 9.83e22

2b. Petro's win margin in 1 sample of 5 voting tables: Petro's win margin in the sample of 5 tables is 0.1996.

2c. Who won? Is this what you would expect? This means Petro won in this sample with a slightly higher win margin than the population. This margin is still within +/- 1.96 SD of the population level margin (for a 95% confidence level), which is what I would expect. This also means that there are many samples which would have a win margin in the favour of Hernandez (0.03- 1.96*0.47), and so we are also quite likely to get some samples in which Hernandez wins the election.

---

3. Now draw a random sample of size 100 voting tables.
    a. How many possible samples of 100 can be drawn?
    b. Again, imagine you were going to declare the winner based on your sample of 100 tables only. Estimate the sample win margin.
    c. Based on your sample of 100 tables, who won the election?

```
## [1] 0.105517
```

Answer:

3a. How many possible samples? If there are N = 103364 voting tables, we can have NC100 voting tables samples. We can also calculate this using choose(103364,100) = 9.83e22

3b. Petro's win margin in 1 sample of 100 voting tables? Petro's win margin in the sample of 100 tables is 0.1055.

3c. Who won? In this sample of 100 tables, Petro won.

---

4. Now compare your estimate of the win margin based on the sample of size 5 with your estimate based on the sample of size 100. Which of the estimates is closest to the population parameter? Is this what you would expect?

Answer:

Insert your answer here.

Win Margin in sample of 100 tables = 0.1055

Win Margin in sample of 5 tables = 0.1996

Win Margin in population = 0.0303

The win margin for sample of 100 tables is closer to the population parameter. This is what I would expect since with a higher n, the sample distribution for the parameter is closer to the expected value of win margin.

MANY SAMPLES

As you may have noticed in questions 2a and 3a, when we draw a random sample we get one out of possibly infinite options. It is not feasible for us to examine all possible samples, but for this exercise you will draw random samples 1000 times, assuming that 1000 is sufficient to approximate the distribution of all possible samples (also known as the '*sampling distribution*').

5. Now draw 1,000 random samples of 5 voting tables each, and for each sample calculate the difference in the proportions of votes received by the two candidates. One way to run this simulation is by using a '`for loop`' with an index (typically named i) to store the results of your different samples[5]. Calculate the average and standard deviation of the 1000 win margins from the simulations.

To help get you started, you can conduct the simulation by running the following code, which will store your simulated results in a new object `'xbar_5'`. Note, the election data in this code is saved as an object named '`co`'. *You may remove the read_csv line and adjust the slice_sample command to match whatever you named the data.* If you would like to understand better what this code does and '`for loops`' more generally, look at details in the Appendix.

---

set.seed(65668)

```
## [1] 0.04981649
```

```
## [1] 0.2268719
```

Answer:

(Note: For consistency, I have removed the 100* term in the code since I have been calculating win margin as a proportion between 0 and 1).

Avg. Win Margin in 1000 samples of 5= 0.0498

SD of Win Margin in 1000 samples of 5= 0.2269

---

6. Now draw 1,000 random samples of 100 voting centers each, and for each of them calculate the difference in the proportions of votes received by the two candidates. Calculate the average and standard deviation of the win margins.

```
## [1] 0.03013495
```

```
## [1] 0.05142623
```

Answer:

Avg. Win Margin in 1000 samples of 100= 0.0301

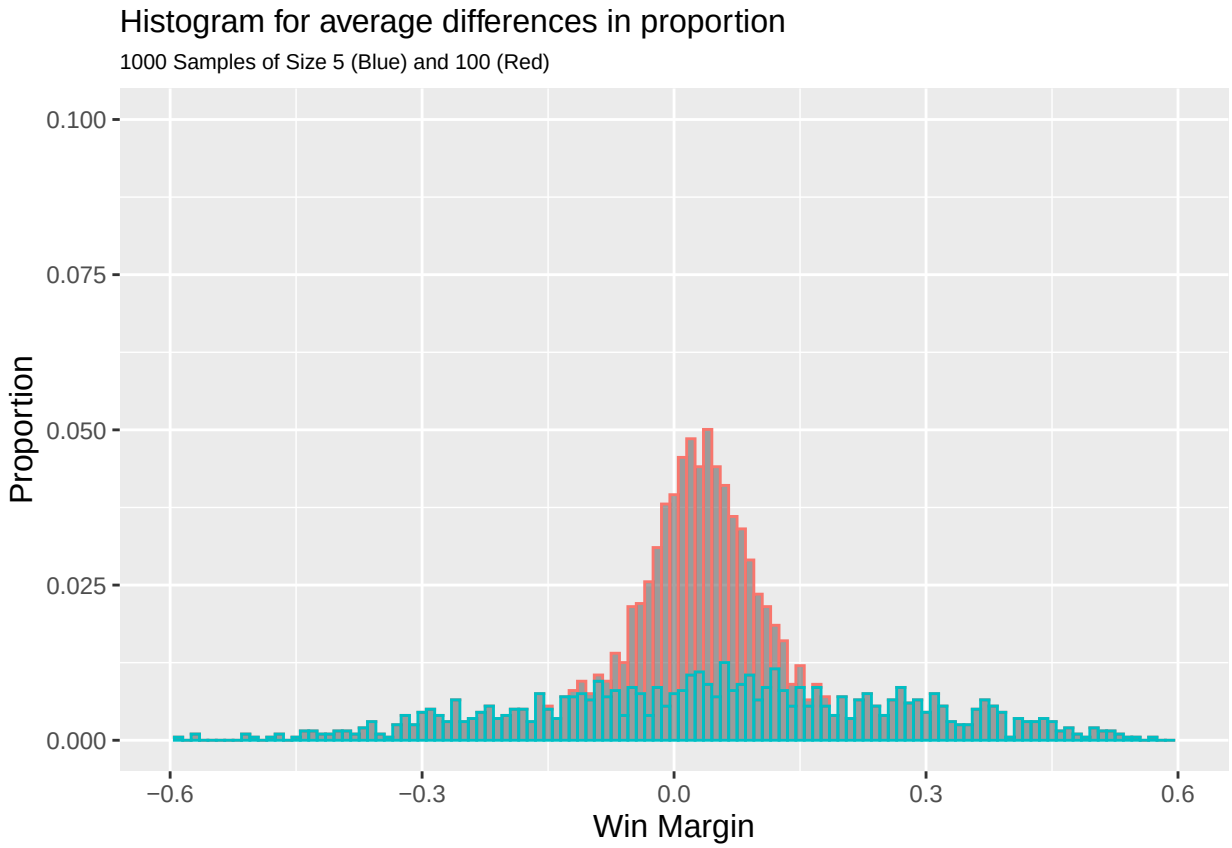
SD of Win Margin in 1000 samples of 100= 0.0514

---

7. Construct a relative frequency histogram (i.e. show proportions on the y-axis, instead of counts) of the average differences in proportions for samples of size 5 and for samples of size 100. Draw these histograms in a way that they are easy to compare visually as covered in math camp (i.e. same scale for the axes). How does the shape of the distribution of average differences in proportions change as the sample size increases (from 5 to 100)?

---

[5] Use the `for` command. If you need help on how to do this, look at the hints in the appendix. You are also free to use the map functions from the `purrr` library, the details of which are also in the appendix.



Answer: The distribution is approximately normal as a consequence of the CLT in both cases. However, as the sample size increases from 5 to 100, these changes happen:

1. The distribution is less spread out, or there are fewer outliers.
2. There is a higher number of samples with the proportion closer to the population estimate.

---

8. Compare the averages from questions 1, 5 and 6. Are the averages of the sampling distributions close to the population parameter? Which average is closer? Is this what you would expect?

Answer:

Insert your answer here.

Q1: Avg Win Margin in Population = 0.0303

Q5: Avg. Win Margin in 1000 samples of 5= 0.0498

Q6: Avg. Win Margin in 1000 samples of 100= 0.0301

The Q6 value is closer, as is expected. This because as we can see in the graph above, the distribution is tighter and we are more likely to get a value that is the closer to the true expected value (population mean).

---

9. Now compare the standard deviations from questions 1, 5 and 6.
    a. Which of these standard deviations is highest? Is this what you would expect?
    b. Are the standard deviations of the sampling distributions higher than, lower than, or approximately the same as the population standard deviation? Explain.

Answer:

Insert your answer here.

Q1: SD in Population = 0.474

Q5: SD in 1000 samples of 5= 0.227

Q6: SD in 1000 samples of 100= 0.051

The SD for population is the highest, and the SD for the sampling distributions of the samples are lesser (lowest for sample size 100).

The SDs of samples are less than the SDs of the population because as we move from a population SD to a sample SD, the potential impact to drive up the SD of any individual extreme value is reduced. The SD of population therefore is most affected by any individual outlier value, but as we move to the SD of samples, this outlier is averaged with all other values in the samples.

10. Looking at all the 1000 samples you drew of 5 and 100 voting tables respectively, examine how many of them called a win for Petro. Calculate and report the following:
    a. Proportion of samples of size 5 in which Petro won the election (i.e. proportion of samples of size 5 in which Petro's win margin is greater than zero).
    b. Proportion of samples of size 100 in which Petro won the election.
    c. Which of the two numbers reported just above is larger? Is this what you would expect? What does your answer to this question imply about the reliability of samples of size 5 vs samples of size 100 to approximate the population parameter?

```
## [1] 600
```

```
## [1] 723
```

Answer:

Insert your answer here.

600/1000 samples of size 5 had Petro winning, 723/1000 samples of size 100 had Petro winning. The second value is more as expected (more values closer to the true population estimate).

Intuitively, I would think that the sample of size 100 has a higher reliability than a sample of size 5 because the distribution of win margin is tighter (there is less variability).

---

11. Repeat these calculations, this time considering how many of the 1000 samples of size 5 and 100 respectively specifically called a **narrow** win for Petro. For the purpose of this question, define narrow win as Petro having as a positive win margin less than 3 percentage points. Calculate and report the following:
    a. Proportion of samples of size 5 in which Petro wins by a narrow margin.
    b. Proportion of samples of size 100 in which Petro wins by a narrow margin.
    c. Which of the two numbers reported just above is larger? Is this what you would expect? What does your answer to this question imply about the reliability of samples of size 5 vs. samples of size 100 to precisely estimate the population parameter?

```
## [1] 54
```

```
## [1] 228
```

Answer:

Insert your answer here.

54/1000 samples of size 5 had a narrow win margin for Petro whereas 228/1000 samples of size 100 had a narrow win margin for Petro.

We know from the brief that the election was a very close call and we can see here that the number of close margin victories in the sample distribution with larger sample sizes is much higher. This is because a larger sample size would have lower variance (or more averaging out of outliers) which can more closely reflect a narrow win margin. But in a smaller sample size, this averaging out of outliers is lesser, resulting in win margins that are too far from the expected value.

This implies that a sample of size 100 has more reliability or less variability when it comes to precisely estimating the population parameter.

---

12. How does your answer to question 9 explain the finding in question 11c?

Answer:

Insert your answer here.

The lower SD for samples of size 100 are what result in the lower variability and the consequently higher proportion of close call wins. In that sense, we could have predicted using the lower SD that we would have a higher proportion of close call wins.

---

13. For this exercise, indicate:

    a. Population parameter of interest (what it is and what value it takes).
    b. Estimator (define it conceptually for this exercise).
    c. Estimates (indicate in words what they are for this specific exercise).

Answer:

Insert your answer here.

The population parameter of interest is Petro's win margin and it takes the value of 0.03

Estimator is the win margin of the sample (ie total votes for petro in the sample - total votes for hernandez in the sample)/(total votes in the sample). Conceptually, it can be thought of as the algorithm we run over different inputs (samples) to get different outputs (estimates).

Estimates are the values we get when we calculate this (or run this algorithm) for the samples of size 5 and size 100 respectively. In this case they are 0.049 and 0.0301 respectively.

---

14. For each of the 1,000 estimates of sample size 100 that you drew in Q6, calculate a 90% confidence interval. For this, assume that the SD of the Win Margin from 1000 estimates that you calculated earlier is the standard error for your confidence intervals[6]. Determine how many of these intervals contain the population parameter of interest. Is this what you would expect? Explain.

[6] Drawing the confidence intervals from first principles in this question requires knowledge beyond the scope of this assignment. Even though this number is not, technically speaking, the standard deviation of the sampling distribution (since it is based only on 1000 estimates), it is a reasonable estimate and hence the conclusions of the Central Limit Theorem (and the formulas for confidence intervals implied by it) hold approximately. If you are familiar with bootstrapping, you will notice the spirit of it in this approximation.

Hint: if you store the calculated upper bound and lower bound estimates of your confidence intervals (CI) in objects named '`x_ub`' and '`x_lb`' respectively, you can combine these into a single data frame with the following code. Once you have this data frame, you can create a new variable, using an appropriate Boolean, to check if a given interval contains the population parameter: `samp_ci <- tibble(x_lb, x_ub)`

```
## [1] 898
```

Answer:

Insert your answer here.

The 90% confidence interval for a sample is (victory margin for the sample - 1.64*SD, victory margin for the sample + 1.64 SD). In other words, if the election were to occur a 100 times, 90 of those occurences would have a resulting win margin that is between these two bounds.

892 of the intervals contain the population parameter, which is very close to 90% and so is expected.

---

15. How would the number you reported just above change under the scenarios below. No R code required to answer this question; you should instead think conceptually.
    a. You did the same exercise but for samples of size 1,000.
    b. You did the same exercise but calculating 95% confidence intervals (instead of 90% confidence intervals)

Answer:

Insert your answer here.

a. If we did the same exercise for samples of size 1000, the standard error would be smaller and therefore the confidence interval smaller or tighter. However, I would still expect close to 90% of the intervals to contain the population parameter, by definition, getting a similar number.

b. If we did the same exercise but by calculating confidence intervals of 95%, I would expect approximately 95% of the intervals to contain the population parameter, which means a higher number (possibly close to 950 instead of the current 892).

QUESTION 2 - RELEVANCE OF SAMPLING DISTRIBUTION TO POVERTY ESTIMATION

Suppose that the true proportion of poor households in your country is 20%. You draw a random sample of 900 households in your country and estimate the proportion of poor households in your sample[7].

1. Specify the following:

Answer:

- Population parameter of interest: Proportion of poor households in the country
- Estimator: Proportion of poor households in the sample
- Sampling Distribution of estimator: An approximately normal distribution of the proportion from a sufficiently large number of samples drawn out of the population
- Shape: Approximately normal
- Mean: 20%
- Standard Deviation: sqrt(p*1-p/n) = 0.013

---

2. Calculate how **likely** is it to draw a random **sample of size 900** from this population in which the **proportion of poor people is 23% or more**. Call this number A.

Hint: you can use the `pnorm()` function in R to calculate the probability the value of a random normal variable is less than a specified level. For example `pnorm(q = 1, mean = 0, sd =1)` calculates the probability a standard normal distribution takes on a value of less than 1.

```
## [1] 0.01050813
```

Answer:

Insert your answer here. This means that there is a 1.05% chance of drawing a sample of size 900 that has 23% or more households that are classified as poor.

---

3. How likely is it to draw a random sample of size 900 from this population in which the proportion of poor people is 20% or less? Is this number greater than A, equal to A, or less than A? Explain your reasoning. No need to do any new calculations.

```
## [1] 0.5
```

Answer:

Insert your answer here.

About half of the samples should have a proportion of poor households that is 20% or less, which is greater than A. This is because 20% is the mean of the distribution.

[7] Assume that you used the exact same definition of poverty that was used to calculate the official poverty rate of 20%.

4. How likely is it to draw a random sample of size 1,600 from this population in which the proportion of poor people is 23% or more? Is this number greater than A, equal to A, or less than A? Explain your reasoning. No need to do any new calculations.

```
## [1] 0.01050813
```

Answer:

Insert your answer here.

It is even more unlikely to draw a random sample of size 1600 from this population that has 23% or more poor households. Since the sample size is bigger, we know that the standard error will be even smaller than 0.013 and the distribution will be more concentrated at the mean.

We know intuitively (or if we visualise the distribution) that fewer values will be present at the outliers of the distribution as a result of the higher sample size.

---

5. **INDIVIDUAL OR SMALL GROUP ACTIVITY**: A policymaker who is intelligent but not well-versed in statistics is confused about which distribution (population distribution, distribution in the sample, or sampling distribution) is used to answer questions (2)-(4) above. Explain to them which one is being used and why. [One short paragraph or a single slide. The slide may include animations.]

   Recall your reflections so far in the course about communicating statistical concepts. Once you/your group have prepared the material, please post it to the "#concepts-samp-dist" Slack channel, mentioning the names of all contributors. And please feel free to learn from and react to other groups' insights. The student/group submitting the best slide might be invited to present it in class.

   Write "Done" below after posting to the "#concepts-samp-dist" Slack channel.

Answer:

Done

QUESTION 3 – ONLINE MODULE ON LOGIC OF HYPOTHESIS TESTING

The goal of this problem set question is to help you prepare you for the class on **Hypothesis Testing Overview** that will be held next week. The idea is to get everyone familiar with the basics of estimators and their key properties so that we can delve deeper in class on this topic than we would be able to do if we had to go through the basics in class.

You will be asked to watch a short module and answer some questions in a quiz. The quiz results will give me information about overall performance of the class that I will use to prepare for class; your individual performance in the quiz will be registered in the system but will not count towards your grade in any way.

To get full credit for this question, you need to watch the module and complete the quiz.

The module is available here.

Answer:

Please enter "Done" in this field once you have completed the quiz.

Done

QUESTION 4 – COURSE FEEDBACK

I would like to take a moment now to get your views on how the course is going. Please complete this brief survey to give me your feedback. Your answers are anonymous, and I would like to encourage you to be both candid and constructive. My goal is to be able to use your feedback to improve the course. I estimate it will take you about 10 minutes to fill it in.

Thanks in advance.

Dan

The survey is available here.

Answer:

Please enter "Done" in this field once you have completed the survey.

Done

TIME USE

Please enter in **this** form the time you spent on each question. This information will only be used for teaching improvements; **please be candid** and report the time (in **MINUTES**) spent in each question.

Please enter "Done" in this field once you have completed the form.

Done

This is a copy of your code.

```
.answer-box {
  background-color: LemonChiffon;
}
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(options(width = 60))
knitr::opts_chunk$set(class.output = "bg-warning")

packages <- c('haven','dplyr', 'ggplot2', 'reshape2', 'tidyverse', 'pracma',
              'lubridate', 'scales', 'ggthemes', 'gt', 'dineq', 'gglorenz')
to_install <- packages[!(packages %in% installed.packages()[,"Package"])]
if(length(to_install)>0) install.packages(to_install,
                                          repos='http://cran.us.r-project.org')
lapply(packages, require, character.only=TRUE)


Last Name:   Chaturvedi
First Name: Shreya
Group members: Sohaib Nasim, Manisha Jha, Kusha
                           [ X ] YES               [   ] NO
#Import data
inputdata <- read.csv("Colombia Election 2022.csv")
# Insert only code here.
total_petro_pop = sum(inputdata$petro, na.rm = TRUE)
total_hernandez_pop = sum(inputdata$hernandez, na.rm = TRUE)
total_votes_cast_pop = sum(inputdata$votes_cast, na.rm = TRUE)
win_margin_pop = (total_petro_pop - total_hernandez_pop)/total_votes_cast_pop


inputdata <- inputdata %>%
  mutate(win_margin_tw = (petro-hernandez)/votes_cast)

win_margin_tw_sd = sd(inputdata$win_margin_tw, na.rm = TRUE)

win_margin_pop
win_margin_tw_sd
# Insert only code here.
set.seed(65668)
sample_5tables <- sample_n(inputdata,5)
total_petro_sample = sum(sample_5tables$petro, na.rm = TRUE)
total_hernandez_sample = sum(sample_5tables$hernandez, na.rm = TRUE)
total_votes_cast_sample = sum(sample_5tables$votes_cast, na.rm = TRUE)
win_margin_sample = (total_petro_sample - total_hernandez_sample)/total_votes_cast_sample

win_margin_sample

# Insert only code here.

set.seed(65668)
sample_100tables <- sample_n(inputdata,100)
total_petro_sample_2 = sum(sample_100tables$petro, na.rm = TRUE)
total_hernandez_sample_2 = sum(sample_100tables$hernandez, na.rm = TRUE)
total_votes_cast_sample_2 = sum(sample_100tables$votes_cast, na.rm = TRUE)
win_margin_sample_2 = (total_petro_sample_2 - total_hernandez_sample_2)/total_votes_cast_sample_2

win_margin_sample_2
# Insert only code here.


set.seed(65668)
#For-loop code
```

```r
xbar_5 <- c()
for (i in 1:1000) {
  samp_5 <- slice_sample(inputdata, n = 5)
  xbar_5[i] <- ((sum(samp_5$petro)-sum(samp_5$hernandez)) / sum(samp_5$votes_cast))
}
mean(xbar_5)
sd(xbar_5)

# Insert only code here.

set.seed(65668)
xbar_100 <- c()
for (i in 1:1000) {
  samp_100 <- slice_sample(inputdata, n = 100)
  xbar_100[i] <- ((sum(samp_100$petro)-sum(samp_100$hernandez)) / sum(samp_100$votes_cast))
}
mean(xbar_100)
sd(xbar_100)
# Insert only code here.
xbar_5_df <- as.data.frame(xbar_5) %>%
  mutate (c = "xbar_5")
colnames(xbar_5_df) <- c("values","set")

xbar_100_df <- as.data.frame(xbar_100) %>%
  mutate (c = "xbar_100")
colnames(xbar_100_df) <- c("values","set")

data <- rbind(xbar_5_df,xbar_100_df)

ggplot(data , aes(x = values, color = set, alpha = 0.9)) +
  geom_histogram(aes(y = stat(count)/sum(stat(count))),
                 binwidth = 0.01) +
  scale_y_continuous(labels = percent) +
  labs(x="Win Margin",
       y = "Proportion",
       title="Histogram for average differences in proportion",
       subtitle="1000 Samples of Size 5 (Blue) and 100 (Red)") +
  xlim(-0.6, 0.6) +
  ylim(0,0.1) +
  theme(plot.title=element_text(size=12, hjust=0),
        plot.subtitle=element_text(size=8,hjust=0),
        plot.caption= element_text(size=7),
        axis.title.x=element_text(size=12,hjust=0.5),
        axis.title.y=element_text(size=12,hjust=0.5)) +
  theme(legend.position = "none",legend.title = element_blank())


# Insert only code here.
sum(xbar_5 > 0)
sum(xbar_100 > 0)

# Insert only code here.
sum(xbar_5 > 0 & xbar_5 < 0.03)
sum(xbar_100 > 0 & xbar_100 < 0.03)

# Insert only code here.
std_error <- 0.051
xbar_100 <- as.data.frame(xbar_100) %>%
  mutate(x_lb = xbar_100 - 1.64*std_error) %>%
  mutate(x_ub = xbar_100 + 1.64*std_error)
```

```r
sum(xbar_100$x_lb < 0.03 & 0.03 < xbar_100$x_ub)

# Insert only code here.
1-pnorm(q = 0.23, mean = 0.2, sd = 0.013)

# Insert only code here.
pnorm(q = 0.2, mean = 0.2, sd = 0.013)
# Insert only code here.

1-pnorm(q = 0.23,mean = 0.2, sd = 0.013)
```