EDUC152, Problem Set #1

Grade:

Overview

add text

Create data and load functions

Run the code in the following chunk, which does the following:

- loads libraries
- loads and creates IPEDS data frame (population)
- creates data frame of generated variables (population)
- creates sample versions of the IPEDS and gnerated data frames

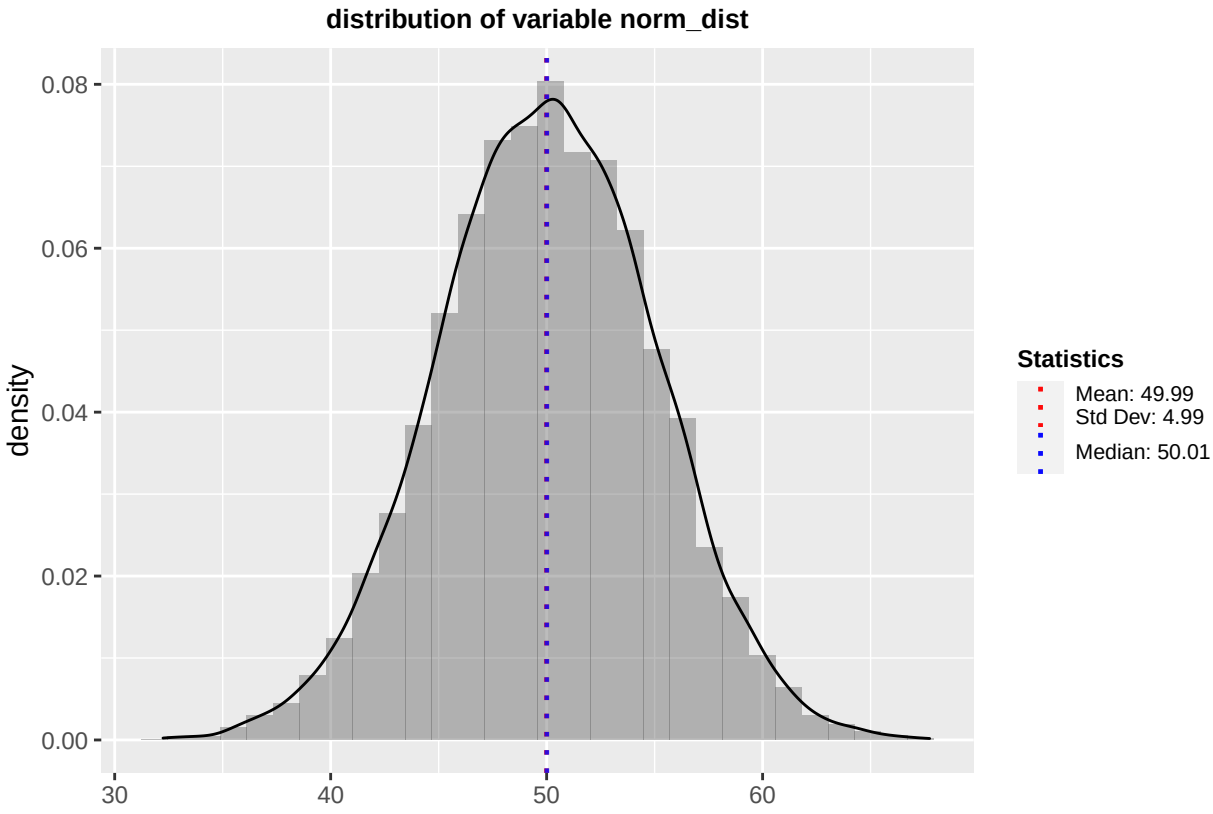
Note: code chunk omitted from PDF document

/1

Distributions and sampling distribution

Use the function we created `plot_distribution()` to plot the distribution of the variable `norm_dist` from the data frame `df_generated_pop`

```
plot_distribution(data_vec = df_generated_pop$norm_dist, plot_title = "distribution of variable norm_di
```



What is the standard deviation and interpret this value in words

- YOUR ANSWER HERE:

Does the distribution above have a normal, left-skewed, or right-skwed shape? Why?

- YOUR ANSWER HERE:

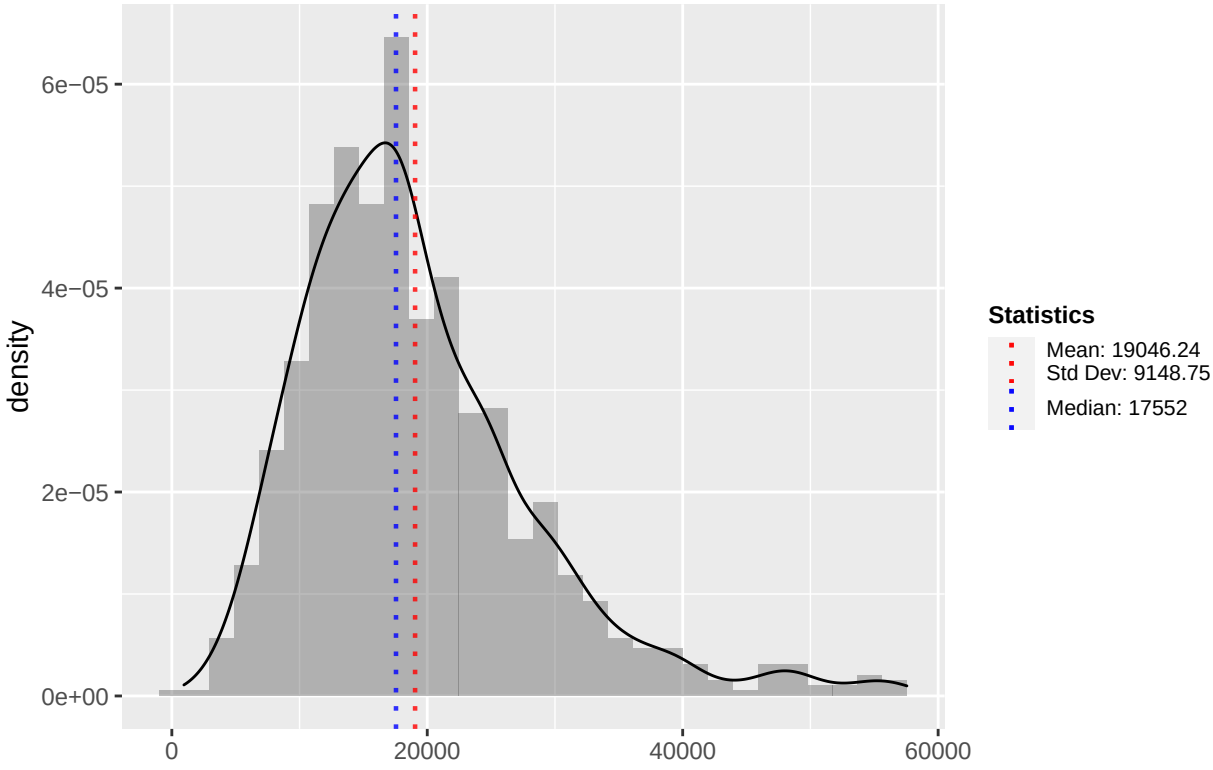
What is the "empirical rule"? Drawing from the empirical rule, what percentage of observations in the above distribution have values between 45 and 55? between 40 and 60? between 35 and 65?

- YOUR ANSWER HERE:

Use the function we created `plot_distribution()` to plot the distribution of the variable `tuitfee_grad_nres` from the data frame `df_ipeds_pop`

- Note: the data frame `df_ipeds_population` contains data on the entire population of research/master's universities, whereas the data frame `df_ipeds_sample` contains data on a random sample of universities from that population

```
plot_distribution(data_vec = df_ipeds_pop$tuitfee_grad_nres, plot_title = "")
```



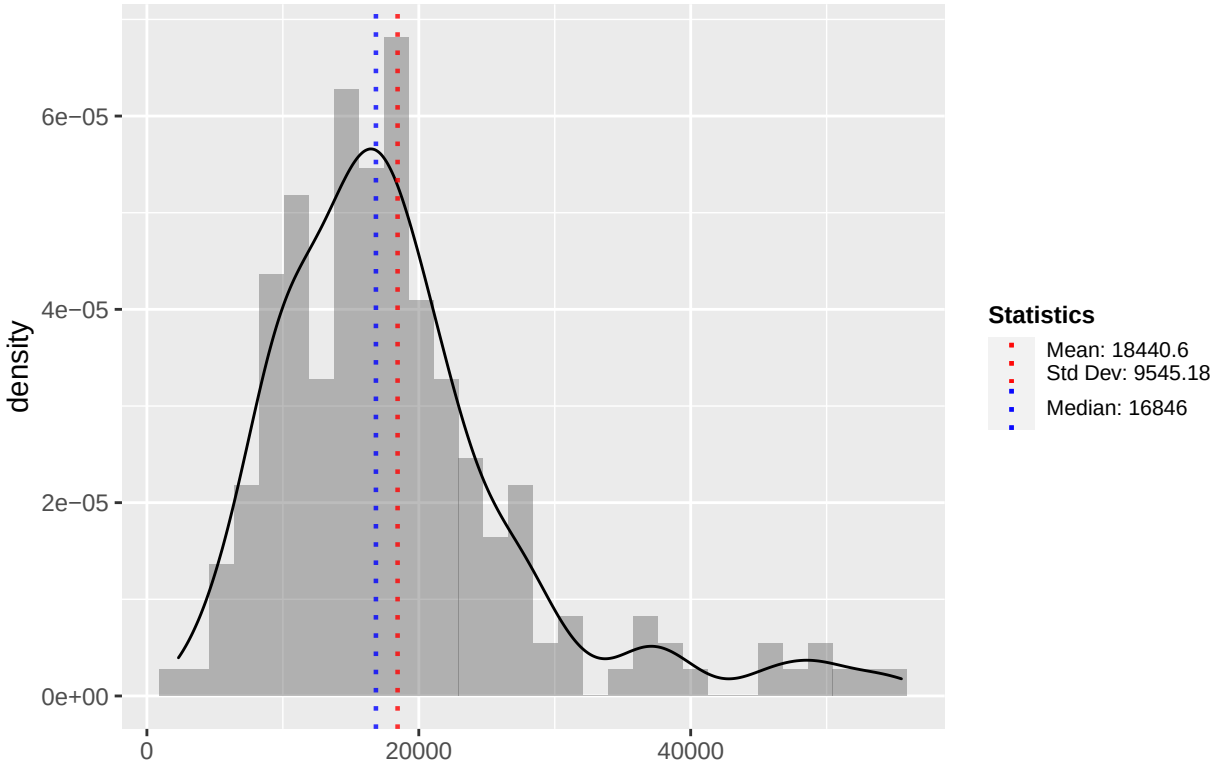
Does this variable appear to have a normal, left-skewed, or right-skewed distribution? why?

- YOUR ANSWER HERE:

Use the function we created `plot_distribution()` to plot the distribution of the variable `tuitfee_grad_nres` from the data frame `df_ipeds_sample`

- Note: the data frame `df_ipeds_population` contains data on the entire population of research/master's universities, whereas the data frame `df_ipeds_sample` contains data on a random sample of universities from that population

```
plot_distribution(data_vec = df_ipeds_sample$tuitfee_grad_nres, plot_title = "")
```



Does this variable appear to have a normal, left-skewed, or right-skewed distribution? why?

- YOUR ANSWER HERE:

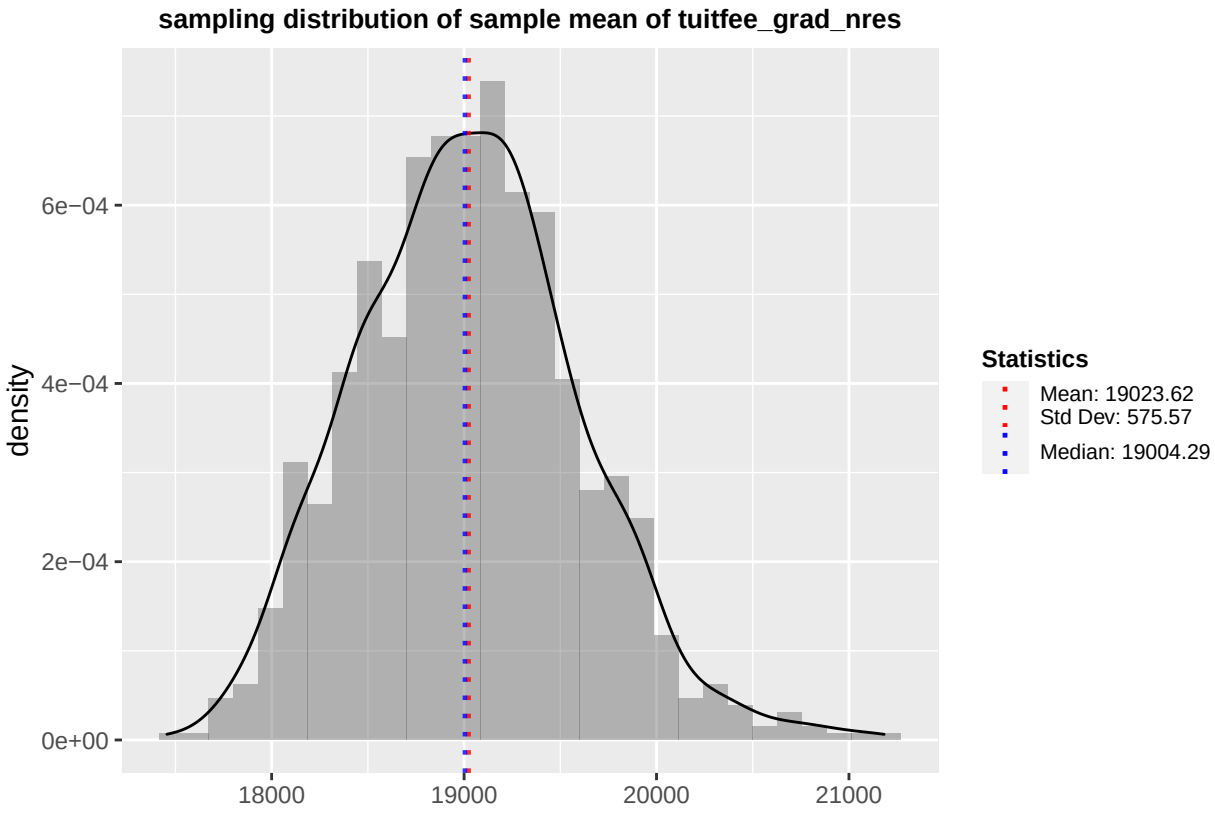
What is a sampling distribution? what is a sampling distribution of a sample mean?

- YOUR ANSWER HERE:

Run the following code, which does the following:

- takes 500 random samples of sample size n=200 from the data frame `df_ipeds_pop`
- for each random sample, calculates the sample mean of variable `tuitfee_grad_nres`
- plots the sampling distribution of the sample mean of variable `tuitfee_grad_nres`

```
set.seed(124)
get_sampling_distribution(data_vec = df_ipeds_pop$tuitfee_grad_nres, num_samples = 1000, sample_size = 1
  plot_distribution(plot_title = "sampling distribution of sample mean of tuitfee_grad_nres")
```




```
#same as above
#plot_distribution(get_sampling_distribution(data_vec = df_ipeds_pop$tuitfee_grad_nres, num_samples = 1
```

Answer the following questions with respect to the above plot (one sentence or less for each answer):

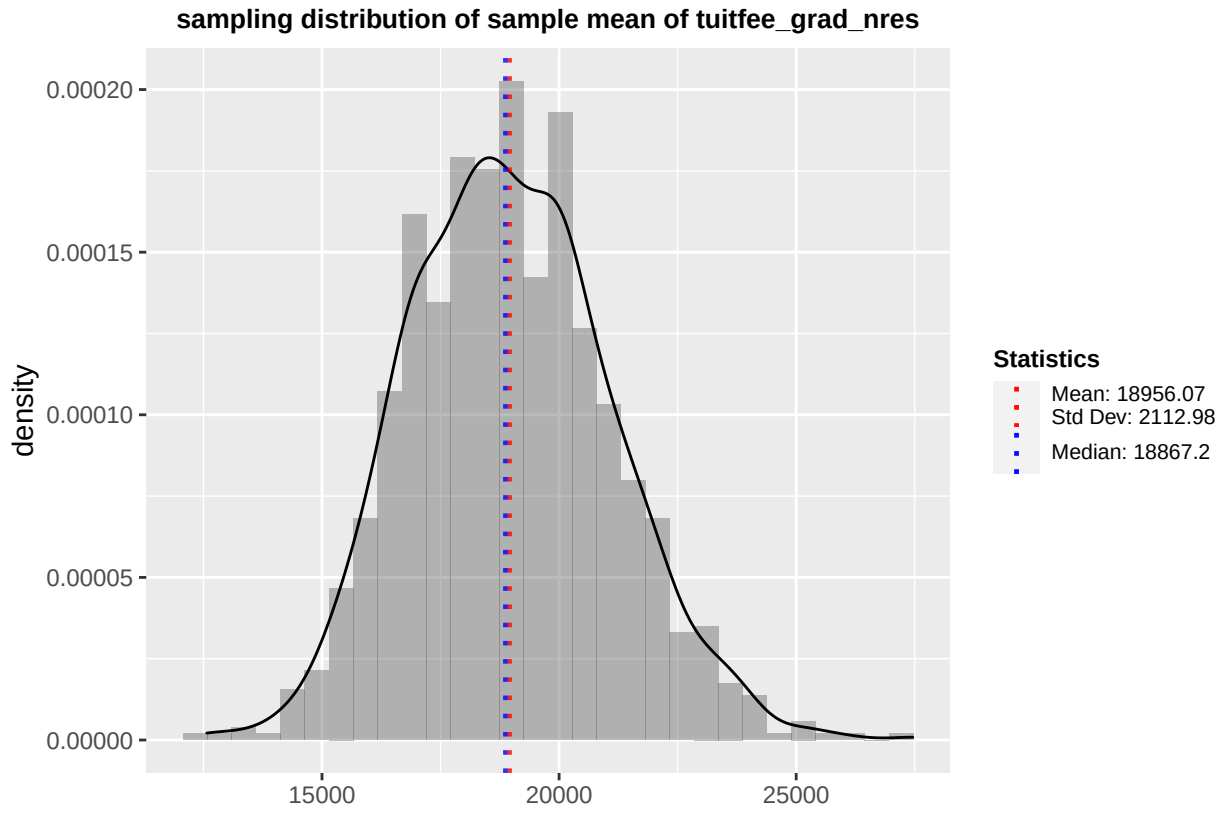
- what does each observation in the above plot represent?
  - ANSWER: a sample mean from one random sample
- would you describe the shape of the above distribtuion as (approximately) normal, left-skwed, or right-skewed?
  - ANSWER:
- Define what thethe concept "standard error" mean (referrring to sampling distribution of sample mean)?
  - ANSWER: standard error refers to a sampling distribution and is the average distance between a sample mean from one random sample and the mean of all sample means
- Why are the concepts "standard error" and "standard deviation of the sampling distribution" equivalent?
  - ANSWER: standard error is standard deviation where each observation is a sample mean as opposed to a single data point
- Interpret the value of standard error in the above plot in words
  - ANSWER: on average a sample mean from one random sample is about 581 away from the mean of all sample means

- Write the formula for sample standard error and state what each component of the formula refers to (e.g., `n` refers to sample size)
  - ANSWER:

Run the following code, which does the following:

- takes 500 random samples of sample size n=20 from the data frame `df_ipeds_pop`
- for each random sample, calculates the sample mean of variable `tuitfee_grad_nres`
- plots the sampling distribution of the sample mean of variable `tuitfee_grad_nres`

```
set.seed(124)
get_sampling_distribution(data_vec = df_ipeds_pop$tuitfee_grad_nres, num_samples = 1000, sample_size = 2
  plot_distribution(plot_title = "sampling distribution of sample mean of tuitfee_grad_nres")
```




```
#,plot_title = 'Sampling distribution')
```

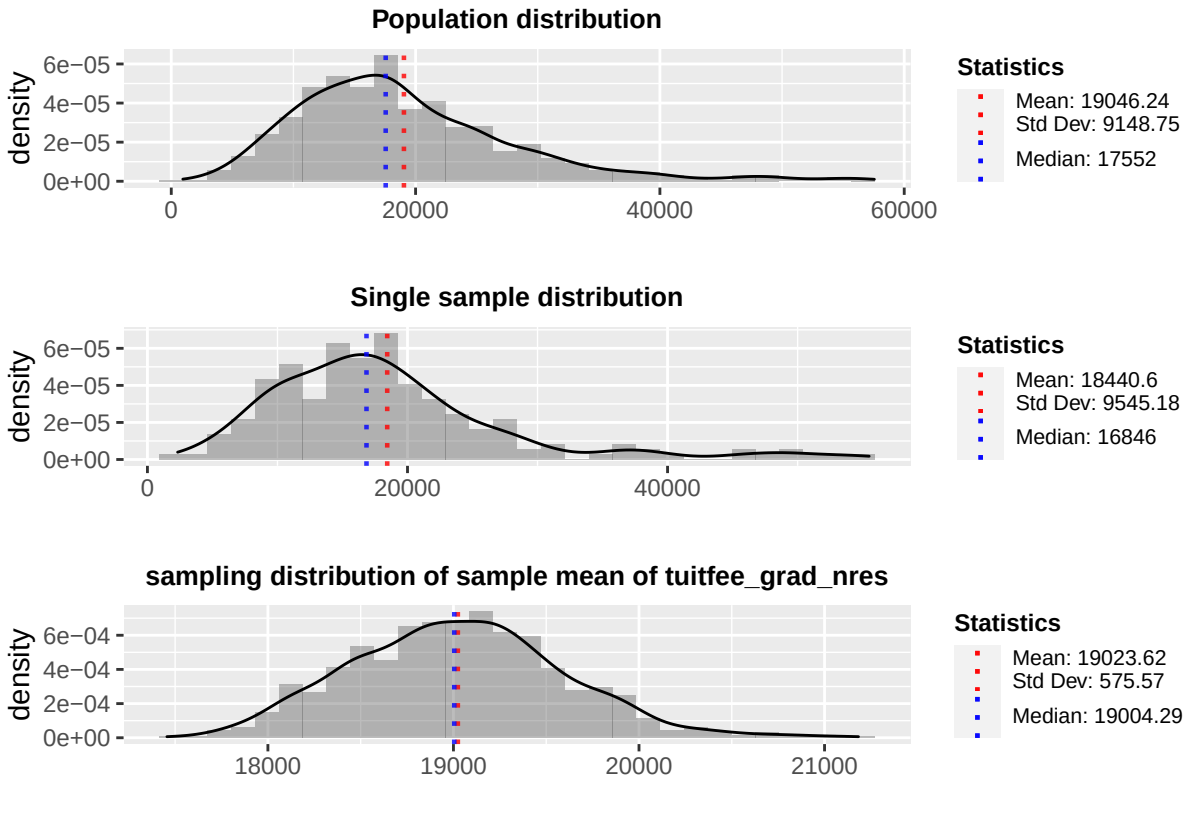
Answer the following questions with respect to the above plot (one sentence or less for each answer):

- Interpret the value of standard error in words
- why is the standard error from this sampling distribution (each sample has sample size n=20) larger than the sampling distribution from the previous example (each sample has sample size n=20)

Run the following code, which does the following:

- plots the population distribution of the variable `tuitfee_grad_nres`
- plots the distribution of the variable `tuitfee_grad_nres` from one sample
- plots the sampling distribution of the sample mean for the variable `tuitfee_grad_nres`

```
set.seed(124)
plot_distribution(df_ipeds_pop$tuitfee_grad_nres, plot_title = 'Population distribution') +
  plot_distribution(df_ipeds_sample$tuitfee_grad_nres, plot_title = 'Single sample distribution') +
  plot_distribution(get_sampling_distribution(data_vec = df_ipeds_pop$tuitfee_grad_nres, num_samples =
  plot_layout(ncol = 1)
```



State the central limit theorem in your own words and explain why it is important for hypothesis testing

- ANSWER:

Hypothesis testing

In this section we will be testing a hypothesis about the variable off-campus room and board (`roomboard_off`)

Here is how IPEDS defines concepts related to room and board and other expenses, frome the IPEDS "Student Charges for Full Academic Year" 2019-20 academic year data dictionary [LINK]:

- "Room charges"
  - The charges for an academic year for rooming accommodations for a typical student sharing a room with one other student.

- "Board charges"
  - The charge for an academic year for meals, for a specified number of meals per week.
- "Other expenses"
  - The amount of money (estimated by the financial aid office) needed by a student to cover expenses such as laundry, transportation, entertainment, and furnishings. (For the purpose of this survey room and board and tuition and fees are not included.)
- Note that most of these variables seem to be defined for an academic year rather than a 12-month calendar year

Here, We have included some code to help you get to know the data. Just run this code and take a look at the output

Print observations for UC campuses

```
df_ipeds_pop %>%
  # keep UC campuses
  filter(unitid %in% c(110398,110635,110644,110653,110662,110671,110680,110699,110705,110714,445188,110
  select(instnm,city,locale,roomboard_off,oth_expense_off) %>% as_factor()
#> # A tibble: 9 x 5
#>   instnm                    city       locale      roomboard_off oth_expense_off
#>   <chr>                     <chr>      <fct>               <dbl>           <dbl>
#> 1 University of California~ Berkeley   City: Mids~         14771            5359
#> 2 University of California~ Davis      Suburb: Sm~         10588            4856
#> 3 University of California~ Irvine     City: Large         12861            5184
#> 4 University of California~ Los Angel~ City: Large         14303            5126
#> 5 University of California~ Riverside  City: Large         10986            4792
#> 6 University of California~ La Jolla   City: Large         13681            4760
#> 7 University of California~ Santa Bar~ Suburb: Mi~         12818            6045
#> 8 University of California~ Santa Cruz City: Small         13216            5442
#> 9 University of California~ Merced     Rural: Fri~          8595            4909
```

The variable `locale` categorizes universities by city/suburb/town/rural and by city size

```
#df_ipeds_pop %>% count(locale)
df_ipeds_pop %>% count(locale) %>% as_factor()
#> # A tibble: 12 x 2
#>    locale              n
#>    <fct>           <int>
#>  1 City: Large       254
#>  2 City: Midsize     142
#>  3 City: Small       147
#>  4 Suburb: Large     199
#>  5 Suburb: Midsize    25
#>  6 Suburb: Small      27
#>  7 Town: Fringe       25
#>  8 Town: Distant      84
#>  9 Town: Remote       66
#> 10 Rural: Fringe      18
#> 11 Rural: Distant      8
#> 12 Rural: Remote       4
```

Average cost of off-campus room & board

```r
mean(df_ipeds_pop$roomboard_off, na.rm = TRUE)
#> [1] 10639.56

#alternative approach for calculating mean room and board
df_ipeds_pop %>% summarize(mean_roomboard_off = mean(roomboard_off, na.rm = TRUE))
#> # A tibble: 1 x 1
#>   mean_roomboard_off
#>                <dbl>
#> 1              10640.
```

Average cost of off-campus room & board, separately for each value of `locale`

```r
df_ipeds_pop %>% group_by(locale) %>%
  summarize(
    sample_size = n(),
    mean_roomboard_off = mean(roomboard_off, na.rm = TRUE)
    ) %>% as_factor()
#> # A tibble: 12 x 3
#>    locale          sample_size mean_roomboard_off
#>    <fct>                 <int>              <dbl>
#>  1 City: Large             254             11821.
#>  2 City: Midsize           142             10166.
#>  3 City: Small             147             10205.
#>  4 Suburb: Large           199             11123.
#>  5 Suburb: Midsize          25             11034.
#>  6 Suburb: Small            27             10597.
#>  7 Town: Fringe             25              9532.
#>  8 Town: Distant            84              8975.
#>  9 Town: Remote             66              9516.
#> 10 Rural: Fringe            18              9405.
#> 11 Rural: Distant            8             10308.
#> 12 Rural: Remote             4              8845
```

What are the five steps in hypothesis testing? for each step, provide a one-sentence description

- ANSWER

In the below questions, you will conduct hypothesis testing steps to answer the research question, "Is the population mean of off-campus room & board equal to $10,000?" You will be using the variable `roomboard_off` from the data frame `df_ipeds_sample`, which is a single random sample from the population data frame `df_ipeds_pop`. You will use a two-sided alternative hypothesis with an alpha level (rejection region) of `.05`

x. State the null and alternative (two-sided) hypothesis

- YOUR ANSWER HERE

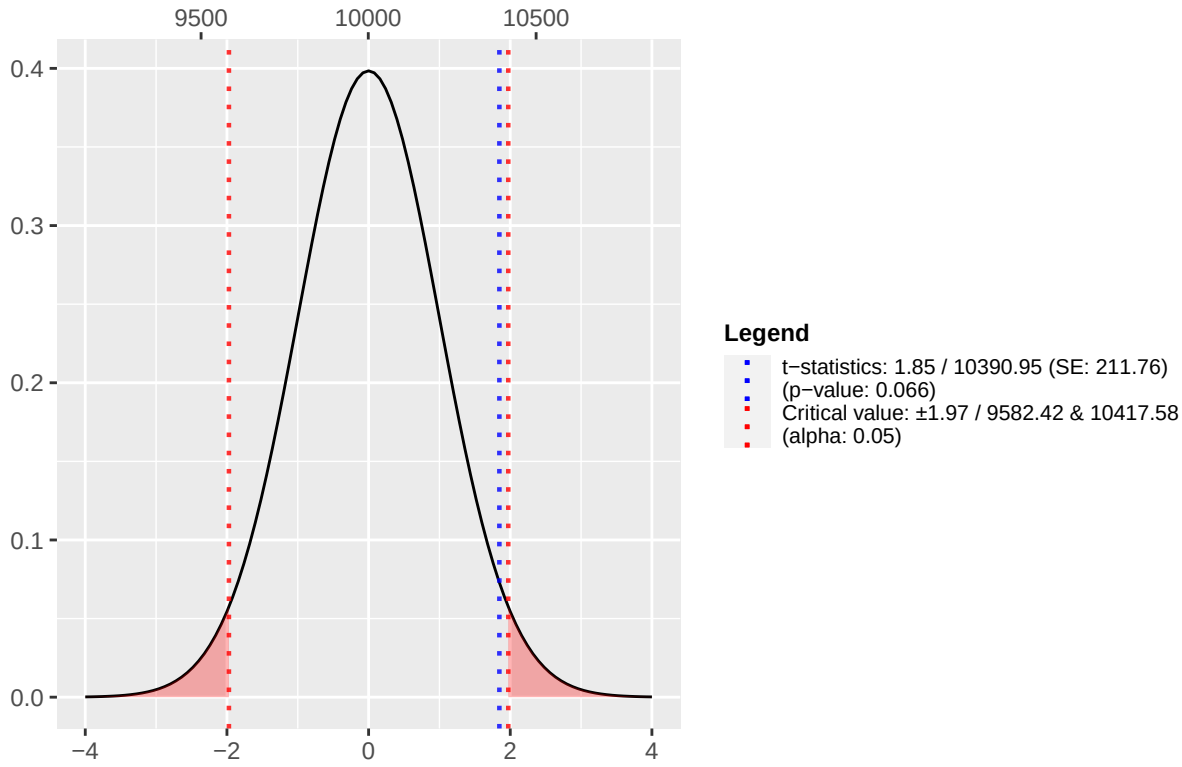
x. use the `t.test()` function to calculate the test statistic

```r
t.test(x = df_ipeds_sample$roomboard_off, mu = 10000)
#>
#>  One Sample t-test
```

```
#>
#> data:  df_ipeds_sample$roomboard_off
#> t = 1.8462, df = 199, p-value = 0.06635
#> alternative hypothesis: true mean is not equal to 10000
#> 95 percent confidence interval:
#>   9973.373 10808.527
#> sample estimates:
#> mean of x
#>  10390.95
```

use function `plot_t_distribution()` we created to plot the sampling distribution under the assumption that H_0 is true

```
plot_t_distribution(df_ipeds_sample$roomboard_off, mu = 10000)
```



X. Interpret the t-value in words and interpret the p-value in words

X. state the conclusion about your hypothesis test

x.

Post a comment/question

PATRICIA - HAVE STUDENTS POST A QUESTION ON PS1 SLACK CHANNEL; SOMETHING THEY LEARNED; OR THEY CAN RESPOND TO COMMENT/QUESTION FROM ANOTHER STUDENT

/2

- Go to the class repository and create a new issue.
- You can either:
  - Ask a question that you have about this problem set or the course in general. Make sure to assign the instructors (@ozanj, @mpatricia01, @cyouh95) and mention your team (e.g., @anyone-can-cook/your_team_name).
  - Share something you learned from this problem set or the course. Please mention your team (e.g., @anyone-can-cook/your_team_name).
- You are also required to respond to at least one issue posted by another student.
- Paste the url to your issue here:
- Paste the url to the issue you responded to here:

Knit to pdf and submit problem set

Knit to pdf by clicking the "Knit" button near the top of your RStudio window (icon with blue yarn ball) or drop down and select "Knit to PDF"

You will submit this problem set by pushing it to your repository. Follow the same steps you used above to add, commit, and push both the `.Rmd` and `.pdf` files.