Social Sciences Intro to Statistics

Week 5.1 Inferential Statistics, about a single variable

Week 5: Learning goal - Formulate hypothesis testing both by hand and with infer commands for a single population mean.

Introduction

Lecture overview:

- Inferential statistics, about single variable
- Hypothesis testing about a (single) population mean (by hand)

Load packages:

```
library(tidyverse)
library(ggplot2)
library(labelled)
library(patchwork)

# Load ipeds dataset from course website
load(url('https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/main/data/ipeds/outpu
```

```
#> Rows: 965
#> Columns: 38
#> $ instnm            <chr> "Alabama A & M University", "University of Alabama a~
#> $ unitid            <dbl> 100654, 100663, 100706, 100724, 100751, 100830, 1008~
#> $ opeid6            <chr> "001002", "001052", "001055", "001005", "001051", "0~
#> $ opeid             <chr> "00100200", "00105200", "00105500", "00100500", "001~
#> $ control           <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 2, ~
#> $ c15basic          <dbl+lbl> 18, 15, 16, 19, 16, 18, 16, 20, 18, 18, 19, 18, ~
#> $ stabbr            <chr+lbl> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", ~
```

```
#> $ city              <chr> "Normal", "Birmingham", "Huntsville", "Montgomery", ~
#> $ zip               <chr> "35762", "35294-0110", "35899", "36104-0271", "35487~
#> $ locale            <dbl+lbl> 12, 12, 12, 12, 13, 12, 13, 12, 23, 43, 21, 13, ~
#> $ region            <dbl+lbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
#> $ tuit_grad_res     <dbl> 10128, 8424, 10632, 7416, 11100, 7812, 10386, 15325,~
#> $ fee_grad_res      <dbl> 1414, 0, 1054, 2740, 690, 766, 1784, 900, 1000, 190,~
#> $ tuit_grad_nres    <dbl> 20160, 19962, 24430, 14832, 31460, 17550, 31158, 153~
#> $ fee_grad_nres     <dbl> 1414, 0, 1054, 2740, 690, 766, 1784, 900, 1000, 190,~
#> $ tuit_md_res       <dbl> NA, 31198, NA, NA, 31198, NA, NA, NA, NA, NA, NA, NA~
#> $ fee_md_res        <dbl> NA, 3464, NA, NA, 0, NA, NA, NA, NA, NA, NA, NA, NA,~
#> $ tuit_md_nres      <dbl> NA, 62714, NA, NA, 62714, NA, NA, NA, NA, NA, NA, NA~
#> $ fee_md_nres       <dbl> NA, 3464, NA, NA, 0, NA, NA, NA, NA, NA, NA, NA, NA,~
#> $ tuit_law_res      <dbl> NA, NA, NA, NA, 24080, NA, NA, 39000, NA, NA, NA, NA~
#> $ fee_law_res       <dbl> NA, NA, NA, NA, 300, NA, NA, 325, NA, NA, NA, NA, 65~
#> $ tuit_law_nres     <dbl> NA, NA, NA, NA, 44470, NA, NA, 39000, NA, NA, NA, NA~
#> $ fee_law_nres      <dbl> NA, NA, NA, NA, 300, NA, NA, 325, NA, NA, NA, NA, 65~
#> $ books_supplies    <dbl> 1600, 1200, 2416, 1600, 800, 1200, 1200, 1800, 998, ~
#> $ roomboard_off     <dbl> 9520, 14330, 11122, 7320, 14426, 10485, 14998, 8020,~
#> $ oth_expense_off   <dbl> 3090, 6007, 4462, 5130, 4858, 4030, 6028, 4600, 3318~
#> $ tuitfee_grad_res  <dbl> 11542, 8424, 11686, 10156, 11790, 8578, 12170, 16225~
#> $ tuitfee_grad_nres <dbl> 21574, 19962, 25484, 17572, 32150, 18316, 32942, 162~
#> $ tuitfee_md_res    <dbl> NA, 34662, NA, NA, 31198, NA, NA, NA, NA, NA, NA, NA~
#> $ tuitfee_md_nres   <dbl> NA, 66178, NA, NA, 62714, NA, NA, NA, NA, NA, NA, NA~
#> $ tuitfee_law_res   <dbl> NA, NA, NA, NA, 24380, NA, NA, 39325, NA, NA, NA, NA~
#> $ tuitfee_law_nres  <dbl> NA, NA, NA, NA, 44770, NA, NA, 39325, NA, NA, NA, NA~
#> $ coa_grad_res      <dbl> 25752, 29961, 29686, 24206, 31874, 24293, 34396, 306~
#> $ coa_grad_nres     <dbl> 35784, 41499, 43484, 31622, 52234, 34031, 55168, 306~
#> $ coa_md_res        <dbl> NA, 56199, NA, NA, 51282, NA, NA, NA, NA, NA, NA, NA~
#> $ coa_md_nres       <dbl> NA, 87715, NA, NA, 82798, NA, NA, NA, NA, NA, NA, NA~
#> $ coa_law_res       <dbl> NA, NA, NA, NA, 44464, NA, NA, 53745, NA, NA, NA, NA~
#> $ coa_law_nres      <dbl> NA, NA, NA, NA, 64854, NA, NA, 53745, NA, NA, NA, NA~
#> Rows: 200
#> Columns: 4
#> $ norm_dist    <dbl> 42.70513, 50.24400, 61.29008, 45.47494, 44.74406, 47.9912~
#> $ rskew_dist   <dbl> 0.34451771, 0.31359906, 0.09375337, 0.05581678, 0.0744584~
#> $ lskew_dist   <dbl> 0.6554823, 0.6864009, 0.9062466, 0.9441832, 0.9255415, 0.~
#> $ stdnorm_dist <dbl> -1.45897348, 0.04880097, 2.25801577, -0.90501164, -1.0511~
#> [1] 32528.35
#> [1] 31620.8
```

Fundamentals of inferential statistics

Inferential statistics is a method to use quantitative data to answer questions we may have about a population that we are observing. Fundamentals of inferential statistics include:

- Hypothesis testing: This is also known as the test of significance, used by testing claims about populations by weighing quantitative evidence for or against a conclusion.
- Confidence intervals: This method uses samples from targeted populations to estimate the true value of a parameter as a range of values.
- Regression Analysis: This method finds trends in data and predict future values of a time series.
- Correlation: This is a statistical test used to determine if there is a statistically significant relationship between two variables.
- T-test: This method compares the values of two groups in order to determine if there is a significant difference between two data sets from the same population.
- ANOVA: This test is used to make inferences about population means for any number of groups and independent variables.
- Chi-squared test: This test is used to help draw conclusions about a population based on a sample, such as whether two variables are related in the population.

As you continue on learning about statistics, you will encounter the rest of these methods. Today we are going to start with hypothesis testing about a single population mean.

Hypothesis testing about a (single) population mean

What and Why hypothesis testing

Quantitative research in social sciences often proceeds as follows:

- Develop a research question (which guides our research)
- Develop one (or more) testable hypothesis based on that research question
- Obtain data necessary to test the hypothesis
- Test the hypothesis by applying an appropriate statistical test to the data

Some examples of research questions co-authors and I have answered over the years:

- What is the relationship between state appropriations and nonresident enrollment at public universities [@RN3753]?

- What is the effect of nonresident enrollment growth on the number of resident students enrolled at public research universities [@RN4290]?
- What is the effect of participation in the Mexican American Studies program on the probability of high school graduation for students in the Tucscon Unified School District [@RN3292]?
- Are high schools with a higher percentage of white students more likely to receive recruiting visits from university admissions officer than high schools with a lower percentage of white students [@RN4450]?

For each of these journal articles, we answered the research question by developing a "testable hypothesis" and testing that hypothesis using some statistical test

Developing testable hypothesis is central to univariate statistical analysis (one variable), bivariate statistical analysis (two variables), and multivariate statistical analysis (3+ variables, usually a regression model)

Example hypotheses for univariate, bivariate, multivariate statistical analyses

- Univariate statistics (hypothesis tests about a single population mean)
  - Hypothesis: the average annual cost of attendance for graduate school (tuition + fees + living expenses) is $50,000
- Bivariate statistics [hypothesis tests about comparing two population means]
  - Hypothesis: the average annual cost of attendance for graduate school (tuition + fees + living expenses) at private universities is higher than public universities
  - Hypothesis: the average annual cost of attendance for graduate school (tuition + fees + living expenses) at universities in urban areas is higher than universities in suburban areas
- Multivariate statistics (usually a regression model with one dependent variable, one independent variable of interest, and one or more "control" variables)
  - where dependent variable (Y) = cost of attendance; independent variable of interest (X) = private or public university; control variable = level of urbanization
  - Hypothesis: cost of attendance for graduate school is higher at private universities than public universities, even after controlling for level of urbanization

Why learn how to do hypothesis testing about a single population mean when this class is supposed to be about regression (and hypothesis tests about regression models)?

- You must learn the general principles/concepts about using point-estimates from sample data to test hypotheses about population parameters
- The simplest practical application of these general principles/concepts is testing hypotheses about the value of a single population mean
- The concepts/steps for hypotheses tests about a single population mean are exactly the same as those for testing hypotheses about regression models

Overview of steps in hypothesis testing

These are the general steps in hypothesis testing:

1. **Hypothesis**
    - formally state your "null" and "alternative" hypothesis
2. **Assumptions**
    - state assumptions that are relied upon by the statistical test you are using to test your hypothesis
3. **Test statistic**
    - Using some appropriate statistical analysis, calculate the "test statistic" necessary to test your hypothesis
4. **p-value (means probability value)**
    - calculate the probability of observing a test statistic as large or larger as the one you calculated
5. **Alpha level/rejection region and conclusion**
    - decide on the "alpha level," the p-value associated with rejection of the null hypothesis
    - compare the p-value you you observed to the alpha level and make a conclusion about your hypothesis test

In real research projects, do researchers always follow these exact steps? In this exact order?

- Yes, they follow these steps
- But researchers do not necessarily follow steps in this exact order
    - e.g., usually, you would decide on an "alpha level" (rejection region) prior to conducting the statistical analysis
- Often, researchers will not write out each step as formally as we will ask you all to do.
    - We ask you to write out each step to give you practice. Later in the quarter, you won't have to write out each step

Example we will use to introduce steps in hypothesis testing

- The population mean cost of attendance (COA) for full-time (resident) graduate students, μ_Y, is `$28,000`

How we will teach you the steps in hypothesis testing in this lecture

- First, Introduce individual steps in detail, so that you develop a deep, conceptual understanding of each step
  - But when thinking about an individual step in detail, it can be hard to remember its relationship to other steps and to hypothesis testing as a whole
- Second, we will do another pratical example, where we work through all steps more quickly
  - so you can get a better sense of the hypothesis testing process as a whole and the relationships between steps

Hypotheses

This section presents a more formal introduction to hypotheses, focusing on univariate statistical analyses rather than bivariate or multivariate

Recall that the goal of inferential statistics is to make statements about a population of interest based on data from a representative sample from the population.

- We make a hypotheses about a population parameter (e.g., population mean of variable Y denotes μ_Y)
- Knowing the true value of the population parameter would require having data on all observations in the population
- Usually, we do not have data on the entire population
- We use sample data to test hypotheses about the population

Definition

- In statistics, a **hypothesis** is a declarative statement about a population.

In univariate statistical analyses, we make a hypothesis about one population paramaeter (e.g., population mean μ_Y) from one population of interest (e.g., all "research" universities and "master's" universities, as defined by the Carnegie Classification)

```
df_ipeds_pop %>% glimpse()
#> Rows: 965
#> Columns: 38
#> $ instnm            <chr> "Alabama A & M University", "University of Alabama a~
#> $ unitid            <dbl> 100654, 100663, 100706, 100724, 100751, 100830, 1008~
#> $ opeid6            <chr> "001002", "001052", "001055", "001005", "001051", "0~
#> $ opeid             <chr> "00100200", "00105200", "00105500", "00100500", "001~
#> $ control           <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 2, ~
#> $ c15basic          <dbl+lbl> 18, 15, 16, 19, 16, 18, 16, 20, 18, 18, 19, 18, ~
#> $ stabbr            <chr+lbl> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", ~
```

```
#> $ city              <chr> "Normal", "Birmingham", "Huntsville", "Montgomery", ~
#> $ zip               <chr> "35762", "35294-0110", "35899", "36104-0271", "35487~
#> $ locale            <dbl+lbl> 12, 12, 12, 12, 13, 12, 13, 12, 23, 43, 21, 13, ~
#> $ region            <dbl+lbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
#> $ tuit_grad_res     <dbl> 10128, 8424, 10632, 7416, 11100, 7812, 10386, 15325,~
#> $ fee_grad_res      <dbl> 1414, 0, 1054, 2740, 690, 766, 1784, 900, 1000, 190,~
#> $ tuit_grad_nres    <dbl> 20160, 19962, 24430, 14832, 31460, 17550, 31158, 153~
#> $ fee_grad_nres     <dbl> 1414, 0, 1054, 2740, 690, 766, 1784, 900, 1000, 190,~
#> $ tuit_md_res       <dbl> NA, 31198, NA, NA, 31198, NA, NA, NA, NA, NA, NA, NA~
#> $ fee_md_res        <dbl> NA, 3464, NA, NA, 0, NA, NA, NA, NA, NA, NA, NA, NA,~
#> $ tuit_md_nres      <dbl> NA, 62714, NA, NA, 62714, NA, NA, NA, NA, NA, NA, NA~
#> $ fee_md_nres       <dbl> NA, 3464, NA, NA, 0, NA, NA, NA, NA, NA, NA, NA, NA,~
#> $ tuit_law_res      <dbl> NA, NA, NA, NA, 24080, NA, NA, 39000, NA, NA, NA, NA~
#> $ fee_law_res       <dbl> NA, NA, NA, NA, 300, NA, NA, 325, NA, NA, NA, NA, 65~
#> $ tuit_law_nres     <dbl> NA, NA, NA, NA, 44470, NA, NA, 39000, NA, NA, NA, NA~
#> $ fee_law_nres      <dbl> NA, NA, NA, NA, 300, NA, NA, 325, NA, NA, NA, NA, 65~
#> $ books_supplies    <dbl> 1600, 1200, 2416, 1600, 800, 1200, 1200, 1800, 998, ~
#> $ roomboard_off     <dbl> 9520, 14330, 11122, 7320, 14426, 10485, 14998, 8020,~
#> $ oth_expense_off   <dbl> 3090, 6007, 4462, 5130, 4858, 4030, 6028, 4600, 3318~
#> $ tuitfee_grad_res  <dbl> 11542, 8424, 11686, 10156, 11790, 8578, 12170, 16225~
#> $ tuitfee_grad_nres <dbl> 21574, 19962, 25484, 17572, 32150, 18316, 32942, 162~
#> $ tuitfee_md_res    <dbl> NA, 34662, NA, NA, 31198, NA, NA, NA, NA, NA, NA, NA~
#> $ tuitfee_md_nres   <dbl> NA, 66178, NA, NA, 62714, NA, NA, NA, NA, NA, NA, NA~
#> $ tuitfee_law_res   <dbl> NA, NA, NA, NA, 24380, NA, NA, 39325, NA, NA, NA, NA~
#> $ tuitfee_law_nres  <dbl> NA, NA, NA, NA, 44770, NA, NA, 39325, NA, NA, NA, NA~
#> $ coa_grad_res      <dbl> 25752, 29961, 29686, 24206, 31874, 24293, 34396, 306~
#> $ coa_grad_nres     <dbl> 35784, 41499, 43484, 31622, 52234, 34031, 55168, 306~
#> $ coa_md_res        <dbl> NA, 56199, NA, NA, 51282, NA, NA, NA, NA, NA, NA, NA~
#> $ coa_md_nres       <dbl> NA, 87715, NA, NA, 82798, NA, NA, NA, NA, NA, NA, NA~
#> $ coa_law_res       <dbl> NA, NA, NA, NA, 44464, NA, NA, 53745, NA, NA, NA, NA~
#> $ coa_law_nres      <dbl> NA, NA, NA, NA, 64854, NA, NA, 53745, NA, NA, NA, NA~

mean(df_ipeds_pop$coa_grad_res, na.rm = TRUE)
#> [1] 32528.35
```

Null and Alternative hypothesis

When developing a hypothesis for quantitative research, we always specify a **null hypothesis** (H_0) AND an **alternative hypothesis** (H_a)

Null hypothesis (H_0)

- In univariate statistics, a null hypothesis (H_0) is a declarative statement that the population parameter has a specific value
- (in words) H_0 : the population mean cost of attendance for for full-time (resident) graduate students, μ_Y, is `$28,000`
- (using symbols) $H_0 : \mu_Y = \mu_{Y0} = \$28,000$
  - where μ_{Y0} refers to the parameter value associated with the null hypothesis
  - when testing a hypothesis about a single population mean, we can refer to μ_{Y0} as the "null population mean"

Alternative hypothesis (H_a)

- An alternative hypothesis (H_a) is a declarative statement that the population parameter falls in some alternative range of values as compared to the value declared by the null hypothesis
- There are two kinds of alternative hypotheses: two-sided; and one-sided
- for a given null hypothesis (H_0), there will always be one two-sided alternative hypothesis and two different one-sided hypotheses

Two-sided alternative hypothesis

- (in words) H_a : the population mean mean cost of attendance for for full-time (resident) graduate students, μ_Y, is not equal to `$28,000`
- (using symbols) $H_a : \mu_Y \neq \$28,000$

One-sided alternative hypothesis (mean is greater than $\$28,000$)

- (in words) H_a : the population mean mean cost of attendance for for full-time (resident) graduate students, μ_Y, is greater than `$28,000`
- (using symbols) $H_a : \mu_Y > \$28,000$

One-sided alternative hypothesis (mean is less than $\$28,000$)

- (in words) H_a : the population mean mean cost of attendance for for full-time (resident) graduate students, μ_Y, is less than `$28,000`
- (using symbols) $H_a : \mu_Y < \$28,000$

Example of null and alternative hypotheses for bivariate statistical analysis

Research question:

- Is the population mean annual cost of attendance for graduate school at public universities ($\mu_{Y_{pub}}$) different from the population mean annual cost of attendance for graduate school at private universities ($\mu_{Y_{priv}}$)?

Null and alternative hypotheses

- null hypothesis (H_0)
  - (in words): H_0 : the population mean annual cost of attendance for graduate school at public universities ($\mu_{Y_{pub}}$) is the same as the population mean annual cost of attendance for graduate school at private universities ($\mu_{Y_{priv}}$)
  - (symbols): $H_0 : \mu_{Y_{pub}} = \mu_{Y_{priv}}$
- Two-sided alternative hypothesis
  - (in words): H_a : the population mean annual cost of attendance for graduate school at public universities ($\mu_{Y_{pub}}$) is different than the population mean annual cost of attendance for graduate school at private universities ($\mu_{Y_{priv}}$)
  - (symbols): $H_a : \mu_{Y_{pub}} \neq \mu_{Y_{priv}}$
- One-sided alternative hypothesis ($pub < priv$)
  - (in words): H_a : the population mean annual cost of attendance for graduate school at public universities ($\mu_{Y_{pub}}$) is less than than the population mean annual cost of attendance for graduate school at private universities ($\mu_{Y_{priv}}$)
  - (symbols): $H_a : \mu_{Y_{pub}} < \mu_{Y_{priv}}$
  - note: this is the same as a one-sided hypothesis where we hypothesize $priv > pub$
- One-sided alternative hypothesis ($pub > priv $)
  - (in words): H_a : the population mean annual cost of attendance for graduate school at public universities ($\mu_{Y_{pub}}$) is greater than than the population mean annual cost of attendance for graduate school at private universities ($\mu_{Y_{priv}}$)
  - (symbols): $H_a : \mu_{Y_{pub}} > \mu_{Y_{priv}}$
  - note: this is the same as a one-sided hypothesis where we hypothesize $priv < pub$

Two-sided or one-sided alternative hypotheses?

In real research projects, we are not usually testing a hypothesis about a single population mean (univariate analysis). Rather, we are usually comparing population means of two different groups (bivariate analysis) or we are examining the relationship between an independent variable and the dependent variable after controlling for other variables (multivariate regression analysis)

Prior to conducting analyses, we usually have an expectation/suspicion about the result

- For most bivariate analyses, we usually suspect that one particular group is has a higher mean value than the other
  - e.g., we suspect that mean cost of attendance at private universities
  - this suggests a one-sided alternative hypothesis H_a

- For most multivariate analyses, we usually suspect the direction of the relationship between X and Y
  - e.g., we expect that "hours spent studying" (X) has a positive relationship with "grade point average" (Y) rather than thinking "the relationship between hours spent studying (X) and grad point average (Y) does not equal zero
  - this suggests a one-sided alternative hypothesis H_a

Should we specify two-sided or one-sided alternative hypothesis, H_a?

- For univariate and bivariate statistical analyses, researchers specify a two-sided alternative hypothesis more often than a one-sided alternative alternative hypothesis
  - often, researchers specify a two-sided alternative hypothesis even when they strongly believe one particular group has a larger population mean then the other
- For multivariate regression analyses, researchers **always** specify and test two-sided alternative hypotheses
  - even when they strongly believe the relationship between X and Y is positive; and even when they strongly believe the relationship between X and Y is positive
- Why this preference for two-sided alternative hypotheses in real research projects?
  - two-sided alternative hypotheses are more "conservative" than one-sided alternative hypotheses;
  - that is, if you specify a two-sided alternative hypothesis, H_a, and reject the null hypothesis, H_0, then it is necessarily true that we would have rejected then null hypothesis, H_0, had we specified a one-sided alternative hypothesis, H_a

Test statistic

restate null, H_0, and alternative (two-sided), H_a, hypothesis for our practical example

- H_0
  - (in words) H_0 : the population mean cost of attendance for for full-time (resident) graduate students, μ_Y, is \$28,000
  - (symbols) $H_0 : \mu_Y = \mu_{Y0} = \$28,000$
- H_a
  - (in words) H_a : the population mean mean cost of attendance for for full-time (resident) graduate students, μ_Y, is not equal to \$28,000
  - (symbols) $H_a : \mu_Y \neq \$28,000$

We must conduct a formal statistical test to decide whether we should reject the null hypothesis

- this is true for testing hypotheses about a population mean from a single population; testing hypotheses about whether two population means are equal; testing hypotheses about a regression coefficient, etc.
- **key to understanding hypothesis testing**: We conduct our test under the assumption that the null hypothesis, H_0, is true

Logic of the test statistic

- What the test statistic calculates
  - if the null hypothesis is true, how unlikely would it be to randomly draw the sample estimate (e.g., sample mean $\bar{Y}$) at least as far away from the null hypothesis value as the one we observed in our single random sample
  - e.g., "if the null hypothesis is true, there is a 1.5% chance of observing a sample mean at least as far away from the null hypothesis value ($\mu_{Y0} = \$28,000$) as the one we observed in our single random sample
- Logic of the test statistic
  - if – under the assumption that the null hypothesis is true – it would be very unlikely to observe the sample estimate we observed, then it is unlikely that the null hypothesis is true

General formula for test statistic (for pretty much any kind of hypothesis test):

$$test\_statistic = \frac{sample\_estimate - value\_associated\_with\_H_0}{sample\_standard\_error}$$

Formula for test statistic about a single population mean

- in words:
  - test-statistic t equals difference between the sample mean $\bar{Y}$ and the population mean associated with the null hypothesis μ_{Y0} divided by the sample standard error of the sample mean $\hat{\sigma}_{\bar{Y}}$
- equation:
  - $t = \frac{\bar{Y} - \mu_{Y0}}{\hat{\sigma}_{\bar{Y}}}$
- where:
  - $\hat{\sigma}_Y$ refers to sample standard deviation of variable Y
  - n refers to sample size
  - sample standard error of the sample mean $= \hat{\sigma}_{\bar{Y}} = \frac{\hat{\sigma}_Y}{\sqrt{n}}$

Calculating t-test statistic for our practical example

$H_0 : \mu_Y = \mu_{Y0} = \$28,000$; $H_a : \mu_Y \neq \$28,000$

Calculate components of t-test (using functions and by hand)

```
# sample size
  length(df_ipeds_sample$coa_grad_res) # assuming no missing observations
#> [1] 200
  df_ipeds_sample %>% summarize(n_non_miss = sum(!(is.na(coa_grad_res)))) # count only numbe
#> # A tibble: 1 x 1
#>   n_non_miss
#>        <int>
#> 1        200

# sample mean of coa_grad_res
  mean(df_ipeds_sample$coa_grad_res, na.rm = TRUE) # using function
#> [1] 31620.8

# sample standard deviation of coa_grad_res
  sd(df_ipeds_sample$coa_grad_res, na.rm = TRUE)
#> [1] 11298.37

# sample standard error of sample mean of coa_grad_res = std_dev/sqrt(n)
  sd(df_ipeds_sample$coa_grad_res, na.rm = TRUE)/sqrt(length(df_ipeds_sample$coa_grad_res))
#> [1] 798.9157
```

Components of t-test:

- sample size, $n = 200$
- sample mean, $\bar{Y} = 3.1620795 \times 10^4$
- Population mean associated with H_0, $\mu_{Y0} = \$28,000$
- sample standard deviation, $\hat{\sigma}_Y = 1.1298374 \times 10^4$
- sample standard error of the sample mean, $\hat{\sigma}_{\bar{Y}} = 798.9157$

Calculating t-test

$$t = \frac{\bar{Y} - \mu_{Y0}}{\hat{\sigma}_{\bar{Y}}} = \frac{30002.74 - 28000}{810.1332} = 2.4721$$

`t.test()` function

- what `t.test()` does
  - "Performs one and two sample t-tests on vectors of data"

  - we use one sample t-test to test hypothesis about single population mean
  - later, we will use two-samle t-test to test hypotheses about whether two population means are equal

- syntax:
  - `t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)`

- selected arguments
  - `x`: vector (variable) you want to calculate t-test for
  - `alternative`: whether you want two-sided or one-sided alternative hypothesis (default is `two.sided`)
  - `mu`: value associated with null hypothesis (default is `0`)

Calculating t-test value (using function and by hand)

```
# t-statistic = (sample_mean - mu_H_0)/(sample std err)

# using function
#?t.test # to see help file for function
t.test(x = df_ipeds_sample$coa_grad_res, mu = 28000)
#>
#>  One Sample t-test
#>
#> data:  df_ipeds_sample$coa_grad_res
#> t = 4.5321, df = 199, p-value = 1.006e-05
#> alternative hypothesis: true mean is not equal to 28000
#> 95 percent confidence interval:
#>  30045.37 33196.22
#> sample estimates:
#> mean of x
#>   31620.8

# by hand
(mean(df_ipeds_sample$coa_grad_res, na.rm = TRUE) - 28000)/(sd(df_ipeds_sample$coa_grad_res,
#> [1] 4.532137
```

Conceptual understanding of test statistic (MOST IMPORTANT!)

The test statistic refers to a **sampling distribution** not the distribution of your single sample

- in particular, the test statistic refers to the sampling distribution under the assumption that the null hypothesis is true, $H_0 : \mu_Y = \mu_{Y0}$

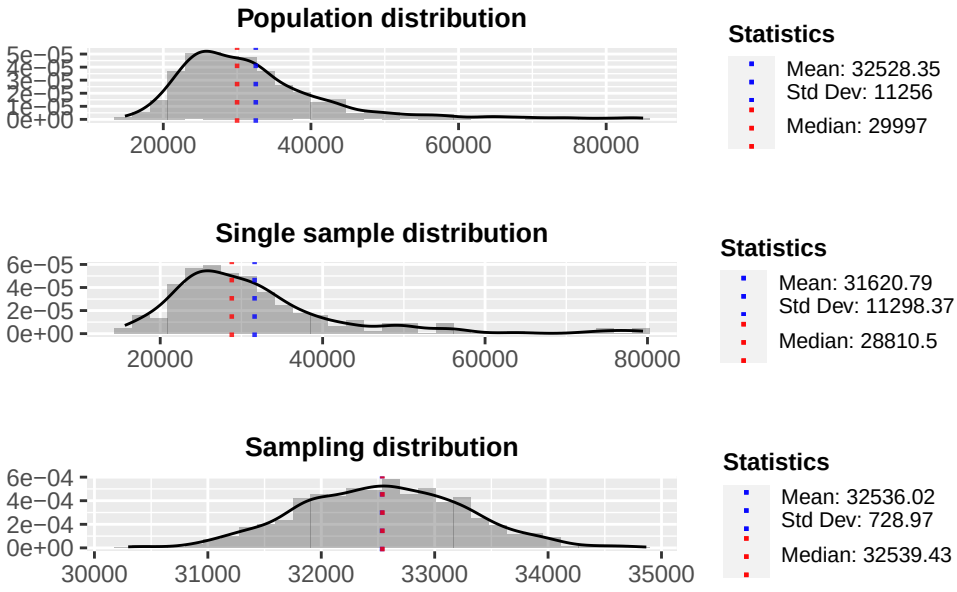
Recall core ideas of sampling distribution (we will refer to sampling distribution of sample mean, $\bar{Y}$)

- The sampling distribution of the sampling mean is a distribution where each observation is a sample mean, $\bar{Y}$, from one random sample taken from the population
- The sampling distribution shows how the value of the sample mean varies from sample to sample
- Standard error (i.e., the standard deviation of the sampling distribution), $\hat{\sigma}_{\bar{Y}}$ is the average distance between one random sample mean, $\bar{Y}$, and the mean of sample means $\bar{Y}_{\bar{Y}}$
  - (Note: we refer to sample standard error of the sample mean $\hat{\sigma}_{\bar{Y}} = \frac{\hat{\sigma}_Y}{\sqrt{n}}$, which can be calculated from sample data, rather than population standard error of the sample mean $\sigma_{\bar{Y}}$, which is a population parameter that is only known if we have data on the entire population)
- Drawing from the central limit theorem, we know that sampling distributions will always be normally distributed so long as sample size is not small
- Therefore, sampling distributions follow the empirical rule:
  - 68% of observations within one standard error
  - 95% of observations wtithin two standard erros
  - 99% of observations within three standard errors

Here we visually stack the following for the variable `coa_grad_res`:

- the population distribution; distribution from a single random sample; and sampling distribution of sample mean

```
plot_distribution(df_ipeds_pop$coa_grad_res, plot_title = 'Population distribution') +
  plot_distribution(df_ipeds_sample$coa_grad_res, plot_title = 'Single sample distribution')
  plot_distribution(get_sampling_distribution(df_ipeds_pop$coa_grad_res), plot_title = 'Sampl
  plot_layout(ncol = 1)
```



Usually we cannot know the sampling distribution because we do not have data on the entire population; we only have data on our single random sample

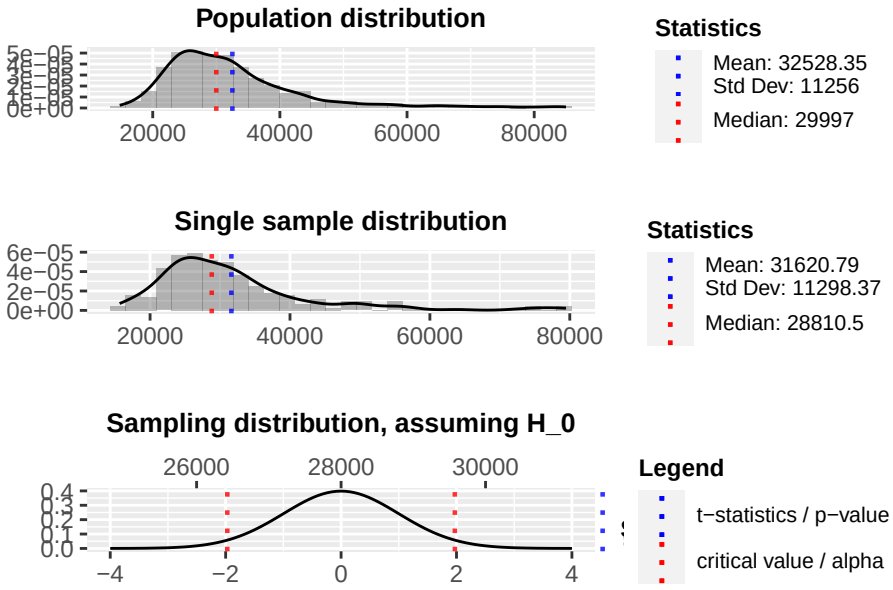
However, hypothesis testing is not based on the true sampling distribution of the sample mean. It is based on the sampling distribution under the assumption that the null hypothesis is correct

- Thanks to the central limit theorem, we have a pretty good idea of the sampling distribution assuming H_0 is true, even when we only have a single random sample!

Here we visually stack the following for the variable `coa_grad_res`:

- the population distribution; distribution from a single random sample; and sampling distribution of assuming that H_0 is true

```
plot_distribution(df_ipeds_pop$coa_grad_res, plot_title = 'Population distribution') +
  plot_distribution(df_ipeds_sample$coa_grad_res, plot_title = 'Single sample distribution')
  plot_t_distribution(df_ipeds_sample$coa_grad_res, mu = 28000,shade_rejection = F, shade_pva
  plot_layout(ncol = 1)
```



The t-test statistic is the distance between the hypothesized H_0 value and the observed sample estimate value $\bar{Y}$ scaled in terms of standard errors

- e.g., it would be unlikely to observe a t-value of greater than `2` or a t-value less than `-2` because we know (from empirical rule and central limit theorem) that 95% of observations fall within two standard deviations of the mean for a normally distributed variable

p-value

"p-value" refers to the probability-value associated with the t-value from your test statistic

Definition

- Under the assumption that H_0 is true, the **p-value** is the probability of observing a sample estimate (and its associated test-statistic) that is at least as far away from the null hypothesis value μ_{Y0} as the one we observed

A small p-value means that it would be unusual to find the sample estimate we observed if the null hypothesis H_0 is find the observed data if 0were true.

Calculating p-value For a two-sided alternative hypothesis ($H_a : \mu_Y \neq \mu_{Y0}$)

- let t be the value of your t-test
- let p by the p-value associated with t

- let $Pr(obs > t)$ is the probability of an observation having a higher value of t than the one you observed
- $p = Pr(obs > t) + Pr(obs < -t)$
- Because the sampling distribution is symmetric (because it is normally distributed):
  - $Pr(obs > t) = Pr(obs < -t)$
- therefore, for a two-sided alternative hypothesis
  - $p = 2 * Pr(obs > t)$

Let's calculate and visualize p-value for a couple different hypothesized values of the population mean μ_{Y0} for the variable `coa_grad_res` (full-time, resident grad school cost of attendance) from the data frame `df_ipeds_sample`

$H_0 : \mu_Y = \mu_{Y0} = \$29,000$ and $H_a : \mu_Y \neq \$29,000$

- Sample mean, $\bar{Y} = 3.1620795 \times 10^4$
- $t = 3.28$
- p-value $= Pr(obs > t) + Pr(obs < -t) = 0.001$
  - $Pr(obs > t) = 6 \times 10^{-4}$
  - $Pr(obs < -t) = 6 \times 10^{-4}$
- below code chunk runs t-test and plots t-value against sampling distribution assuming H_0 is true

```
mean(x = df_ipeds_sample$coa_grad_res)
#> [1] 31620.8
t.test(x = df_ipeds_sample$coa_grad_res, mu = 29000)
#>
#>  One Sample t-test
#>
#> data:  df_ipeds_sample$coa_grad_res
#> t = 3.2804, df = 199, p-value = 0.001224
#> alternative hypothesis: true mean is not equal to 29000
#> 95 percent confidence interval:
#>  30045.37 33196.22
#> sample estimates:
#> mean of x
#>   31620.8
plot_t_distribution(df_ipeds_sample$coa_grad_res, mu = 29000,shade_rejection = F, shade_pval
```