Social Sciences Intro to Statistics

Final: Due MONTH, DATE, YEAR at 11:59pm

Belle Lee

2024-06-15

Overview

You made it to the end! Congratulations on working hard the last ten weeks in this course. You're so close to the finish line. Similar to your previous psets, this final use the datasets we have been using in the course. However, unlike the psets, you will work on the final on your own.

- Note: Change the values of the YAML header above to your name and the date.

Question 1: Creating an R project

Create an R project

- Create a folder where you want to save files associated with the final. Let's call that folder "Final", but you can name it whatever you want.
  - For instance, it could be SSS » Final.
- In RStudio, click on "File" » "New Project" » "Existing Directory" » "Browse".
- Browse to find and select your Final folder.
- Click on "Create Project".
  - An R project file has the extension ".Rproj".
  - The name of the file should be "final.Rproj", or whatever you named the folder.

Save this final.Rmd file anywhere in the folder named final.

- At the top of this .qmd file, type in your first and last name in the appropriate place in the YAML header (e.g. "Belle Lee").
- in the date field of the YAML header, insert the date within quotations (any date format is fine).

- Now click the "Render" button near the top of your RStudio window (icon with blue arrow sign) or drop down "File" and select "Render Document".
  - Alternatively you can use the shortcut: **Cmd/Ctrl + Shift + k**.

Question 2: Definitions

Define the following terms: 1. Standard deviation: **ANSWER KEY:**

2. p-value: **ANSWER KEY:**
3. Mean: **ANSWER KEY:**
4. alpha-level: **ANSWER KEY:**
5. Standard error: **ANSWER KEY:**
6. What are the measures of central tendency? Provide an explanation of each. **ANSWER KEY:**
7. Explain what is the Central Limit Theorem and why is it important? **ANSWER KEY:**
8. Covariance: **ANSWER KEY:**
9. Correlation: **ANSWER KEY:**
10. What are the OLS (Ordinary Least Squares) Assumptions? **ANSWER KEY:**

Question 3: Conduct a two group hypothesis test

1. Load the necessary package(s): `tidyverse`, `ggplot2`, `labelled`, `patchwork`
2. Load the `ipeds` dataframe within this code chunk below.

```
# ANSWER KEY
library(tidyverse)
library(ggplot2)
library(labelled)
library(patchwork)

#load netflix data
load(url('https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/main/data/ipeds/outpu
```

3. Run the following code chunk (similar to what was provided in past lectures and psets to create ipeds data frame with fewer variables.

```
#> 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
```

4. We want to know from the `df_ipeds_pop` data frame if students who attend schools in New England area (region 1 in Bureau of Economic Analysis, or the New England Region) have higher books and supplies cost (`books_supplies`) than students who attend schools in the Far West area (region 8 of Bureau of Economic Analysis, or the Far West Region).

Please write out H_0 null hypothesis and H_a alternative hypothesis:

#ANSWER KEY

$H_0 : \mu_{region_1} = \mu_{region_8}$ - (in words): H_0 : the population mean books and supplies cost for students attending school in the New England area, μ_{region_1}, is the same as the population mean books and supplies cost for students attending school in the Far West area (μ_{region_8}).

Two-sided alternative hypothesis (H_a)

- $H_a : \mu_{region_1} \neq \mu_{region_8}$
- (in words): H_a : the population mean books and supplies cost for students attending school in the New England area, μ_{region_1}, is different than the population mean books and supplies cost for students attending school in the Far West area (μ_{region_8}).

5. First conduct the two sample t-test by hand

```r
# ANSWER KEY
# 1. Calculate the Means
region_1 <- df_ipeds_pop %>% filter(region == 1)
region_8 <- df_ipeds_pop %>% filter(region == 8)

mean_r1 <- mean(region_1$books_supplies)
mean_r8 <- mean(region_8$books_supplies)

# 2. Calculate the Variance
var_r1 <- var(region_1$books_supplies)
var_r8 <- var(region_8$books_supplies)

# 3. Calculate the Standard Errors
n_r1 <- length(region_1$books_supplies)
n_r8 <- length(region_8$books_supplies)
```

```r
SE <- sqrt(var_r1/n_r1 + var_r8/n_r8)
SE
#> [1] 39.53825

# 4. Calculate the t-statistic
t_stat <- (mean_r1 - mean_r8) / SE
t_stat
#> [1] 0.6191829

# 5. Calculate degrees of freedom
# This can be done using the formula for unequal variances (Welch's t-test)
df <- (var_r1/n_r1 + var_r8/n_r8)^2 /
      ((var_r1^2 / (n_r1^2 * (n_r1 - 1))) + (var_r8^2 / (n_r8^2 * (n_r8 - 1))))
df
#> [1] 158.0155

# 6. Calculate the p-value
# We can find the p-value of the calculated t-statistic by using the pt() function in R
p_value <- 2 * pt(-abs(t_stat), df)
p_value
#> [1] 0.5366875
```

6. Now, perform with the `t.test()` function in R

```r
# ANSWER KEY
# Perform the two-sample t-test
t_test_result <- t.test(region_1$books_supplies, region_8$books_supplies, var.equal = FALSE)

# Display the result
t_test_result
#>
#>  Welch Two Sample t-test
#>
#> data:  region_1$books_supplies and region_8$books_supplies
#> t = 0.61918, df = 158.02, p-value = 0.5367
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  -53.61022 102.57304
#> sample estimates:
#> mean of x mean of y
#>  1172.933  1148.452
```

7. What is the p-value? What can we say about it in relation to an alpha level of 0.05?

ANSWER KEY:

The p-value was `0.537` which is much greater than the typical alpha level or significance level of 0.05.

8. What can we conclude from our two sample t-test?

Since our p-value of `0.537` is greater than `0.05`, we fail to reject the null hypothesis. Which means, there is no significant different between the population mean books and supplies cost for students attending school in the New England area, μ_{region_1} and the population mean books and supplies cost for students attending school in the Far West area (μ_{region_8}).

Render to pdf and submit problem set

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

- Go to the [class website] (Need to fill in classwebsite) and under "Final", click on the "Final submission link"
- Submit both .qmd and pdf files
- Use this naming convention "lastname_firstname_ps#" for your .Rmd and pdf files (e.g. lee_belle_final.qmd & lee_belle_final.pdf)