Social Sciences Intro to Statistics

Week 6.2 Comparing Two Groups (Continued)

Week 6: Learning goal - Evaluate two groups with hypothesis testing and comparing population means.

Introduction

Lecture overview:

- Fundamental concepts in causal inference
- Hypothesis testing comparing population means of two groups

Load packages:

```
#> 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
#> Rows: 3,745
#> Columns: 11
#> $ id         <dbl+lbl>  943,  986, 1915, 2020, 2130, 3053, 3219, 3455, 3821, 4~
#> $ grade      <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
#> $ star       <dbl+lbl> 2, 2, 1, 2, 1, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 1, 1~
#> $ read       <int> 447, 450, 448, 447, 431, 451, 478, 455, 430, 474, 424, 439,~
#> $ gender     <dbl+lbl> 2, 2, 1, 1, 1, 2, 2, 2, 1, 1, 1, 1, 2, 2, 1, 1, 1, 2, 1~
#> $ ethnicity  <dbl+lbl> 1, 2, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1~
#> $ lunch      <dbl+lbl> 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 2, 1, 1~
#> $ school     <dbl+lbl> 3, 2, 3, 3, 3, 3, 3, 3, 3, 2, 3, 2, 2, 3, 3, 3, 4, 4, 3~
#> $ degree     <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 2, 1, 2, 1, 2, 2, 1, 2~
#> $ experience <int> 7, 21, 16, 5, 8, 3, 11, 10, 13, 0, 6, 13, 1, 8, 13, 13, 14,~
```

```
#> $ treatment  <dbl> 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0,~
```

Fundamental concepts in causal inference

Causal inference is the process of determining the independent, actual effect of a particular phenomenon that is a component of a larger system.

- More simply, the process of identifying the effect of an independent variable on an outcome of interest

Here's the difference between descriptive research questions and causal research questions:

Descriptive research questions

- Can investigate the magnitude of a problem (univariate):
  - What percentage of high school graduates attend college?
- Investigate correlational relationship between variables (sometimes called "associational" relationships):
  - Relationship between buying felt furniture pads and credit score?
  - Relationship between avg. income at a high school and the number of off-campus recruiting visits by universities?

Causal research questions

- Want to know the "causal effect" of independent variable (X) on outcome (Y); If you change value of X, causal effect is the change in Y due to the change in X
- Have the form "what is effect of X on Y?" Examples:
  - What is the effect of class size on math scores?
  - What is effect of grant aid on graduation?

We will use the following example research questions to explain causal inference concepts:

What is the effect of having a "small" class size ($T_i = 1$) vs. "large" class size ($T_i = 0$) on reading test score (Y) for elementary school students? - This research question is from the Tennessee Student Teacher Achievement Ratio (STAR) experiment - We have data from this experiment - LINK to information and data on the Tennessee STAR project

Hypothesis testing comparing population means of two groups

Investigate STAR data

We want to know if Kindergarten students randomly assigned to the "small" class size (variable `treatment` equals `1`) have higher values of reading test score than students randomly assigned to the "large" class size (variable `treatment` equals `0`)

So let's state H_0 null hypothesis and H_a alternative hypothesis:

null hypothesis (H_0)

- $H_0 : \mu_{Y_{treated}} = \mu_{Y_{control}}$
- (in words): H_0 : the population mean reading score for kindergarten students assigned to the treatment (small class size), $\mu_{Y_{treated}}$, is the same as the population mean reading score for kindergarten students assigned to the control (big class size) ($\mu_{Y_{control}}$)

Two-sided alternative hypothesis (H_a)

- $H_0 : \mu_{Y_{treated}} \neq \mu_{Y_{control}}$
- (in words): H_0 : the population mean reading score for kindergarten students assigned to the treatment (small class size), $\mu_{Y_{treated}}$, is different than the population mean reading score for kindergarten students assigned to the control (big class size) ($\mu_{Y_{control}}$)

Remember the general formula for (any) test statistic

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

Formula for test statistic about two population means

$t = \frac{(\bar{Y}_2 - \bar{Y}_1) - 0}{\hat{\sigma}_{\bar{Y}_2 - \bar{Y}_1}} = \frac{sample\_estimate - value\_associated\_with\_H_0}{sample\_standard\_error}$

- $\bar{Y}_2$ is sample mean of group 2
- $\bar{Y}_1$ is sample mean of group 1
- $\hat{\sigma}_{\bar{Y}_2 - \bar{Y}_1}$ is the sample standard error of $\bar{Y}_2 - \bar{Y}_1$
- *Note*: doesn't matter which group is group 1 and which is group 2
- $\hat{\sigma}_{\bar{Y}_2 - \bar{Y}_1} = \sqrt{\frac{\hat{\sigma}^2_{Y_1}}{n_1} + \frac{\hat{\sigma}^2_{Y_2}}{n_2}}$
  - n_1 is the sample size of group 1
  - n_2 is the sample size of group 2
  - $\hat{\sigma}^2_{Y_1}$ is the sample standard deviation of group 1, squared

- $\hat{\sigma}^2_{Y_2}$ is the sample standard deviation of group 2, squared

Calculating t-statistic for Tennessee STAR example, by hand Let's start first by calculating by hand to get more practice understanding the different steps to finding the t-statistic

```
df_stark %>% group_by(treatment) %>%
  summarize(
    n = n(),
    n_nonmiss = sum(!is.na(read)),
    mean = mean(read, na.rm = TRUE),
    sd = sd(read, na.rm = TRUE)
  )
#> # A tibble: 2 x 5
#>   treatment     n n_nonmiss  mean    sd
#>       <dbl> <int>     <int> <dbl> <dbl>
#> 1         0  2006      2006  435.  30.9
#> 2         1  1739      1739  441.  32.5

# point estimate - H_0_value
  (440.5474 - 434.7323) - 0
#> [1] 5.8151

# standard error
  sqrt(30.93590^2/2006 + 32.49738^2/1739)
#> [1] 1.041333

# t-statistic = (point estimate - H_0_value)/(standard error)
  (440.5474 - 434.7323)/sqrt(30.93590^2/2006 + 32.49738^2/1739)
#> [1] 5.584283
```

Interpreting values

- point estimate = `5.815`
  - the sample mean reading score for treated (small class) students is 5.815 larger than the sample mean reading score for untreated (large class) students
- standard error = `1.041`
  - on average the point estimate $\bar{Y}_{treated} - \bar{Y}_{control}$ a single random sample is 1.04 away from the mean of point estimates from an infinite number of random samples
- t-statistic = `5.584`

  - under the assumption that H_0 is true (i.e., $\mu_{Y_{treated}} - \mu_{Y_{control}} = 0$) the observed point estimate is 5.584 standard errors away from the value associated with H_0 (i.e., 0)

Calculating t-statistic for Tennessee STAR example using `t.test()` function

`t.test()` function, using the "formula" method

- syntax:
  - `t.test(formula, data, subset, mu=0, alternative = c("two.sided", "less", "greater"))`
- selected arguments
  - `formula`: follows form `formula = y_var ~ x_var`
    - where `y_var` is the outcome variable and `x_var` is the variable that defines the two groups you are comparing
  - `data`: data frame that contains variables named in `formula`
  - `alternative`: whether you want two-sided or one-sided alternative hypothesis (default is `two.sided`)
  - `subset`: subset of data to test (useful if `x_var` has more than two groups and you want to restrict test to those two groups)
  - `mu`: value associated with null hypothesis (default is `0`)

```
#?t.test
t.test(formula = read ~ treatment, mu = 0, data = df_stark)
#>
#>  Welch Two Sample t-test
#>
#> data:  read by treatment
#> t = -5.5843, df = 3610.1, p-value = 2.519e-08
#> alternative hypothesis: true difference in means between group 0 and group 1 is not equal
#> 95 percent confidence interval:
#>  -7.856798 -3.773477
#> sample estimates:
#> mean in group 0 mean in group 1
#>        434.7323        440.5474
```

p-value and Conclusion

Definition of p-value (same as for hypothesis test about a single population mean)

- Under the assumption if null hypothesis is true, the **p-value** is the probability of observing a point estimate as far away from the null hypothesis value as the one we observed

Above, we observed a p-value of less than `0.01`, indicating it would be very unlikely to observe the point estimate we observed ($\bar{Y}_{treated} - \bar{Y}_{control} = 5.82$) under the assumption that H_0 is true

Let's use an alpha level of `.05` to determine whether to reject H_0

Conclusion:

- The p-value of 0.00 is less than the alpha level of `.05`. Therefore, we reject H_0 and accept H_a
- In words, the population mean reading score for kindergarten students assigned to the treatment (small class size), $\mu_{Y_{treated}}$, is different than the population mean reading score for kindergarten students assigned to the control (big class size) ($\mu_{Y_{control}}$)
- Furthermore, we can say that the population mean reading score for students in the treatment is larger than the population mean reading score for students in the control group.