Social Statistics

Differences In Means

October 31, 2023

Assignment 4 Review - Question 1

Does the mean number of days of poor mental health differ from 3 at the 99% confidence level?

```r
t.test(gss_subset$mntlhlth, mu = 3, conf.level = .99)
```

```
	One Sample t-test

data:  gss_subset$mntlhlth
t = 3.0631, df = 1407, p-value = 0.002232
alternative hypothesis: true mean is not equal to 3
99 percent confidence interval:
 3.088947 4.037473
sample estimates:
mean of x 
  3.56321 
```

Assignment 4 Review - Question 1

- Reject the null hypothesis
  - → test statistic of 3.0631 is more extreme than 2.58
  - → p-value is less than .01
  - → 99 percent confidence interval does not include null hypothesis value

Assignment 4 Review - Question 2

2. Among respondents reporting any days of poor mental health, does the mean number of days differ from 8.5 at the 95% confidence level?

```r
t.test(gss_subset$mntlhlth[gss_subset$mntlhlth>0], mu = 8.5)
```

```
    One Sample t-test

data:  gss_subset$mntlhlth[gss_subset$mntlhlth > 0]
t = -1.2017, df = 619, p-value = 0.23
alternative hypothesis: true mean is not equal to 8.5
95 percent confidence interval:
 7.425064 8.758806
sample estimates:
mean of x
 8.091935
```

Assignment 4 Review - Question 2

- Do not reject the null hypothesis
  - → test statistic of -1.2017 is not more extreme than -1.96
  - → p-value is not less than .05
  - → 95 percent confidence interval includes the null hypothesis value

Assignment 4 Review - Question 3

3. Consider respondents who have not entered or completed college. Among these respondents, does the proportion with any mental health days differ from .40 at the 99% confidence level?

```
# Start with binary variable
# identifying respondents with any mental health days

gss_subset <- gss_subset |>
  mutate(mntlhlth_any = ifelse(mntlhlth > 0, 1, 0))
```

Assignment 4 Review - Question 3

```
# Don't need binary variable for education
# But having one might be faster and cleaner

gss_subset <- gss_subset |>
mutate(college = ifelse(degree == "Bachelor" |
                                 degree == "Graduate", "4-Year Degree",
                        ifelse(degree == "Junior College", "2-Year Degree",
                               "None"))) |>
mutate(college = factor(college,
                        levels = c("None",
                                   "2-Year Degree",
                                   "4-Year Degree")))
```

Assignment 4 Review - Question 3

```r
# Find frequencies

addmargins(table(gss_subset$college,
                 gss_subset$mntlhlth_any))
```

```
                  0    1  Sum
None            451  340  791
2-Year Degree    67   59  126
4-Year Degree   270  221  491
Sum             788  620 1408
```

We need the number with a “None” for degree and a 1 for any mental health days. And we need the total number with a “None” for degree.

Assignment 4 Review - Question 3

```r
prop.test(340, 791, p = .4, conf.level = .99)
```

```
	1-sample proportions test with continuity correction

data:  340 out of 791, null probability 0.4
X-squared = 2.8108, df = 1, p-value = 0.09363
alternative hypothesis: true p is not equal to 0.4
99 percent confidence interval:
 0.3846459 0.4762089
sample estimates:
        p 
0.4298357 
```

Assignment 4 Review - Question 3

- Do not reject the null hypothesis
  - → p-value is not less than .01
  - → 99 percent confidence interval includes null hypothesis value
  - → Don't look at the X-squared test statistics for prop.test()

Assignment 4 Review - Question 4

4. Consider respondents who have completed college or more. Among these respondents, does the proportion with any mental health days differ from .40 at the 95% confidence level?

```
# Wording is a little confusing.
# Could or could not include respondents with 2-year degree.
# We will include them here...

addmargins(table(gss_subset$college, gss_subset$mntlhlth_any))
```

```
                 0   1  Sum
None           451 340  791
2-Year Degree   67  59  126
4-Year Degree  270 221  491
Sum            788 620 1408
```

Assignment 4 Review - Question 4

```r
prop.test(59 + 221, 126 + 491, p = .4)
```

```

    1-sample proportions test with continuity correction

data:  59 + 221 out of 126 + 491, null probability 0.4
X-squared = 7.221, df = 1, p-value = 0.007205
alternative hypothesis: true p is not equal to 0.4
95 percent confidence interval:
 0.4141319 0.4940689
sample estimates:
        p
0.4538088
```

Assignment 4 Review - Question 4

- Reject the null hypothesis
  - → p-value is less than .05
  - → 95 percent confidence interval does not include the null hypothesis value
  - → Don't look at the X-squared test statistics for prop.test()

Comparing Samples

- Assignment questions measured the difference between an estimated mean and a null hypothesis value in terms of standard errors
- This week, we will measure the difference between two estimated means, and then measure that distance from a null hypothesis value in terms of standard errors

Comparing Samples

- Basics are the same: we need means, standard errors, and null hypotheses but we estimate them slightly differently
- Assumptions are also the same
  - → Distribution of differences between means is normally distributed
  - → For large sample sizes, t-distribution still approximates z-distribution

Comparing Samples

- Significance tests define groups (not datasets) as *samples*
- Samples are *independent* if the observations are random
  - → Coin flips are independent of each other.
  - → Across years, cross-sectional surveys (like GSS) are independent
- Samples are *dependent* if observations are matched
  - → Can be the same observations in a long-term panel (PSID, NLSY, etc.) or multiple measures in a short-term study (scores from two exams)
  - → Can be different observations if respondents' answers could be correlated (partners, siblings, etc.)

CIs for Comparing Means

- From CI for mean to CI for difference in means
- CI formula for difference in means is similar to what we used for means: $\mathbf{CI} = (\bar{y}_2 - \bar{y}_1) \pm t(se)$
- SE is still the first step, but now want SE *of the difference*: $se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- When coding, easier to replace numerators with variances: $se = \sqrt{\frac{var_1}{n_1} + \frac{var_2}{n_2}}$

CIs for Comparing Means

CIs for Comparing Means

- Example using `gss_week8` data. We want to only keep observations with non-missing values for the `memnum` variable:

```
gss_week8 <- gss_week8 |>
  filter(!is.na(memnum))
```

- We want to compare mean memberships across two degree categories. Options?

CIs for Comparing Means

- Could use binary variables or indexing. For this example, create a binary variable called `college` where everyone with at least a college degree gets a 1 and everyone else gets a 0. Try using `str_detect()` here.

```
gss_week8 <- gss_week8 |>
  mutate(college = ifelse(str_detect(degree, "Degree"), 1, 0))
```

CIs for Comparing Means

- For the standard error formula, we'll need the number of respondents in each category of `college`:

```r
table(gss_week8$college)
```

```
   0    1 
1035  430 
```

CIs for Comparing Means

- We also need the mean number of memberships for college degree holders and non college degree holders. Options here?

```
mean(gss_week8$memnum[gss_week8$college==0])
```

```
[1] 1.218357
```

```
mean(gss_week8$memnum[gss_week8$college==1])
```

```
[1] 2.551163
```

CIs for Comparing Means

This also works:

```r
gss_week8 |>
  group_by(college) |>
  summarise(mean_memnum = mean(memnum, na.rm = TRUE))
```

```
# A tibble: 2 × 2
  college mean_memnum
    <dbl>       <dbl>
1       0        1.22
2       1        2.55
```

CIs for Comparing Means

- We want to know if the difference between these two means is significant
- In the language of hypothesis testing, we want to know if we can reject the null hypothesis that the true difference between these two sample means is zero
- Start with the difference:

```
diff <- 2.551163 - 1.218357
diff
```

```
[1] 1.332806
```

CIs for Comparing Means

- Then find the standard error of the difference:

$$se = \sqrt{\frac{var_{memnum,college=0}}{n_{college=0}} + \frac{var_{memnum,college=1}}{n_{college=1}}}$$

```
diffse_college0 <- var(gss_week8$memnum[gss_week8$college==0]) /
                                1035

diffse_college1 <- var(gss_week8$memnum[gss_week8$college==1]) /
                                430

diffse <- sqrt(diffse_college0 + diffse_college1)

diffse
```

```
[1] 0.1147038
```

CIs for Comparing Means

- Construct the 95% confidence interval for the difference
- Starting value is the difference, rest of the formula is the same:
  - → difference ± 1.96*(standard error of difference)

```
diff_ll95 <- diff - 1.96*diffse
diff_ul95 <- diff + 1.96*diffse
diff_ci95 <- c(diff_ll95, diff, diff_ul95)

diff_ci95
```

```
[1] 1.107987 1.332806 1.557625
```

CIs for Comparing Means

- In sampling distribution, 95% of the time the difference in mean memberships between those with college degrees and those without will fall between 1.108 and 1.558
- We can be 95% confident that the difference in the population will fall within the range
- We calculated difference as `memnum[college==1] - memnum[college==0]` so positive value tells us the mean is higher for college degree holders

Significance of Differences in Means

- At 95% confidence level, can we say that the mean memberships differs between these two groups?

$$t = \frac{\text{observed} - \text{expected}}{\text{se}}$$

- Observed = Difference; Expected = Null Hypothesis Value, Standard Error = standard error *of the observed difference*
- With means, usually $H_0 : \mu = 0$ and $H_A : \mu \neq 0$

Significance of Differences in Means

```
# Test Statistic:
((2.551163 - 1.218357) - 0) / diffse
```

```
[1] 11.61955
```

- Test statistic tells us observed difference is 11.62 standard errors away from null hypothesis' expected difference of 0
- 11.62 > 1.96, so we can reject the null hypothesis

Comparing Means - Shortcut!

- Place the two means you want to compare in the `t.test()` function:

```r
t.test(gss_week8$memnum[gss_week8$college==1],
            gss_week8$memnum[gss_week8$college==0])
```

```
	Welch Two Sample t-test

data:  gss_week8$memnum[gss_week8$college == 1] and gss_week8$memnum[gss_week8$college
== 0]
t = 11.62, df = 639.01, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.107563 1.558047
sample estimates:
mean of x mean of y 
 2.551163  1.218357 
```

Comparing Means - Shortcut

Comparing Means - Exercise

- Is the difference in mean memberships between those in the “Some College” degree category and those in the “HS Diploma” degree category significant at the .01 alpha level?

```r
t.test(gss_week8$memnum[gss_week8$degree == "Some College"],
       gss_week8$memnum[gss_week8$degree == "HS Diploma"],
       conf.level = .99)
```

```
    Welch Two Sample t-test

data:  gss_week8$memnum[gss_week8$degree == "Some College"] and
gss_week8$memnum[gss_week8$degree == "HS Diploma"]
t = 2.2388, df = 137.46, p-value = 0.02678
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -0.06942738  0.90219305
sample estimates:
mean of x mean of y
 1.687500  1.271117
```

Comparing Means - Exercise

```
        Welch Two Sample t-test

data:  subset$memnum[subset$degree == "Some College"] and subset$memnum[subset$degree == "HS Diploma"]
t = 2.2388, df = 137.46, p-value = 0.02678
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -0.06942738  0.90219305
sample estimates:
mean of x mean of y
 1.687500  1.271117
```

- Cannot reject null hypothesis
  - → Test statistic is less extreme than 2.58 (red box)
  - → p-value greater than .01 (blue box)
  - → 99% confidence interval includes null hypothesis value of zero (green box)

Group Exercises

- Some of the differences we have been waiting to test!
  - → Age at first birth (`agekdbrn`) by race (`racehisp`)
  - → Age (`age`) by self employment status (`wrkslf`)
  - → Number of political actions (`polactions`) by sex (`sex`)
  - → Number of political actions (`polactions`) by class (`class`)
- Give a sociological hypothesis for why you expect the difference you are testing will or will not be statistically significant
- Test the significance of the difference in means between two categories at the 95% confidence level