Problem set 7: Inference

Due November 13, 2023, at 9pm

(Your name here)

Note: Start with the file `ps7_2023_inference.qmd` (available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F23/tree/main/assignments). Modify that file to include your answers. Make sure you can "render" the file (e.g. in RStudio by clicking on the `Render` button). Submit both the qmd file and the PDF via Canvas.

For questions 1 and 2, you will use the `oregon_subset.csv` dataset (available under `data` on the course github).

This is a subset of the data from the Oregon Health Insurance Experiment (OHIE). Here is a description of the data from the replication archive link:

> In 2008, a group of uninsured low-income adults in Oregon was selected by lottery to be given the chance to apply for Medicaid. This lottery provides an opportunity to gauge the effects of expanding access to public health insurance on the health care use, financial strain, and health of low-income adults using a randomized controlled design.

Medicaid is a state-administered health insurance program that is available to adults and children with limited economic resources. At the time of the experiment, Oregonians could enroll in Medicaid through two programs. OHP Plus served "categorically eligible" people, which includes children, pregnant women, and disabled people in low-income families. OHP Standard offered Medicaid to low-income adults not eligible for OHP Plus. Due to budget shortfalls OHP Standard was closed to new enrollment in 2004. In 2008 the state opened OHP Standard to new enrollment, and due to surplus demand it decided to allocate spots in the program using a lottery (i.e. random assignment).

The variables you will use are:

- `lottery_winner`: 1 if participant received the opportunity to enroll in Medicaid, otherwise 0
- `enrolled`: 1 if participant enrolled in Medicaid at some point between joining the study and the follow-up survey collecting post-treatment outcomes

- `borrow`: response on the follow-up survey to “In the last 12 months, have you had to borrow money, skip paying other bills, or pay other bills late in order to pay health care bills?” – 1 if yes, 0 if no

Question 1: Inference for a difference in sample means

For this question, your estimand is the effect of winning the OHIE lottery (and thus being eligible to enroll in Medicaid) on the probability of needing to borrow money to pay medical bills.

Your (plug-in) estimator of this effect is $\overline{Y}_1 - \overline{Y}_0$, where $\overline{Y}_1$ is the proportion of lottery winners who needed to borrow (i.e. the sample mean of `borrow` among those with `lottery_winner == 1`) and $\overline{Y}_0$ is the proportion of lottery losers who needed to borrow (i.e. the sample mean of `borrow` among those with `lottery_winner == 0`).

(1a) Report the value of this estimator in the dataset. Interpret this number, i.e. explain in a sentence what it means about the relationship between Medicaid eligibility and borrowing.

Answer:

First we load the dataset:

```
datt <- read.csv("./../data/oregon_subset.csv")
```

The estimate is:

```
(itt_y <- mean(datt$borrow[datt$lottery_winner == 1], na.rm = T) -
   mean(datt$borrow[datt$lottery_winner == 0], na.rm = T))
```

```
[1] -0.03814682
```

The probability of needing to borrow to cover medical expenses was .038 lower among lottery winners than among non-winners.

(1b) Suppose that the data we have is an iid sample from a larger population of Oregonians who participated in the lottery. The following is a derivation of the sampling variance of your estimator. Explain each step.

Letting n_1 be the number of lottery winners and n_0 be the number of lottery losers,

$$
\begin{aligned}
\mathrm{V}\left[\overline{Y}_1 - \overline{Y}_0\right] &= \mathrm{V}\left[\overline{Y}_1\right] + \mathrm{V}\left[\overline{Y}_0\right] - 2\mathrm{Cov}[\overline{Y}_1, \overline{Y}_0] && \text{(Step 1)}\\
&= \mathrm{V}\left[\overline{Y}_1\right] + \mathrm{V}\left[\overline{Y}_0\right] && \text{(Step 2)}\\
&= \frac{\mathrm{V}\left[Y_1\right]}{n_1} + \frac{\mathrm{V}\left[Y_0\right]}{n_0} && \text{(Step 3)}
\end{aligned}
$$

Answer

- Step 1 uses the "Variance Rule" (Theorem 2.2.3 in A&M) [Note there was a mistake on the problem set – it should be a minus sign instead of a plus sign before the covariance term]
- In Step 2 we eliminate the covariance term because the sample mean among lottery winners is independent of the sample mean among lottery losers across samples
- In Step 3 we apply the variance of the sample mean (Theorem 3.2.4)

(1c) Using the result from (1b) and a normal approximation, report the 95% confidence interval for your estimate.

Answer:

```
sampling_var_1 <- var(datt$borrow[datt$lottery_winner == 1], na.rm = T)/
  sum(datt$lottery_winner == 1 & !is.na(datt$borrow), na.rm = T)
sampling_var_0 <- var(datt$borrow[datt$lottery_winner == 0], na.rm = T)/
  sum(datt$lottery_winner == 0 & !is.na(datt$borrow), na.rm = T)

combined_standard_error <- sqrt(sampling_var_1 + sampling_var_0)

itt_y + 1.96*c(-1, 1)*combined_standard_error
```

```
[1] -0.05299780 -0.02329584
```

(1d) Using the result from (1b) and a normal approximation, report a two-sided p-value, where the null hypothesis is that the effect of eligibility is zero.

Answer:

```
t_stat <- abs(itt_y)/combined_standard_error
```

```
2*(1 - pnorm(t_stat))
```

```
[1] 4.790139e-07
```

Question 2: Bootstrap inference for the Wald estimator

For this question, we focus on a different estimand. Define *compliers* as subjects who enroll in Medicaid if and only if they gain eligibility through the lottery. Your estimand is the effect of enrolling in Medicaid on needing to borrow money to pay medical bills, specifically for compliers. This is known as the Local Average Treatment Effect or the Complier Average Treatment Effect; you will learn about it in a Causal Inference course.

The estimand can be written as

$$\tau_{\text{LATE}} = \mathrm{E}\left[Y_1 - Y_0 \mid D_1 - D_0 = 1\right]$$

where

- Y_1 is the value of `borrow` if one wins eligibility in the lottery,
- Y_0 is the value of `borrow` if one does not win eligibility in the lottery,
- D_1 is the value of `enrolled` if one wins eligibility in the lottery,
- D_0 is the value of `enrolled` if one does not win eligibility in the lottery.

The Wald estimator is

$$\hat{\tau}_{\text{Wald}} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{D}_1 - \overline{D}_0},$$

where

- $\overline{Y}_1$ is the average value of `borrow` among lottery winners in the sample,
- $\overline{Y}_0$ is the average value of `borrow` among lottery non-winners in the sample,
- $\overline{D}_1$ is the average value of `enrolled` among lottery winners in the sample,
- $\overline{D}_1$ is the average value of `enrolled` among lottery non-winners in the sample.

Under assumptions that you will learn about in Causal Inference, $\mathrm{E}\left[\hat{\tau}_{\text{Wald}}\right] = \tau_{\text{LATE}}$.

It is possible to derive an estimator for the sampling variance of the Wald estimator, but it's not easy. We'll use the bootstrap instead.

(2a) Report the value of the Wald estimator in the dataset. Interpret this number, i.e. explain in a sentence what it means about the relationship between Medicaid enrollment and borrowing.

Answer:

```r
wald_numerator <- mean(datt$borrow[datt$lottery_winner == 1], na.rm = T) -
  mean(datt$borrow[datt$lottery_winner == 0], na.rm = T)
wald_denominator <- mean(datt$enrolled[datt$lottery_winner == 1], na.rm = T) -
  mean(datt$enrolled[datt$lottery_winner == 0], na.rm = T)

(wald_estimate <- wald_numerator/wald_denominator)
```

```
[1] -0.1590592
```

For compliers (i.e. those who enroll in Medicaid if and only if they win eligibility via the lottery), enrolling in Medicaid reduces the probability of having to borrow money to pay medical bills by about .16.

(2b) Use the bootstrap ("naive bootstrap") to estimate the standard error of the Wald estimator and use that standard error (and a normal approximation) to report the 95% confidence interval for τ_{LATE}.

Answer:

We generate 500 bootstrap resamples:

```r
m <- 500
walds <- rep(NA, m)
# make a smaller version of data to speed up analysis a bit
dattX <- datt[, c("lottery_winner", "enrolled", "borrow")]
for(i in 1:m){
  dattY <- dattX[sample(1:nrow(dattX),
                        size = nrow(dattX), replace = T), ]
  wald_numerator <- mean(dattY$borrow[dattY$lottery_winner == 1], na.rm = T) -
    mean(dattY$borrow[dattY$lottery_winner == 0], na.rm = T)
  wald_denominator <- mean(dattY$enrolled[dattY$lottery_winner == 1], na.rm = T) -
    mean(dattY$enrolled[dattY$lottery_winner == 0], na.rm = T)
  walds[i] <- wald_numerator/wald_denominator
}
```

```r
wald_std_error <- sd(walds)

(conf_int <- wald_estimate + 1.96*c(-1, 1)*wald_std_error)
```

```
[1] -0.22105743 -0.09706094
```

The confidence interval is approximately $[-.22, -.10]$.

(2c) Using your standard error from (2b) and a normal approximation, report a two-sided p-value, where the null hypothesis is that the average effect of medicaid enrollment on compliers is zero.

Answer:

```r
t_stat <- abs(wald_estimate)/wald_std_error
2*(1 - pnorm(t_stat))
```

```
[1] 4.94421e-07
```

This is a very small p-value, which casts significant doubt on the null hypothesis.

Question 3: Randomization inference

Taylor Swift, the musician, is reportedly dating an American football player named Travis Kelce. Swift has attended some of Kelce's games this season. Recently a broadcaster observed that Kelce produced better results when Swift was in attendance, referring to a "Taylor Swift effect".

The dataset `kelce.csv` (available on the course github under `data`) contains the following variables about the first seven games of the season:

- `yards`: receiving yards by Kelce, a measure of his performance (higher is better)
- `swift`: 1 if Swift attended the game, 0 otherwise
- `home`: 1 if it was a home game, 0 otherwise

(3a) Report and store the difference in Kelce's average receiving yards when Swift attended vs did not attend.

Answer:

First we load the data:

```
dat <- read.csv("./../data/kelce.csv")
```

Here is the difference in Kelce's yards with and without Taylor Swift in attendance:

```
(actual_diff <- mean(dat$yards[dat$swift == 1]) -
   mean(dat$yards[dat$swift == 0]))
```

```
[1] 57.66667
```

(3b) Using randomization inference, simulate the randomization distribution of the "Taylor Swift effect" under the sharp null hypothesis that Kelce's yardage in each game did not depend at all on Swift's attendance, i.e. that Kelce's yardage in each game that Swift **did not** attend would have been the same if she **had attended**, and Kelce's yardage in each game that Swift **did attend** would have been the same if she **had not attended**. That is, you should generate the distribution of the estimated effect of Swift's attendance (under the sharp null) across possible combinations of four games that Swift could have chosen to attend. Plot the histogram of this randomization distribution and include a vertical line for the observed value.

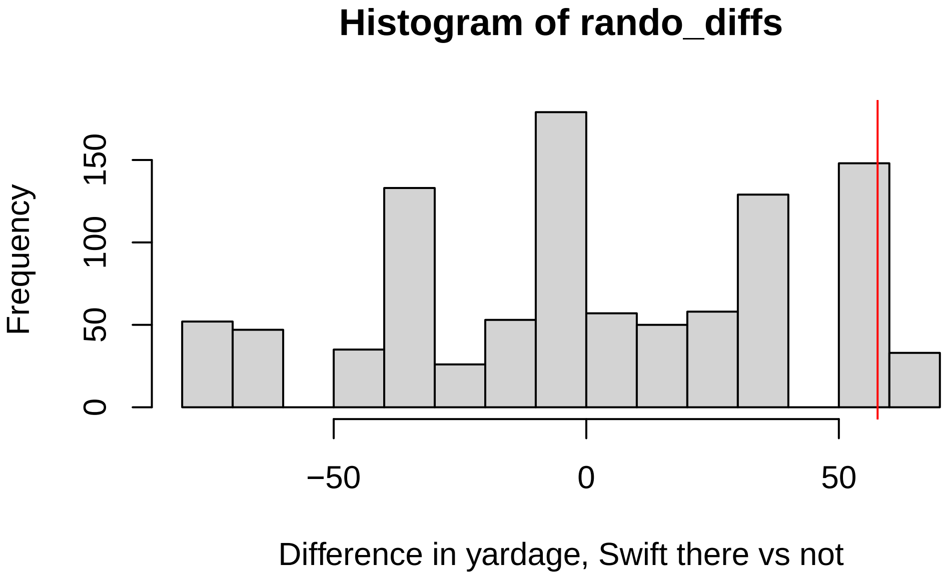
Answer:

This code runs the randomization inference:

```
m <- 1000
rando_diffs <- rep(NA, m)
for(i in 1:m){
  dat$fake_swift <- sample(dat$swift)
  rando_diffs[i] <- mean(dat$yards[dat$fake_swift == 1]) -
    mean(dat$yards[dat$fake_swift == 0])
}
```

And here is the requested plot:

```r
hist(rando_diffs, xlab = "Difference in yardage, Swift there vs not")
abline(v = actual_diff, col = "red")
```



(3c) Again using randomization inference (not a normal approximation), compute the upper one-tailed p-value of the observed "Taylor Swift effect", i.e. the proportion of randomizations producing a yardage difference bigger than the one we observed.

Answer

```r
mean(rando_diffs > actual_diff)
```

```
[1] 0.033
```

(3d) Again using randomization inference (not a normal approximation), compute the two-tailed p-value of the observed "Taylor Swift effect".

Answer

```r
mean(abs(rando_diffs) > actual_diff)
```

```
[1] 0.132
```

(3e) **Bonus**: Is this convincing evidence that Taylor Swift's presence affects Kelce's performance? Discuss.

Answer: The low *p*-values indicate that we would be unlikely to see such a large difference in performance between games Swift attended and those she didn't attend if her presence had no effect *and* she chose which games to attend randomly. But Swift did not choose to attend games randomly, and it's possible that the factors she used to determine which games to attend also affect Kelce's performance. The `home` variable in the dataset indicates that Swift mostly attended home games. One possibility is that Kelce does better in home games (e.g. because he is energized by the crowd or doesn't like staying in hotels), in which case Swift's attendance may not be the cause of the difference we observe.

In short, the evidence for a "Taylor Swift effect" would be stronger if Swift's attendance at Kelce's games had been randomly determined.