Announcements

Make sure to sign in on the google form (I send a list of which section questions are useful for which pset questions afterwards)

Pset 11 due Friday 4/28

Causal inference intricacies

Mark whether the following statements are true or false. Explain your answers.

1. $$E(Y|W=0) = E(Y(0)|W=0)$$

True: $E(Y|W=0) = E(Y(1)W + Y(0)(1-W)|W=0) = E(Y(0)|W=0)$.

2. $$E(Y(0)|W=0) = E(Y(0))$$

False: by LOTP,

$$E(Y(0)) = E(Y(0)|W=0)P(W=0) + E(Y(0)|W=1)P(W=1)$$

so if $E(Y(0)|W=1) \neq E(Y(0)|W=0)$, the original statement will not be true. For example, this could happen if only people who were going to recover from an illness anyway were given a particular treatment.

3. In an RCT,
$$E(Y(0)|W=0) = E(Y(0))$$

True: in an RCT, W is independent of $(Y(0), Y(1))$, so the conditioning adds nothing. With the equations above, $E(Y(0)|W=1) = E(Y(0)|W=0)$ must be true because $Y(0)$ is independent of W.

4. In an RCT, $Y \perp\!\!\!\perp W$.

False: $Y = Y(0)(1-W) + Y(1)W$, and it could be the case that everyone without the treatment ($W=0$) stays ill while everyone with the treatment ($W=1$) recovers, so W would perfectly explain Y. This extreme explanatory power holds whenever $Y(0)$ and $Y(1)$ are constants.

5. $Y = Y(0) + W\tau$ where $\tau = Y(1) - Y(0)$.

True:

$$Y = Y(0)(1-W) + Y(1)W = Y(0) + W(Y(1) - Y(0))$$

6. $$E(Y|W=w) = E(Y(0)) + wE(\tau)$$

False: from above,

$$\begin{aligned} E(Y|W=w) &= E(Y(0) + W\tau|W=w) \\ &= E(Y(0)|W=w) + wE(\tau|W=w) \\ &= E(Y(0)|W=w) + wE(Y(1) - Y(0)|W=w) \\ &= (1-w)E(Y(0)|W=w) + wE(Y(1)|W=w) \\ &\neq (1-w)E(Y(0)) + wE(Y(1)) \\ &= E(Y(0)) + wE(\tau) \end{aligned}$$

where the unequality comes from the same idea as in 2 that $E(Y(0)|W=1)$ is not necessarily equal to $E(Y(0)|W=0)$ if the treatments are not independent of the potential outcomes.

7. In an RCT,
$$E(Y|W=w) = E(Y(0)) + wE(\tau)$$

True:

$$\begin{aligned} E(Y|W=w) &= E(Y(0)+W\tau|W=w) \\ &= E(Y(0)|W=w) + wE(Y(1)|W=w) - wE(Y(0)|W=w) \\ &= E(Y(0)) + wE(Y(1)) - wE(Y(0)) \\ &= E(Y(0)) + wE(\tau) \end{aligned}$$

8. $E(\tau|W=1) = \frac{E(W\tau)}{E(W)}$

True:

$$E(W\tau) = E(WY(1)) - E(WY(0)) = E(Y(1)|W=1)P(W=1) - E(Y(0)|W=1)P(W=1)$$

Noting that $P(W=1) = E(W)$, we divide across to get

$$\frac{E(W\tau)}{E(W)} = E(Y(1)|W=1) - E(Y(0)|W=1) = E(\tau|W=1)$$

9. Assuming unconfoundedness with another random variable X (conditional on X, $(Y(0), Y(1)) \perp\!\!\!\perp W$),

$$E(WY) = E(E(W|X)Y(1))$$

True:

$$E(WY) = E(E(WY(1)|X)) = E(E(W|X)E(Y(1)|X)) = E(E(E(W|X)Y(1)|X)) = E(E(W|X)Y(1))$$

where the first and last equalities used Adam's law, the second used unconfoundedness, and the third used the fact that $E(W|X)$ is a function of X to "put back what's known." Note that $E(W|X)$ is also called the propensity score if W can only be 0 or 1.

10. Not necessarily assuming unconfoundedness, with $E(\tau|X=x) > 0$ for all x, it is possible to have $E(Y(1)) < E(Y(0))$.

False: by Adam's law, $E(\tau) = E(E(\tau|X))$, and $E(\tau|X) > 0$, so the LOTUS integral will be strictly positive and $E(E(\tau|X)) > 0$ as well. Thus, $E(\tau) > 0 \implies E(Y(1)) > E(Y(0))$, so the statement above is not possible.

Free Distribution or Cost-Sharing?

This set of questions will be looking at the 2010 paper "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." Before this paper, many development economists argued that cost-sharing, charging a much-reduced but non-zero price for healthcare resources, was necessary to avoid wasting the resources on people who did not need them. In this paper, Jessica Cohen and Pascaline Dupas claim this is not the case. Instead, with a randomized control trial for insecticide treated net (ITN) distribution, they show there is no evidence that cost-sharing reduces wastage, but cost-sharing does significantly decrease demand for ITNs.

The study involved randomizing the cost of ITNs at rural Kenyan health clinics for pregnant women and (1) tracking ITN sales and (2) following up with women to see whether they were using the nets. Originally, four prices were used ($0, $0.15, $0.30, and $0.60), which represent 100% to 90% subsidies from the original price of the ITN. For simplicity, we will be grouping these into $0 and non-$0 groups. The original data is available here.

1. For this first part, for woman i, let W_i be 0 if the woman received a free ITN and 1 if the woman purchased a net. Let Y_i be the indicator of whether the net is hanging when the researchers visit the woman. Suppose we use a finite sample model, so our treatment effects are $\tau_i = y_i(1) - y_i(0)$ and we condition on $y_i(1)$ and $y_i(0)$. The parameter of interest is $\bar{\tau}$, the average treatment effect. Our method of moments estimator is:
$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i W_i}{E(W_i)} - \frac{Y_i(1-W_i)}{E(1-W_i)}$$
where W_i (and therefore Y_i) is the source of randomness. Show that this can be rewritten in terms of n_0 (the number of women who received a free net), n_1 (the number of women who purchased a net), S_0 (the number of women who used a free net), and S_1 (the number of women who used a purchased net).

By symmetry, since each woman is equally likely to receive a net, $E(1-W_i) = n_0/n$. Applying the same logic with $E(W_i)$,
$$\begin{aligned}
\hat{\tau} &= \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i W_i}{E(W_i)} - \frac{Y_i(1-W_i)}{E(1-W_i)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i W_i}{n_1/n} - \frac{Y_i(1-W_i)}{n_0/n} \\
&= \frac{1}{n}\left(\frac{S_1}{n_1/n} - \frac{S_0}{n_0/n}\right) \\
&= \frac{S_1}{n_1} - \frac{S_0}{n_0}
\end{aligned}$$
where $\sum_{i=1}^{n} Y_i W_i = S_1$ since $Y_i W_i$ will only be 1 if the woman both purchases a net and uses it.

2. State the Fisher null and explain how to conduct a randomization test. Suppose we use a one-sided test and the usual randomization test procedure. Justify why the reported p-value will never underestimate the true p-value.

The Fisher null is $H_0 : \tau_i = 0$ for all i versus $H_a : \tau_i \neq 0$ for at least one i. The null $\tau_i = 0$ for all i implies $y_i(0) = y_i(1) = Y_i$, so we can leave the Y_i and sample many vectors of W_i with n_1 1s, compute $\hat{\tau}^*$ for each sample, and then find the proportion that are larger in magnitude than the observed $\hat{\tau}$.

However, since we only care about testing whether women who purchase nets are more likely to use them, we will use the null $H_0 : \tau_i \leq 0 \iff y_i(1) \leq y_i(0)$ for all i and $H_a : \tau_i > 0$ for at least one i. When still testing against a null of equality, if the null $y_i(1)$ are less than $y_i(0)$, our observed $\hat{\tau}$ would be even more extreme, so the p-values we obtain are at least the true p-values.

```
nsims <- 10^5
sim_store <- vector(length = nsims)
n_free <- sum(followup$FOL_pricepaid == 0)
```

```r
for (i in 1:nsims) {
  # Sample n_0 random indices
  indices <- sample(1:nrow(followup), n_free, replace = F)

  # Compute tau star
  sim_store[i] <- mean(followup$FOL_hanging[-indices]) -
    mean(followup$FOL_hanging[indices])
}

# Compute tau hat
tau_obs <- mean(followup$FOL_hanging[followup$FOL_pricepaid != 0]) -
  mean(followup$FOL_hanging[followup$FOL_pricepaid == 0])

# P-value
mean(sim_store >= tau_obs)
```

```
## [1] 0.6222
```

3. Because both the W_i and Y_i are binary in this example, we can find the exact distribution of $\hat{\tau}$ under the null. Find this exact distribution.

Consider white marbles as Y_is such that $Y_i = 1$ and black marbles as Y_is such that $Y_i = 0$, and take samples of n_1 Y_is. The number of Y_is we sample that equal 1 has the distribution $T \sim \text{HGeom}(S_0 + S_1, n - (S_0 + S_1), n_1)$ with S_1 and S_0 as defined in question 1 (note that we treat S_1 and S_0 as known and fixed now). Each $\hat{\tau}^*$ can then be written as $\frac{T}{n_1} - \frac{S_1 + S_0 - T}{n_0}$.

4. Show how to find the p-value from this distribution.

The probability of $\hat{\tau}^*$ being at least some t is:

$$\begin{aligned}
P(\hat{\tau}^* \geq t) &= P\left(\frac{T}{n_1} - \frac{S_1 + S_0 - T}{n_0} \geq t\right) \\
&= P\left(Tn_0 - (S_1 + S_0 - T)n_1 \geq tn_1n_0\right) \\
&= P\left(T(n_0 + n_1) - (S_1 + S_0)n_1 \geq tn_1n_0\right) \\
&= P\left(T \geq \frac{tn_1n_0 + (S_1 + S_0)n_1}{n_0 + n_1}\right) \\
&= 1 - F\left(\frac{tn_1n_0 + (S_1 + S_0)n_1}{n_0 + n_1} - 1\right)
\end{aligned}$$

where the last CDF is of a $\text{HGeom}(S_0 + S_1, n - (S_0 + S_1), n_1)$ random variable. The p-value can then be found as

$$P(\hat{\tau}^* \geq \hat{\tau}) = 1 - F\left(\frac{\hat{\tau}n_1n_0 + (S_1 + S_0)n_1}{n_0 + n_1} - 1\right)$$

By plugging in $\hat{\tau}$, we actually can get

$$P(\hat{\tau}^* \geq \hat{\tau}) = 1 - F\left(\frac{S_1n_0 - S_0n_1 + (S_1 + S_0)n_1}{n_0 + n_1} - 1\right) = 1 - F\left(S_1 - 1\right)$$

This turns out to be equivalent to Fisher's exact test!

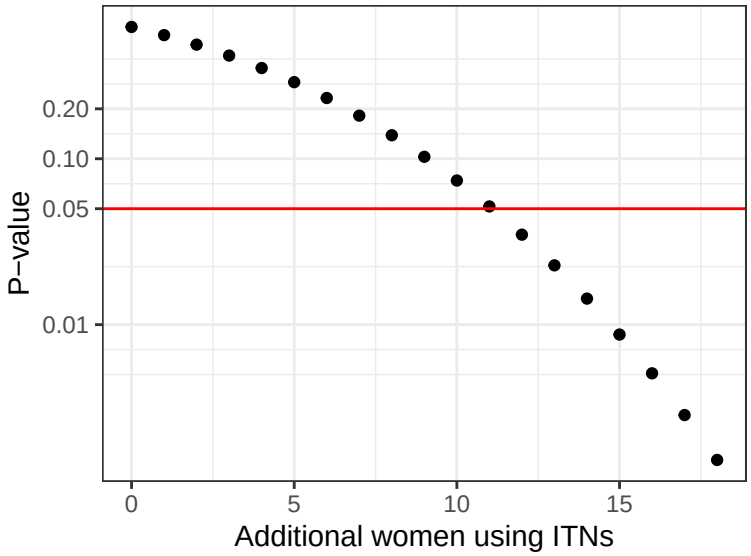
```r
S_0 <- sum(followup$FOL_hanging & followup$FOL_pricepaid == 0)
S_1 <- sum(followup$FOL_hanging & followup$FOL_pricepaid != 0)
n_0 <- sum(followup$FOL_pricepaid == 0)
n_1 <- sum(followup$FOL_pricepaid != 0)
```

```
# Exact p-value
1 - phyper(S_1 - 1, S_0 + S_1, n_0 + n_1 - (S_0 + S_1), n_1)
```

```
## [1] 0.6231785
```

5. One possible concern whenever implicitly seeking to retain a null is that the desired conclusion can follow from simply not obtaining enough data. Suggest a method to test this.

One simple test is to find the critical value above which we would have rejected the null and compare that to the observed value. That is, we can find how many more women who purchased nets would have had to use them for the results to become significant. This number turns out to be 12, which out of the $n_1 = 83$ women who already purchased a net and the $S_1 = 47$ who are already using nets, is quite a lot.



6. We can follow a similar process to test whether non-zero prices are associated with reduced ITN purchases. Explain why we cannot use the exact test from above. Interpret the results.

```
sim_store <- vector(length = nsims)
n_free <- sum(sales$cost == 0)
for (i in 1:nsims) {
  # Randomize and compare means
  indices <- sample(1:nrow(sales), n_free, replace = F)
  sim_store[i] <- mean(sales$weeklynetsales[indices]) -
    mean(sales$weeklynetsales[-indices])
}

# Observed value
tau_obs <- mean(sales$weeklynetsales[sales$cost == 0]) -
  mean(sales$weeklynetsales[sales$cost != 0])

# P-value
mean(sim_store >= tau_obs)
```

```
## [1] 0.04112
```

The results from above only apply to binary Y_i. We obtain a p-value of $0.04 < 0.05$, so we reject the null and conclude that significantly fewer ITNs are distributed when they require a non-zero payment.

One last thing: When reanalyzing the data to create this set of questions, we made two simplifying assumptions. First, we collapsed the continuous cost variable into a binary. Second, we looked at raw net sales rather than

net sales normalized to clinic popularity. Performing this same analysis with normalization actually yields insignificant differences for both the effect of cost on usage and the effect of cost on sales. However, the better way to perform this analysis is to use regression on the continuous cost range, and such analysis yields the conclusions found in the paper whether the net sales are normalized to clinic popularity or not.

Feedback

If you didn't get a chance to do so last week, please take a moment to provide some feedback on section this semester.