Problem Set #1

BST 258: Causal Inference – Theory and Practice

Daniel Xu

2/11/24

Question 2

a) $P(A = 1) = m/n$

b) For two units $i \neq j$,

$$\begin{aligned}
P(A_i = 1, A_j = 1) &= P(A_j = 1 | A_i = 1) P(A_i = 1) \\
&= \frac{(m-1)}{(n-1)} \frac{m}{n} \\
P(A_i = 1, A_j = 0) &= P(A_j = 0 | A_i = 1) P(A_i = 1) \\
&= \frac{(n-1-m)}{(n-1)} \frac{m}{n} \\
P(A_i = 0, A_j = 1) &= P(A_j = 1 | A_i = 0) P(A_i = 0) \\
&= \frac{m}{(n-1)} \frac{(n-m)}{n} \\
P(A_i = 0, A_j = 0) &= P(A_j = 0 | A_i = 0) P(A_i = 0) \\
&= \frac{(n-1-m)}{(n-1)} \frac{(n-m)}{n}
\end{aligned}$$

c) For two units $i \neq j$,

$$\begin{aligned}
Var(A_i) &= E(A_i^2) - E(A_i)^2 \\
&= P(A_i = 1) \left(1 - P(A_i = 1)\right) \\
&= \frac{m}{n} \left(1 - \frac{m}{n}\right) \\
Cov(A_i, A_j) &= E(A_i A_j) - E(A_i) E(A_j) \\
&= P(A_i = 1, A_j = 1) - P(A_i = 1) P(A_j = 1) \\
&= \frac{(m-1)m}{(n-1)n} - \left(\frac{m}{n}\right)^2
\end{aligned}$$

d) Assuming $Y_i(1)$ and $Y_i(0)$ fixed,

$$\begin{aligned} E(\theta^{ATT}) &= E\left(\frac{1}{m}\sum_{i=1}^{n} A_i\left(Y_i(1) - Y_i(0)\right)\right) \\ &= \frac{1}{m}\sum_{i=1}^{n}\left(Y_i(1) - Y_i(0)\right)E(A_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i(1) - Y_i(0)\right) \\ &= \theta^{SATE} \end{aligned}$$

Question 3

Given that θ is fixed,

$$Var(Y_i(1)) = Var(Y_i(0) + \theta) = Var(Y_i(0))$$

and by linearity of covariances,

$$\begin{aligned}
\rho(Y_i(1), Y_i(0)) &= \frac{Cov(Y_i(1), Y_i(0))}{\sqrt{Var(Y_i(1))Var(Y_i(0))}} \\
&= \frac{Cov(Y_i(0) + \theta, Y_i(0))}{Var(Y_i(0))} \\
&= \frac{Var(Y_i(0)}{Var(Y_i(0))} \\
&= 1
\end{aligned}$$

Question 4

Assuming that you guess completely at random, let X be the number of correct guesses. Clearly, $X \sim Hypergeometric(N = 8, K = 4, n = 4)$ where N is the total number of cups, K is the total number of cups with tea poured first, and n is the number of cups you guess have tea poured first. Hence, we know the distribution of X and can easily calculate

$$P(X = 0) = \frac{1}{70}$$
$$P(X = 1) = \frac{16}{70}$$
$$P(X = 2) = \frac{36}{70}$$
$$P(X = 3) = \frac{16}{70}$$
$$P(X = 4) = \frac{1}{70}$$

Question 5

a) There is confounding of the treatment and success relationship by stone size. Large stones are harder to deal with (lead to worse outcomes) and physicians were more likely to assign patients with large stones to treatment A.

b) For small stones:

.	Treatment A	Treatment B
Male	94.7% (71/75)	95.0% (19/20)
Female	83.3% (10/12)	86.0% (215/250)

For large stones:

.	Treatment A	Treatment B
Male	74.5% (181/243)	75.0% (15/20)
Female	55.0% (11/20)	66.7% (40/60)

c) This phenomenon where a trend appears when stratified but then reverses when aggregated is known as Simpson's paradox. It can be problematic if the person analyzing the data attempts to blindly draw causal conclusions. For instance, in this case, without stratifying by stone size, the analyst may naively conclude that treatment B is more effective. However, especially because the data was not collected from a randomized experiment, we cannot in general attribute marginal associations to causal effects. In fact, here it seems that once we stratify by the confounding variable (stone size), we see that if anything, Treatment B seems more effective than Treatment A.

Question 6

Clearly, we see that across all sample sizes, we have higher power rejecting the strong null than the weak null. This coincides with the statement that if there is evidence against Neyman's null, then there should also be evidence against Fisher's – in other words, it is "easier" to reject Fisher's strong null. We also see that as the sample size increases, so does the power against both nulls (approaching 1), which is again to be expected as both tests are consistent.

```r
n1 <- c(10, 25, 50, 100, 250)
n0 <- n1
n <- n1 + n0
nsim <- 1000
B <- 10000

strong_null_reject <- matrix(data = NA, ncol = length(n1), nrow = nsim)
weak_null_reject <- matrix(data = NA, ncol = length(n1), nrow = nsim)
set.seed(1)
for (k in 1:length(n1)) {
  Y1 <- rnorm(n[k], mean = 1/10, sd = sqrt(1/16))
  Y0 <- rnorm(n[k], mean = 0, sd = sqrt(1/16))
  for (i in 1:nsim) {
    A <- sample(c(rep(1, times = n1[k]), rep(0, times = n0[k])),
                size = n[k], replace = FALSE)
    Y <- A * Y1 + (1 - A) * Y0
    dif_mean <- (1/n1[k]) * sum(A * Y) - (1/n0[k]) * sum((1 - A) * Y)
    var_dif_mean <- (1/n1[k]) * (1/(n1[k] - 1)) *
      sum((Y - (1/n1[k]) * sum(A * Y))^2) +
      (1/n0[k]) * (1/(n0[k] - 1)) *
      sum((Y - (1/n0[k]) * sum((1 - A) * Y))^2)
    # generate null distribution
    dif_mean_null <- rep(NA, times = B)
    for (j in 1:B) {
      A <- sample(c(rep(1, times = n1[k]), rep(0, times = n0[k])),
                  size = n[k], replace = FALSE)
      dif_mean_null[j] <- (1/n1[k]) * sum(A * Y) -
        (1/n0[k]) * sum((1 - A) * Y)
    }
    strong_null_reject[i,k] <-
      mean(abs(dif_mean_null) >= abs(dif_mean)) <= .05
    weak_null_reject[i,k] <-
      abs(dif_mean) / sqrt(var_dif_mean) >= 1.96
  }
}

results <-
  data.frame(n,
             "strong" = apply(strong_null_reject, MARGIN = 2, FUN = mean),
             "weak" = apply(weak_null_reject, MARGIN = 2, FUN = mean))
```

```r
library(knitr)
kable(results, booktabs = T, caption = "Power for rejecting weak vs. strong null hypotheses f
```

Table 3: Power for rejecting weak vs. strong null hypotheses for 1,000 simulated datasets.

n	Power against strong null	Power against weak null
20	0.268	0.000
50	0.448	0.032
100	0.644	0.143
200	0.540	0.121
500	1.000	0.992

Question 7

a) We can set derivatives with respect to α and β to 0 and solve. Clearly

$$U(\alpha) = -\frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha - \beta A_i) = 0$$

which implies that

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta} A_i).$$

Likewise,

$$U(\beta) = -\frac{1}{n}\sum_{i=1}^{n} A_i(Y_i - \alpha - \beta A_i) = 0$$

which implies (by substitution) that

$$\begin{aligned}
\frac{\beta}{n}\sum_{i=1}^{n} A_i &= \frac{1}{n}\sum_{i=1}^{n}(A_i Y_i - \alpha A_i) \\
\frac{\beta m}{n} &= \frac{1}{n}\sum_{i=1}^{n} A_i Y_i - \frac{\alpha m}{n} \\
\beta &= \frac{1}{m}\sum_{i=1}^{n} A_i Y_i - \alpha \\
\beta\left(\frac{n-m}{n}\right) &= \bar{Y_1} - \bar{(Y)} \\
\hat{\beta} &= \frac{n}{n-m}\left(\bar{Y_1} - \bar{Y}\right)
\end{aligned}$$

and we can resubstitute this expression for $\hat{\alpha}$.

b) We can show that $\hat{\beta}$ is "unbiased" for the SATE, which in some sense suggests validity.

$$\begin{aligned}
E(\hat{\beta}) &= \frac{n}{n-m}E(\bar{Y_1} - \bar{Y}) \\
&= \frac{n}{n-m}\left(\frac{1}{m}\sum_{i=1}^{n} E(A_i)Y_i(1) - E\left(\frac{1}{n}\sum_{i=1}^{n}(A_i Y_i + (1-A_i)Y_i)\right)\right) \\
&= \frac{n}{n-m}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{m}{n^2}\sum_{i=1}^{n} Y_i(1) - \frac{n-m}{n^2}\sum_{i=1}^{n} Y_i(0)\right) \\
&= \frac{n}{n-m}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)) \\
&= \theta^{SATE}
\end{aligned}$$