Homework 4

Lee Kennedy-Shaffer

Course Week 10

Show your work and express all answers in lowest terms. You are welcome to use results and definitions from the posted document on the Moodle, but be sure to explain why a result or property can be used if it is not obvious (e.g., if certain conditions must be met, how do we know they are?).

Note: If you use R to get results, please also write the answers in context or make clear which number is which result.

Question 1: Variance and Power for Stratified and Matched RCTs.

We again will look at power and sample size for Kalla and Broockman (2016). Assume that, under the control condition, the probability of getting a meeting with a DC-based staffer is $\pi_0 = 0.30$ (similar to Table 1 results) and we want power to detect a minimum effect (risk difference) of 10 percentage points ($\theta = 0.10$), so to detect $\pi_1 = 0.40$. Assume throughout that there are n total study units ($n_1 + n_0 = n$, where n_1 is the number in the intervention condition and n_0 the number in the control condition), a significance level of $\alpha = 0.05$, and a two-sided alternative hypothesis. The estimator is the difference in sample proportions: $\hat{\theta} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i} - \frac{1}{n_0}\sum_{i=1}^{n_0} Y_{0i}$.

A. In terms of n_0 and n_1, what is the variance of the estimator?

B. Assuming that $n = 192$ and we randomize equally ($n_0 = n_1 = n/2$), how much power will an un-stratified, un-matched study have for this MDE?

C. We have the option of a randomization ratio other than 1:1. Under the constraint that $n_1 + n_0 = n$ is fixed, find the ratio of n_1/n_0 that gives the minimum possible variance. Using $n = 192$ and rounding the number of participants in each group, what is the design effect for that randomization ratio compared to 1:1?

D. Suppose we go back to a 1:1 randomization, but now we stratify the legislators into two groups: those with an above-average record on environmental bills (stratum 1) and those with a below-average record on environmental bills (stratum 2). Assume that the strata are equal in size. If, under the control condition, the probability of a meeting would be 40% in stratum 1 and 20% in stratum 2, what is the pooled (weighted average) variance and standard deviation of the outcome across all four stratum-arm combinations? What are the variance and standard error of the estimator? Assume that the intervention effect is still a difference of 10 percentage points in each stratum.

E. Using the stratified trial, how much power does this study have? **Hint:** You can use `power.t.test` with 96 observations per sample, using the weighted-average SD for a single outcome as the SD parameter.

F. Now suppose we pair-match legislators. If the variance of the differences between paired legislators is 0.2, find the variance and standard error of the estimator.

G. Using the pair-matched trial, how much power does this study have? **Hint:** You can use `power.t.test` with 96 observations and option `type="paired"`, using the SD of a pair difference as the SD parameter. Again, continue to use 0.1 as the true effect.

H. Find the design effects of both the stratified and matched trials compared to the unstratified 1:1 trial.

Question 2: Sample Size Calculations for Single-Arm, Multi-Arm, and Factorial RCTs

In this question, we're going to calculate the sample size for a COVID-19 vaccine trial like the one in Baden et al., *New England Journal of Medicine*, 2021. That trial was going to be analyzed using a survival analysis method. To simplify, we'll assume the analysis will be for the difference between two proportions using a pooled variance two-sample t-test. The result will be fairly similar.

A. First, we need an estimated variance of the observations. Assume that there will be about 180 symptomatic infections among 30,000 participants over the course of the trial. If we use this ratio ($p = 180/30000$) as the overall probability of infection, then what is the variance of the binary random variable representing whether one randomly-chosen individual is infected?

B. As we have seen, effect sizes for vaccines are often represented as the vaccine efficacy, which is equal to $VE = 1 - \frac{p_1}{p_0}$, where p_1 is the probability of symptomatic infection in the vaccine arm and p_0 is the probability of infection in the control arm. For the t-test, however, we need the effect size given by the absolute difference between the mean outcome (i.e., proportion of participants with symptomatic infections) in the control and vaccine arm. Keeping the overall probability of infection among all participants in the trial fixed at the value p in A, and assuming that we will randomize half of the participants ot the vaccine arm and half to the control arm, find values of p_1 and p_0 that give $VE = 0.60$, or 60% vaccine efficacy.

C. Use the results of part C to find the absolute difference between means for which we will power this trial. Use this and the result from A to find the Cohen's d (standardized effect size), the absolute difference between means divided by the pooled standard deviation.

D. Find the number of individuals needed (both in each arm and total) to have 90% power to detect this effect size with a 5% significance level. You can calculate by hand or use `R`.

E. Suppose we want to make this a multi-arm trial comparing four different candidate vaccines (Pfizer, Moderna, Johnson and Johnson, and AstraZeneca) to the same control arm. Using the same parameters as above and adjusting for the fact that we will conduct four hypothesis tests, how many individuals will we need in each arm and in the trial as a whole? Compare this to if we conducted four separate trials, each with the size found in D.

F. Finally, suppose we also want to test, for each of the four vaccines and the control group, three different regimens: a single shot, two shots three weeks apart, and two shots three months apart. For the control group, they will get placebo shots at each relevant time point. Explain how we could use a factorial design for this study and how many different groups there would be. Give one advantage and one disadvantage (statistical, scientific, logistical, or ethical) of this factorial design compared to first conducting a study to identify the best single-shot vaccine and then conducting a study comparing different regimens using only that vaccine.

Question 3. Multiple Testing: Probability

Suppose we want to test two new interventions vs. a control arm. Assume first that we are running two independent two-arm trials each comparing one new intervention to the control. We design each to have significance level α and power $1 - \Delta$ for the desired MDE.

A. Suppose that neither new medicine works (so the null hypothesis is true for both trials). What is the probability that at least one trial will find a significant result?

B. If both new medicines work at the MDE level, what is the probability that at least one trial will find a significant result?

C. If we use a Bonferroni correction to account for two hypothesis tests, then what will the probability of at least one trial being significant be if neither medicine works? Show that this is less than whatever original α is used.

D. Explain why the correction is necessary (with reference to the overall risk of Type I or Type II Errors) and what effect it will have (directionally) on the power of the tests.

E. In a multi-arm trial, instead of being independent, the two tests are positively correlated (because they share a control group). How does this affect the probability of a significant result occurring on at least one test under the null hypothesis compared to your result from A? (You won't be able to get a specific formula, but can figure out the direction of the difference). Does this make the Bonferroni correction more or less conservative? *Hint:* Remind yourself how correlation is calculated for a binary random variable and think of each test being significant as a binary RV (1=significant, 0=not significant).

Question 4. Multiple Testing: Simulation

Continue the previous question, but now using the migraine medicine RCT from HW3, Q3. Recall that under the control condition, the outcome follows $Y_0 \sim Pois(2)$. If either new intervention works, the outcome in that arm will follow $Y_1 \sim Pois(1.5)$ (same for both new interventions). So our $\theta_{MDE} = 2 - 1.5 = 0.5$. Ignore missing data/loss to follow-up, etc.

A. For both studies to have 90% power at $\alpha = 0.05$ to detect $\theta = 0.5$, they each need 148 people in each arm. Simulate 1000 times both studies under the null hypothesis. For each of the 1000 simulations, conduct the t-test testing whether there is a significant difference (without corrections), and save whether each study is significant and whether at least one of the studies is significant. Find the empirical Type I Error of each study individually and of whether either study was significant. Does this match your results from Q3A? *You can use the suggested code at the end of the assignment.*

B. Do the same thing with the Bonferroni correction. Does this match your results from Q3C?

C. Do the same thing if the medicines work, both with and without the correction. What empirical powers do you find? Which version matches your prediction from Q3B? Why is that?

Question 5: Final Project Topic and Past Study Review

With your final project group, determine the topic (overarching scientific question of interest) for your project. Describe this in a few sentences: what is the field your study will relate to, and what question will it seek to answer? *Feel free to write this together and copy-paste for all members of your group.*

In addition, your group should identify several studies (reported in peer-reviewed, academic articles) on related topics in the field. You should each pick a separate one of these studies, give a full citation for it here, and write one paragraph summarizing the design and results of that study, and another paragraph describing its limitations in answering the question that your group hopes to answer. *You may (and should) discuss this part with your group, but each member must choose a different study and write up these paragraphs separately.*

Suggested R Code for Question 4

```
OneSim <- function(lambda0, lambda1, nPerArm, alpha) {
  S1_Y0 <- rpois(n=nPerArm, lambda=lambda0)
  S1_Y1 <- rpois(n=nPerArm, lambda=lambda1)
  S2_Y0 <- rpois(n=nPerArm, lambda=lambda0)
  S2_Y1 <- rpois(n=nPerArm, lambda=lambda1)
  S1_p <- t.test(x=S1_Y0, y=S1_Y1, alternative="two.sided")$p.value
  S2_p <- t.test(x=S2_Y0, y=S2_Y1, alternative="two.sided")$p.value
  return(c(S1_p < alpha, S2_p < alpha, S1_p < alpha | S2_p < alpha))
}

set.seed(XX)

NumSims <- XX
mean0 <- XX
mean1 <- XX
SSPerArm <- XX
sig_level <- XX

Results <- t(replicate(n=NumSims, expr=OneSim(lambda0=mean0, lambda1=mean1,
                                              nPerArm=SSPerArm, alpha=sig_level)))
colnames(Results) <- c("Study1_Sig","Study2_Sig","Either_Sig")
Results <- as_tibble(Results)
```