Homework 2

Lee Kennedy-Shaffer

Course Week 5

Do **all** of the parts that are labelled with a letter. You are not required to complete any optional problems this time.

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: Combining Surveys Continued

We are going to return to the idea of question 4 from HW 1 and show a more general result.

A. In general (any study design, any estimand, any type of estimand), if we have two independent, unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, we can make a general weighted-average estimator $\hat{\theta}_{avg} = w_1\hat{\theta}_1 + w_2\hat{\theta}_2$. Show that if $w_1 + w_2 = 1$, then $\hat{\theta}_{avg}$ is an unbiased estimator of θ.

B. Denote the variances of these two estimators by v_1^2 and v_2^2, respectively. We do not control these variances, so we can treat them as unknown constants. Given the constraint from part A ($w_1 + w_2 = 1$), prove that the variance of $\hat{\theta}_{avg}$ is minimized subject to this constraint by $w_1^* = \frac{1/v_1^2}{1/v_1^2+1/v_2^2}$ and $w_2^* = \frac{1/v_2^2}{1/v_1^2+1/v_2^2}$. This is the *inverse-variance weighting* approach. *Hint:* You can optimize this function by Lagrange multipliers or using substitution. Note that these are equivalent to $w_1^* = \frac{v_2^2}{v_1^2+v_2^2}$ and $w_2^* = \frac{v_1^2}{v_1^2+v_2^2}$.

C. For any finite number K of independent, unbiased estimators, find a condition on the weights $w_1, \ldots, w_K$ such that $\sum_{k=1}^{K} w_k\hat{\theta}_k$ is unbiased.

D. In terms of the weights from part C, assuming that the estimators have variances $v_1^2, \ldots, v_K^2$, find the variance of the weighted estimator. Your answer should be a function of v_k and w_k for $k = 1, \ldots, K$. *Note:* you can assume that the weights are fixed, not random, variables.

OPTIONAL: The variance-minimization result from part B can be generalized to K as well. You're welcome (but not required) to try to prove it using the Lagrange multipliers approach.

Question 2: Biased Surveys and MSEs

Suppose we are going to conduct a survey to estimate the average number of hours watched on streaming services by Vassar students. Assume that the population distribution is Poisson-distributed with expectation λ. So if Z is a randomly-chosen student's response, $Z \sim Pois(\lambda)$, where Z and λ are in hours/week.

A. One approach is to take a simple random sample of 30 students, get their responses $Z_1, \ldots, Z_{30}$, and use as estimator 1 the sample mean: $\hat{\lambda}_1 = \frac{1}{30}\sum_{i=1}^{30} Z_i$. Show that $\hat{\lambda}_1$ is unbiased for λ and find its variance and sampling distribution (it is sufficient to find the distribution of $30\hat{\lambda}_1$).

B. Another approach would be to post an ad on some streaming sites inviting people to take the survey. Suppose that if we do that, 60 students will take the survey, but they have a slightly different distribution. Students who never watch streaming TV would not click on the ad, so among students who do click on the ad, our responses follow a Zero-truncated Poisson distribution (see Wikipedia or the updated Notes/Review document for details). Denote the responses from this survey by $Y_1, \ldots, Y_{60}$. What are the expected value, bias, and variance of $\hat{\lambda}_2 = \frac{1}{60}\sum_{i=1}^{60} Y_i$? Simplify the bias formula as much as possible. *Hint:* first, use the properties of the zero-truncated Poisson to find the expectation and variance of a single observation Y. Then you can think of this as a SRS from a population with that distribution.

C. Find and compare the Biases, Variances, and MSEs of the two estimators, assuming that the true expected number of hours watched per week is $\lambda = 5$.

D. In a short paragraph, describe some of the advantages and disadvantages of each of these sampling approaches. Which do you prefer and why? Be sure to consider statistical, practical, and ethical trade-offs.

E. If n_1 people are sampled in the SRS and 60 in the advertised opt-in design, find an inequality with respect to n_1 equivalent to $MSE(\hat{\lambda}_1) \geq MSE(\hat{\lambda}_2)$. You can continue to use $\lambda = 5$. Explain what this means in words about sample sizes and MSEs.

F. Do you think it's fair in this case to assume that the only bias in the opt-in design will be removing people who watch 0 streaming? What will be the consequences for bias and MSE if it also removes a lot of people who watch a small but non-zero amount of streaming?

Question 3: Stratified Sampling

Suppose we are conducting a survey to find out about the costs of groceries. We will ask households how much they spent on groceries in the past week. We are interested in the estimand $\mu = E[X]$, the expected weekly cost of groceries for a U.S. household in dollars, where X is the weekly cost for a randomly-chosen household.

A. We are considering stratifying based on whether the household has kids or does not have kids. What other variables do you think might be predictive of the outcome? Are they feasible to stratify on?

B. We are going to stratify into two groups: $J = 1$ for households without kids, and $J = 2$ for households with kids. Assume that the distribution of X given $J = 1$ is $N(150, 40^2)$ and the distribution of X given $J = 2$ is $N(270, 60^2)$. Finally, assume that 67% of households do not have kids and 33% do. What is the marginal (overall) expectation and variance of X? *Hint:* Use the law of total expectation and the law of total variance.

C. If we did a SRS on the whole population, sampling 100 households, what would the bias, variance, and standard error of the estimator be? Call it $\hat{\mu}_{SRS}$.

D. If we conduct the stratified sample using a self-weighted sample with 67 households without kids and 33 with, what are the bias, variance, and standard error of that estimator? Call it $\hat{\mu}_{Strat}$.

E. What is the design effect comparing $\hat{\mu}_{Strat}$ to $\hat{\mu}_{SRS}$? In words, what does this mean?

F. If we got the weights wrong, and the real population has 60% of households without kids and 40% with, what are the bias, variance, and MSE of $\hat{\mu}_{Strat}$? *Note:* you'll have to re-calculate the marginal expectation of X using the true weights, but the weights in $\hat{\mu}_{Strat}$ are still 0.67 and 0.33.

G. In your opinion, are the benefits of stratified sampling in this case worth the trade-offs?

Question 4: Comparing Probabilities

The Moderna COVID-19 vaccine trial is reported in Baden et al., *New England Journal of Medicine*, 2021, posted on the Moodle site. The actual study used a time-to-event (also known as survival) analysis, but we can simplify it to an analysis of proportions of patients who had a certain outcome by 3 months after their second dose. There were 14,073 participants followed-up in the placebo/control group and 14,134 in the vaccine group. Of these, 185 in the placebo group had symptomatic COVID and 11 in the vaccine group did. 30 in the placebo group had severe COVID and 0 in the vaccine group did. And 1 in the placebo group died while 0 in the vaccine group did.

A. For each of the three outcomes, find the estimated risk difference, risk ratio, and odds ratio comparing the new intervention (vaccine) arm to the placebo arm. *Recall:* the odds of an event are the probability of the event divided by one minus the probability.

B. If these estimates are correct, how many people, on average, would you need to switch from the placebo to the vaccine in order to prevent 1 symptomatic COVID-19 case? This is known as the **number needed to treat**.

C. What are the ranges of possible values for these three contrasts (not just in this study, but in general)? What values correspond to no effect of treatment?

D. In vaccine studies, we usually report instead the vaccine efficacy, which is 1 - the odds ratio. What is the value of VE against symptomatic COVID in this study? What value would represent a perfect vaccine (prevents all cases) and what would represent an ineffective vaccine (no different from placebo)?

E. Which of these three contrasts would you find most useful in recommending to someone whether to get the vaccine or not? Why? Does your answer change if you're deciding whether to implement a policy to use a specific treatment on all patients? Why or why not?

Question 5: Reporting Studies and Public Communication of Statistics

Read one of two *New York Times* articles reporting on studies we've looked at:

- **Zimmer and Grady, Nov. 16, 2020**, on the Moderna vaccine results; or
- **Blum, Nov. 11, 2023**, on the semaglutide (Wegovy) cardiovascular outcomes study results.

Note: you should be able to access both articles even without an *NYT* account. Let me know if you have any trouble.

In 1–2 paragraphs, answer the following questions:

- How does the article explain the goals, scientific questions, and study design of the described study?
- How does it report the estimate and the uncertainty of the estimate?
- How does the article discuss the statistical challenges of study design and the statistical properties we have covered in class?
- Do you think the article does a good job of communicating the results of this study and the advantages and limitations of its design to the public? Why or why not? Specifically, what would you like to see more or less of?