Homework 1

Lee Kennedy-Shaffer

Due Course Week 3

Do **all** of the parts that are labelled with a letter. In addition, **choose at least one** of the parts labeled OPTION (there's one at the end of each question 1–4) and do that. The first is probability-focused, the second is practically-focused, the third is coding/inference-focused, and the fourth is simulation-focused.

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: Sensitivity Analyses for Sample Sizes to Estimate Proportions

Consider the example from weeks 1/2 of class: a simple random sample survey to estimate the proportion of likely voters who will vote Democratic in the next presidential election. The estimator is $\hat{\theta} = \bar{Y}_n$, the sample average of the n responses, where responses are coded as 1 for Democrat and 0 for other. The true (unknown) proportion (i.e., the estimand) is $\theta = P[Y = 1]$.

In class, we found that to have a 95% confidence level margin of error less than 5 percentage points, we needed to sample 400 people, assuming the true proportion was 0.5.

A. If the true proportion is 0.6, how many people do we need to sample to get the same MOE? Explain why this is higher/lower/the same as the MOE if the true proportion was 0.5.

B. If the true proportion is 0.4, how many people do we need to sample to get the same MOE?

C. Create a graph of the required sample size as the true proportion θ varies in the open interval $(0, 1)$. Note: you can use R, Excel, or any other method to get the graph.

D. Prove that for any desired margin of error, the required sample size is maximized when $\theta = 0.5$. Hint: leave the margin of error as an unknown variable, write the sample size as a function of θ and that variable, and use some calculus.

E. From a design standpoint, if we want to make conservative assumptions about our parameters, what should we use for the unknown θ in our sample size calculations?

OPTION 1: Think about other probability distributions you've seen. Where else is there a relationship between the expectation parameter and the variance? How will that affect operating characteristics of the sample mean?

Question 2: Surveys Estimating a Mean

Consider a similar SRS, but now we want to estimate the mean household income in the U.S. population. Let Z be the income of a randomly-chosen household. Assume that household income is normally-distributed with mean μ and variance σ^2:

$$Z \sim N(\mu, \sigma^2)$$

A. Just based on your own background knowledge, do you think a normal distribution is reasonable here? Why or why not?

B. Our estimator will be $\hat{\mu} = \bar{Z}_n$, the average of the observed data points $Z_1, \ldots, Z_n$. Explain in a sentence or two why this is a "reasonable" estimator.

C. Find the expectation and variance of the estimator under our given assumptions.

D. Find the sampling distribution of the estimator under our given assumptions.

E. If we will calculate a 95% confidence interval for the estimand, and we want the width of the CI to be no more than 2ℓ (so a margin of error less than or equal to ℓ), find a formula for the required sample size n. What parameters does this size depend on and what doesn't it depend on?

F. Assume that the mean household income is \$100,000 and the standard deviation is \$40,000. If we sample 130,000 households, what will the expectation, sampling variance, and standard error (square root of the sampling variance) of the estimator be? What will the 95% margin of error width be?

OPTION 2: The mean and sample size in F are drawn from the Census's Current Population Survey for 2021 (https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-01.html). But the standard error you calculated should be smaller than the one shown there by a fair amount. Why might that be?

Question 3: Simulating Operating Characteristics

Let's continue with the SRS from the previous problem and the assumed parameters in F: $Z \sim N(100000, 40000^2)$, and a sample size of $n = 130,000$. Another approach to determining operating characteristics is to use simulation.

In each simulation, we are going to simulate the experiment of sampling 130,000 households from this population and recording their income. Then we take the mean and variance of those incomes and record those for each simulation as our estimate and sample variance. We repeat this 500 times to get the results from a set of 500 simulations.

Note: Some code snippets and hints are given on the last page. If you're familiar with R functions, though, give it a try yourself! To get the same randomized results each time you run it, include the line `set.seed(XX)` with some number of any length in place of `XX` before the randomizing part of your code.

A. Give the mean of the estimates of the simulations. What operating characteristic does this represent?

B. Give the variance and SD of the estimates of the simulations. What operating characteristics do these represent?

C. Give the square root of the mean of the sample variances. What parameter does this represent?

D. For each simulation, find the estimated 95% confidence interval using that simulation's estimate and sample variance. What is the mean width of these CIs? How does this relate to some answer in 2F? Hint: add columns to the tibble with the lower bound of the CI and the upper bound of the CI. Then make another that calculates the width.

OPTION 3: Add a column that records whether the true value of μ is within the CI for each simulation. How often should this occur across the simulations? Is that what happens?

Question 4: Combining Two Surveys

Let's go back to a situation where we are trying to estimate a proportion θ. Suppose two different surveys are conducted, each using simple random sampling. They each use the estimator that's the simple average. Survey 1 samples 100 people and survey 2 samples 300 people. Denote the estimate from survey 1 as $\hat{\theta}_1$ and the estimate from survey 2 as $\hat{\theta}_2$. Assume the results of the two surveys are independent.

We are going to combine the two surveys to see if that improves our estimator.

A. Find the expectations and variances of $\hat{\theta}_1$ and $\hat{\theta}_2$.

B. One combined estimator could be $\hat{\theta}_3 = \frac{1}{2}\hat{\theta}_1 + \frac{1}{2}\hat{\theta}_2$. Find the expectation and variance of this. Compare its operating characteristics to the two individual estimators.

C. Another approach is called *inverse-variance weighting*. The combined estimator is a weighted average of the individual estimators, with weights proportional to their variances and summing to 1. In this case, the combined estimator would be $\hat{\theta}_4 = \frac{100}{400} \cdot \hat{\theta}_1 + \frac{300}{400} \cdot \hat{\theta}_2$. Find the expectation and variance of this. Compare its operating characteristics to the two individual estimators and $\hat{\theta}_3$.

D. Finally, if we can access the individual data, we could combine the samples and use our usual sample mean estimator. Call this $\hat{\theta}_5$ and find its expectation and variance. What do you notice?

OPTION 4: If you want more practice doing simulations, you can simulate this as well. Simulate the two experiments (note that you'll need to choose a specific value of θ for the simulations; I use 0.5). For each simulation, you should save the estimators from parts B, C, and D (using the mean of the observations simulated in each sample). Then compare the empirical operating characteristics (empirical expectation and variance/SE) of each of these three estimators.

Question 5: Interpretations

Read the Jan. 11, 2023 *STAT News* article by Brittany Trang at https://www.statnews.com/2023/01/11/wastewater-data-biobot-health-covid19/. In a paragraph or so, describe the scientific question of interest, estimator, and estimand of a wastewater surveillance sampling survey. How is the sample taken? If we wanted to estimate the COVID rate in Poughkeepsie as a whole and we sampled randomly from the Vassar wastewater system, what would be possible sources of bias in this approach?

Sample Code for Question 3

Note: you can put this into a code chunk in a .Rmd file by adding `{r chunk-name}` after the three back-ticks at the top.

```
# This function simulates SS observations from a normal distribution
## with expectation mu and variance sigma^2.
## It returns a vector with the sample mean and variance of those SS observations.
OneSim <- function(SS, mu, sigma) {
  observations <- rnorm(n=SS, mean=mu, sd=sigma)
  estimate <- mean(observations)
  sample.var <- var(observations)
  return(c(Est=estimate,SVar=sample.var))
}

set.seed(XXXX) # Put some number in here so that you get the same results each time you run it.

## These next lines will run the above function n times (note this is the number
### of simulations to do, which is different from the sample size SS)
### and save the results as rows in a tibble.
### Make sure to put in the appropriate parameters.
Sim.Results <- replicate(n=, OneSim(SS=, mu=, sigma=))
Sim.Results.tidy <- tibble(Estimate=Sim.Results["Est",],
                           Sample.Var=Sim.Results["SVar",])

## Now you can use mutate, summarize, and other functions on that tibble
### to get the desired summaries across the n simulated data sets.
## For example, to get the mean, variance, and SD of both columns, use:
Op.Chars <- Sim.Results.tidy %>% summarize(across(all_of(c("Estimate","Sample.Var")),
                                                  list(mean=mean, var=var, sd=sd),
                                                  .names="{.col}.{.fn}"))
```