SurvMeth 740: Fundamentals of Inference

Homework 4: Randomization Inference (Max: 100 pts.)

DUE: October 31, 2025, 12:59PM (on Canvas)

1. Sampling Distributions. Using software of your choosing, generate a population of size N=100,000 with random draws from a normal distribution with mean=15 and variance=3. For your random draws, set a seed of 41279 in R to ensure replication.

a. (5 points) Give the finite target population mean and population variance. Is there superpopulation here, if so, what is it?

b. (5 points) Draw a single simple random sample, without replacement (SRSWOR), of size n=1,000. Calculate the estimated mean, the estimated standard error of the estimated mean, and a 95% CI for the population mean.

c. (10 points) Draw a single simple random sample of size n=1,000. Store the estimated mean of the variable and the estimated variance of the variable calculated from this sample. Repeat this process 2,000 times. Although the estimates from this process will not yield the full sampling distributions, these should be a good approximations. Plot the approximation of the sampling distribution of the estimated mean. Plot the approximation of the sampling distribution of the estimated variance. Extract the 2.5 and 97.5 percentiles of the sampling distribution of the estimated mean.

d. (5 points) How do the 95% CI and the percentiles in 1a and 1b compare? Explain the difference between these quantities in 1a and 1b.

e. (10 points) When evaluating alternative estimation and inferential procedures, we are often concerned with whether or not a computed confidence interval following a particular procedure has the stated level of coverage (i.e., 95% of the intervals computed in the same way across repeated samples cover the true population parameter). Compute the coverage rate of the 95% CI for the population mean, computed for each of the 2,000 repeated samples. Do these intervals have the stated level of coverage?

2. Confidence Interval for Transformation.

a. (5 points) Using the example for the mean no. of housing units **renting** in each block on slide 10 Module 8, compute the 95% confidence interval for the transformation logit ($Q(Y)$) using the results of slide 12.

b. (5 points) Using the results in a.) now compute the 95% confidence interval for $Q(Y)$.

3. Design Effects and Effective Sample Sizes.

a. (3 points) Suppose that the data from **Module 8, Slide 10** represent a **two-stage cluster sample**, rather than a simple random sample, where a simple random sample of 20 clusters (blocks) was selected at the first stage, and each of the exactly 60 housing units in each block, i.e., no subsampling within sampled blocks, was asked whether they own or rent at the second stage. We therefore have a sample of size n = 20 x 60 = 1,200 housing units, selected from a population of size N = 270 x 60 = 16,200 housing units. What is the estimated proportion, p, of housing units in this population that are rentals?

b. (4 points) What is the estimated variance of this estimated proportion? Hint: for a two-stage cluster sample with equal-size clusters at the first stage,

$$var(p) = (1 - f)\frac{s^2}{a},$$

$$w\square ere\ a = \#\ of\ sampled\ clusters\ and\ s^2 = \frac{1}{a-1}\sum_{i=1}^{20}(p_i - p)^2$$

Given this variance, compute a 95% CI for this proportion.

c. (5 points) Assume instead that this was a simple random sample of the same size (n = 1,200), and compute the estimated variance of the estimated proportion under this design. What is the design effect for this estimated proportion due to the two-stage cluster sampling? What is the effective sample size in the case of this two-stage cluster sample? Write a plain English definition of this effective sample size.

d. (8 points) Which of the two sampling plans would you select (for a sample of size n = 1,200) based on the design effect and the comparison of these estimated variances? What would be the implication for the frequentist coverage rate of the 95% confidence intervals if you assumed a simple random sample of size n = 1,200, when in fact a two-stage cluster sample was selected?

4. Consider a population of size N=4 given by $\boldsymbol{Y} = \{5, 10, 35, 100\}$. Suppose the goal is to estimate the finite population mean $\bar{Y} = \frac{1}{4}\sum_{i=1}^{4} Y_i$ based on a simple random sample of size $n = 2$. Consider three estimating procedures: 1) the arithmetic mean defined as $\frac{1}{2}\sum_{i=1}^{2} y_i$, 2) the harmonic mean defined as $\left(\frac{1}{2}\sum_{i=1}^{2}\frac{1}{y_i}\right)^{-1}$, and 3) the geometric mean, defined as $\left(\prod_{i=1}^{2} y_i\right)^{1/2}$

a. (5 points) For all 6 possible simple random samples of size n=2 (Hint: you can derive the entire sampling distribution based on all possible samples under this sampling plan!), compute the arithmetic, harmonic, and geometric means.

b. (3 points) Compute the bias of each estimating procedure.

c. (3 points) Compute the variance of each estimating procedure.

d. (4 points) Compute the mean squared error of each estimating procedure.

e. (5 points) Based on your results in a)-d), briefly comment on the strengths and weaknesses of each procedure.

f. (10 points) Repeat a)-d) for a different population $\boldsymbol{Y} = \{5, 10, 16, 18\}$. Briefly discuss how these results update your comments in e).

4. (5 points) Provide one important benefit (pro) and one important drawback (con) for inferences based on each of the three primary statistical viewpoints: Frequentist and Bayesian.