# Friday, December 12

## Information about the Final Examination

1. The final examination is on Monday, December 15th, from 12:45 to 2:45 PM. It will be in the same room as lecture.

2. The format of the final examination is like that of the other examinations. The only difference is that it is a "cumulative" examination (i.e., including material covered over the entire semester), and you will have 120 minutes rather than 50 minutes to complete it. It will not have more questions than the other examinations, and is worth the same amount (25%) as the other examinations. As usual the final examination is open-notes.

3. The final examination will focus more on concepts. There will be fewer computational problems.

## Foundations

1. Understand the distinction between a *population* and a *sample*, between an *element* and a *sampling unit*, between a *parameter* and *estimator*, and between a *target* and *auxiliary* variable.

2. What do we mean by a *sampling frame*?

3. What do we mean by a *(probability) sampling design*?

4. What are *inclusion probabilities* and what are *selection probabilities*? What are *first-order* versus *second-order* inclusion probabilities?

5. What do we mean by the *sampling distribution* of an estimator?

6. How do we talk about the "accuracy" of an estimator in terms of its *bias*, *variance*, and *mean squared error*?

7. What is the relationship between the *standard error*, the *variance of an estimator*, the *bound on the error of estimation*, and a *confidence interval*?

## Sampling Designs

1. From a short description of a sampling design (assuming you are given adequate information) you should be able to identify whether the design is a *simple random sampling design*, a *stratified random sampling design*, a *one-stage cluster sampling design*, or a *two-stage cluster sampling design* (assuming it is one of these). Think about the defining characteristics of each design.

2. Relative to a simple random sampling design with the same sample size (i.e., the same number of sampled elements), what advantages and disadvantages (if any) do we have by using *stratified random sampling* or *cluster sampling*? Are there any advantages to *two-stage* cluster sampling over *one-stage* cluster sampling? What?

3. Both stratified random sampling and cluster sampling (both one-stage and two-stage) use the fact that the elements in a population have been partitioned into groups that we call either *strata* or *clusters*. What then distinguishes a stratified random sampling design from a cluster sampling design, and what distinguishes a one-stage cluster sampling design from a two-stage cluster sampling design?

4. If the goal is to reduce the variance of an estimator of $\tau$ or $\mu$, what is the best way to assign elements to *strata*? What is the best way to assign elements to *clusters*? Think about *when* the variance will be relatively small in terms of how elements are stratified/clustered.

5. How can simple random sampling be used as a component of various *complex* sampling designs, including stratified random sampling and cluster sampling designs? Also what kinds of sampling designs might we use to sample *clusters* in a cluster sampling design?

6. What is meant by *optimum allocation* in the context of stratified random sampling. What is the basic approach? Which strata should get the largest/smallest allocation?

7. What is the distinction between sampling *with* versus *without* replacement?

8. Have some understanding of some of the specialized sampling designs that we discussed including *ranked set sampling*, *systematic sampling*, *Poisson sampling*, *line-intercept sampling*, *fixed-area plot sampling*, *adaptive cluster sampling*, and *balanced sampling*.

## Estimation/Inference

1. Several times we have seen that there were a couple of different estimators that could be used to estimate a total or mean. What factors would we consider in evaluating the advantages/disadvantages of each estimator relative to the other? Given a pair of such estimators you should be able to list their relative advantages and disadvantages (if any).

2. What do we mean by *variance estimation*?

3. In what situations is a ratio or regression estimator best suited?

4. In one-stage and two-stage cluster sampling designs, what is the *auxiliary variable* that we are using when we use the "ratio estimator" of $\tau$ or $\mu$?

5. How do we specify the target variable if we want to estimate *abundance* (i.e., the number of elements in a population)? How do we specify it if we want to estimate the abundance *for a domain* (i.e., the number of elements in a particular sub-population)?

6. How do we specify the target variable if we want to estimate the mean or total *in a given domain*? How do we specify the target variable if we want to estimate *abundance* (either overall or in a domain)?

7. What does it mean to say that a sample is *calibrated*?

8. What is meant by *post-stratification*? How is it different from stratified random sampling?

9. Many estimators used in survey sampling (and in statistics in general) are *biased*. Why is this sometimes tolerable?

10. Given the values of the target variable and the inclusion probability for the elements within a sample, how do we compute the *Horvitz-Thompson estimator* and the *Hájek estimator* for $\tau$ and $\mu$? And for $\mu$ how is this done when the number of elements in the population is known versus when it is not?

11. How/when do we use a *Hansen-Hurwitz estimator*?

12. If we can choose our own selection probabilities when sampling with replacement, or our own inclusion probabilities when sampling without replacement, how should we specify them so as to reduce the variance of an estimator of $\tau$ or $\mu$?

13. What determines *survey weights*? How are survey weights used in estimation?

14. Be familiar with why and how we would use *double sampling* to (a) estimate $\mu$ or $\tau$ when we have strata and (b) are using a ratio estimator of $\mu$ or $\tau$.

15. How does detectability and non-response affect inferences? How do we conceptualize these issues in terms of inclusion probabilities?