NMSA603 – exam topics

I. Simple random sampling

Question: How to deal with units where the value of the study variable is not available (dropouts)?

Data: BCI dataset, study variable = "Zn".

- assume that the study area is divided into a set of pixels (finite population). However, for some pixels, the value of the study variable is not available (inaccessible locations, built-up areas, etc.). We will call these "white pixels."

- assume we know the total number of pixels N but not the number of white pixels. This means that we do not know the inclusion probabilities. The SSwR book, Section 3.1, provides a way to estimate the population mean or the population total.

- using the simple random sampling without replacement, consider the following approaches to estimation of the population mean:

1) if a white pixel is sampled, value from the nearest available pixel is used instead. If more available pixels have the same distance from the sampled white pixel, you need to decide which value to use (randomly chosen value? mean of the nearest available values? something else?). This produces biased estimators, according to the SSwR book.

2) when choosing the sample, take a larger sample size using the same sampling design, creating a back-up list of units to be sampled in case some of the sampled pixels turn out to be white. This produces biased estimators, too, because a ratio estimator is used.

- using simulations, investigate the empirical bias of the first approach, depending on the desired sample size, the fraction of white pixels in the study area (you choose a method to assign the white color to some pixels), and the pixel size (resolution).

- in all cases, compute also the "better" estimates using the second approach and determine the empirical bias, too.

- please describe your experiments, summarize your findings and provide your interpretations and conclusions in a pdf report.

II. Stratified simple random sampling

Question: What is the effect of stratification with respect to a covariate (correlated with the study variable) compared to geographical stratification?

Data: BCI dataset, study variable = "Ca", covariate = "Fe".

- assume that the study area is divided into a set of pixels (finite population) and that there are no "white pixels" in it.

- use proportional allocation of sample size to strata.

- using simulations, investigate the sampling distribution (most importantly, the sampling variance) of the estimated mean using geographical stratification and stratification based on a covariate, for different sample sizes and different number of strata. Compare the results with the simple random sampling design (no stratification).

- please describe your experiments, summarize your findings and provide your interpretations and conclusions in a pdf report.

III. Systematic sampling

Question: What are the inclusion probabilities in systematic random sampling when the study area has a non-rectangular shape and/or contains "white pixels"?

Data: Voorst (from the SSwR book).

- assume that the study area is divided into a set of pixels (finite population). However, for some pixels, the value of the study variable is not available (inaccessible locations, built-up areas, etc.). We will call these "white pixels."

- consider the square sampling grid and systematic random sampling, where a primary unit is selected, which then determines all the corresponding pixels to be sampled (if they contain a value of the study variable, i.e. they are not white).

- consider two cases: A) the primary unit is sampled from all possible pixels, both white and non-white; B) the primary unit is sampled only from non-white pixels. Compare the results for both cases.

- determine the expected sample size E(n) by simulation, see Section 5.1 in the SSwR book.

- determine the inclusion probabilities \pi_k for each population unit, first by the theoretical arguments discussed in the lecture and then by simulation.

- decide if \pi_k = E(n)/N or not (perform a test for each pixel, do not forget about the multiple comparison problem).

- decide if \pi_k follow the theoretical arguments (again, perform a test for each pixel, do not forget about the multiple comparison problem).

- please describe your experiments, summarize your findings and provide your interpretations and conclusions in a pdf report.

IV. One-stage cluster random sampling

Question: Which way of sampling the clusters in one-stage cluster random sampling is the most efficient?

Data: BCI dataset, study variable = "Mg".

- assume that the study area is divided into a set of pixels (finite population). However, for some pixels, the value of the study variable is not available (inaccessible locations, built-up areas, etc.). We will call these "white pixels." You can choose the positions of the white pixels in any way you like.

- you define the clusters in some way (possibly inspired by the Voorst dataset in the SSwR book), the same for the strata of clusters.

- compare the sampling variances when estimating the population mean in the following cases:

A) clusters selected with probabilities proportional to size, without replacement (Section 6.2 of the SSwR book),

B) clusters selected by simple random sampling, without replacement (Section 6.3 of the SSwR book),

C) clusters selected by stratified cluster random sampling (Section 6.4 of the SSwR book, but in each stratum, use sampling without replacement to be consistent with the other cases),

D) simple random sampling of population units, without replacement.

- investigate the dependence of the results on the desired sample size and the fraction of white pixels.

- please describe your experiments, summarize your findings and provide your interpretations and conclusions in a pdf report.

V. Two-stage cluster random sampling

Question: Which way of sampling the clusters in two-stage cluster random sampling is the most efficient?

Data: BCI dataset, study variable = "Cu".

- assume that the study area is divided into a set of pixels (finite population). However, for some pixels, the value of the study variable is not available (inaccessible locations, built-up areas, etc.). We will call these "white pixels." You can choose the positions of the white pixels in any way you like.

- you define the clusters in some way (possibly inspired by the Voorst dataset in the SSwR book), the same for the strata of clusters.

- in the second stage (subsampling of clusters), use simple random sampling without replacement.

- compare the sampling variances when estimating the population mean in the following cases:

A) clusters selected with probabilities proportional to size, with replacement (see the formulas from the lecture and Section 7.1 of the SSwR book),

B) clusters selected by simple random sampling, without replacement (see the formulas from the lecture and Section 7.3 of the SSwR book),

C) clusters selected by stratified cluster random sampling, in each stratum select the clusters by simple random sampling without replacement (Section 7.4 of the SSwR book),

D) simple random sampling of population units, without replacement.

- investigate the dependence of the results on the desired sample size and the fraction of white pixels.

- please describe your experiments, summarize your findings and provide your interpretations and conclusions in a pdf report.

VI. Probabilities proportional to size (systematic sampling)

Question: How is the sampling variance influenced by the ordering of population units in systematic sampling with probabilities proportional to size?

Data: you can create artificial datasets for illustration or use whatever data you like (does not need to have a spatial structure).

- consider the problem of estimating the population mean.

- find examples (or construct artificial datasets) where ordering the units by size decreases the sampling variance a lot, compared to random ordering of units, and where ordering the units by size does not affect the sampling variance. Does the gain in precision depend on the sample size?

- as a compromise between the two extreme cases above (ordering all units by size vs. completely random ordering), consider the following approach. Divide the population units into two groups according to their size, i.e. group A consists of the N/2 largest units, group B consists of N/2 smallest units. To arrange all units in some order, first shuffle randomly the units in group A and then continue with randomly shuffled units from group B. Verify (by simulation and possibly some theoretical arguments) that the joint inclusion probabilities \pi_{ij} are positive in this case (and hence an unbiased estimator of the sampling variance can be developed). Investigate the sampling variance in this case and compare it to the results obtained above. Is there a noticeable gain in precision compared to a completely random ordering?

- repeat the same with three groups instead of two.

- compare your results with simple random sampling without replacement (ignoring the size variable). In which situations is using sampling with probabilities proportional to size the most beneficial compared to simple random sampling?

- please describe your experiments, summarize your findings and provide your interpretations and conclusions in a pdf report.

VII. Balanced sampling

Question: How much improvement can be achieved by including more covariates in the balancing procedure?

Data: BCI dataset, study variable = "Ca", covariates = other variables in the file.

- consider the problem of estimating the population mean of the study variable using balanced sampling.

- you have one variable of interest (the study variable) and a handful of auxiliary variables (covariates). Propose a methodology for selecting which covariates to use as balancing variables (and how many) in the given sampling problem, with the aim to minimize the sampling variance. Distinguish two cases: A) ideal situation in which we know all the values of the study variable before sampling, B) realistic situation where we do not know the values of the study variable or just a few values from a small pilot study, together with the corresponding covariate values.

- whenever useful, perform simulation experiments to help you develop the methodology.

- apply the two proposed methodologies (developed for cases A and B, respectively) to the given dataset and compare the results in terms of bias and sampling variance.

- please describe your experiments, summarize your findings and provide your interpretations and conclusions in a pdf report.