

Overview: The Estimation of Subpopulation (or Domain) Parameters in Finite Population Sampling.

In virtually all surveys interest will not be only in the estimation of overall population parameters like the population mean, total, proportion but in the corresponding parameters for subpopulations of special interest. This material is discussed in detail in Section 5.6 in Thompson.

Some examples

1. In a consumer survey one could look all consumers as well as various subpopulations of consumers. For example, consumers of different income levels or regions of the country.
2. A political survey based on a registered voter roll as the frame would have a focus on all registered voters but it might also have a focus on the subpopulation of likely voters.
3. In an angler survey on fishing expenditures the total expenditure for the whole population may be a parameter of special interest but there would also be interest in subpopulations of interest such as males and females or different age groups.

The concept of estimation of subpopulation parameters is related to but distinct from the topic of stratification that we will cover later (Chapter 11). However, *here the subpopulation sample sizes are random* whereas under stratification they are fixed by design. (This makes the concept basically identical to post stratification.).

A key result is that the distribution of n_k is *hypergeometric* with

$$E(n_k) = \frac{N_k n}{N}$$

We have to consider two situations. First is when the subpopulation sizes (N_k) are known (that is the frame contains information on the subpopulation variable say age) and the second is when we have to consider them unknown (that is the frame does not include information on the subpopulation variable). We will come back to this point a bit later on

In the case where the subpopulation sizes are unknown we replace $\frac{N_k - n_k}{N_k}$ by $\frac{N - n}{N}$ in various variance equations. This approximation comes from replacing n_k by its expectation.

In lecture I will present the various estimators with their properties. A key result used in proofs is for conditional expectations and variances and converting them to unconditional expectations and variances using the following equations.

$$E(Y) = E[E(Y|X)]$$

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)]$$

The estimator for the subpopulation total is the only one that requires any real thought and I will summarize the results for the two estimators depending on whether the subpopulation sizes are known or not in the lecture.

Frame Issues and Subpopulation Estimation

Lets go back and consider the frame in a bit more detail and its relationship to subpopulations a bit further. To run a probability sampling based survey one needs to have a frame. A frame is simply a list of all the elements of the finite population.

$$U = \{1, 2, 3, \dots, N\}$$

The key assumptions about frames that we have already discussed is that the list or frame is complete and without duplications. This may or may not be true in practice but it is

what all our theory is based on. The list information must be such that each element in our universe is uniquely identified. In human surveys this is usually the name of the individual.

I have not yet discussed what other information may or may not be included with the basic frame information (ie the identifier). This gets to the basis of why there are complexities in studying subpopulations. Sometimes a lot of information on subpopulations is available with the frame information (that is before we start sampling) whereas often the information is not available and can only be obtained when the sample is taken.

Let us suppose we want to carry out a mail survey of licensed freshwater anglers in North Carolina and that in addition to the whole population estimates we would like to study subpopulations based on age. The agency may be concerned that they are not reaching enough young people for example. The basic information on the frame would be names and addresses but there may or may not be age information on the license which is used to form the frame. Let us consider both cases in turn.

If age information had been provided on the frame we would know the sizes of the subpopulations (N_k) in advance. This profoundly affects our inference as we discussed earlier. It is especially important in the estimation of the subpopulation total. Here we can use

$$\hat{t}_k = N_k \bar{y}_k .$$

If age information had not been provided on the frame we would not know the sizes of the subpopulations (N_k) in advance. In this case we have to use the

approximation that $\frac{N_k}{n_k} \approx \frac{N}{n}$. This is especially true for the estimation of the

subpopulation total where we have to use a different estimator

$$\hat{\tau}'_k = (N/n) \sum_{i=1}^{n_k} y_{ki} .$$

Stratification and its Relationship to studying Subpopulation Inference.

If age information had been provided on the frame we would know the sizes of the subpopulations (N_k) in advance. This means that instead of taking a simple random sample from the whole population we could have taken a simple random sample from each subpopulation. This is called a *stratified random sample* and it has great advantages if we have a heterogeneous universe. We will study this kind of sampling in great detail later in the semester.

If age information had not been provided on the frame we would not know the sizes of the subpopulations (N_k) in advance. In this case we can't use stratified random sampling. However, we can employ what is called the "*post stratification*" approach. See Chapter 11.6 in Thompson.