Monday, Apr 11

Survey Sampling Designs

In **experiments** and **observational studies** the population is typically a hypothetical construct (i.e., the infinite set of all possible observations), but in **surveys** it is usually a real thing (i.e., and a finite set of all possible observations). For this reason in surveys we often define the population as the set of all things or "elements" that we observe, rather than the observations themselves. Similarly in a survey we often define the *sample* is a subset of elements from a population. The goal of the survey is to use the information in the sample (a statistic or what we call an *estimator*) to infer the value of a parameter which describes the population.

Examples:

1. *Quality control.* A warehouse contains thousands of "widgets" produced by a company. Our goal is to estimate the proportion of defective widgets in the warehouse (p) based on the number of defective widgets in a sample of a few hundred widgets.
2. *Education.* Suppose we want to know the mean test score (μ) for several hundred students at a high school. But rather than testing every student we select a *sample* of a hundred students at the school and test them.
3. *Forestry.* A forest contains hundreds of thousands of trees. We want to estimate the mean volume of all the trees in the forest (μ). To do this we select a sample of a few hundred trees and compute their volume.

The sampling design can be formally defined as *the set of all possible samples and the probability of each possible sample.* But we often describe a sampling design in terms of the *process* of selecting a sample, which then implies which samples are possible and their corresponding probabilities.

Simple Random Sampling

A **simple random sampling** design is one in which *every possible sample of n elements* has an equal probability of being selected.

Example: Suppose we have a population of $N = 5$ elements: A, B, C, D, E. The table below shows a *simple random sampling* design for a sample size of $n = 3$.

Sample	Probability
A, B, C	0.1
A, B, D	0.1
A, B, E	0.1
A, C, D	0.1
A, C, E	0.1
A, D, E	0.1
B, C, D	0.1
B, C, E	0.1
B, D, E	0.1
C, D, E	0.1

Example: Suppose we have $N = 4$ elements: Frodo, Merry, Pippin, Sam. The table below shows a *simple random sampling* design for a sample size of $n = 2$.

Sample	Probability
Frodo, Merry	1/6
Frodo, Pippin	1/6
Frodo, Sam	1/6
Merry, Pippin	1/6
Merry, Sam	1/6
Pippin, Sam	1/6

What would the *process* be for selecting a simple random sampling?

Comments about Simple Random Sampling

1. Simple random sampling is relatively *mathematically* simple. Inferences based on a sample obtained using simple random sampling are relatively straight forward. We have already learned that we can estimate μ with $\bar{x}$ and how to compute the margin of error.[1]
2. Simple random sampling is not as *efficient* as some other designs in several ways. Some sampling designs can produce *more representative samples, producing smaller margins of error*, while other sampling designs can be *cheaper and easier to implement*.
3. Simple random sampling is often used as a "building block" for what are called *complex sampling designs*. Any sampling design that is not simple random sampling is a complex sampling design. The other sampling designs discussed below are all complex sampling designs.

Stratified Random Sampling

A **stratified random sampling** design involves two steps:

1. The elements in the population are divided into two or more groups called *strata*.
2. Simple random sampling is applied to *each* strata and then the samples are combined.

Example: Suppose we have the population of $N = 7$ elements, A, B, C, D, E, F, G.

1. We divide these into two strata: the stratum A, B, C, D and the stratum E, F, G.
2. We apply simple random sampling to each strata. Suppose we select $n_1 = 3$ elements from the first stratum so that the possible samples from that stratum are as follows.

Sample	Probability
A, B, C	1/4
A, B, D	1/4
A, C, D	1/4
B, C, D	1/4

   And suppose we select $n_2 = 2$ elements from the second stratum so that the possible samples from that stratum are as follows.

   The possible combined samples and their probabilities are as follows.

   The probabilities are found by multiplying the probabilities of the samples from each strata.

Each sample has an equal chance of being selected. So why is this *not* a simple random sampling design? Hint: What are the probabilities of the samples A, B, C, D, E or A, B, C, D, F?

[1] The margin of error is $t\frac{s}{\sqrt{n}}\sqrt{1 - n/N}$ as simple random sampling is *without replacement*.

Sample	Probability
E, F	1/3
E, G	1/3
F, G	1/3

Sample	Probability
A, B, C, E, F	1/12
A, B, C, E, G	1/12
A, B, C, F, G	1/12
A, B, D, E, F	1/12
A, B, D, E, G	1/12
A, B, D, F, G	1/12
A, C, D, E, F	1/12
A, C, D, E, G	1/12
A, C, D, F, G	1/12
B, C, D, E, F	1/12
B, C, D, E, G	1/12
B, C, D, F, G	1/12

Advantages of Stratified Random Sampling

1. Administrative convenience. It may be easier to conduct several smaller simple random sampling designs than coordinate one larger simple random sampling design.
2. Interest in individual strata. The design ensures samples from *all* strata. A simple random sampling design might sample few or no elements from a stratum of interest.
3. Smaller margin of error. By assuring samples from each strata, the combined sample tends to be more representative of the population, resulting in a smaller margin of error.

Stratification and Allocation

Planning a stratified random sampling design can involve a couple of decisions that can greatly benefit the design.

1. **Stratification** is how the elements are assigned to strata.
2. **Allocation** is how we distribute the total sample size (n) over the two or more strata to determine the number of elements to sample *for each stratum* (i.e., $n_1, n_2, \ldots$).

One-Stage Cluster Sampling

A one-stage **cluster sampling** design involves two steps:

1. The elements in the population are divided into groups called *clusters*.
2. Simple random sampling is applied to select a *sample of clusters*.

How is this different from stratified random sampling?

Example: Suppose we have a population of $N = 9$ elements: A, B, C, D, E, F, G, H, I.

1. The elements are divided into four clusters: {A, B}, {C, D}, {E, F, G}, {H, I}.
2. We use simple random sampling to select a sample of *two clusters*. The possible samples and their probabilities are as follows.

Sample	Probability
A, B, C, D	1/6
A, B, E, F, G	1/6
A, B, H, I	1/6
C, D, E, F, G	1/6
C, D, H, I	1/6
E, F, G, H, I	1/6

Advantages and Disadvantages of Cluster Sampling

The advantages of cluster sampling are that (a) it can be less expensive than simple or stratified random sampling and (b) it can be used when a *sampling frame* is unavailable (a **sampling frame** is a list of all the elements in the population).

A disadvantage of cluster sampling is that the margin of error is often larger than what it would be for simple random sampling or stratified random sampling.

Two-Stage Cluster Sampling

A two-stage cluster sampling can be described as follows.

1. The elements in the population are divided into groups called *clusters*.
2. Simple random sampling is applied to select a *sample of clusters*.
3. Simple random sampling is applied to *each sampled cluster*.

Advantages and Disadvantages of Two-Stage Cluster Sampling

Two-stage cluster sampling has the same advantages and disadvantages of one-stage cluster sampling. But it often has a smaller margin of error compared to one-stage cluster sampling, because we can control *two* sample sizes: the number of clusters to sample, and the number of elements to sample from each sampled cluster.

Systematic Sampling

Systematic sampling attempts to select a sample of elements that is "spread out" (often temporally or spatially). It is technically a special kind of cluster sampling.

Advantages of Systematic Sampling

1. Sometimes useful when there is no sampling frame available.
2. Lower margin of error than simple random sampling and some cluster sampling designs.