Sampling

In most statistical applications, one *should* select a sample from a population. The idea is to infer, or make educated guesses about, the population from information in the sample.

Typically a sample is just a small portion of the population.

Census: inspection of the entire population

Why sample rather than do a census?

- Reduced cost
- Greater speed
- Greater scope: personnel and equipment needed for census may be prohibitive
- Greater accuracy (sometimes)
Assume there are $N$ units in the population.

Frame – complete list of all the units in the population

1-1 correspondence between $\{1, 2, \ldots, N\}$ and frame.

**Probability sampling**

1. One can define all the possible samples that could be taken. Label the distinct samples $S_1, S_2, \ldots, S_k$. Each $S_i$ is a subset of $\{1, 2, \ldots, N\}$.

2. Each possible sample has a given probability of being selected. Call the probability of selecting $S_i$ $P(S_i)$.

3. Sample is taken by using a random process in which the samples have probabilities $P(S_1), P(S_2), \ldots, P(S_k)$. Of course,

\[ P(S_1) + P(S_2) + \cdots + P(S_k) = 1. \]
The reason for using probability sampling is that it allows an objective assessment of the accuracy of inferences made about the population.

Example of non-probability sampling:

Judgement sample – an expert chooses “typical” or “representative” members of population

Such a method does not admit a mathematical, or probabilistic, assessment of accuracy.

Let the frame be \(\{1, 2, \ldots, N\}\), and suppose there is one measurement of interest corresponding to each item. Call these measurements \(Y_1, Y_2, \ldots, Y_N\).

Relevant population characteristics for what follows:

\[
\mu = \frac{1}{N} \sum_{i=1}^{N} Y_i \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \mu)^2
\]
\( \mu \) is the population mean and \( \sigma^2 \) the population variance.

**Simple random sampling (SRS)**

SRS is a method of selecting a sample of \( n \) distinct units such that each of the possible samples has the same chance of being chosen.

Total number of ways of selecting a sample of \( n \) distinct units is

\[
\binom{N}{n} = \frac{N!}{n!(N-n)!},
\]

where \( j! = j(j-1)(j-2)\cdots(2)(1) \) for any whole number \( j \), with \( 0! \) defined to be 1.

This number is usually enormous! For example, suppose \( N = 100 \) and \( n = 20 \).

\[
\frac{100!}{20!80!} = \frac{100(99)(98)\cdots81}{20!} = 535,983,400,000,000,000,000
\]

214
According to the probability sampling scheme mentioned previously, we would need to enumerate all \( k = \binom{N}{n} \) samples and then randomly select one.

Fortunately, this is not necessary. An equivalent way to select the sample is as follows:

- Randomly select 1 of the \( N \) units. (This means that each item in the population has probability \( 1/N \) of being selected.)

- Randomly select one of the remaining \( N - 1 \) items.

- Now we have two items in our sample. Randomly select one of the remaining \( N - 2 \) items.

- Continue in this way until we have \( n \) items in our sample.
Let \( y_1, \ldots, y_n \) denote the measurements in the sample, and suppose we are interested in estimating the population mean \( \mu \). We may estimate \( \mu \) by the sample mean

\[
\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n).
\]

In 651 you learned that the standard error of a sample mean is \( \sigma / \sqrt{n} \). This is actually a bit of a fib unless you use sampling with replacement.

The SRS scheme described on the previous page is sampling without replacement. For SRS, the actual standard error of \( \bar{y} \) is

\[
\frac{\sigma}{\sqrt{n}} \sqrt{\text{fpc}},
\]

where

\[
\text{fpc} = 1 - \frac{n - 1}{N - 1}.
\]

Fpc stands for finite population correction.

Note that if \( N = \infty \), then \( \text{fpc} = 1 \) and we get the standard error we are used to. Otherwise, \( \text{fpc} \) is smaller than 1 and adjusts for the fact that the population is finite in size.
Other methods of sampling

We'll discuss two methods: stratified random sampling (StRS) and cluster sampling.

*Stratified random sampling*

The population is divided into $L$ subpopulations or *strata*.

The strata are nonoverlapping and contain $N_1$, $N_2$, $\ldots$, $N_L$ units, respectively. So, in this case

\[ N = N_1 + N_2 + \cdots + N_L. \]

Suppose simple random samples of sizes $n_1, n_2, \ldots, n_L$ are drawn independently from the strata. This sampling procedure is known as *stratified random sampling*.

Often the sample sizes are taken to be proportional to the strata sizes. In other words, if $n$ is the total sample size, then

\[ \frac{n_i}{n} \approx \frac{N_i}{N} \quad \text{for each stratum}. \]
Reasons for using StRS

1. More precise estimates within subpopulations.

2. Administrative convenience. Field offices take samples locally.

3. Sampling problems may be different for different parts of the population.

4. Gains in precision when the strata distributions are well-separated, as pictured below:
The precision of simple random sampling is adversely affected by discrepancies between strata, while that of stratified sampling is not.

**Single-stage cluster sampling**

We have \( N \) units, or clusters, and we still want to estimate a population mean.

Each unit is divided into \( M_i \) smaller units called elements. The population consists of \( M_1 + M_2 + \cdots + M_N = M^* \) elements.

Let \( Y_{ij} \) denote the measurement associated with element \( j \) within cluster \( i \). The cluster or unit total is

\[
Y_i = Y_{i1} + Y_{i2} + \cdots + Y_{iM_i}.
\]
In a single stage cluster sample, a SRS of \( n \) clusters is chosen.

For convenience, suppose \( M_i = M \) for every cluster. We may then ask the question

“What advantage is there to taking a cluster sample of \( nM \) elements as opposed to a SRS of \( nM \) elements?”

It may be shown that such a cluster sample is never more precise (for purposes of estimating the population mean) than a SRS, and is often much less precise than SRS.

Main motivation for cluster sampling is administrative convenience. It may be much easier to obtain a frame of all clusters than of all elements.

Suppose an element is a household in a certain city. Define a cluster as a city block. We could take a SRS of \( n \) blocks, and then sample each house within each of the blocks.
**Two-stage cluster sampling**

Another common practice is to select a SRS of clusters, and instead of sampling every element within each cluster, take a SRS of elements from each of the clusters.

**Example:** Suppose a study of farms in “rural” Texas counties is being done. An element is a farm and a unit (or cluster) is a rural county. Randomly select \( n \) rural counties, and then randomly select \( M \) farms from each of the selected counties.

In this scenario, no frame of all farms is needed. We only need frames for the farms within each of the selected counties. Also, because there may be a large number of farms within each county, it is too expensive to consider all farms within a county, as in single-stage cluster sampling.