Wednesday, Aug 21

A **simple random sampling** design with a sample size of n is one in which *every possible sample of n distinct elements has the same probability of being selected.*

Example: Suppose we had the population $\mathcal{P} = \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4\}$. What is the simple random sampling design if $n = 2$?

$$
\begin{aligned}
\mathcal{S}_1 &= \{\mathcal{E}_1, \mathcal{E}_2\} & P(\mathcal{S}_1) &= 1/6 \\
\mathcal{S}_2 &= \{\mathcal{E}_1, \mathcal{E}_3\} & P(\mathcal{S}_2) &= 1/6 \\
\mathcal{S}_3 &= \{\mathcal{E}_1, \mathcal{E}_4\} & P(\mathcal{S}_3) &= 1/6 \\
\mathcal{S}_4 &= \{\mathcal{E}_2, \mathcal{E}_3\} & P(\mathcal{S}_4) &= 1/6 \\
\mathcal{S}_5 &= \{\mathcal{E}_2, \mathcal{E}_4\} & P(\mathcal{S}_5) &= 1/6 \\
\mathcal{S}_6 &= \{\mathcal{E}_3, \mathcal{E}_4\} & P(\mathcal{S}_6) &= 1/6
\end{aligned}
$$

What if $n = 3$?

$$
\begin{aligned}
\mathcal{S}_1 &= \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3\} & P(\mathcal{S}_1) &= 1/4 \\
\mathcal{S}_2 &= \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_4\} & P(\mathcal{S}_2) &= 1/4 \\
\mathcal{S}_3 &= \{\mathcal{E}_1, \mathcal{E}_3, \mathcal{E}_4\} & P(\mathcal{S}_3) &= 1/4 \\
\mathcal{S}_4 &= \{\mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4\} & P(\mathcal{S}_4) &= 1/4
\end{aligned}
$$

Why *simple* random sampling?

1. A good place to start for didactic purposes.
2. A "baseline" sampling design against which to compare "complex" sampling designs.
3. A "building block" for *complex* sampling designs.

A **complex sampling design** is essentially any probability sampling design other than simple random sampling!

Methods of Selecting a Simple Random Sample

1. Complete enumeration. Create a table of all possible samples and select one at random. Impractical unless the number of possible samples is small.

2. Draw-by-draw selection. Select this first element at random such that each element has a probability of $1/N$ of being selected. Then select the second element at random from the remaining $N - 1$ elements such that each of these has a probability of $1/(N - 1)$ of being selected. And so on.

3. Random sort. Assign to each element a random number between, say, 0 and 1. Sort the elements by the value of that random variable and select the n elements with the largest (or smallest) value of that random variable.

There are many others.

Properties of Simple Random Sampling

1. Let N denote the number of sampling units (elements) in the population. For a simple random sampling design with sample size n the number of possible samples is
$$\binom{N}{n} = \frac{N!}{n!(N-n)!}.$$
This is the number of combinations. The factorial is defined as $x! = x \times (x-1) \times (x-2) \times \cdots \times 1$ where x is any positive integer, and $0! = 1$.[1]

2. The probability of each sample is
$$P(\mathcal{S}_i) = \frac{1}{\binom{N}{n}}.$$
for every sample.

3. Simple random sampling is a type of **element sampling** meaning that the sampling units are the elements so that $\mathcal{U}_1 = \{\mathcal{E}_1\}, \mathcal{U}_2 = \{\mathcal{E}_2\}, \dots, \mathcal{U}_N = \{\mathcal{E}_N\}$.

4. Sampling is done *without replacement* meaning that each element can only appear in the sample once. If it is possible for the same element to appear in the sample more than once then we are sampling *with replacement* (more on that later).

5. Every element has the same *inclusion probability* of $\pi_i = \frac{n}{N}$.[2] The **inclusion probability** of an element, denoted as π_i, is the *probability that the element will be included within the selected sample.*

[1] Many scientific calculators and computing programs include a function to compute the number of combinations. In R the function is `choose`. For example, the number of samples of size $n = 3$ from a population of $N = 20$ for a simple random sampling design is computed as `choose(20,3)` which gives 1140.

[2] To see why consider that the number of samples that include the i-th element equals the total number of samples minus the number of samples that *exclude* the i-th element, which is
$$\binom{N}{n} - \binom{N-1}{n}.$$
Each of these samples has a probability of
$$\frac{1}{\binom{N}{n}},$$
and so the probability of observing any one of those samples is obtained by summing the probability of each of these samples across the samples that include the i-th element, or simply multiplying the two terms above together since all samples have the same probability. Doing this and simplifying by writing out the factorials and canceling-out common factors gives n/N.

Inference for Population Totals and Means

Assume simple random sampling design with a population of N elements. Let $y_1, y_2, \dots, y_N$ be the values of the target variable for elements $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_N$ respectively.

The *population total* (τ) is

$$\tau = y_1 + y_2 + \cdots + y_N = \sum_{i=1}^{N} y_i,$$

and the *population mean* (μ) is

$$\mu = \frac{y_1 + y_2 + \cdots + y_N}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

Clearly the two parameters are clearly related in that $\mu = \tau/N$ and $\tau = N\mu$.

Note: Quantities that directly describe the population of elements are sometimes called **parameters**.

To estimate μ we *might* use the sample mean which can be written as

$$\bar{y} = \frac{1}{n}\sum_{i \in \mathcal{S}} y_i,$$

where $i \in \mathcal{S}$ denotes the set of indices of the elements included in the sample (sometimes called an *index set*). An example might be $\mathcal{S} = \{2, 3, 9\}$ denoting the second, third, and ninth elements in the population.

An alternative notation assumes that we rearrange the indices so that the first n elements are the elements in the sample so that the values of the target variable for the sampled elements are $y_1, y_2, \dots, y_n$. In the example above y_2 would become y_1, y_3 would become y_2, and y_9 would become y_3. Then we can write the sample mean as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

In the example above the mean would then be computed as

$$\bar{y} = \frac{1}{3}(y_1 + y_2 + y_3).$$

An *estimator* of τ is

$$\hat{\tau} = \frac{N}{n}\sum_{i \in \mathcal{S}} y_i \quad \text{or} \quad \hat{\tau} = \frac{N}{n}\sum_{i=1}^{n} y_i,$$

depending on what system of notation we are using. This is sometimes called the “expansion estimator” because we “expand” the sum for the sample to estimate the sum for the population.[3] Note that because

$$\bar{y} = \frac{1}{n}\sum_{i \in \mathcal{S}} y_i \quad \text{or} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

then we can also write $\hat{\tau} = N\bar{y}$. One way to motivate this estimator is to use the fact that $\tau = N\mu$ and replace μ with its estimator $\bar{y}$.

Note: Quantities computed from a sample of observations are sometimes called **statistics**.

[3] Note that we would expect the ratio of the sample and population totals to approximately equal the ratio of the number of elements contributing to the total, so that

$$\frac{\sum_{i \in \mathcal{S}} y_i}{\tau} \approx \frac{n}{N} \Rightarrow \tau \approx \frac{N}{n}\sum_{i \in \mathcal{S}} y_i,$$

and so we “expand” (i.e., or maybe “inflate”) the sample total by a factor of N/n to attempt to (approximately) match the population total.

Note: It would make sense to write $\hat{\mu}$ for the estimator of μ rather than as $\bar{y}$ because we write $\hat{\tau}$ as the estimator for τ, but $\bar{y}$ is a common convention when the estimator for μ is the sample mean.

Note: In statistics we make a distinction between an **estimator** and an **estimate**. The *estimator* is basically the formula like $\hat{\tau} = \frac{N}{n} \sum_{i \in \mathcal{S}} y_i$ whereas an *estimate* is a specific value based on a sample of observations such as $\hat{\tau} = 3$.

Sampling Distributions

The **sampling distribution** of an estimator is its probability distribution (i.e., the possible estimates produced by the estimator and their probabilities).

Example: Suppose $\mathcal{P} = \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4\}$, $y_1 = 1$, $y_2 = 2$, $y_3 = 2$, and $y_4 = 7$, and assume a simple random sampling design with $n = 2$. What is the sampling distribution of $\hat{\tau}$? What if $n = 3$?