Monday, Aug 19

Introduction

What this class *is* about:

1. *Sampling design.* How can we obtain a "useful" sample of things from a population of things? How can we do this effectively while also controlling survey cost? How can we use outside information to create a design that might produce a more useful sample?
2. *Inference.* How can we characterize the "usefulness" of a particular sampling design? How do we use the information in the sample? How can we use information outside the sample to improve inferences?

This class is *not* about *measurement* (e.g., questionnaire design, disease diagnosis, quantification of foliage cover), although that is *very* important.

Example applications:

How many Hobbits living in the Shire have foot lice? (epidemiology)

What is the average lead content of pottery shards at a site? (archaeology)

What is the average algebra skills test score at a high school? (education)

What proportion of undergraduates at UI use an Android phone? (marketing)

How many otter dens are there along the cost of Scotland? (ecology)

Advantages and Disadvantages of Sampling

What are the advantages and disadvantages relative to a complete census?

Advantages:

1. reduced cost
2. faster
3. greater scope
4. greater accuracy

Disadvantages:

1. loss of accuracy
2. requires technical expertise

Elements, Sampling Units, Samples, and Populations

An **element** is the fundamental *observational unit* (i.e., the "thing" from which we get a value of a variable of interest). Let $\mathcal{E}_i$ denote the i-th element. For each element we can observe the value of a *target variable* (y_i).

Example: Suppose we have 12 elements: $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_{12}$.

A **population** is the set of *all* elements of interest. We could denote this as $\mathcal{P}$.

Example: The population is $\mathcal{P} = \{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_{12}\}$.

A **sampling unit** is a set of *one or more elements* such that if one of the elements within the sampling unit are in a sample, then all of the units within that sampling unit are in the sample (i.e., the elements in a sampling unit are always included in a sample *together*). The sampling units partition the population of units in such a way that each element appears in one and only one sampling unit. The i-th sampling unit will be denoted as $\mathcal{U}_i$.

Example: The sampling units *could be* as follows:

$$\begin{aligned}
\mathcal{U}_1 &= \{\mathcal{E}_1, \mathcal{E}_2\} \\
\mathcal{U}_2 &= \{\mathcal{E}_3\} \\
\mathcal{U}_3 &= \{\mathcal{E}_4, \mathcal{E}_5, \dots, \mathcal{E}_9\} \\
\mathcal{U}_4 &= \{\mathcal{E}_{10}, \mathcal{E}_{11}, \mathcal{E}_{12}\}
\end{aligned}$$

A **sampling frame** is a list (literal or conceptual) of all of the *sampling units* in the population.

A **sample** is a set of sampling units, and thus also a subset of the elements in a population. These are the only elements for which we observe the value of the target variable. Samples will be denoted as $\mathcal{S}$ or $\mathcal{S}_i$ if we need to refer to more than one sample.

Example: The sample

$$\mathcal{S} = \{\mathcal{U}_1, \mathcal{U}_4\} = \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_{10}, \mathcal{E}_{11}, \mathcal{E}_{12}\}$$

contains two sampling units and five elements. The sample

$$\mathcal{S} = \{\mathcal{U}_1, \mathcal{U}_2\} = \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3\}$$

contains two sampling units and three elements.

Sampling Designs

A (probability) **sampling design** is (a) the list of all possible samples and (b) the probability of each possible sample.

Example: Here is a sampling design for the population $\mathcal{P} = \{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_{12}\}$ based on the sampling units defined earlier.

$$\begin{aligned}
\mathcal{S}_1 &= \{\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3\} = \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4, \mathcal{E}_5, \mathcal{E}_6, \mathcal{E}_7, \mathcal{E}_8, \mathcal{E}_9\} \\
\mathcal{S}_2 &= \{\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_4\} = \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_{10}, \mathcal{E}_{11}, \mathcal{E}_{12}\} \\
\mathcal{S}_3 &= \{\mathcal{U}_1, \mathcal{U}_3, \mathcal{U}_4\} = \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_4, \mathcal{E}_5, \mathcal{E}_6, \mathcal{E}_7, \mathcal{E}_8, \mathcal{E}_9, \mathcal{E}_{10}, \mathcal{E}_{11}, \mathcal{E}_{12}\} \\
\mathcal{S}_4 &= \{\mathcal{U}_2, \mathcal{U}_3, \mathcal{U}_4\} = \{\mathcal{E}_3, \mathcal{E}_4, \mathcal{E}_5, \mathcal{E}_6, \mathcal{E}_7, \mathcal{E}_8, \mathcal{E}_9, \mathcal{E}_{10}, \mathcal{E}_{11}, \mathcal{E}_{12}\}
\end{aligned}$$

$$\begin{aligned}
P(\mathcal{S}_1) &= 0.2 \\
P(\mathcal{S}_2) &= 0.3 \\
P(\mathcal{S}_3) &= 0.4 \\
P(\mathcal{S}_4) &= 0.1
\end{aligned}$$

Probability Sampling

Why *probability sampling*?

1. To help avoid intentional or unintentional biases in our inferences from the sample to the population.
2. Probability is mathematically tractable — we may not know what the sample will be, but we can speak volumes to how certain (or uncertain) we can be about what it *might* be.