CP Statistics Chapter 4 – Designing Studies

4.1: Surveys and Samples

Population, Census and Sample

- The **population** in a statistical study is the entire group of individuals we want information about. For example, all registered voters in a given county.
- A **census** collects data from every individual in the population.
- A **sample** is a subset of individuals in the population from which we actually collect data.

Bias

The design of a statistical study shows **bias** if it would consistently underestimate or consistently overestimate the value you want to know.

Convenience Sampling

A **convenience sample** chooses the individuals easiest to reach. This will typically result in a biased sample of like-minded individuals.

Voluntary Response Sample

A **voluntary response** sample consists of people who choose themselves by responding to a general invitation. Voluntary response samples show bias because people with strong opinions (often in the same direction) are most likely to respond.

Simple Random Sample

A **simple random sample** (*SRS*) of size *n* consists of *n* individuals from the population chosen in such a way that every set of *n* individuals has an equal chance to be the sample actually selected. An SRS is the most basic form of random sampling.

Random Digits

A **table of random digits** is a long string of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with these two properties:

1. Each entry in the table is equally likely to be any of the 10 digits 0 through 9.
2. The entries are independent of each other. That is, knowledge of one part of the table gives no information about any other part.

Choosing an SRS

Choose an **SRS** in two steps:

Step 1: Label. Assign a numerical label to every individual in the population.

Step 2: Table. Use Table B to select labels at random.

Stratified Random Sample

To get a **stratified random sample**, start by classifying the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the sample.

Systematic Random Sample

A **systematic random sample** requires an ordered list of the sampling frame (say alphabetical) and involves three steps:

1. Determine the sampling interval, which is symbolized by "*k*" (it is the population size "*N*" divided by the desired sample size "*n*"). So $k = N / n$.
2. Randomly select a number between 1 and *k*, and include that person in your sample.
3. Also, include each *kth* element in your sample.

For example, suppose you wish to survey 100 people out of a population of 1000. So *k* would be 1000 / 100 = 10. Now you get a random number between 1 and 10 (say it was 7). So your sample starts with person #7 and then every 10th person after that: 7, 17, 27, 37, 47, etc....

Under this systematic method, when you get to the end of your sampling frame you will have all the people to be included in your sample.

Cluster Sample

To get a **cluster sample**, start by classifying the population into groups of individuals that are located near each other, called *clusters*. Then choose an SRS of the **clusters**. All individuals in the chosen clusters are included in the sample.

Forms of Bias in Surveys and Samples

- **Undercoverage** occurs when some members of the population cannot be chosen in a sample.
- **Nonresponse** occurs when an individual chosen for the sample can't be contacted or refuses to participate.
- A systematic pattern of incorrect responses in a sample survey leads to **response bias**.
- The **wording of questions** is the most important influence on the answers given to a sample survey.

4.2: Experiments

Observational Study vs Experiment

- An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses.
- An **experiment** deliberately imposes some treatment on individuals to measure their responses.
- When our goal is to understand cause and effect, experiments are the *only* source of fully convincing data. The distinction between observational study and experiment is one of the most important in statistics.

Confounding occurs when two variables are associated in such a way that their effects on a response variable cannot be distinguished from each other. Observational studies often fail to provide valid causal links between variables due to confounding

The Language of Experiments

A specific condition applied to the individuals in an experiment is called a **treatment**. If an experiment has several explanatory variables, a treatment is a combination of specific values of these variables.

The **experimental units** are the smallest collection of individuals to which treatments are applied. When the units are human beings, they often are called **subjects**.

Principles of Experimental Design

The basic principles for designing experiments are as follows:

1. **Comparison**. Use a design that compares two or more treatments.
2. **Random assignment**. Use chance to assign experimental units to treatments. Doing so helps create roughly equivalent groups of experimental units by balancing the effects of other variables among the treatment groups.
3. **Control**. Keep other variables that might affect the response the same for all groups.
4. **Replication**. Use enough experimental units in each group so that any differences in the effects of the treatments can be distinguished from chance differences between the groups.

Completely Randomized Design

- In a **completely randomized design**, the treatments are assigned to all the experimental units completely by chance.
- Some experiments may include a **control group** that receives an inactive treatment or an existing baseline treatment.
- The response to a dummy treatment is called the **placebo effect**.
- In a **double-blind experiment**, neither the subjects nor those who interact with them and measure the response variable know which treatment a subject received.

Block Design

A **block** is a group of experimental units that are known before the experiment to be similar in some way that is expected to affect the response to the treatments.
In a **randomized block design**, the random assignment of experimental units to treatments is carried out separately within each block.

Matched Pairs Design

- A **matched pairs design** is a randomized blocked experiment in which each block consists of a matching pair of similar experimental units.
- Chance is used to determine which unit in each pair gets each treatment.
- Sometimes, a "pair" in a matched-pairs design consists of a single unit that receives both treatments. Since the order of the treatments can influence the response, chance is used to determine with treatment is applied first for each unit.