Chapter 8. Producing Data: Sampling

Note. Every statistical test involves some type of sampling of a population. This chapter discusses proper and improper ways to produce a sample.

Observation versus Experiment

Definition. An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses. The purpose of an observational study is to describe some group or situation. An **experiment** deliberately imposes some treatment on individuals in order to observe their responses. The purpose of an experiment is to study whether the treatment causes a change in the response.

Definition. Two variables are **confounded** when their effects on a response variable cannot be distinguish from each other.

Example. Exercise 8.2 page 192.

Sampling

Definition. The **population** in a statistical study is the entire group of individuals about which we want information. A **sample** is a part of the population from which we actually collect information. We use a sample to draw conclusions about the entire population. A **sampling design** describes exactly how to choose a sample from the population.

Example S.8.1. Hey Moe!
A Stoogeologist wonders if Moe hit Curly more than he hit Shemp. He does not have time to watch all of the 97 Moe-Larry-Curly films and the 77 Moe-Larry-Shemp films. He decides to watch 10 Moe-Larry-Curly films and 10 Moe-Larry-Shemp films and to count the number of times Moe hits either Curly or Shemp per film. What are the populations? What are the samples? What might the sampling design be?

How to Sample Badly

Definition. A sample selected by taking the members of the population that are easiest to reach is called a **convenience sample**.

Example S.8.2. Bad Slap Count.
In Example S.8.1, the Stoogeologist goes to his DVD collection and chooses to watch the first 10 episodes he can find from each population. This is an example of a convenience sample.

Definition. The design of a statistical study is **biased** if it systematically favors certain outcomes.

Note. A common type of biased sample is one in which people with opinions call in or log on to register a response.

Definition. A **voluntary response sample** consists of people who choose themselves by responding to a broad appeal. Voluntary response samples are biased because people with strong opinions are most likely to respond.

Simple Random Samples

Definition. 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.

Note. All of the inference done in this part of the book depends on the randomness of the samples we use. In order to produce a random sample, we must have some type of random number generator. The text describes the use of a table of random digits. Table B on page 686 is such a table.

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

Note. Table B on page 686 contains randomly generated digits. The digits appear in blocks of five to make it easier to read.

Note. Using Table B to Choose a SRS.

Step 1. Label. Give each member of the population a numerical label of the *same length*.

Step 2. Table. To choose a SRS, read from Table B successive groups of digits of the length you used as labels. Your sample contains the individuals whose labels you find in the table.

Two things can go wrong in Step 2: (1) There may be a number produced by Table B which is not a label in the population, and (2) Table B may produce the same label of a member of the population more than once. In both cases, simply ignore these numbers.

Example S.8.3. Simple Random Slaps.

Use Table B to generate a simple random sample of size 10 for the 97 Moe-Larry-Curly films.

Solution. Label the films as $01, 02, 03, \ldots, 97$. Notice that we must label the first nine with a starting digit of '0' to keep label length two for all members of the population. We can label the films chronologically (though any labeling scheme will be sufficient). So we have: 01 = Women Haters, 02 = Punch Drunk, 03 = Men in Black, . . . , 97 = Half Wits' Holiday. Now we read numbers from Table B two at a time. If we start in the first row (labeled 101) then we get: 19 22 39 50 34 05 75 62 87 13. Therefore our simple random sample would be:

05	Horses' Collars
13	Movie Maniacs
19	Slippery Silks
22	Three Dumb Clucks
34	Three Missing Links
39	Yes, We Have No Bonanza
50	No Census, No Feeling
62	What's The Matador?
75	Phoney Express
87	Micro-Phonies

Example S.8.3. Simple Random Slaps 2.
Use an online random number generator to produce a simple random sample of size 10 for the 77 Moe-Larry-Shemp films. Notice that it is easy to find online lists of the Stooges' films which are numbered chronologically. This means that the first Shemp film is labeled 98 (Fright Night) and the last one is labeled 174 (Commotion on the Ocean). It would be easiest to find a random number generator that allows you to randomly choose 10 numbers from 98 to 174.

Other Sampling Designs

Definition. A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.

Definition. To select a **stratified random sample**, first classify the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

Example S.8.4. Stratified Stooges.
Suppose a Stoogeologist wants to know how many slaps per film there are in the Three Stooges' films. She knows that there are three "generations" of Stooges' films: Moe-Larry-Curly films (97 films), Moe-Larry-Shemp (77 films), and Moe-Larry-Joe (16 films). Describe how she might take a stratified random sample from this population.

Solution. The three strata would be the three "generations" of films. She could take a simple random sample from each strata. However, since the strata are different sizes, the samples should reflect this difference. For example, she might take a SRS of size 10 from the Moe-Larry-Curly strata, a SRS of size 8 from the Moe-Larry-Shemp strata, and a SRS of size 2 from the Moe-Larry-Joe strata.

Note. A **multistage sample** is a stratified sample in which the strata themselves are divided.

Cautions about Sample Surveys

Note. Many statistical studies are the results of surveys. Some of you may have to create and analyze a survey during your academic career. As faculty members, we occasionally get such surveys from students (undergraduate and graduate). There are potential problems with surveys, though.

Note. Some things to remember when surveying are:

1. **Undercoverage** occurs when some groups in the population are left out of the process of choosing the sample.
2. **Nonresponse** occurs when an individual chosen for the sample can't be contacted or refuses to participate.
3. **Response bias** is the result of someone filling out a survey (or in an interview) who gives a response that they may think the surveyor (or interviewer) want to hear instead of answering truthfully.

4. The **wording of questions** is the most important influence on the answers given to a sample survey.

Example. Exercise 8.15 page 204.

Inference about the Population

Note. We are ultimately interested in drawing conclusions about a population based on a sample. That is, we are interested in *inferring* a population parameter based on the corresponding parameter in the sample. We will explore this much more quantitatively later, but for now we comment that larger random samples give more accurate results than smaller samples.

Example. Exercise 8.50a page 212.

rbg-12-28-2008