13. Sampling

One of the first steps in research design to define the population you wish to study. There are then two possible strategies for learning about this population:

- **Census** – gather complete information on everyone; or
- **Sample** – gather information on only part of the population.

Most of the time you will necessarily gather a sample. Because a sample is *incomplete information*, your data and your conclusions will necessarily be subject to error.

This unit is about sampling. The incomplete information contained in a sample means that there will be *error*. Thus reducing *systematic error* or *bias* becomes critical.
Sampling has two competing goals:

- Reduce costs; and
- Minimize error.

Error is minimized by:

- making the sample as representative of the population as possible and hence reducing *bias* and
- increasing sample size (the *replication* principal of experimental design.

There are a number of different kinds of samples. Some of the various types are:

- Convenience Samples;
- Self-selected Samples;
- Simple random samples;
- Systematic Samples;
- Stratified random samples;
- cluster samples;
- multistage samples.

The first two are inexpensive and easy to produce. They are also very prone to error. Self-selected samples are especially prone to bias, since you tend to sample the respondents who have strong motivations to participate.

Systematic samples often sound appealing, but also can lead in unexpected ways to error. For example, suppose you were to conduct an “occupancy survey" of commercial properties in a particular neighborhood based on sampling, say, every tenth lot. It is possible that every tenth lot could turn out to be a corner lot, hence more desirable and less likely to be vacant. This would result in *systematic error* or *bias* in your sample.

Simple Random Samples.

Because samples necessarily contain error, the goal of the research is minimize that error. Random samples are a technique designed to reduce bias in the sample.

In a random sample, every member of the population has an equally likely chance of being a selected for the sample.

The random character of the error introduced means that it is not *systematic* and hence the resulting error (which must necessarily be present since a sample has incomplete information) will not be *systematic error*, i.e., the sample will not be *biased*.

Simple Random Samples. In a simple random sample, every member has an equally likely chance of being selected for the sample – much like a huge lottery. In order to construct a simple random sample, you must:

- Construct a list of all members of the population;
- Randomly select members from the list.

To do the random selection, think of rolling dice or flipping a coin for each member on the list to decide if they are in the sample or not. That way, whether or not any particular member of the population has an equally likely chance of “winning" the roll and being in the sample. Tools like Excel include random number generators that let you assign random

numbers to each member of the population to facilitate sampling.

Selecting a sample based on who happens to be available at the time of collection is not a random sample. It’s a convenience sample and is susceptible to bias.

Notice that your sample is only as good as your listing of the population. If your list is biased, then so is your sample. Generating the list can be expensive, difficult, and sometimes even impossible.

Non-respondents. Some people who are selected to participate may choose not to do so. You need to have at least a 50% response rate in order to use your sample. You should strive for an 80% or higher response rate. The Nielson organization (the company that does television ratings) routinely gets response rates of over 95%

Stratified Random Samples.

Because a simple random sample is often difficult to construct, another strategy is called the *stratified random sample*. This technique tries to make the sample representative of the Population by identifying traits important to the study and assuring that the sample and the population have a similar distribution of those traits.

The technique is similar to blocking, except that blocking is systematically applied after the sample is selected; stratification is applied as part of the sample design prior to selecting the sample.

When to Stratify. Sometimes you can identify attributes that are important to the response you are studying.

If you are studying voter preferences, then, for example, political affiliation is an important attribute of a subject. You would want your sample to have about the same proportion of Democrats, Republicans and Independents as the population you were studying.

Random sampling might result, by chance, in more of one party or another in the sample. This would result in bias. Stratified sampling is a strategy to avoid this.

Steps in Stratified Random Sampling.

1. Divide the population into Strata which are homogeneous with respect to attributes important to your study. (Don't use shoe size as a stratum in a presidential preference poll!)
2. Do a random sample from each stratum.
3. Pool the strata together to obtain the overall sample.

13.1. Example.

Suppose you are doing an opinion poll in Oklahoma and that you want to use gender and ethnicity as strata. Since there are two genders and six ethnicities (White, Black, Asian, Hispanic, Indian and Other) this results in $2 \times 6 = 12$ strata. Suppose in addition that you want a sample of size 700. The problem is to find the "right" number for each strata to make the

sample representative of the population:

Cell Distributions		
	Male	*Female*
White		
African-American		
Asian-American		
Hispanic		
Native American		
Other		
Total	*350*	*350*

Here is the strata distribution in the Oklahoma Population:

Cell Distributions		
	Male	*Female*
White	*29%*	*29%*
African-American	*7%*	*7%*
Asian-American	*2%*	*2%*
Hispanic	*4%*	*4%*
Native American	*7%*	*7%*
Other	*1%*	*1%*

Solution.

Cell Computations Compute the cell sizes according to the population percentages (recall the sample was to be 700).

For example,

$$\begin{aligned}\text{\# of white males} &= 29\% \times 700\\ &= 203\end{aligned}$$

or

$$\begin{aligned}\text{\# of Hispanic females} &= 4\% \times 700\\ &= 28\end{aligned}$$

Repeating these computations for each cell gives sample sizes for each stratum:

Cell Distributions		
	Male	Female
White	203	203
African-American	49	49
Asian-American	14	14
Hispanic	28	28
Native American	49	49
Other	7	7
Total	350	350

∎

Sampling Fractions. This example assumes *proportionate sampling Fractions*. Unequal sampling fractions can also be done, although then the computations the sample statistics are slightly more complex. For example, means are then computed for each strata and a weighted av-

erage of the strata means – using the population proportions – is computed.

Unequal sampling fractions are sometimes more appropriate. For example, there are only 4 female admirals in the US Navy, so a stratified random sample would be unreasonable.

Other Considerations Suppose the Oklahoma researchers added three income levels to their study:

- low (say under $25K/year);
- middle (say $25001-$75000 / year); and
- high (say higher than $75K/year).

This results in $2 \times 6 \times 3 = 36$ strata, clearly a more complex sampling problem. To preserve confidentiality, each cell in the sample should have at least three subjects. All of this could result in large samples – often beyond a reasonable budget.

Summary.

> In stratified random sampling the heterogeneity of the sample is obtained by combining internally homogeneous strata.

This approach is time consuming and expensive, leading a search for other alternatives

Cluster Sampling.

13.2. Example.

Suppose you are doing a dietary survey of fourth graders in Oklahoma. Some control variables might be

- *gender of the child;*
- *family income;*
- *residency (urban, suburban or rural);*

- *ethnicity.*

A stratified random sample using these criteria is probably not possible – even getting a list of all fourth graders in Oklahoma might not be possible.

If you think of the child as the "sampling unit," then this problem is difficult. There is another possible “sampling unit:”

- The school might be a sampling unit!

It would certainly be possible to obtain a list of all schools in Oklahoma. One could then randomly select from this list of "clusters," then do a census in each school.

With this technique, *the heterogeneity of the sample is obtained by combining internally heterogeneous clusters.*

Multistage Sampling.

This more modern approach combines stratified random sampling and cluster sampling techniques. For example, in our dietary survey we might:

> Stratify the schools (urban, suburban, rural) before sampling students within in each school. When sampling the students within the schools, we could further stratify according to income, ethnicity and gender.

At the school level this latter stratification is more manageable. This is how modern polling organizations operate.

More Examples.

13.3. Example.

In 1936 the American Mercury Magazine polled over 10,000 household on the presidential election that year. They used a carefully selected random sample based on phone book listings. The poll predicted a landslide victory for Alf Landon.
Of course, Franklin Roosevelt won by the largest margin in US history up to that time.
What did they do wrong?

13.4. Example.

The Hite Report. Shere Hite mailed out over 100,000 surveys to a very carefully designed stratified random sample of US women. When she analyzed her nearly 9,000 replies she found stunning results. For example,

- *70% reported extra-marital affairs;*
- *80% reported their marriages were a mistake.*

Other researchers were unable to repeat these results. Why?

Conclusions
- Sampling inevitably results in error.
- Sampling techniques are designed to minimize bias by making the sample as representative of the population as possible.
- Sampling has two competing goals – minimizing costs and maximizing accuracy.
- Poorly designed samples are the cause of many flawed research conclusions.