Data Collection

Reading

Sections 1.3, 1.4, 1.5

Practice Problems

1.9.4 (Page 60) 1.17, 1.18, 1.20, 1.26, 1.27

1.9.5 (Page 63) 1.31, 1.32, 1.33, 1.34, 1.35, 1.36

Notes

Data Collection

The proper collection of data is as important as its analysis. We discuss in this section the various questions related to data collection.

Population Every research question invariably refers to a target population. This population is typically massive and collecting answers from all individuals in the population is infeasible.

Sample Most of the times a sample is selected from the population, and data is collected on that sample. Some of the questions that arise in this context are:

1. What can we infer for the entire population from the information in our sample?
2. How should the sample be selected and data from the sample collected, and how does that affect our further work?
3. What effect does the size of the sample have on our inferences?
4. What biases are inherent in our sampling method and how can we minimize their effect?

The selection of a sample is crucial. For instance many years ago a national poll to determine who would win the elections provided completely wrong estimates, because they targeted a sample (those individuals that were well off during a depression) that was not representative of the population as a whole.

> No amount of clever analysis can compensate for badly collected data.
>
> To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.
>
> Sir Ronald A. Fisher

There are broadly speaking four different sources of data:

Anecdotal Evidence The worst kind of evidence is that collected haphazardly based on the information we immediately have access to. Reading or hearing a statement one person made and using it to extrapolate that everyone would be of the same opinion, is an example of anecdotal evidence.

Survey A more controlled system for collecting information involves surveying a representative sample from a population. The actual survey can take various forms, from directly asking them questions to recording their behavior as they navigate a web site.

Observational Study Observational studies follow a number of individuals, recording information about them, often over an extended period of time. Typically in observational studies there is little direct involvement with and control over the subjects.

Observational studies come in two forms:

1. **Prospective studies** identify individuals and collect information about them as events unfold.
2. **Retrospective studies** on the other hand identify and collect information after the events have occurred, e.g. by examining medical records.

Controlled Experiment The golden standard of data collection, and not always practical. In a controlled experiment, the researcher directly sets the values for the predictor variables, and measures the response. For example if we wanted to investigate whether smoking causes cancer, a very unethical way to do this would be via a controlled experiment where we force people to smoke for extended periods of time and witness whether they develop lung cancer or not.

Controlled Experiments differ from Observational Studies in that they are more intrusive, controlling the experiment parameters rather than simply monitoring.

Sampling

Sampling from a population is a very common approach to collecting information. It however contains many risks. The various risks can be summarized by saying that the end result is to select a sample that is not representative of the population as a whole. And if the sample does not represent the population, then there is very little we can do to draw reliable conclusions.

A key step in ensuring a proper sample is to start by selecting a **random sample**, by which we mean that every individual in the population has an equal chance of being selected for our sample. This is a great start as it ensures for example that the researcher doesn't just contact people they know, who are likely to share the researcher's views and therefore not really represent the population.

Even within the question of sampling however, there are a couple of different possibilities:

Simple Random Sample The simplest of random sampling methods, we simply pick a number of individuals from the whole in such a way that everyone is equally likely to be selected (e.g. by essentially assigning to each person a number, then drawing numbers at random).

Stratified Random Sample A stratified random sample aims to ensure that we have sufficient represntation from the various geographical locations or groups that we study. The goal in general is to group together in the same **stratum** cases that are in some way similar.

For example if we wanted to investigate the difference between greeks and non-greeks, we could consider dividing our population into 4 **strata**: male greeks, female greeks, male non-greeks and female non-greeks. Then we draw simple random samples within each stratum, and this ensures sufficient representation of all the different groups. This may however result in non-equal representation overall: We may end up with just as many male greeks as male non-greeks, even though perhaps there are very few male non-greeks.

Cluster Sample In a cluster sample we divide our population into lots of small clusters, then select some clusters at random and sample everyone in those clusters. For example in order to do a demographic study, we could choose at random some counties around the country, then collect data on (almost) everyone living in that county.

The difference from stratified samples is that we expect individuals within clusters to be different from each other, but that clusters overall are similar to each other. In stratified sampling on the other hand, we expect the individuals within a stratum to be similar to each other, but that strata would be different from each other.

Multistage Sample This is an extension of the cluster sampling method. After we divide the entire population in clusters and randomly select some clusters, we further randomly select a number of individuals within each cluster, rather than sampling everyone within the cluster.

Here is a brief summary of the methods:

Stratified Random Sample
- Individuals within a stratum similar to each other.
- Sampling from *each* stratum.
- Random sample within each stratum.

Cluster Sample
- Clusters similar to each other.
- Random sample of the clusters.
- Sampling *everyone* in the selected clusters.

Multistage Sample
- Random sample of the clusters.
- Random sample within each selected cluster.

Principles of Experiment Design

As there is a greater degree of control on experiments, a number of principles have emerged to ensure proper experimentation and minimize biases.

Controlling In order to measure the effect of a treatment or intervention, we also require a group that did not receive that treatment. This is called a **control group**, as opposed to the **treatment group**. Researchers must aim for the two groups to be as similiar in all other respects as possible. For example putting all males in one group and females in the other group would result in a useless experiment: We would have no way of knowing if the desired effect was due to the treatment or due to the person's gender.

Randomization An effort must be made to randomize the assignment of individuals in groups. This again serves to make the treatment group and control group as similar as possible, without introducing unconcious biases of the researcher.

Replication The experiment must be repeated on multiple individuals, in order to be able to accurately estimate the treatment effect and to differentiate it from chance occurence.

Blocking In many cases there are key variables (like gender) that the researcher expects might have an effect on the response. In that case it is important to first separate the individuals in blocks according to those variables, then produce random samples and control groups within each block. For example a study on a new drug might first group patient into a low risk block and a high risk block.