2 Data collection

Data collection is one of the most important parts of engineering statistics. If collected properly, data can make formal inferences easy to complete and easy to understand. On the other hand, if data is collected poorly, it can become nearly impossible to salvage a badly designed study and gain insights.

This chapter covers the general principles of data collection and effective experimentation.

2.1 Sampling

Q: The most common question engineers ask about data collection is

A: The answer depends on the variation in response that one expects.

Often we want to answer a question (conduct a study) about an identifyable, concrete population of items, but we want to use a **sample** to represent this (typically) much larger population.

Why?

Example 2.1. Measuring some characteristics of a sample of 20 electrical components (note: this is one sample with 20 units; the sample size is 20) from an incoming lot of 200.

If a sample is to be used to stand for a population, how that sample is chosen becomes very important.

A sample should

2.1.1 Systematic and judgement based methods

Definition 2.1. In *systematic sampling*, create a list of every member of the population. From the list, randomly select the first sample element from the first k elements on the population list. Thereafter, we select every k^{th} element on the list.

Disadvantage:

Definition 2.2. In *judgement-based sampling*, select based on the opinion of an expert.

Disadvantage:

2.1.2 Simple random sampling

Definition 2.3. A *simple random sample of size n* from a population is a sample selected in such a manner that every collection of n items in the population is a priori equally likely to compose the sample.

Example 2.2. A statistics instructor wanted to know how many hours per week her students spend watching cat videos on YouTube. Rather than asking each one of them, she puts all of their names in a hat and draws out 10. This is a simple random sample of size 10.

Steps to randomly sample mechanically:

1. Let M be the number of digits in the number N, where N is the population size.
2. Give each member of the population an M-digit label.
3. Move through the table of random digits from left to right, top to bottom, selecting population members for the sample when you encounter their indices (ignoring indices that have already been chosen) until you have selected n units for the sample.

Table B.1
Random Digits

12159	66144	05091	13446	45653	13684	66024	91410	51351	22772
30156	90519	95785	47544	66735	35754	11088	67310	19720	08379
59069	01722	53338	41942	65118	71236	01932	70343	25812	62275
54107	58081	82470	59407	13475	95872	16268	78436	39251	64247
99681	81295	06315	28212	45029	57701	96327	85436	33614	29070

Example 2.3. Take a simple random sample of 12 units of pig iron out of a shipment of 90 units.

Alternatively: Use a computer.

2.2 Effective experimentation

Purposefully changing a system and observing what happens as a result is a principled way of learning how a system works.

A typical experimental situation:

Example 2.4 (Chemical purity)**.** Suppose you want to know about the effect of two different reactants (A and B) on the purity of a chemical for a given mixing speed and batch size. Reactant A has 2 levels (a_1 and a_2) and reactant B also has 2 levels (b_1 and $b2$).

2.2.1 Taxonomy of variables

Planning an experiment is complicated. There are typically many different charcteristics of the system an engineer is interested in improving and many variables that might influence them. Some terminology is needed.

Definition 2.4. A *response variable* in an experiment is one that is monitored as characterizing system performance/behavior.

Definition 2.5. A *supervised (or managed) variable* in an experiment is one over which an investigator exercises power, choosing a setting or settings for use in the study. When a supervised variable is held constant (has only one setting), it is called a *control variable*. When a supervised variable is given several settings in a study, it is called an *experimental variable*.

Definition 2.6. A *concomitant (or accompanying) variable* in an experiment is one that is observed but is neither a primary response variable nor a managed variable. Such a variable can change in relation to either experimental or unobserved causes and may or may not itself have an impact on a response variable.

Example 2.5 (Chemical purity, cont'd)**.** What are the response variables, controlled variables, experimental variables, and concomitant variables?

Example 2.6 (Wood joint strength, pg. 39)**.** Dimond and Dix experimented with three different woods and three different glues, investigating joint strength properties. Their primary interest was the effects of wood type and glue type on joint strength in a tension test and joint strength in a shear test. In addition, they found that the strengths were probably related to the variables drying time and pressure, so they hold these two variables constant. They also observed that variation in strengths could also have originated in properties of the particular specimens glued, such as moisture content although they haven't utilized this variable in the analysis of the data.

What is a full/complete factorial study for this experiment? What are the response variables, controlled variables, experimental variables, and concomitant variables?

2.2.2 Extraneous variables

Definition 2.7. *Extraneous variables* are undesirable variables that influence the relationship between the variables that an experimenter is examining. Extraneous variables that vary with the levels of the independent variable are the most dangerous type in terms of challenging the validity of experimental results. These types of extraneous variables have a special name, *confounding variables*.

There are three basic ways to handle extraneous variable:

1.

2.

3.

Definition 2.8. A *block* of experimental units, experiental times of observation, experimental conditions, etc. is a homogeneous group within which different levels of primary experimental variables can be applied and compared in a relatively uniform environment.

Definition 2.9. *Randomization* is the use of a randomizing device at some point where experimental protocol is not already dictated by the specification of the supervised variables. Often it means that assigning experimental units to the experimental conditions at random.

Example 2.7 (Chemical purity, cont'd)**.** Assume time of day is an extraneous variable.

Completely randomized design:

Randomized complete block design:

2.2.3 Some key issues of data collection

1. Comparative study

2. Replication

3. Allocation of resources