CHAPTER 11: REDUCING THE ERROR, A TEMPLE FOR IDEAL DATA

EVIDENCE

11

You presumably wouldn't mistake a junker car for a Porsche. But can you tell junker data from Porsche data? The quality of the data has a huge impact on the conclusions you can draw.

SECTION 1

Introduction

Unless your job requires you to design experiments, you may rarely have the opportunity to gather data using the ideas we discuss in this chapter. Hence you may view this exercise as being pointless. But even if you don't use these methods, others that manipulate your behavior and opinion do. Advertising agencies design experiments using these rules to determine which of several ads is most effective in manipulating your behavior, and politicians use polls to adjust their positions. Furthermore, the ideas we discuss should be used (but often aren't used) to gather scientific evidence introduced in rape and murder trials, and to evaluate the safety and effectiveness of new drugs. In short, even if you never conduct an experiment using the principles we discuss, your life has been and will continue to be influenced by these principles. Because data are so important, your knowledge of the quality and limitations of data is also important.

There are five features that affect the accuracy of data:

THE IDEAL DATA TEMPLATE
REPLICATES
STANDARDS
RANDOM
BLIND
EXPLICIT PROCEDURES

Our goal in this chapter is to explain each of these features briefly in the context of an example. The goal in this example is the hypothetical one of assessing the fraction of the Texas A&M student body whose blood-alcohol levels exceed the legal intoxication threshold the night of the A&M-Texas football game (at midnight). In particular, we want to know how to avoid errors in making this measurement. Each of the four features will be shown to be relevant.

SECTION 2

Replicates

What are replicates?	Replicates are multiple observations taken under similar conditions.
Why use replicates?	They reduce sampling error, and reduce or allow detection of some human and technical error.
When to use replicates?	Any time variation is expected to arise from these 2 kinds of errors.

The most basic requirement of any data set is that the data be replicated -- doing things more than once. Replication includes any of the following-- taking the same measurement more than once, using more than one subject, using multiple groups, undertaking multiple studies.

Our hypothetical sampling of A&M students for intoxication levels should be based on a large number of students. If we surveyed only 10 students, then we could expect to be no closer than within 5% of the true value (precision error), and because of sampling error, we could be even further off. With 100 students, we could expect to get much closer to the true value, and 1000 students would get us still closer.

As hinted at above, replication applies to many aspects of a study. Consider a study to test the effectiveness of a medication. Replication can be achieved by enlisting several patients, but replication can also be achieved by testing different batches of the same medication, by performing the study at different times of the year and in different years, and so on. For reasons that are not always clear, results of two replicates of the entire study do not always agree, so multiple levels of replication are usually required before we fully accept any result. In our A&M example, we might get different results from year to year, depending on the outcome of the game and attitudes in College Station about student drinking. No matter how much a study is replicated, however, there are always countless ways in which it is not replicated.

SECTION 3

Standards

What are standards?	They are observations offering a known point of comparison for a measurement.
Why use standards?	To verify a measurement and thereby detect or avoid technical and human errors (e.g., to establish that a machine or person is working correctly).
When to use standards?	To verify a measurement; whenever there is any reasonable possibility of human or technical error.

The most basic requirement of any data set is that the data be replicated -- doing things more than once. Replication includes any of the following-- taking the same measurement more than once, using more than one subject, using multiple groups, undertaking multiple studies.

Our hypothetical sampling of A&M students for intoxication levels should be based on a large number of students. If we surveyed only 10 students, then we could expect to be no closer than within 5% of the true value (precision error), and because of sampling error, we could be even further off. With 100 students, we could expect to get much closer to the true value, and 1000 students would get us still closer.

As hinted at above, replication applies to many aspects of a study. Consider a study to test the effectiveness of a medication. Replication can be achieved by enlisting several patients, but replication can also be achieved by testing different batches of the same medication, by performing the study at different times of the year and in different years, and so on. For reasons that are not always clear, results of two replicates of the entire study do not always agree, so multiple levels of replication are usually required before we fully accept any result. In our A&M example, we might get different results from year to year, depending on the outcome of the game and attitudes in College Station about student drinking. No matter how much a study is replicated, however, there are always countless ways in which it is not replicated.

The standards needed to verify measurements taken in a variety of settings.

STANDARD	UNKNOWN
Weight of a known mass on a scale	Weight of any other object on the same scale
Thermometer reading of boiling water at sea level	Thermometer reading of other substances
The reading of a sober person on a breathalyzer test	The reading of a suspected drunk on the test

A **proficiency test** is a test involving standards. A proficiency test is simply the submission of known samples (standards) to an individual or agency that takes measurements. A proficiency test enables one to measure the error rate in data.

Reference databases also represent standards for a population. Strictly speaking, a reference database is a collection of known values from different individuals in a population. Reference databases are especially important when measuring characteristics that differ from person to person (e.g., fingerprints, hair samples, odors, blood types). The reference database enables you to know, for example, how common each characteristic is in the human population. For example, a reference database would tell you whether a particular DNA type, fingerprint, or hair type was rare or common. Reference databases are not typically used to detect human and technical error, however.

Standards are similar to controls (in the Evaluation section). The only difference is that a standard is a type of control used to verify that measurements are being taken accurately. When we introduce controls in the Evaluation section, we will indicate that we are using them to evaluate a model, such as whether a treatment is having an effect (does Y change if we change X). You may think of a standard as a control for data measurement.

Randomization

What is randomization?	It is the process of making choices according to some random process.
Why use randomization?	It destroys unwanted associations in the data and thereby eliminates many kinds of bias.
When to use randomization?	Any time a choice is made between two or more equivalent options.

It would not be possible for a limited task force to sample all A&M students on the night in question. Choices would thus have to be made about which students would be tested (we'll assume that we have access to all of them, even if they go home or hide out in their dorm room). The only acceptable method, if we want accurate data, is to choose randomly - to literally use random numbers or flip a coin and base the choice on these random numbers. (Random is not the same as haphazard.) Other methods of choosing the sample risk the possibility of biases. For example, were we to sample just fraternity members, we would likely get a different result than if we sampled dorms or the library. There are lots of methods that may seem to be random (closing your eyes and choosing a name from a phone book), which are not truly random, so in this class, we consider something to be chosen randomly only when it is stated as being chosen by a coin flip, roll of a die, using a random number table, or drawn (blindly) from a hat.

SECTION 5

Blind Data

What is meant by blind?	It is the gathering of data when the subjects and/or observer do not know the treatment each individual received.
Why use blind methods?	It prevents certain kinds of biases.
When to use blind methods?	Blind observations should be taken when there is any possibility of subjectivity in gathering or interpreting the results; blind subjects should be used when they can influence the results by knowing the treatment they have received.

Blind designs may have several dimensions to them:

- Blind observers The person gathering the data is unaware of the treatment of each subject
- Blind subjects (applies only when the subjects are humans)
    - Subjects are unaware that an experiment is being conducted
    - Subjects are aware of the experiment but unaware of the group to which they have been assigned. In this case, a “placebo” is used to fool the subjects, so that no one is sure which group they belong to.

Blind Observers: Protocols employing blind observers are needed when there is a large element of subjectivity in gathering the data. They prevent the observer's preconceptions from influencing the data gathered. For example, if you wanted to determine whether children fed candy were more hyperactive than children fed apples, you would not want the observer to know which children were fed candy and which apples; it would be too easy to unintentionally over-interpret the activities of candy-fed children. Doing the experiment blind prevents the observer's preconceptions from influencing the data. In our hypothetical study of A&M students, we would want blind subjects (students should not know who was going to be tested, and even better, should not know that the study was being conducted). We would also want the choice of students for testing to be done blind (random is even better), to avoid selecting a sample unrepresentative of the student body .

BLIND DATA
Testing new drugs. Half the patients in the experiment receive the new drug and the other half receive a placebo. The doctor evaluating the patients does not know which patients received the new drug and which the placebo. (blind observer)
Student evaluations of professors. The professor does not know which student wrote any particular evaluation. (blind observer)

DATA NOT COLLECTED IN A BLIND MANNER
Expert witnesses in criminal cases. An expert witness knows what those who have hired him (either the defense or prosecution) want to show.
Endangered species surveys. The Endangered Species Act sometimes requires a landowner to hire a private biological consultant to determine if there are endangered species on his land. Before doing the survey, the consultant in all likelihood will know what his client wants to hear.
Grading exams by hand. The person grading a paper knows who wrote it. For numerical answers this is not a major concern, but with the more subjective grading of essay questions, there is the possibility that the graders preconceived notions about the students abilities could influence the grade assigned.

These tables offer several examples of observations made blind and others lacking this feature. Blind studies prevent the researcher's preconceived notions from influencing the data collected.

Blind Subjects:

One further variation to the concept of a blind experimental design using humans is that the patient is not informed of the treatment being received. To render a study blind in which medicines are tested, the subjects in the control group receive a placebo (inert pill). Although a placebo may seem like an unnecessary precaution, experience teaches us that it is not superfluous. Patient attitude has a major effect on the recovery period for some illnesses and surgeries, and a convincing body of data shows that patient who know they are in a control group often do not recover as quickly as patients unaware that they are in the control group.

In some cases, it is impossible to conceal the treatment from the subjects but it may be possible to conceal from them that they are part of an experiment. This kind of design would arise in studying people's responses to some kind of experience, such as something they read or felt. For example, we might want to test the effect of having students listen to 30 minutes of soothing music versus hard rock prior to an exam. There would be no way to conceal from the person the type of music they were exposed to, but it would be possible to conduct the study without telling them that an experiment was involved. In this type of blind design, the subjects are prevented from modifying their response in anticipation of the outcome of the study, since they are not aware that a study is being conducted.

Double Blind:

Designs with blind observers and patients are known as double-blind. As noted, there are circumstances in which some of these features are not relevant to a design. How does one tell when a feature is relevant or irrelevant? The relevance depends on the goal of the study and the model being tested. Once those dimensions have been specified, it is possible to indicate whether each feature of ideal designs is relevant.

Explicit Protocols (Explicit Procedures)

What is an explicit protocol? A protocol is a procedure -- ANY procedure or set of methods. An explicit protocol is simply a formalized procedure for gathering data -- a set of specific rules.
Why use one? An explicit protocol enables different observations to be taken uniformly and specifies which of the other features of ideal data will be applied.
When to use explicit protocols? In all data gathered for some important purpose.
What are the consequences of failing to use an explicit protocol? The data may be gathered inconsistently and be unreliable or unrepeatable.

In any serious attempt to gather data, a formal protocol should be used to specify how the data are to be gathered. The simplest step to take towards creating an explicit protocol is decide to record the data in a systematic fashion (often merely by writing down the observations). For example, consider how a reporter would record the events at a city league softball game versus the way a casual spectator would record the events. The reporter would record specific items such as final score, winning and losing pitchers, inning-by-inning history of hits, errors critical to the final outcome, and so forth. By contrast, the casual spectator would probably remember the final score and possibly the pitchers, but many other details would go unrecorded. The protocol will minimally indicate which of the preceding four features are present (blind, random, replication, standards).

Use of a written protocol has two effects. First, it enables subsequent data to be gathered in similar fashion. It is essential that all observations be repeatable (for they are otherwise useless), and an explicit protocol allows data to be gathered consistently from one time to the next. Second, an explicit protocol is itself a model that can be subjected to evaluation and improvement using the scientific method. That is, formalizing a protocol is the first step in improving it.

The field sobriety test that police officers give suspected drunk drivers is based on a protocol. It consists of having the suspect,

- walk heel-to-toe,
- extend their arms and then touch their nose,
- balance on one foot,

and so on. Certain types of data gathered using this protocol are consistent with the hypothesis that the driver is drunk (e.g., failing these simple coordination tests), while other data are not. The protocol makes it easy to compare data gathered by different police officers, and it ensures that all relevant data are gathered from each suspect. Additionally, the protocol makes a police officer's case against a suspected drunk driver stronger in court than it would otherwise be - the officer can cite specific tasks that the suspect was unable to accomplish, instead of testifying that the suspect "looked and acted drunk."

PROTOCOL	MODEL FOR WHICH DATA ARE GATHERED
Job application and interview	Applicant's skills and suitability
Takeoff checklist	Safety of the flight
Car repair manual	Car function
Balancing a checkbook	Current account balance

More Complicated Protocols

Corporations use protocols to prevent fraud and similar activities, as well as to gather data about their financial condition. These protocols cover everything from how clerks operate the cash registers, to how the money is transported from the store to the bank, to more abstract problems such as how depreciation and good will are treated on the company's books. The care with which the accountants develop and implement their protocol determines the quality of the data the corporation has for making financial decisions.

The rules that determine what kind of evidence can be admitted in a criminal trial also constitute a protocol. Although this protocol may appear to be completely different than the corporation's protocol for gathering financial data, it is similar in that both protocols systematize the gathering of data. The rules governing admissibility of evidence allow certain types of data to be presented to the court but prohibit others:

- no hearsay evidence,
- no evidence gathered contrary to the U.S. constitution and laws,
- all witnesses are subject to cross-examination.

One may argue about the desirability of particular features of this protocol (indeed, a fair amount of political debate does). But at least everyone involved knows exactly what types of data are permitted, and how the data can and can't be gathered.

To appreciate the importance of explicit protocols, consider the day-care workers who were accused and convicted of sexually-abusing the children under their care (and are now being released). No records were kept of the psychological interviews with the children, so it was not possible to know whether the allegations came unsolicited from the children or were instead suggested by the psychologist conducting the interviews. More recently, it has been suggested that psychologists are capable of inspiring false memories in people about earlier events in their lives. Explicit protocols are thus obviously vital in evaluating these two different models of the source of the children's accusations.

Limitations of Explicit Protocols:

A protocol is a model, and like all models, it has limitations. First, a protocol never contains all information about data gathering. That which is left out may be trivial or important, and the fact that a protocol contains lots of detail does not mean that all important features have been included. Second, it is usually impossible to follow any explicit protocol exactly, unless the protocol is worded so vaguely as to admit many different ways of gathering the data. Whenever reading a protocol to assess how data were gathered, there are two questions which should be asked in understanding the protocol's limitations:

- Was the protocol followed?
- What is omitted from the protocol?

Protocols for Interpreting and Analyzing Data

The raw data are rarely presented. At the least, averages and standard errors are given. In DNA evidence, the lab may declare a match; they may even give the actual numbers obtained from a DNA sample, but the raw data in the database used for comparison are not presented.

More importantly, samples may be analyzed multiple times. Labs and investigators sometimes throw out data (sometimes with good reason), or they may downplay some data in preference to others. Every study has its unique points, and how the data are handled can have profound effects on the conclusion reached. In one of the early trials involving DNA evidence in the U.S. (New York vs. Castro), the lab declared a match between the victim and a blood spot found on the defendant's watch. Yet inspection of the raw data from the lab indicated that the two samples did NOT match (according to the lab's published protocols); there were many other cases in which their analysis of the data ignored discrepancies that should have caused them to reject a match.

The interpretation and analysis of data is thus an important step in the presentation of data. The explicit protocol should describe how the data were analyzed, as well as which data were omitted (and why). In some cases, we can make a clear distinction between errors in recording data versus errors in interpreting or analyzing data. For example, incorrectly reporting the result of a coin flip is clearly a mistake in gathering the data, whereas the discrepancy between the true proportion of heads and the observed proportion heads (in data properly recorded) is an error that affects interpretation.

Adherence to Protocol has Become a Substitute for Data

The goals in specifying a protocol are (i) to minimize error, (ii) understand what types of errors may still be present, and (iii) allow others to gather data in a similar way. In the long run, an explicit protocol enables us to develop even better ways of gathering data (yet another realm of progress). In various bureaucratic and legal settings, however, the protocol assumes an even more important role: it becomes a surrogate for data quality. That is, a company, agency, or person is evaluated strictly on whether they are following the protocol, independently of the quality of data they produce. A recent audit of the FBI DNA crime lab, for example, was limited entirely to whether the proper documentation was being maintained (as specified by protocol). Thus, it did not matter if the quality of DNA typing was good or bad, only whether the lab was following procedure and filling out the requisite paperwork. As an analogy, you could imagine evaluating a company assembling computers. The company documents the fact that all parts get assembled correctly, but does not actually check on the quality of the parts it uses or whether the end product works the way it should. You can imagine just how "useful" such an evaluation procedure might be.

SECTION 9

Exercise

These 5 descriptions are merely introductions to complicated aspects of data gathering. Even in cases in which you do not understand the full complexities, you should be able to inspect the description of data gathering for the 5 elements in our template. Popular articles on topics that have been approached from a scientific perspective provide material that can be inspected for the presence or absence of these 5 features; in many cases, the descriptions of studies are not clear about certain features of data, in which case the information is ambiguous.