CHAPTER 10: WHY ERROR IS UNAVOIDABLE

EVIDENCE

10

Data are models, and as such, are never perfect. But there are a few standard types of errors to watch out for.

SECTION 1

Why Data Are Important

Whether it be the advance warning of hurricanes, an impending military conflict, or customer response to a new product, we have no difficulty appreciating the importance of good data. Good data do not solve all problems, but they help in making decisions. Had the designers of the Titanic understood the ship's true limitations, they would likely have provisioned the ship with an adequate number of lifeboats. The crucial data were obtained on the ship's first - and last - voyage. And knowing that a hurricane is 24 hours from destroying your dwelling may not allow you to save the dwelling, but you can at least escape personal injury.

Data constitute our only link to the world we study - they are the best models of nature we have. As a consequence, they hold a supreme position in all applications of the scientific method. When predictive models are found to be at odds with good data, we keep the data and discard the predictive models. Any set of data is necessarily a model, with all the inherent limitations of models, but data comprise very special models. Ultimately, progress in science, business, and industry rests on the data.

Much of progress in science is simply the gathering of better and better data. Compared to even twenty years in the past, we live in an information age, and we now have access to data on countless phenomena. Satellites give us information about weather and climate, opinion polls indicate how the public thinks on all sorts of topics, journals and newspapers give us information about the latest products, various statistics on the economy and social events impact our optimism and pessimism about the future. The information we have today is more extensive, more detailed, and arrives faster than at any time in the past. We refer to the changes leading to this increase in information as technology, and the scientific method has been the basis for much of the improved technology we enjoy today. But improved technology does not guarantee better data. Improved technology does indeed enable us to gather better data, if we know how, but we can just as easily use improved technology to gather poor data.

The quality of data is important for a simple reason. As we noted in an earlier chapter, science does not prove models to be true. The ultimate reason behind this conclusion is that any set of data is consistent with many models. What determines the quality of a data set is the number of models that can be rejected by it. Poor data sets are consistent with many different models, hence are of little use.

There are several issues that concern data. The one addressed here is simply how to gather accurate data -- to get the measurements as exact as needed and to minimize the error associated with the data. When we address the topic of Evaluation in subsequent chapters, we address the further problem of gathering data that enables a researcher to test particular models. This latter dimension of data concerns the interpretation of data rather than their accuracy and is relevant to experimental design.

SECTION 2

Data As Models -- With the Usual Problems

Data represent a special type of model, one that is central to the scientific method. We use data to tell us about the phenomenon we are studying. Abstract models, such as theories and hypotheses are models that help us simplify nature. But data are our surrogates of reality.

Like all models, data are "false." No matter how hard we may try, the data will never exactly match what we think they represent. Instead of referring to data as being "false," however, we say that data are "measured with error." In this context, "error" does not imply blunder (as in baseball), rather it means "variation."

Another way to look at data is this. There is one fundamental issue that underlies data collection in all applications of the scientific method: if the data were to be gathered more than once (and by someone else), would they turn out the same each time? We say that data are measured with error to describe the extent that attempts to record the same phenomena differ. That is, any variation that causes our measurement of something to be inexact is error.

A universal goal when using the scientific method is to reduce the error/variation so that you know what the data represent (as closely as you need). This claim may seem to contradict the statement above that error is unavoidable. But, in fact, there are ways to reduce the error. Understanding how this error can arise is the first step in reducing it.

SECTION 3

Four Types of Error in Data, plus one in Analysis

When someone makes a claim that you find hard to believe or at least need to know whether you can trust it, there are three critical questions to ask:

1. What's the evidence? (what are the data?)
2. How was it obtained? (are the data any good?)
3. Who obtained it? (can the source be trusted?)

To some extent, all 3 questions pertain to this section of the course ("data"), although (3) is also addressed by the last part of the book. This chapter focuses specifically on (2), doing our best to ensure data quality.

The basic problem is that several things can "go wrong" with data. Another way of saying this is that there are different sources of error in data. Although error cannot be completely eliminated -- a single coin flip can only be 100% heads or tails, even though the probability of heads may be 50% -- there are safeguards and precautions that can reduce many types of errors. However, different types of error require different safeguards. Understanding the different types of error is thus the first step in understanding those precautions.

We will recognize 4 types of error in data collection: (i) rounding, precision and accuracy (RPA), (ii) sampling, (iii) human and technical (H&T), and (iv) bias (chiefly unintentional bias). In addition, error may arise *after* the data are collected, in their analysis and processing -- (v) faulty data analysis.

A fundamental point to keep in mind is that **not all variations in data represent error**. Data from different sources may differ because the sources are truly different -- because 'nature' is different between the sources. We try to identify and eliminate or reduce errors in data to make it easier to discover when the processes generating the data are themselves different. Differences due to nature are what we care about -- as when a drug truly improves an outcome, or a hazardous chemical really does harm us. Errors in data complicate the detection of differences arising from natural causes -- as when patients treated with a useless drug nonetheless show an improved outcome either by chance or merely because they felt the drug 'should' improve their outcomes. Only when we know how to detect errors in data can we make the distinction.

Rounding, Precision, and Accuracy Error

Some kinds of measurements can never be made exactly, so we have to "round off" the value at some quantity that is less than exact. When a machine fails to provide a value beyond a fixed number of decimal places, we call it precision error. Consider the weight (mass) of a penny. To the nearest tenth, a penny is 2.5gm. To the nearest 0.01, it is 2.54. Using the finest balances, we could measure the mass to many more decimal places. But at some point, we would reach the limit of precision for our scale and thus be left with a rounded-off value. Or we would reach the point that the scale was no longer accurate enough to consistently give us the correct weight (accuracy error). We can never measure the mass exactly - not even to millions of decimal places, much less to an infinity of decimal places. In many branches of science, this type of error is specifically included in measurements by providing a measurement ± (plus or minus) some smaller value, such as 101 ± 0.23 meters. The number behind the ± indicates the level of error that can be expected in the number preceding the ±.

Precision and rounding error apply to many kinds of measurements - those in which we are not simply counting numbers of things: time, speed, weight, energy, volume, distance, and many others. For most non-technical applications of the scientific method, however, this kind of error is unimportant because we don't care about the value beyond a few decimal places. In economics, for example, a company is not likely concerned about the cost to produce an item to the fraction of a cent. And our monetary system forces each of us to accept rounding error because we cannot pay in fractions of a cent. Rounding error even applies to the estimation of percents and probabili-

Sampling Error

random deviation from an average

Another source of error comes from sampling only some of the data in which we are interested. Consider again a coin toss. If the probability of heads was exactly 1/2, and we tossed the coin 4 times, there is only a 3/8 chance that we would get 2 heads and 2 tails (1/8 of the time we would obtain all heads or all tails). The reason is sampling error. As a second example, we might be interested in the percent student attendance in lecture. The average attendance might be 60%, but attendance on some days would certainly be higher than on other days. Again, we would attribute this variation to sampling error. In both cases, the data we gather in one trial would not generally match exactly the data we gathered in other trials. The issue here is not in our ability to count accurately -- we know how many heads and tails we got or how many people attended class. Rather, the error lies in the fact that what actually happens one time is not the same as what happens the next, even though the underlying rules or probabilities are the same.

Sampling error is a widespread phenomenon that is often ascribed to random "noise" and unmeasured variables. In the case of a coin toss, the outcome of the toss is usually attributed to random noise. In the case of student attendance, there would undoubtedly be reasons why each non-attending student missed class, but the reasons would be too diverse to measure and thus be attributed to unmeasured variation.

Sampling error is universal, although its importance may vary greatly from case to case. The way to reduce sampling error (discussed in the next chapter) is to make many observations and to obtain an average that swamps out most of the sampling error made in each observation. Sampling error is a big problem in studies of environmental hazards (e.g., cancer-causing agents), because only a low percentage of people develop any specific kind of health problem, so we need large samples to overcome the sampling error. For example, if we observe 1 excess case of cancer in one million people who eat bacon and 0 excess cases of cancer in people who avoid bacon, we can't infer that the cancer rates differ between the two groups because sampling error would give us this result 50% of the time if there was no difference between the groups. We would need a sample size about 10 times larger than this to overcome sampling error.

Technical and Human Error

Our machines and our abilities to record data are not foolproof. Technicians handling hundreds of tubes, loading samples, and labeling samples can and do make mistakes. A common example occurs in televised football games, in which an official misreads a play and inappropriately assigns penalties. And a machine which has been calibrated wrong or whose calibration has drifted will also give erroneous data - the Hubble space telescope gave fuzzy pictures during the first few years of its operation due to faulty assembly.

Some machines and people are obviously less error-prone than others, and indeed, some technicians may never actually make any mistakes in their career. But there is always the possibility of error, and no amount of observations on any machine or person can show that a mistake is impossible (recalling our points about sampling error above).

While some instances of RPA, sampling error, and unintentional bias (next) may be errors caused by humans, our category of human and technical error is used here to describe errors that do not fall into those other categories.

Unintentional Bias

Biases are consistent differences between the data gathered and what the data are thought to represent. In particular, bias is a tendency of the data to fall more on one side of the average than the other, which distinguishes it from sampling error. Whereas sampling error tends to balance itself out in the average as more observations are gathered, bias persists -- when data are biased, gathering bigger samples means that the average of the data is certain to differ from the expected average (or the true average). For example, opinion polls are often conducted over the telephone. Data gathered in these surveys do not represent people who lack telephones, and those data would be biased if people lacking phones had consistently different opinions than people with phones. Or consider the frequency of people carrying the AIDS virus. At this time, the frequency in the U.S. population is thought to be something like 1 in 200. But the frequency of people with this virus would be much higher in some groups than in others (prostitutes versus nuns, for example). The data for one subgroup would then be a biased model of the population at large; this bias would be important when calculating the chance of acquiring the virus from a sexual encounter with someone who had lots of other sexual partners, for example. Another, similar example comes from a kit once marketed to allow couples to "choose" the sex of their child. The kit involved instructions on the timing of sex as well as some buffers to supposedly influence the relative success of sperm with an X versus sperm with a Y chromosome (the chromosomes of the one lucky sperm indeed determines the sex of the child). The evidence in support of this method was a collection of letters from parents who wrote to the physician who developed the method, and a majority of letters reported success. It does not take much imagination to figure out that this sample of letters was likely biased -- parents whose baby was not the chosen sex were undoubtedly less inclined (and perhaps even reluctant to) write about their "failure." The FDA was not fooled, however, and the kit was withdrawn soon after it was marketed.

Unintentional bias is easy to confuse with sampling error. Remember that bias represents a deviation consistently to one side. As an analogy, think of sighting-in a rifle. If the rifle sights are mis-aligned, the average of the bullets will consistently lie to one side of the bull's-eye, no matter how many shots are fired. This is analogous to bias. Where-ever the sights are set, however, bullets will lie in a cluster around the average point of impact; this scatter around the average is akin to sampling error.

Biases may occur by mistake or deliberately. In this chapter, we restrict attention to accidental or unintentional bias. In a subsequent module, we deal with the problem of deliberate bias, as when people intentionally attempt to deceive.

Faulty Data Professing and Analysis

The four categories described above apply to the data themselves, what are often called raw data. The raw data are the most basic type of data used, and everything rests on those. But it is rare that the raw data are used or published directly. More commonly, the raw data are processes or analyzed in some fashion, and those analyses are published. For example, a large set of data may be summarized by the average and variance. They may be subjected to statistical tests. Tables will often be used to present interesting features of the data. These analyses may be simple or highly complicated, and there are many, many ways to analyze most data sets.

Even when the data are collected in a reasonable manner, such that most types of errors have been avoided or minimized, a faulty analysis can mislead in the same way that faulty data can be misleading. The main difference, however, and a critical one, is that faulty analysis can be corrected without gathering better data -- whenever the original data are available for a subsequent analysis. Faulty analysis may be corrected by the individuals who did the original analysis (if they recognize their mistake) or, more commonly, by others doing the analysis on the same data. Thus the key difference between faulty analysis and errors in data is that the analysis applies **after** the data are gathered. Errors in data apply to the gathering of the data themselves (**during the gathering of data)**. Thus a faulty analysis is not something that arises from human and technical error, from sampling error, from RPA error, or from bias. It is its own type of error.

There are countless ways to conduct a faulty analysis. Instead of attempting to classify them, we will merely consider faulty data processing and analysis as a single type of error, but again, it is a type of error arising after the data are gathered.