PART I

FOUNDATIONS

1

SOURCES OF ERROR

Don't think—use the computer.

—Dyke (tongue in cheek) [1997].

Statistical procedures for hypothesis testing, estimation, and model building are only a *part* of the decision-making process. They should never be used as the sole basis for making a decision (yes, even those procedures that are based on a solid deductive mathematical foundation). As philosophers have known for centuries, extrapolation from a sample or samples to a larger incompletely examined population must entail a leap of faith.

The sources of error in applying statistical procedures are legion and include all of the following:

- Using the same set of data to formulate hypotheses and then to test those hypotheses.
- Taking samples from the wrong population or failing to specify in advance the population(s) about which inferences are to be made.
- Failing to draw samples that are random and representative.
- Measuring the wrong variables or failing to measure what you intended to measure.
- Failing to understand that *p*-values are statistics, that is, functions of the observations, and will vary in magnitude from sample to sample.
- Using inappropriate or inefficient statistical methods.
- Using statistical software without verifying that its current defaults are appropriate for your application.
- Failing to validate models.

Common Errors in Statistics (and How to Avoid Them), Third Edition. Edited by P. I. Good and J. W. Hardin

But perhaps the most serious source of error is letting statistical procedures make decisions for you.

In this chapter, as throughout this book, we first offer a preventive prescription, followed by a list of common errors. If these prescriptions are followed carefully, you will be guided to the correct and effective use of statistics and avoid the pitfalls.

PRESCRIPTION

Statistical methods used for experimental design and analysis should be viewed in their rightful role as merely a part, albeit an essential part, of the decision-making procedure.

Here is a partial prescription for the error-free application of statistics:

1. Set forth your objectives and your research intentions *before* you conduct a laboratory experiment, a clinical trial, or a survey or analyze an existing set of data.
2. Define the population about which you will make inferences from the data you gather.
3. List all possible sources of variation. Control them or measure them to avoid confounding them with relationships among those items that are of primary interest.
4. Formulate your hypotheses and all of the associated alternatives. (See Chapter 2.) List possible experimental findings along with the conclusions you would draw and the actions you would take if this or another result proves to be the case. Do all of these things *before* you complete a single data collection form and *before* you turn on your computer.
5. Describe in detail how you intend to draw a representative sample from the population. (See Chapter 3.)
6. Use estimators that are impartial, consistent, efficient, robust, and minimum loss. (See Chapter 5.) To improve the results, focus on sufficient statistics, pivotal statistics, and admissible statistics and use interval estimates. (See Chapters 5 and 6.)
7. Know the assumptions that underlie the tests you use. Use those tests that require the minimum number of assumptions and are most powerful against the alternatives of interest. (See Chapters 5, 6, and 7.)
8. Incorporate in your reports the complete details of how the sample was drawn and describe the population from which it was drawn. If data are missing or the sampling plan was not followed, explain why and list all differences between the data that were present in the sample and the data that were missing or excluded. (See Chapter 8.)

FUNDAMENTAL CONCEPTS

Three concepts are fundamental to the design of experiments and surveys: variation, population, and sample.

A thorough understanding of these concepts will forestall many errors in the collection and interpretation of data.

> If there were no variation—if every observation were predictable, a mere repetition of what had gone before—there would be no need for statistics.

Variation

Variation is inherent in virtually all of our observations. We would not expect the outcomes of two consecutive spins of a roulette wheel to be identical. One result might be red, the other black. The outcome varies from spin to spin.

There are gamblers who watch and record the spins of a single roulette wheel hour after hour, hoping to discern a pattern. A roulette wheel is, after all, a mechanical device, and perhaps a pattern will emerge. But even those observers do not anticipate finding a pattern that is 100% predetermined. The outcomes are just too variable.

Anyone who spends time in a schoolroom, as a parent or as a child, can see the vast differences among individuals. This one is tall, that one is short, though all are the same age. Half an aspirin and Dr. Good's headache is gone, but his wife requires four times that dosage.

There is variability even among observations on deterministic formula-satisfying phenomena such as the position of a planet in space or the volume of gas at a given temperature and pressure. Position and volume satisfy Kepler's laws and Boyle's law, respectively, but the observations we collect will depend upon the measuring instrument (which may be affected by the surrounding environment) and the observer. Cut a length of string and measure it three times. Do you record the same length each time?

In designing an experiment or a survey, we must always consider the possibility of errors arising from the measuring instrument and from the observer. It is one of the wonders of science that Kepler was able to formulate his laws at all given the relatively crude instruments at his disposal.

Population

> The population(s) of interest must be clearly defined before we begin to gather data.

From time to time, someone will ask us how to generate confidence intervals (see Chapter 8) for the statistics arising from a total census of a population. Our answer is that we cannot help. Population statistics (mean, median, 30th percentile) are not

estimates. They are fixed values and will be known with 100% accuracy if two criteria are fulfilled:

1. Every member of the population is observed.
2. All the observations are recorded correctly.

Confidence intervals would be appropriate if the first criterion is violated, for then we are looking at a sample, not a population. And if the second criterion is violated, then we might want to talk about the confidence we have in our measurements.

Debates about the accuracy of the 2000 United States Census arose from doubts about the fulfillment of these criteria.[1] "You didn't count the homeless" was one challenge. "You didn't verify the answers" was another. Whether we collect data for a sample or an entire population, both of these challenges or their equivalents can and should be made.

Kepler's "laws" of planetary movement are not testable by statistical means when applied to the original planets (Jupiter, Mars, Mercury, and Venus) for which they were formulated. But when we make statements such as "Planets that revolve around Alpha Centauri will also follow Kepler's laws," we begin to view our original population, the planets of our sun, as a sample of all possible planets in all possible solar systems.

A major problem with many studies is that the population of interest is not adequately defined before the sample is drawn. Don't make this mistake. A second major problem is that the sample proves to have been drawn from a different population than was originally envisioned. We consider these issues in the next section and again in Chapters 2, 6, and 7.

Sample

A sample is any (proper) subset of a population.

Small samples may give a distorted view of the population. For example, if a minority group comprises 10% or less of a population, a jury of 12 persons selected at random from that population fails to contain any members of that minority at least 28% of the time.

As a sample grows larger, or as we combine more clusters within a single sample, the sample will resemble more closely the population from which it is drawn.

How large a sample must be drawn to obtain a sufficient degree of closeness will depend upon the manner in which the sample is chosen from the population.

Are the elements of the sample drawn at random, so that each unit in the population has an equal probability of being selected? Are the elements of the sample drawn independently of one another? If either of these criteria is not satisfied, then even a very large sample may bear little or no relation to the population from which it was drawn.

[1] *City of New York v. Department of Commerce*, 822 F. Supp. 906 (E.D.N.Y, 1993). The arguments of four statistical experts who testified in the case may be found in Volume 34 of *Jurimetrics*, 1993, 64–115.

An obvious example is the use of recruits from a Marine boot camp as representatives of the population as a whole or even as representatives of all Marines. In fact, any group or cluster of individuals who live, work, study, or pray together may fail to be representative for any or all of the following reasons [Cummings and Koepsell, 2002]:

1. Shared exposure to the same physical or social environment.
2. Self-selection in belonging to the group.
3. Sharing of behaviors, ideas, or diseases among members of the group.

A sample consisting of the first few animals to be removed from a cage will not satisfy these criteria either, because, depending on how we grab, we are more likely to select more active or more passive animals. Activity tends to be associated with higher levels of corticosteroids, and corticosteroids are associated with virtually every body function.

Sample bias is a danger in every research field. For example, Bothun [1998] documents the many factors that can bias sample selection in astronomical research.

To forestall sample bias in your studies, before you begin, determine all the factors that can affect the study outcome (gender and lifestyle, for example). Subdivide the population into strata (males, females, city dwellers, farmers) and then draw separate samples from each stratum. Ideally, you would assign a random number to each member of the stratum and let a computer's random number generator determine which members are to be included in the sample.

Surveys and Long-term Studies

Being selected at random does not mean that an individual will be willing to participate in a public opinion poll or some other survey. But if survey results are to be representative of the population at large, then pollsters must find some way to interview nonresponders as well. This difficulty is exacerbated in long-term studies, as subjects fail to return for follow-up appointments and move without leaving a forwarding address. Again, if the sample results are to be representative, some way must be found to report on subsamples of the nonresponders and the dropouts.

AD HOC, POST HOC HYPOTHESES

Formulate and write down your hypotheses before you examine the data.

Patterns in data can suggest but cannot confirm hypotheses unless these hypotheses were formulated *before* the data were collected.

Everywhere we look, there are patterns. In fact, the harder we look, the more patterns we see. Three rock stars die in a given year. Fold the U.S. $20 bill in just the right way, and not only the Pentagon but also the Twin Towers in flames are

revealed.[2] It is natural for us to want to attribute some underlying cause to these patterns. But those who have studied the laws of probability tell us that more often than not, patterns are simply the result of random events.

Put another way, there is a greater probability of finding at least one cluster of events in time or space than finding no clusters at all (equally spaced events).

How can we determine whether an observed association represents an underlying cause-and-effect relationship or is merely the result of chance? The answer lies in our research protocol. When we set out to test a specific hypothesis, the probability of a specific event is predetermined. But when we uncover an apparent association, one that may have arisen purely by chance, we cannot be sure of the association's validity until we conduct a second set of controlled trials.

In the International Study of Infarct Survival [1988], patients born under the Gemini or Libra astrological birth signs did not survive as long when their treatment included aspirin. By contrast, aspirin offered an apparent beneficial effect (longer survival time) to study participants with all other astrological birth signs.

Except for those who guide their lives by the stars, there is no hidden meaning or conspiracy in this result. When we describe a test as significant at the 5% or 1-in-20 level, we mean that 1 in 20 times we'll get a significant result even though the hypothesis is true. That is, when we test to see if there are any differences in the baseline values of the control and treatment groups, if we've made 20 different measurements, we can expect to see at least one statistically significant difference; in fact, we will see this result almost two-thirds of the time. This difference will not represent a flaw in our design but simply chance at work. To avoid this undesirable result—that is, to avoid attributing statistical significance to an insignificant random event, a so-called Type I error—we must distinguish between the hypotheses with which we began the study and those which came to mind afterward. We must accept or reject these hypotheses at the original significance level while demanding additional corroborating evidence for those exceptional results (such as dependence of an outcome on an astrological sign) that are uncovered for the first time during the trials.

No reputable scientist would ever report results before successfully reproducing the experimental findings twice, once in the original laboratory and once in that of a colleague.[3] The latter experiment can be particularly telling, as all too often some overlooked factor not controlled in the experiment—such as the quality of the laboratory water—proves responsible for the results observed initially. It's better to be found wrong in private than in public. The only remedy is to attempt to replicate the findings with different sets of subjects, replicate, then replicate again.

Persi Diaconis [1978] spent some years investigating paranormal phenomena. His scientific inquiries included investigating the powers linked to Uri Geller (Fig. 1.1), the man who claimed he could bend spoons with his mind. Diaconis was not surprised

[2]A website with pictures is located at http://www.foldmoney.com/.

[3]Remember "cold fusion?" In 1989, two University of Utah professors told the newspapers that they could fuse deuterium molecules in the laboratory, solving the world's energy problems for years to come. Alas, neither those professors nor anyone else could replicate their findings, though true believers abound; see http://www.ncas.org/erab/intro.htm.

Figure 1.1. Photo of Uri Geller. *Source*: Reprinted with permission of Aquarius 2000 of the German Language Wikipedia.

to find that the hidden "powers" of Geller were more or less those of the average nightclub magician, down to and including forcing a card and taking advantage of ad hoc, post hoc hypotheses.

When three buses show up at your stop simultaneously, or three rock stars die in the same year, or a stand of cherry trees is found amid a forest of oaks, a good statistician remembers the Poisson distribution. This distribution applies to relatively rare events that occur independently of one another. The calculations performed by Siméon-Denis Poisson reveal that if there is an average of one event per interval (in time or in space), then while more than a third of the intervals will be empty, at least a quarter of the intervals are likely to include multiple events (Fig. 1.2).

Anyone who has played poker will concede that one out of every two hands contains something interesting (Table 1.1). Don't allow naturally occurring results to fool you or lead you to fool others by shouting "Isn't this incredible?"

The purpose of a recent set of clinical trials was to see if blood flow and distribution in the lower leg could be improved by carrying out a simple surgical procedure prior to the administration of standard prescription medicine.

The results were disappointing on the whole, but one of the marketing representatives noted that the long-term prognosis was excellent when a marked increase in blood flow was observed just after surgery. She suggested that we calculate a

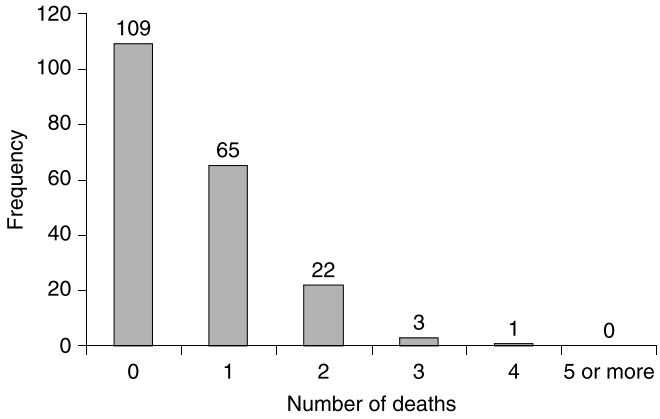

Figure 1.2. Frequency plot of the number of deaths in the Prussian army as a result of being kicked by a horse (200 total observations).

TABLE 1.1. Probability of Finding Something Interesting in a Five-Card Hand

Hand	Probability
Straight flush	0.0000
Four of a kind	0.0002
Full house	0.0014
Flush	0.0020
Straight	0.0039
Three of a kind	0.0211
Two pairs	0.0475
Pair	0.4226
Total	0.4988

p-value[4] for a comparison of patients with improved blood flow after surgery versus patients who had taken the prescription medicine alone.

Such a *p*-value is meaningless. Only one of the two samples of patients in question had been taken at random from the population (those patients who received the prescription medicine alone). The other sample (those patients who had increased blood flow following surgery) was determined after the fact. To extrapolate results from the samples in hand to a larger population, the samples must be taken at random from, and be representative of, that population.

The preliminary findings clearly called for an examination of surgical procedures and of patient characteristics that might help forecast successful surgery. But the

[4]A *p*-value is the probability under the primary hypothesis of observing the set of observations we have in hand. We can calculate a *p*-value once we make a series of assumptions about how the data were gathered. Today statistical software does the calculations, but it's still up to us to validate the assumptions.

generation of a *p*-value and the drawing of any final conclusions had to wait on clinical trials specifically designed for that purpose.

This doesn't mean that one should not report anomalies and other unexpected findings. Rather, one should not attempt to provide *p*-values or confidence intervals in support of them. Successful researchers engage in a cycle of theorizing and experimentation so that the results of one experiment become the basis for the hypotheses tested in the next.

A related, extremely common error whose correction we discuss at length in Chapters 13 and 14 is to use the same data to select variables for inclusion in a model and to assess their significance. Successful model builders develop their frameworks in a series of stages, validating each model against a second independent data set before drawing conclusions.

TO LEARN MORE

On the necessity for improvements in the use of statistics in research publications, see Altman [1982, 1991, 1994, 2000, 2002], Cooper and Rosenthal [1980], Dar, Serlin, and Omer [1994], Gardner and Bond [1990], George [1985], Glantz [1980], Goodman, Altman, and George [1998], MacArthur and Jackson [1984], Morris [1988], Thorn, Pilliam, Symons, and Eckel [1985], and Tyson, Furzan, Reisch and Mize [1983].