WILEY

Common Errors in Statistics

(and How to Avoid Them)

SECOND EDITION

PHILLIP I. GOOD

JAMES W. HARDIN

COMMON ERRORS IN STATISTICS (AND HOW TO AVOID THEM)

Second Edition

Phillip I. Good
Huntington Beach, CA

James W. Hardin
Columbia, SC

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

ISBN-13: 978-0-471-79431-8
ISBN-10: 0-471-79431-7

Printed in the United States of America.

10 9 8 7 6 5 4 3 2

Preface

ONE OF Dr. Good's very first statistical applications was an analysis of leukemia cases in Hiroshima, Japan after World War II; on August 7, 1945 this city was the target site of the first atomic bomb dropped by the United States. Was the high incidence of leukemia cases among survivors the result of exposure to radiation from the atomic bomb? Was there a relationship between the number of leukemia cases and the number of survivors at certain distances from the atomic bomb's epicenter?

To assist in the analysis, Dr. Good had an electric (not an electronic) calculator, reams of paper on which to write down intermediate results, and a prepublication copy of Scheffé's *Analysis of Variance*. The work took several months, and the results were somewhat inclusive, mainly because he could never seem to get the same answer twice—a consequence of errors in transcription rather than the absence of any actual relationship between radiation and leukemia.

Today, of course, we have high-speed computers and prepackaged statistical routines to perform necessary calculations. Yet access to statistical software will no more make one a statistician, than access to a chainsaw will make one a lumberjack. Allowing these tools to do our thinking for us is a sure recipe for disaster—just ask any emergency room physician.

Pressed by management or by funding needs, too many research workers have no choice but to go forward with data analysis regardless of the extent of their statistical training. Alas, although a semester or two of undergraduate statistics may suffice to develop familiarity with the names of some statistical methods, it is not enough to ensure awareness of all the circumstances under which these methods may be applicable.

The purpose of the present text is to provide a mathematically rigorous but readily understandable foundation for statistical procedures. Here for the second time are such basic statistical concepts as null and alternative hypotheses, *p*-value, significance level, and power. Reprints from the statistical literature provide illustration as we reexamine sample selection, linear regression, the analysis of variance, maximum likelihood, Bayes' theorem, meta-analysis and the bootstrap.

For the second edition, we've added material from online courses we offer at statistics.com. This new material is devoted to unbalanced designs, report interpretation, and alternative modeling methods.

More good news for technophobes: Dr. Good's articles on women's sports have appeared in the *San Francisco Examiner*, *Sports Now*, and *Volleyball Monthly*, Twenty-two of his short stories are also in print. If you can read the sports page, you'll find the presentation of material in this text easy to read and to follow. Lest the statisticians among you believe this book is too introductory, we point out the existence of hundreds of citations in statistical literature calling for the comprehensive treatment we have provided. Regardless of past training or current specialization, this book will serve as a useful reference; you will find applications for the information contained herein whether you are a practicing statistician or a well-trained scientist who just happens to apply statistics in the pursuit of other science.

The primary objective of the opening chapter is to describe the main sources of error and provide a preliminary prescription for avoiding them. The hypothesis formulation—data gathering—hypothesis testing and estimate cycle is introduced, and the rationale for gathering additional data before attempting to test after-the-fact hypotheses is detailed.

Chapter 2 places our work in the context of decision theory. We emphasize the importance of providing an interpretation of each and every potential outcome in advance of consideration of actual data.

Chapter 3 focuses on study design and data collection, for failure at the planning stage can render all further efforts valueless. The work of Berger and his colleagues on selection bias is given particular emphasis.

Desirable features of point and interval estimates are detailed in Chapter 4 along with procedures for deriving estimates in a variety of practical situations. This chapter also serves to debunk several myths surrounding estimation procedures.

Chapter 5 reexamines the assumptions underlying testing hypotheses. We review the impacts of violations of assumptions and detail the procedures to follow when making 2- and *k*-sample comparisons. In addition, we cover the procedures for analyzing contingency tables and 2-way experimental designs if standard assumptions are violated.

Chapter 6 is devoted to the value and limitations of Bayes' theorem, meta-analysis, and resampling methods.

Chapter 7 lists the essentials of any report that will utilize statistics, debunks the myth of the "standard" error, and describes the value and limitations of *p*-values and confidence intervals for reporting results. Practical significance is distinguished from statistical significance and induction is distinguished from deduction. Chapter 8 covers much the same material, but the viewpoint is that of the report reader rather than the report writer. Of particular importance is a section on interpreting computer output.

Twelve rules for more effective graphic presentations are given in Chapter 9 along with numerous examples of the right and wrong ways to maintain reader interest while communicating essential statistical information.

Chapters 10 through 13 are devoted to model building and to the assumptions and limitations of a multitude of regression methods and data mining techniques. A distinction is drawn between goodness of fit and prediction, and the importance of model validation is emphasized. Seminal articles by David Freedman and Gail Gong are reprinted.

Finally, for the further convenience of readers, we provide a glossary grouped by related but contrasting terms, an annotated bibliography, and subject and author indexes.

Our thanks to William Anderson, Leonardo Auslender, Vance Berger, Peter Bruce, Bernard Choi, Tony DuSoir, Cliff Lunneborg, Mona Hardin, Gunter Hartel, Fortunato Pesarin, Henrik Schmiediche, Marjorie Stinespring, and Peter A. Wright for their critical reviews of portions of this text. Doug Altman, Mark Hearnden, Elaine Hand, and David Parkhurst gave us a running start with their bibliographies. Brian Cade, David Rhodes, and, once again, Cliff Lunneborg helped us complete the second edition.

We hope you soon put this text to practical use.

Sincerely yours,

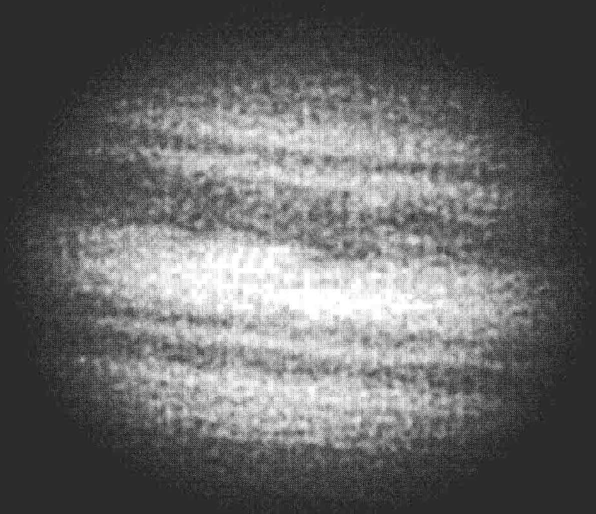
Phillip Good
brother_unknown@yahoo.com
Huntington Beach CA.

James Hardin
jhardin@gwm.sc.edu
Columbia, SC.
July 2003/2005

Contents

Part I
FOUNDATIONS

Don't think—use the computer.
Dyke (tongue in cheek) (1997)



Chapter 1

Sources of Error

STATISTICAL PROCEDURES FOR HYPOTHESIS TESTING, ESTIMATION, AND MODEL building are only a *part* of the decision-making process. They should never be quoted 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 both to formulate hypotheses and to test them**
- **Taking samples from the wrong population or failing to specify the population(s) about which inferences are to be made in advance**
- **Failing to draw random, representative samples**
- **Measuring the wrong variables or failing to measure what you'd hoped to measure**
- **Using inappropriate or inefficient statistical methods**
- **Failing to validate models**

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

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

Common Errors in Statistics (and How to Avoid Them), 2e, by Phillip I. Good and James W. Hardin.

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 the use you plan to make of your research *before* you conduct a laboratory experiment, a clinical trial, or survey or analyze an existing set of data.**
2. **Define the population to which you will apply the results of your analysis.**
3. **List all possible sources of variation. Control them or measure them to avoid their being confounded with relationships among those items that are of primary interest.**
4. **Formulate your hypothesis 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 should prove 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 4.) To improve results, focus on sufficient statistics pivotal statistics, and admissible statistics, and use interval estimates. (See Chapters 4 and 5.)**
7. **Know the assumptions that underlie the tests you use. Use those tests that require the minimum of assumptions and are most powerful against the alternatives of interest. (See Chapter 5.)**
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 data that were present in the sample and data that were missing or excluded. (See Chapter 7.)**

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 our observations. We would not expect 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% deterministic. 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, today, that one short. Half an aspirin and Dr. Good's headache is gone, but his wife requires four times that dosage for relief.

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 on 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 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 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 7) for the statistics arising from a total census of a population. Our answer is no, 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 from a sample or an entire population, equivalents of both of the previously mentioned challenges 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," then 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 source of error is that the sample proves to have been drawn from a different population than was originally envisioned. We consider this problem in the next section and again in Chapters 2, 5, and 6.

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 grow to more closely resemble the population from which it is drawn.

How large a sample must be to obtain a sufficient degree of closeness will depend on 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.

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,

[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.

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, determine before you begin 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 only 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 (those who never participate) and the dropouts (those who stop participating at some point).

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 United States twenty-dollar bill in just the right way and not only the