Statistical Rules of Thumb

Second Edition

Gerald van Belle
University of Washington
Department of Biostatistics
and Department of Environmental and Occupational Health Sciences
Seattle, WA

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

This Page Intentionally Left Blank

Statistical Rules of Thumb

Statistical Rules of Thumb

Second Edition

Gerald van Belle
University of Washington
Department of Biostatistics
and Department of Environmental and Occupational Health Sciences
Seattle, WA

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

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

Library of Congress Cataloging-in-Publication Data:

van Belle, Gerald.
Statistical rules of thumb / Gerald van Belle. — 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-14448-0 (pbk.)
1. Mathematical statistics. I. Title.
QA276.12.V36 2008
519.5—dc22 2008004431

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To Johanna

"Our science is like a store filled with the most subtle intellectual devices for solving the most complex problems, and yet we are almost incapable of applying the elementary principles of rational thought. In every sphere, we seem to have lost the very elements of intelligence: the ideas of limit, measure, degree, proportion, relation, comparison, contingency, interdependence, interrelation of means and ends."

Simone Weil, from "The Power of Words," in *Simone Weil: An Anthology*, by Sian Miles (editor), Grove Press, New York, page 222 [1986].

Contents

Preface to the Second Edition

Regularly re-reading one's own book (or its reviews) is either narcissistic or masochistic. I have resisted the temptation. Re-reading the book in preparation for the second edition both pleased me and made me cringe. I have tried to keep the pleasing and to remove the cringy. Almost all of the reviews of the first edition were positive, they also contained many useful suggestions which I have tried to incorporate in this edition.

The second edition contains two new chapters on observational studies and evidence-based medicine. These two areas are in constant and complementary development.

Some of the rules were dropped and others substituted. I tried to update the material with new references and sources. The WEB is dominating more and more. I used various search engines to check for more recent references and examples.

The field of statistics is too big for me to get my arms around. The day of the polystat is probably gone. But there is still room for basics–as I see them.

The WEB site is in the process of revision. The Rule of the Month has been dropped–it seemed a good idea but generating a rule per month turned out to be infeasible. The URL address is still the same: http://www.vanbelle.org.

I was helped by many people in preparing the second edition. Larissa Stanberry helped with Latex, R, and general computer savvy. Corinna Mar was very helpful with graphics in R. The graph that graces the front cover is her work. Finding these two new colleagues and friends has been inspiring.

My colleagues in the department of biostatistics continue to be the most stimulating crew in the world–I always look forward to the coffee time at 10 am. Lloyd Fisher is a source of wisdom and knowledge and, above all, a friend. Thomas Lumley gave very useful advice. Mary Lou Thompson read the two new chapters and made very many useful comments.

The two new chapters were written in Newport, OR. I am indebted to Kathy Howell (Oregon State University) for letting us use her cottage for extended stays. My colleague, Paul Crane who helped shape the first edition read the two new chapters with the result that they were rewritten and recast.

This preface is being written exactly six years after the writing of the preface for the first edition.

As with the first edition, my wife, Johanna, was supportive and with a great deal of will-power stopped nagging me about finishing the book. As a very experienced physician, she also provided free medical advice on the chapter on evidence-based medicine.

None of the shortcomings of this book are attributable to any of the above.

GERALD VAN BELLE

Seattle, Washington
May 21, 2008

Preface to the First Edition

This book presents statistical rules of thumb for practitioners of statistical science. The intent is stimulate thinking about designing and conducting a study, analyzing its data, or communicating its results. The book should be accessible to those who have had a basic course in statistics. It should be useful to researchers, graduate students in one of the quantitative sciences, and scientists who want to understand the statistical underpinnings of research efforts.

The book begins with basic statistical considerations about inference, assumptions and statistical processes. The next two chapters deal with sample size and covariation. Chapters 4 and 5 deal with particular areas: epidemiology and environmental studies. The choice of these topics is based on my lifelong association with these areas. Chapter 6 deals with the design of studies and Chapter 7 with the presentation of data. These two chapters are self-contained; start with them to get quick ideas about design, analysis and presentation of research results. Chapter 8 completes the book by dealing with statistical consulting. In a sense, this brings us back to the topic of Chapter 1. The chapters can be read independently of each other with the exception of Chapter 2 which provides a basis for much of the remainder of the book. Most of the equations in the book are numbered, though they are not always referenced. This was done in order to facilitate discussion of the rules on the web site.

One of the goals I set for the book was to provide a justification for every rule. This could be based on the implications of a statistical model, or practical experience. Every rule is discussed under five headings: introduction, statement of the rule, illustration, basis of the rule, and discussion and extensions.

A book begins as a gleam in an author's eye and, like starlight, may take years to reach the viewer. This book is no exception. It goes back at least 10 years to a conversation with Bea Shube, then an editor at Wiley. In a letter to me from 1992 she said, "I want it, I want it, I want it!" Since then I have collected statistical rules of thumb in a nonsystematic and nonrandom fashion. It's a little late, but here it is.

It was the enthusiastic encouragement of Steve Millard that started a serious attempt at writing. He also suggested the format for each rule and produced a preliminary draft of Chapter 1. Tom Louis, a statistician's statistician, made many suggestions and shrewd observations. Jim Hughes has been a colleague for more than 20 years providing suggestions and comments on many aspects of the book. Mathematics graduate student Shawn Cokus helped me with the intricacies of LaTeX–giving the stuff the manuals don't tell: how to make it work on your machine. He made it look so easy that I had to ask him twice in some cases. Lance Jolley generated the initial graphs for Chapter 2; he did proofreading and provided overall support.

The National Research Center for Statistics and the Environment (Peter Guttorp, director) provided partial support for a year. As a visitor to NRCSE during 1998-1999 I worked on this book (among other things). Working with Peter Guttorp, Paul Sampson, and Mary Lou Thompson was always, and continues to be, a pleasure.

Ed Gehan provided the suggestion and reference to the rule about estimating the standard error from the range of the observations. Scott Emerson convinced me about the virtues of a power of 0.975. Clarice Weinberg shared a preprint of an article on p-values. I suspect that in spite of her wise remarks it won't be the end of the controversy.

Several classes of students in statistical consulting were given parts of the book. It's always exciting to share with students and get their creative and unjaded comments.

Roger Higdon and Bud Kukull read Chapter 4 making insightful comments. Margaret Pepe was extremely helpful with comments on screening, and other comments on Chapter 4. Jackie Benedetti and Nancy Temkin read and commented on Chapter 8. Michael Levitz helped me during the most crucial part of the writing: when graphs just would not come out right. He did most of the graphs and worked on the indexing. One of my sons, Louis, provided one of the graphs in Chapter 7 and also helped me set up the Web site.

The editors at Wiley, Steve Quigley, Heather Haselkorn, and Lisa Van Horn were helpful and straightforward. Amy Hendrickson provided excellent LaTeXadvice and produced the final camera ready copy. It's been a pleasure to work with them.

My colleagues in the departments of biostatistics, environmental health, statistics, and epidemiology provide a concentration of expertise and commitment that is phenomenal. It is at national meetings that you realize what an extraordinary group of faculty have been assembled at the University of Washington. I particularly want to single out Lloyd Fisher. Although he did not comment on the book directly he has been my life-long collaborator and friend. He represents for me the ideal combination of scholarship and integrity.

A person who has been my thought-companion all my professional life is Daniel B. DeLury (now deceased). I took my first statistics courses from him at the University of Toronto. His aphoristic statements still ring in my ears: "A frequency distribution

is a quality of permanence not inherent in the items themselves," and "Randomization puts systematic effects into the error term." Already in 1963 he was worried about "statistical analyses untouched by the human mind," surely even more of a worry forty years later.

Paul Crane read the manuscript after it was in "final" form. His suggestions for improvement were so perceptive and valuable that I spent several weeks implementing them. There are few pages that do not bear his constructive and critical markings.

Having acknowledged tremendous assistance and encouragement, it should be clear that I am responsible for the final content of the book. Mistakes are mine, not the friends listed above.

I expect that many readers will look for their favorite rule of thumb and not find it. Please accept my apologies. You may quarrel with a particular rule and feel you have a better rule. In both cases, send me an e-mail and I will incorporate it on my Web site (http://www.vanbelle.org). I plan to post a Rule of Thumb of the Month on this site. Perhaps a second volume can be written; a deep lode of statistical wisdom can be mined, and it is my hope that this book will stir a collective desire to bring this wisdom to the surface.

This book would not have seen the light of day without the loving, quiet but firm, encouragement of my wife, Johanna. This book is dedicated to her.

GERALD VAN BELLE

Seattle, Washington
January 15, 2002

Acronyms

Acronym	**Definition, page**
AIDS	Acquired Immunodeficiency Syndrome, 191
AIRS	Aerometric Information Retrieval System, 100
ANOVA	Analysis of Variance, 17
AR(1)	First Order Autoregressive Process, 174
ARE	Asymptotic Relative Efficiency, 138
ARR	Absolute Risk Reduction, 150
ASA	American Statistical Association, 229
CAST	Cardiac Arrhythmia Suppression Trial, 132
C	Cost of a study, 47
CD4	Cluster of Differentiation (on T4 cells), 191
CV	Coefficient of Variation, 32
DLR	Diagnostic Likelihood Ratio, 136
EBM	Evidence-Based Medicine, 143
EPA	Environmental Protection Agency, 87
EPV	Events Per Variable, 126
FDA	Food and Drug Administration, 9

FEV_1	Forced Expiratory Volume after 1 second, 91
GEE	Generalized Estimating Equations, 190
GLMM	Generalized Linear Mixed Models, 190
HEI	Health Effects Institute, 88
HIV	Human Immunodeficiency Virus, 131
HMO	Health Maintenance Organization, 156
IMS	Institute of Mathematical Statistics, 218
IRT	Item Response Theory, 82
ISI	International Statistical Institute, 229
ITT	Intent To Treat, 156
MEGO	My Eyes Glaze Over, 224
NMMAPS	National Morbidity and Mortality Air Pollution Study, 96
NGO	Non-Governmental Organization, 87
NNH	Number Needed to Harm, 152
NNT	Number Needed to Treat, 151
NPV	Negative Predictive Value, 135
NRC	National Research Council, 88
NUATRC	(Mickey Leland) National Urban Air Toxics Research Center, 88
PBPK	Physiologically Based Pharmacokinetic, 11
PC	Proportionate Change, 33
ppb	parts per billion, 213
ppm	parts per million, 177
PCA	Principal Component Analysis, 126
PCB	Polychlorinated biphenyl, 93
PPV	Positive Predictive Value, 135
PM	Particulate Matter, 97
PREV	Prevalence, 134
PVC	Premature Ventricular Contraction, 148
RCT	Randomized Clinical Trial, 144
ROC	Receiver-Operator Characteristic, 136
RR	Relative Risk, 150
RRR	Relative risk reduction, 150
SAS	Statistical analysis system (original name), 170

SE	Standard error, 35
SIDS	Sudden infant death syndrome, 116
SENS	Sensitivity, 134
SPEC	Specificity, 134
STAR	Science to achieve results, 87
TR	Treatment received: 156
TSP	Total suspended particulates, 103
URL	Universal (Uniform) Resource Locator, 232
WEB	Short version of "World Wide Web," 231
WHI	Women's Health Initiative, 7
WHO	World Health Organization, 54

1

The Basics

This chapter discusses some fundamental statistical issues dealing with variation, statistical models, calculations of probability and the connection between hypothesis testing and estimation. These are basic topics that need to be understood by statistical consultants and those who use statistical methods. The selection of these topics reflects the author's experience and practice.

There would be no need for statistical methods if there were no variation or variety. Variety is more than the spice of life; it is the bread and butter of statisticians and their expertise. Assessing, describing and sorting variation is a key statistical activity. But not all variation is the domain of statistical practice, it is restricted to variation that has an element of randomness to it.

Definitions of the field of statistics abound. See a sampling in van Belle et al. (2004). For purposes of this book the following characteristics, based on a description by R.A. Fisher (1935) will be used. Statistics is the study of populations, variation, and methods of data reduction. He points out that "the same types of problems arise in every case." For example, a population implies variation and a population cannot be wholly ascertained so descriptions of the population depend on sampling. The samples need to be reduced to summarize information about the population and this is a problem in data reduction.

1.1 FOUR BASIC QUESTIONS

Introduction

R.A. Fisher's definitions provide a formal basis for statistics. It presupposes a great deal that needs to be made explicit. For the researcher and the statistical colleague there is a broader program that puts the Fisher material in context.

Rule of Thumb

Any statistical treatment must address the following questions:

1. What is the question?
2. Can it be measured?
3. Where, when, and how will you get the data?
4. What do you think the data are telling you?

Illustration

Consider the question, "does air pollution cause ill health"? This is a very broad question that was qualitatively answered with the London smog episodes of the 1940s and 1950s. Lave and Seskin (1970), among others, tried to assess the quantitative effect and this question still with us today. That raises the non-trivial questions whether "air pollution" and "ill health" can be measured. Lave and Seskin review measures of the former such as sulfur in the air and suspended particulates. In the latter category they list morbidity and mortality. The third question of data collection was addressed by considering data from 114 Standard Metropolitan Statistical Areas in the U.S. which contained health information and other government sources for pollution information. The fourth question was answered by running multiple regressions controlling for a variety of factors that might confound the effect, for example age and socioeconomic status.

A host of questions can be raised but in the end this was a landmark study that anticipated and still guides research efforts today.

Basis of the Rule

The rule essentially mimics the scientific method with particular emphasis on the role of data collection and analysis.

Discussion and Extensions

The first question usually deals with broad scientific issues which often have policy and regulatory implications. Another example is global warming and its cause(s). But not all questions are measurable, for example, how do we measure human happiness or wisdom? In fact, most of the important questions of life are not measurable (a reason for humility). "Measurability" implies that there are "endpoints" which address the basic question. Frequently we need to take a short-cut. For example, income as a summary of socio-economic status. Given measurable values of the question we can then test whether one set of values differs from another. So testability implies measurability.

This raises the question whether the difference in the endpoints reflect an important difference in the first question. An example of this kind of question is the difference between *statistical significance* and *clinical significance* (It may be better to say *clinical relevance*–statistical significance may point to a very important mechanistic framework). In this context there also needs to be careful considerations of measurements that are not taken. This issue will be addressed in more detail in the chapter on observational studies.

If it is agreed that the question is measurable the issue of data selection or data creation comes up. The three subquestions that focus the discussion. They locate the data selection in space and time, and context. The data can range from administrative data bases to experimental data, they can be retrospective or prospective. The "how" subquestion deals with the process that will actually be used. If sampling is involved, the sampling mechanism must be carefully described. In studies involving animals and humans this, especially, requires careful attention to ethics (but not restricted to these, of course). Broadly speaking there are two approaches to getting the data: observational studies and designed experiments.

The next step is analysis and interpretation of the data which, it is hoped, answers questions 1 and 2. Questions 1-3 focus on design, ranging from collecting anecdotes to doing a survey sample to conducting a randomized experiment. Question 4 focuses on analysis–in which statisticians have developed particular expertise (and sometimes ignore questions 1–3 by saying. "Let X be a random variable . . ."). But is is clear that the answers to the questions are inextricable interrelated. Other issues implied by the question include the statistical model that is used, the robustness of the model, missing data, and an assessment of the many sources of variability.

The ordering reflects the process of science. Data miners who address only question 4 do so at their own risk.

1.2 OBSERVATION IS SELECTION

Introduction

The title of this rule is from Whitehead (1925)–so the idea is not new. This is perhaps the most obvious of rules; and is not taken into account the majority of the time.

Rule of Thumb

Observation is selection.

Illustration

The observation may be straightforward but the selection process not. An example that should be better known (selection?) is the vulnerability analysis of planes returning from bombing missions during Word War II. Aircraft returning from missions had been hit in various places. The challenge was to determine which parts of the plane to reinforce to decrease their vulnerability. The naive approach started with figuring out where the hits had occurred. A second and improved approach was to adjust the number of hits by the area of the plane. The third step recommended by the statistician Abraham Wald: reinforce the planes where they had not been hit! His point was that the observations were correct, but not the selection process. What was of primary interest were the planes that did not return. Using an insightful statistical model he showed that the engine area (showing the fewest hits in returning planes) was the most vulnerable. This is one of those aha! situations where we immediately grasp the key role of the selection process. See Mangel and Samaniego (1984) for the technical description and references.

Basis of the Rule

To observe one thing implies that another is not observed, hence there is selection. This implies that the observation is taken from a larger collective, the statistical "population."

Discussion and Extensions

Often the observation is of interest only in so far as it is representative of the population we are interested in. For example, in the vulnerability analysis, the plane that provided the information about hits might not have been used again and scrapped for parts.

Selection can be subconscious as when we notice Volvo cars everywhere after having bought one. Thus it is important to be able to recognize the selection process.

The selection process in humans is very complicated as evidenced by contradictory evidence by witnesses of the same accident. Nisbett and Ross (1980) and Kahneman, Slovic, and Tversky (1982) describe in detail some of the *heuristics* we use in selecting information. The bottom line is "know your sample."

1.3 REPLICATE TO CHARACTERIZE VARIABILITY

Introduction

A fundamental challenge of statistics is to characterize the variation that we observe. We can distinguish between *systematic variation* and non-systematic variation which sometimes can be characterized as *random variation*. An example of systematic variation is the mile-marker on the highway or the kilometer-marker on the autobahn.

This kind of variation is predictable. Random variation cannot be described in this way. In this section we are concerned with random variation.

Rule of Thumb

Replicate to characterize random variation.

Illustration

Repeated sampling under constant conditions tends to produce replicate observations. For example, the planes in the previous illustration have the potential of being consider replicate observations. The reason for the careful wording is that many assumptions need to be made such as the planes have not been altered in some way that affects their vulnerability, the enemy has not changed strategy, and so on.

At a more mundane level, the baby aspirin tablets we take are about as close to replicates as we can imagine. But even here, there are storage requirements and expiration dates that may make the replications invalid.

Basis of the Rule

Characterizing variability requires repeatedly observing the variability since the it is not a property inherent in the observation itself.

Discussion and Extensions

The concept of replication is intuitive but difficult to define precisely. The idea of constant conditions is technically impossible to achieve since time marches on. Marriott (1999), defines replication as "execution of an experiment or survey more than once so as to increase precision and to obtain a closer estimation of sampling error." He also makes a distinction between replication and repetition, reserving the former for repetition at the same time and place.

In agricultural research the basic replicate is called a plot. Treatments can be compared by using several plots to each treatment so that the variability within a treatment is replicate variability.

There is one method that ensures replication: randomization of observational units to two or more treatments. More will be said about in the chapter on design.

1.4 VARIABILITY OCCURS AT MULTIPLE LEVELS

Introduction

As soon as the concept of variability is grasped it becomes clear that there are many sources of variability. Again, here the sources may be systematic or random. The emphasis here is, again, on random variability.

Rule of Thumb

Variability occurs at multiple levels.

Example

In education there is clearly variation in talents from student to student, from classroom to classroom, from school to school, from district to district, from country to country. In this example there is a hierarchy with students nested within schools and so on.

Basis of the Rule

The basis of the rule is the recognition that there are levels of variation.

Discussion and Extensions

Each level of an observational hierarchy has its own units and its own variation. Suppose that the variable is expenditure per student. This could be expanded to expenditure per classroom, school or district. In order to standardize the expenditure per student could be used but for other purposes it may be useful to compare expenditure at the district level. However, if district are compared then the number of students served is usually considered. The number of students would be a *confounder* in comparison of districts. More will be said about confounders in Chapter 3.

1.5 INVALID SELECTION IS THE PRIMARY THREAT TO VALID INFERENCE

Introduction

The challenge is to be able to describe the selection process–a fundamental problem for applied statisticians. *Selection bias* occurs when the sample is not *representative* of the population of interest; this usually occurs when the sampling is not random. For example, a telephone survey of voters excludes those without telephones. This