Reading Materials
for
Epid 9510a

Biostatistical Research Methods

Complied by

GY Zou & Yun-Hee Choi

July 28, 2009

Department of Epidemiology & Biostatistics • Schulich School of Medicine & Dentistry
The University of Western Ontario
Kresge Building • Room K201 • London, Ontario • N6A 5C1 • Canada
Telephone: (519) 661-2162 • Fax: (519) 661-3766 • www.schulich.uwo.ca/epidem

Epid 9510: Biostatistical Research Methods

Course Description: This course is meant to give graduate students in the biostatistical collaborated program an introduction to the necessary skills and knowledge for biostatistical research, focusing on the activities involved in writing thesis. The ultimate goal is to enable students to fully understand the strengths and limitations of new biostatistical methods developed by others, or develop new methods on their own.

Instructors: GY Zou (gzou@robarts.ca) and
YH Choi (yun-hee.choi@schulich.uwo.ca).

Textbook: No textbooks. Reading materials provided.

Class meetings: Tuesday 10:30 to 12:30, Thursday 10:30 to 12:00, Kresge 116.

Grade: Assignment 1: 20%, manuscript 70%, Participation: 10%.

Course outline

Sept 15: Introduction
Sept 17: Intro SAS

Sept 22: Intro R (Choi)
Sept 24: Intro LaTeX (Choi)

Sept 29: Stat Inference (Choi)
Oct 1: Stat Inference (Choi)

Assignment 1: Derive var(ρ_A), var(ρ_{FC}), and var(ρ_P) in Zou and Donner (2004, Biometrics 60: 807-811). Use LaTeX for typing.

Sept 6: Stat Inference (Choi)
Oct 8: Stat Inference (Choi)

Oct 13: How to select topic and Read
Oct 15: How to Search

Assignment 2: select a topic and literature list
Assignment 1 due on Oct 9.

Oct 20: Proportion as an example for reading and searching
Oct 22: Reading Classical statistical papers

Oct 27: Bootstrap
Oct 29: Bootstrap

Assignment 2 due on Oct 23.

Oct 3: Simulation by SAS
Oct 5: Simulation by R (Choi)

Nov 10: Writing paper using LaTeX
Nov 12: Writing thesis using LaTeX

Nov 17: Writing in English (with examples of Zou)
Nov 19: Writing in English (cont)

Nov 24: Writing slides with LaTeX
Nov 26: Effective presentation

Dec 1: Student presentation
Dec 3: Student presentation

Dec 17: Manuscript due (to be Graded by Zou and Choi before Dec 20)

Special Section: Statistical Training and Curricular Revision

What Is Statistics?

Emery N. BROWN and Robert E. KASS

We use our experience in neuroscience as a source of defining issues for the discipline of statistics. We argue that to remain vibrant, the field must open up by taking a less restrictive view of what constitutes statistical training.

KEY WORDS: Cross-disciplinary statistical research; Statistical paradigm; Statistical thinking.

1. SHORT SUPPLY

Our field faces fundamental challenges. The statistical needs of science, technology, business, and government are huge and growing rapidly, producing a shortfall in statistical workforce production. In their summary of an National Science Foundation workshop, *The Future of Statistics*, Lindsay, Kettenring, and Siegmund (2004) reported that

> Workshop participants pointed repeatedly to shortages in the pipeline of students and unmet demand from key industries and government laboratories and agencies. . . . The shortage may prove quite damaging to the nation's infrastructure.

The growth in demand for data analysis may be attributed in large part to the exponential increase in computing power and data collection capabilities. At the same time, there is a worrisome tendency for quantitative investigators or technical staff to attack problems using blunt instruments and naive attitudes. Our discipline as a whole has been gloriously productive, making available a wide variety of tools. But we have been less successful in producing easy-to-master operating instructions and training programs. We have effectively created a supply side of the problem: Statistical education has not been sufficiently accessible. Curricula in statistics have been based on a now-outdated notion of an educated statistician as someone knowledgeable about existing approaches to handling nearly every kind of data. Degrees in statistics have emphasized a large suite of techniques, and introductory courses too often remain unappetizing. The net result is that at every level of study, gaining statistical expertise has required extensive coursework, much of which appears to be extraneous to the compelling scientific problems students are interested in solving.

We also must acknowledge that some of the most innovative and important new techniques in data analysis have come from researchers who would not identify themselves as statisticians. Computer scientists have been especially influential in the past decade or so. The influx of methodology from outside the discipline is not new; indeed, the field of statistics itself is relatively young, with much foundational achievement predating the advent of departments of statistics. But an undeniable fear lurks in the hearts of many statistics professors: As others leap daringly into the fray, attempting to tackle the most difficult problems, might statistics as we know it become obsolete?

The two of us recently co-organized the fourth international workshop on statistical analysis of neural data. This series of conferences has brought together quantitatively oriented experimenters and cutting-edge data analysts working in the field of neuroscience, offering new challenges for statistical science in the process. We and others have found the high quality of statistical application gratifying and the articulation of new ideas very stimulating. One of the reactions from readers of our grant proposal to the National Science Foundation took us by surprise, however. Only a relatively small minority of our speakers and participants came from departments of statistics, and as a result, some reviewers questioned whether the Division of Mathematical Sciences should be supporting this activity. Luckily, the program officers handled this issue adeptly, in part by getting cosponsorship from Computational Neuroscience. But the issue is an aspect of the existential identity crisis; the reviewers were grappling with the vexing question, raised by institutional structures, of who should be counted as a statistician.

The participation in neuroscientific research of many nonstatisticians doing sophisticated data analysis is not surprising. The brain is considered a great scientific frontier. Studying it creates many technological challenges, and because neuronal networks form electrical circuits, fundamental contributions to neurophysiology have been made by physical arguments, in the

Emery N. Brown is Warren M. Zapol Professor of Anaesthesia, Harvard Medical School, Department of Anesthesia and Critical Care, Massachusetts General Hospital, Boston, MA 02114 and Professor of Computational Neuroscience and Health Sciences and Technology, Department of Brain and Cognitive Sciences, MIT-Harvard Division of Health Science and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139 (E-mail: *enb@neurostat.mit.edu*). Robert E. Kass is Professor, Department of Statistics, Center for the Neural Basis of Cognition, and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15217 (E-mail: *kass@stat.cmu.edu*). The thoughts herein have resulted from many extended discussions with colleagues, especially in the Department of Statistics at Carnegie Mellon, where Kass was Department Head from 1996 to 2005 and Brown serves on the external advisory board. Brown's research was partially supported by Grants DP1 OD003646, R01 MH59733, and R01 MH071847. Kass' research was partially supported by Grants RO1 MH064537, RO1 EB005847, and R90 DA023426.

 DOI: 10.1198/tast.2009.0019

form of differential equations. Furthermore, brain science is where artificial neural network models arose, not as machines for nonparametric multiple regression, but rather as descriptors of cognitive mechanisms. For these reasons, neuroscience has attracted many researchers trained in quantitative disciplines, especially physics and engineering. Although their activities might make some statisticians nervous when it comes to federal grants and other resources, a more serious threat is a disciplinary attitude that contrasts strikingly with what we see among many statisticians. Physicists and engineers very often become immersed in the subject matter. In particular, they work hand in hand with neuroscientists and often become experimentalists themselves. Furthermore, physicists and engineers (and likewise computer scientists) are ambitious; when faced with problems, they tend to attack, sweeping aside impediments stemming from limited knowledge about the procedures that they apply. In seeing this, we often shudder, and we criticize this cavalier attitude later in this article. But there is a flip side to our reaction; in contrast, we find that graduate students in statistics often are reticent to the point of inaction. Somehow, in emphasizing the logic of data manipulation, teachers of statistics are instilling excessive cautiousness. Students seem to develop extreme risk aversion, apparently fearing that the inevitable flaws in their analysis will be discovered and pounced upon by statistically trained colleagues. Along with communicating great ideas and fostering valuable introspective care, our discipline has managed to create a culture that often is detrimental to the very efforts it aims to advance.

We are worried. While we expect that in many institutions—perhaps most—there may exist specific courses and programs that are exemplary in certain respects, in the aggregate, we are frustrated with the current state of affairs. The concerns that we have articulated here are not minor matters to be addressed by incremental improvement; rather, they represent deep deficiencies requiring immediate attention.

2. CHANGING TIMES

In making critical comments, we hope to stir discussion and debate. We do not wish to be misunderstood, however; our most fundamental loyalty is to the discipline of statistics. We appreciate its role in technical advances over the past century, and see even greater opportunities for essential contributions in the future, as scientific investigations rely on more massive and intricate data sets to examine increasingly complex phenomena. Furthermore, besides utility, there is great beauty in the subject. We have spent considerable effort learning and trying to advance neuroscience. But even after substantial exposure to one of the most exciting and rapidly developing areas of science, we still believe that statistics, with its unique blend of real-world mathematics, epistemology, and computational technique, is the most deeply interesting and rewarding of all intellectual endeavors. There are strong arguments to suggest that much of cognition is based on pattern learning, and that humans have well-developed neural machinery for making inferences implicitly, without conscious recognition. Perhaps part of the pleasure that we get from statistical reasoning comes from bringing a harmonious coherence to otherwise unappreciated brain processes. Regardless of its biological explanation, however, there is certainly an inspiring aesthetic of statistics driven in part by the emotional overlay of trying to tame uncertainty. The problem is not with the nature of the discipline. There are compelling reasons to love statistics and to pass on to others both knowledge of its methods and appreciation of its powerful logic.

So where have things gone wrong? We believe that the primary source of the current difficulties is an anachronistic, yet pervasive conception of statistics. The problem is that departments of statistics often act as if they are preparing students to be short-term consultants, able to answer circumscribed methodological questions based on limited contemplation of the context. This short-term consultant model relegates the statistician to a subsidiary position, and suggests that applied statistics consists of handling well-formulated questions, so as to match an accepted method to nearly any kind of data. This notion may have developed partly because—at least in the United States—statistics evolved from mathematics with its lone investigator, and partly because a qualified statistician could know the entire field. The large majority of senior statisticians began their academic careers as math majors. Within statistics departments, mathematical thinking influenced both research and infrastructure, whereas the mathematics involved was relatively limited, so that Ph.D. statisticians could master the technical details in diverse areas of statistics. Graduate programs thus emphasized mathematically thorough knowledge of multiple branches of the field. At one time, this served a useful purpose. But statistics has expanded and deepened, so that individuals rarely have state-of-the-art, rigorous expertise in more than a few well-developed subdomains. Furthermore, in today's dynamic and interdisciplinary world, success in confronting new analytical issues requires both substantial knowledge of a scientific or technological area and highly flexible problem-solving strategies. In neuroscience, for example, a statistician will have far more impact once he or she is able to generate ideas for scientific investigation. In other fields, the situation is surely analogous. The discipline of statistics needs to recognize our new situation and act accordingly. We suggest two overarching principles of curricular revision.

3. A FOCUS ON STATISTICAL THINKING

According to syllabi and lists of requirements, statistics courses and degree programs tend to emphasize mastery of technique. But statisticians with advanced training and experience do not think of statistics as simply a collection of methods; like experts in any field, they consider their subject highly conceptual. This deserves emphasis, because it distinguishes a disciplinary approach from efforts that might be disparaged as the work of amateurs. In neuroscience, we have seen many highly quantitative researchers trained in physics and engineering, but not in statistics, apply sophisticated techniques to analyze their data. These often are appropriate and sometimes are inventive and interesting. In the course of perusing many, many articles over the years, however, we have found ourselves critical of much published work. Starting with vague intuitions, particular algorithms are concocted and applied, from which strong scientific statements are made. Our reaction too often is negative;

we are dubious of the value of this approach, believing that alternatives are preferable. Or we may concede that a particular method possibly may be a good one, but the authors have done nothing to indicate that it performs well. In specific settings, we often come to the conclusion that the science would advance more quickly if the problems were formulated differently—in a manner more familiar to trained statisticians. As an example, neuroscientists developed the highly intuitive "spike-triggered average" to identify an association between a neural spike train, which may be considered a point process, and a continuous stimulus. Point process analysis by a member of Columbia's Department of Statistics (Paninski 2003) has shown that spike-triggered averaging can be inconsistent in some realistic settings, but that consistent estimators may be constructed using generalized linear (or nonlinear) regression models, an approach first championed by Brillinger. (For related references and other examples, see Brown, Kass, and Mitra 2004; Kass, Ventura, and Brown 2005.)

The statistician's perspective, missing from much analysis of neural data, is the most important thing that we can provide. Once students have it, they will be empowered in diverse situations. Thus, we suggest that the primary goal of statistical training at all levels should be to help students develop *statistical thinking.*

What exactly do we mean by this? Different statisticians would use somewhat different words to describe what defines the essential elements of our discipline's approach, but we believe there is general consensus about the substance, which can be stated quite concisely. Statistical thinking uses probabilistic descriptions of variability in (1) inductive reasoning and (2) analysis of procedures for data collection, prediction, and scientific inference. For instance, a prototypical description of variability among data pairs $(x_1, y_1), \ldots, (x_n, y_n)$ is the nonparametric regression model

$$Y_i = f(x_i) + \varepsilon_i,$$

in which each ε_i is a random variable. This may be used to suggest methods of smoothing the data and to express uncertainty about the result [both of which are part of item (1)] and also to evaluate the behavior of alternative smoothing procedures [item (2)]. One can dream up a smoothing method, and apply it, without ever referencing a model—indeed, this is the sort of thing that we witness and complain about in neuroscience. Meanwhile, among statisticians there is no end of disagreement about the details of a model and the choice among methods (What space of functions should be considered? Should the ε_i random variables enter additively? Independently? What class of probability distributions should be used? Should decision-theoretic criteria be introduced, or prior probabilities?). The essential component that characterizes the discipline is the introduction of probability to describe variation in order to provide a good solution to a problem involving the reduction of data for a specified purpose. This is not the only thing that statisticians do or teach, but it is the part that identifies the way they think. We provide a bit more discussion of this notion in the Appendix.

Currently, statistical thinking is internalized as a byproduct of extensive statistical training. Elevating it to an overarching goal allows curricula to be assessed according to the way in which statistical thinking is engendered.

4. FLEXIBLE CROSS-DISCIPLINARITY

Contemporary students see before them a world dominated by "big science," with a host of exciting paths to participate in progress. Many students recognize a fundamental role for statistics, and most see great value in learning statistical methods, but they are increasingly motivated by a desire to solve important problems. In this context, the very best quantitatively oriented students often come from other quantitative disciplines, including computer science, physics, and engineering, and they have many options.

As an example, because of his involvement in computational neuroscience at Carnegie Mellon, one of us (Kass) became aware of an outstanding senior undergraduate, a young woman majoring in computer science at one of the top liberal arts colleges, with nearly perfect GPA and GRE score. She was very interested in computational aspects of neuroimaging and wanted to pursue a Ph.D. However, she had never taken a statistics course, and in fact had taken only one math course beyond calculus. It had not occurred to her that statistics might be a good option, and, from the standpoint of admission to a graduate program in statistics, she presented logistic complications; it was not clear exactly what she would study, or how many years it would take to complete her degree. We must make room for students like this and recruit them.

To attract students with nontraditional quantitative backgrounds, statistics programs must guide these students toward making important contributions in a timely manner. Cross-disciplinary projects will have to play a major role. Once a department accepts as its primary mission helping students develop an ability to think like statisticians, it is freed from the constraints of excessive content and can recognize alternative ways that students can demonstrate their abilities and achievements. On the one hand, we see cross-disciplinary work as essential to anyone with any kind of statistical credentials—and thus to statistical training at every level. On the other hand, we view cross-disciplinary research as an opening to students of varied backgrounds—a way of welcoming them into the fold and a mechanism for streamlining training, making programs more manageable and the discipline more inviting.

To satisfy different kinds of students, programs also must allow multiple pathways toward degrees. Increasing the emphasis on cross-disciplinarity goes hand in hand with reducing the importance of particular courses and thereby decreases programmatic rigidity. Flexibility is paramount. We do not wish to remove theoreticians from our midst; indeed, many nonmathematicians will blossom in theoretical directions. Rather, our aim is to allow a broader notion of who counts as a statistician.

5. IMPLICATIONS

If someone is able to (i) appreciate the role of probabilistic reasoning in describing variation and evaluating alternative procedures and (ii) produce a cutting-edge cross-disciplinary analysis of some data, should we feel comfortable calling that person a statistician? We think so, and we would like to see our profession broaden its perspective to a sufficient degree to make this possible.

We further believe that it is consequential to declare (i) and (ii) to be defining goals for a training program. In applying this at the graduate level, however, we presume that to do "cutting edge" work, along the way a trainee would have had to have learned something about classical techniques, such as linear regression, some area of modern statistics (e.g., nonparametric regression, dimensionality reduction, graphical models), and also general inferential tools, such as the bootstrap and Bayesian methods. Furthermore, appreciation of probabilistic reasoning comes from repeated exposure to it in varied contexts. Both of these require mathematical and computational skills. Thus, we are proposing variations on what is currently in place in training programs throughout the country; each training program formulates (explicitly or implicitly) a list of skills and units of knowledge that are truly essential, and figures out how the items on the list are to be taught and evaluated. What constitutes inculcation of statistical thinking may be in the eye of the beholder—in this case, the departmental training program. On the other hand, we have argued that the status quo is unacceptable. Here are four recommendations.

1. *Minimize prerequisites to research*. There are continual disagreements about the stage at which trainees should do research. We strongly favor making cross-disciplinary projects widely available, even to those with minimal backgrounds. Although advanced trainees will have more tools at their disposal, talented quantitatively oriented students can quickly learn how to apply and interpret statistical techniques without formal coursework—indeed, we witness this repeatedly in neuroscience. There has been a tendency in statistics to have students first understand, then do. But this sequence can be reversed, giving a statistical faculty supervisor the opportunity to demonstrate in practice the value of knowing the theoretical underpinnings of methodology. Perhaps most importantly, as we stated earlier, students who want to solve real problems will be attracted to cross-disciplinary research. At both the graduate and undergraduate levels, exciting research opportunities are likely to be among the best recruitment tools.

2. *Identify ways of fostering statistical thinking*. How should we help our students internalize a principled approach to data collection, prediction, and scientific inference? Appreciation of statistical thinking should begin in introductory courses. Each instructor of a first course in statistics grapples with ideas behind reasoning from data, and much effort has gone into texts for such classes. Although we recognize the many great strides taken by textbook authors, we are not entirely satisfied with the typical content of introductory courses. For example, in teaching young neurobiologists, we have found it helpful to stress the value of probabilistic reasoning through propagation of uncertainty via simulation methods—as in bootstrap confidence intervals or Bayesian inference—and to emphasize "principles" by including explicit discussion of mean squared error. Both topics seem more advanced than what is usually found in elementary texts. To be attracted to the subject, however, the most gifted students must see it as deep, with serious theoretical content. Courses tend to be categorized as either theoretically oriented for math/statistics majors or method-oriented "service courses" for other disciplines, and we find too little similarity between the two. The main point here is that the first college-level exposure to statistics matters. Although for pedagogical purposes, central ideas must remain simple and approachable, we believe that it is important to represent the discipline as being rich in profound concepts. More fundamentally, one goal of every first course in statistics for quantitatively capable students should be to interest some of the students in further study.

At the graduate level, existing curricula succeed in getting students to think like statisticians, but focus on this goal is necessary if programs are to be streamlined. Students will still need exposure to statistical reasoning in multiple diverse settings, together with emphasis on (a) the roles of heuristics, computational considerations, and/or generative models in producing procedures and (b) theoretical performance, balanced by convenience, computational efficiency, and interpretability. Many excellent books on such topics as nonparametric regression, density estimation, time series analysis, and Bayesian methods offer very good comparative discussions combining both theoretical and practical concerns. The only problem we see is that they are designed for full-semester courses, whereas in many cases the modern student may wish to devote only a couple of weeks to each within formal course work. We believe that there is an important place for courses, and texts, that give quick impressions while reinforcing underlying principles.

We also take it for granted—but nonetheless believe it worth mentioning— that training programs at every level should include many opportunities for trainees to interact with experienced statisticians (in, e.g., journal clubs, informal seminars, social events), partly to see how they think about problems, but also to have role models reinforce the joys and benefits of pursuing statistics.

3. *Require real-world problem solving*. Experienced statisticians spend much of their collaborative time trying to understand the nature of the data collection process and its relationship to scientific or technological issues. Some students, especially those with backgrounds in experimental science, tend to be well prepared in this dimension, asking appropriate questions, digging up background material, and readily grasping the big picture. Many others, however, have difficulty making connections among scientific ideas, the resulting data, and appropriate analytic strategies. Having recognized this basic skill for applied statistics, we must help our students develop it. Several methods for doing so exist. Project courses, especially at the undergraduate level, can be helpful. Extended research projects—learning by doing—can of course be among the best ways to develop problem-solving skills. An important caveat, however, is that some projects are so well formulated that execution becomes straightforward, and little effort toward big-picture comprehension is needed. We come across students who in the course of doing statistical analyses exhibit remarkably little curiosity about the material they are analyzing. Most likely this is because they have not been taught a systematic approach to problem solving and do not appreciate the payoff from pursuing it.

4. *Encourage deep cross-disciplinary knowledge*. In neuroscience, as elsewhere, statistical training can shape how data lead to useful knowledge. Once the information obtainable from an experiment is clearly understood, a new aspect of the scientific landscape may come into view. Consequently, statisticians

can make major contributions by redefining problems and redirecting data-collection efforts.

In this regard, we distinguish two alternative roles. The first role has been played by both of us; like other senior statisticians in varied domains, we have spent many years learning scientific principles and methods and building collaborations with colleagues, so that our suggestions for research problems and approaches are taken seriously and often followed. The second role requires a deeper commitment to cross-disciplinary training, however. One of us (Brown) became a practicing anesthesiologist in addition to being a statistician. As a result of his extensive physiological knowledge and expertise, he has been able to create a laboratory and is undertaking a series of experiments on brain activity to describe how anesthetic drugs produce the state of anesthesia. Many others in the profession play a similar "principal investigator" role. Two examples are John Quackenbush in the Biostatistics Department in the Harvard School of Public Health and the Dana Farber Cancer Institute, who formulates and executes experiments that use genomic and computational approaches to study networks and pathways in cancer development and progression, and Wing Wong in the Department of Statistics at Stanford University, who conducts experiments on developmental genomics and signal transduction that are informed by statistical considerations.

Faculty who run extradisciplinary experiments and contribute to disciplinary methodology are becoming fairly common in engineering and physics, but not in statistics. The change in attitude that we advocate should in time produce more such people in departments of statistics. In addition to accepting the desirability of these appointments, however, more joint training programs are needed. As models in neuroscience, we can point to our own institutions. The Harvard/MIT Health Sciences and Technology Ph.D. program trains students in quantitative subjects while also having them take substantial medical school courses and serve on rotations in the hospital as a medical student would. Carnegie Mellon's Ph.D. Program in Neural Computation is similar, requiring mastery of a technical discipline (e.g., computer science or statistics) together with multiple courses in the brain sciences, and rotation through experimental laboratories. Again, to attract large numbers of students, course requirements in interdisciplinary programs must be stripped down to manageable essentials. We would like to see more such joint programs that offer credentials in statistics.

6. DISCUSSION

The report by Lindsay, Kettenring, and Siegmund (2004) was aimed at the general community of mathematical scientists. Our discussion has been inward-looking, and critical. Although there is much to be admired in statistical training programs throughout the world, we accuse them of harboring obsolete attitudes about the nature of statistics. Statistics is a wonderful field, but the way in which statisticians view it must evolve. We have suggested defining what our discipline brings to the table, labeling the perspective that we believe to be so fundamentally valuable "statistical thinking." We also have advocated greater encouragement of cross-disciplinary training. Deepening cross-disciplinary involvement and welcoming more experimentalists and other practitioners into the clan of statisticians need not diminish the importance of the theoretical core. Quite the contrary; those with hands-on knowledge of context-driven issues can help identify methodological problems, prodding theory to advance in productive new directions.

Our first main message is that training programs should have a clearer notion of what they intend do. The second message is that these programs generally need to strengthen and deepen their commitment to cross-disciplinary work. In this, we follow many others. We have emphasized the contrast between short-term consulting and deeper, long-term engagement, which require different attitudes and skills. We are sympathetic to the promise made by Birnbaum (1971) that "each student of statistics working with me at any level shall also work systematically with another study adviser representing a scientific or technological research discipline of interest to the student," and we agree with Gnanedesikan (1990) that training should focus less on defining the appropriate encompassing content and more on instilling a relevant sense of cross-disciplinary curiosity: "We need a switch turned on, a value established, for impelling statisticians to be challenged intellectually and through a desire to contribute to solving major problems in other fields."

The worth of cross-disciplinary work, and its essential role in stimulating new statistical theory and methods, seems to be much more widely appreciated now than in the past. We want to push harder, partly because we feel that curricular ramifications have not been given sufficient attention, but also because the world needs more statistically oriented scientific principal investigators. Such scientific leadership is, again, not just a recent development. As one example, in the mid-1970s, Fred Mosteller, a master at initiating interdisciplinary collaborations on topics he deemed scientifically important, became interested in the benefits of surgical therapies, which typically are not studied using randomized controlled clinical trials. This led to his formulation of a large research effort involving statisticians, surgeons, anesthesiologists, and public health specialists to investigate the costs, risks, and benefits of surgery (Bunker, Barnes, and Mosteller 1977). Mosteller was not trained in surgery, but he was clearly the intellectual leader of the project. This kind of leadership is not limited in any way to areas in which "principal investigator" has a literal meaning in a biomedical context. As emphasized by Keller-McNulty (2007), many of today's big challenges throughout society are tackled by large teams, and these teams are in desperate need of statistical thinking at the very top levels of management. We suggest that a way forward begins with a focus on the fundamental goals of training, combined with a broad vision of the discipline of statistics.

APPENDIX: WHAT IS STATISICAL THINKING?

Snee (1990) noted that "many of us talk about statistical thinking but rarely define it." Although the field is so broad that a single notion of statistical thinking cannot possibly be universally applicable, we provided above a succinct definition coming from our own experience that we believe articulates a widely held consensus. We are, at least, in line with Ru-

bin (1993) when he said that

> The special training statisticians receive in mapping real problems into formal probability models, computing inferences from the data and models, and exploring the adequacy of these inferences, is not really part of any other formal discipline, yet is often crucial to the quality of empirical research.

Similarly, Mallows (1998) wrote that

> Statistical thinking concerns the relation of quantitative data to a real-world problem, often in the presence of variability and uncertainty. It attempts to make precise what the data has to say about the problem of interest.

In combining these points of view, we wished to recognize the centrality of probabilistic reasoning while distinguishing two roles for it. First, there is the inductive movement from description of variation to expressions of knowledge and uncertainty. A probabilistic description of variation would be "the probability of rolling a 3 with a fair die is 1/6," whereas an expression of knowledge would be "I'm 90% sure that the capital of Wyoming is Cheyenne." These two sorts of statements, which use probability in different ways, are sometimes considered to involve two different kinds of probability, called "aleatory probability" and "epistemic probability." Bayesians merge these, applying the laws of probability to go from quantitative description to quantified belief, but in every form of statistical inference, alleatory probability is used somehow to make epistemic statements. This is the first role of probabilistic reasoning. The second role is in evaluating procedures. We understand statistical thinking to be based on these two roles for probabilistic reasoning. This allows us to elaborate our definition of statistical thinking by stating that it involves two principles:

1. Statistical models of regularity and variability in data may be used to express knowledge and uncertainty about a signal in the presence of noise, via inductive reasoning.
2. Statistical methods may be analyzed to determine how well they are likely to perform.

The downside of spelling out a definition is that it can be easy to get sidetracked on the details. For starters, we intend "signal" to denote general underlying phenomena and relationships of interest, whereas "noise" refers to sources of variation that are being separated from the signal. We find these terms helpful partly because the nonparametric regression model, where they become explicit, is a useful archetype. Furthermore, we believe that there is at least some modest historical evidence to support the importance of such a basic dichotomy. Stigler (1999) considered why psychology adopted statistical methods so much earlier than economics or sociology, and why astronomy did do so even earlier. His answer was that the theory of errors, arising in astronomy, was based on a conceptualization encapsulated by "observation = truth + error," and that psychophysics was able to introduce this to psychology via careful experimental design. Using our words, this suggests that the idea of considering data to be generated by combining signal and noise was essential to the historical development of statistical thinking.

A related detail is that, just as there are disagreements about the subtleties of the nonparametric regression model and its application, there are important issues surrounding the role of modeling in statistics. We intend to use "statistical model" very broadly, with the only restriction being that probability is involved, so that the notion covers models with relatively weak assumptions, as in a two-sample permutation test, or strong assumptions, as in many Bayesian multilevel hierarchical models. Our formulation cannot accommodate the perspective of Breiman (2001), but we believe that it is entirely consistent with the views given in discussions of that article by Cox (2001) and Efron (2001). Here we are also remaining agnostic about the extent to which a model may be "explanatory" or "empirical," as discussed by Cox (1990) and Lehmann (1990), recognizing that "[these descriptions] represent somewhat extreme points of a continuum" (Kruskal and Neyman 1956). Rather, we believe that when Box (1979) stated that "all models are wrong, but some are useful," he was expressing a quintessential statistical attitude.

[Received September 2008. Revised September 2008.]

REFERENCES

Birnbaum, A. (1971), "A Perspective for Strengthening Scholarship in Statistics," *The American Statistician*, 25, 14–17.

Box, G. E. P. (1979), "Robustness in the Strategy of Scientific Model Building," in *Robustness in Statistics*, eds. R. L. Launer and G. N. Wilkinson, New York: Academic Press.

Breiman, L. (2001), "Statistical Modeling: The Two Cultures" (with discussion), *Statistical Science*, 16, 199–231.

Brown, E. N., Kass, R. E., and Mitra, P. (2004), "Multiple Neural Spike Train Analysis: State-of-the-Art and Future Challenges," *Nature Neuroscience*, 7, 456–461.

Bunker, J. P., Barnes, B. A., and Mosteller, F. (1977), *Costs, Risks, and Benefits of Surgery*, Oxford: Oxford University Press.

Cox, D. R. (1990), "Role of Models in Statistical Analysis," *Statistical Science*, 5, 169–174.

——— (2001), Comment on "Statistical Modeling: The Two Cultures," by L. Breiman, *Statistical Science*, 16, 216–218.

Efron, B. (2001), Comment on "Statistical Modeling: The Two Cultures," by L. Breiman, *Statistical Science*, 16, 218–219.

Gnanedesikan, R. (1990), "Looking Ahead: Cross-Disciplinary Opportunities for Statistics," *The American Statistician*, 44, 121–125.

Kass, R. E., Ventura, V., and Brown, E. N. (2005), "Statistical Issues in the Analysis of Neuronal Data," *Journal of Neurophysiology*, 94, 8–25.

Keller-McNulty, S. (2007), "From Data to Policy: Scientific Excellence Is Our Future," *Journal of the American Statistical Association*, 102, 395–399.

Kruskal, W., and Neyman, J. (1956), "Stochastic Models and Their Applications to Social Phenomena," unpublished lecture at Joint Statistical Meetings, Detroit; referenced by E. L. Lehmann (1990).

Lehmann, E. L. (1990), "Model Specification: The Views of Fisher and Neyman, and Later Developments," *Statistical Science*, 5, 160–168.

Lindsay, B. G., Kettenring, J., and Siegmund, D. O. (2004), "A Report on the Future of Statistics," *Statistical Science*, 19, 387–413.

Mallows, C. (1998), "The Zeroth Problem," *The American Statistician*, 52, 1–9.

Paninski, L. (2003), "Convergence Properties of Three Spike-Triggered Analysis Techniques," *Network: Computation in Neural Systems*, 14, 437–464.

Rubin, D. R. (1993), "The Future of Statistics," *Statistics and Computing*, 3, 204.

Snee, R. D. (1990), "Statistical Thinking and Its Contribution to Total Quality," *The American Statistician*, 44, 116–121.

Stigler, S. M. (1999), *Statistics on The Table: The History of Statistical Concepts and Methods*, Cambridge, MA: Harvard University Press, Chap. 10.

Transfer of Technology From Statistical Journals to the Biomedical Literature

Past Trends and Future Predictions

Douglas G. Altman, Steven N. Goodman, MD, PhD

Objective.—To investigate the speed of the transfer of new statistical methods into the medical literature and, on the basis of current data, to predict what methods medical journal editors should expect to see in the next decade.

Design.—Influential statistical articles were identified and the time pattern of citations in the medical literature was ascertained. In addition, longitudinal studies of the statistical content of articles in medical journals were reviewed.

Main Outcome Measures.—Cumulative number of citations in medical journals of each article in the years after publication.

Results.—Annual citations show some evidence of decreasing lag times between the introduction of new statistical methods and their appearance in medical journals. Newer technical innovations still typically take 4 to 6 years before they achieve 25 citations in the medical literature. Few methodological advances of the 1980s seem yet to have been widely cited in medical journals. Longitudinal studies indicate a large increase in the use of more complex statistical methods.

Conclusions.—Time trends suggest that technology diffusion has speeded up during the last 30 years, although there is still a lag of several years before medical citations begin to accrue. Journals should expect to see more articles using increasingly sophisticated methods. Medical journals may need to modify reviewing procedures to deal with articles using these complex new methods.

(*JAMA*. 1994;272:129-132)

THE INFLUX of statistical methods into the medical literature has increased over more than 60 years. Over the same period, statistics itself has undergone major changes, so that not only is the use of statistics in medical research much more common, but the methods used have become progressively more complex. Although some of the methods being introduced in medical research were developed in other contexts, many statistical advances have arisen as solutions to problems arising in medical research. Changes in the type of statistical methods being used in medical articles have implications for editors, referees, and readers.

We report herein a study of citations to investigate the transfer of new statistical methods into the medical literature. We predict some new methods that medical journal editors should expect to see in the next decade.

From the Medical Statistics Laboratory, Imperial Cancer Research Fund, London, England (Mr Altman), and Oncology Center, Division of Biostatistics, The Johns Hopkins University, Baltimore, Md (Dr Goodman).

Presented in part at the Second International Congress on Peer Review in Biomedical Publication, Chicago, Ill, September 10, 1993.

Reprint requests to Medical Statistics Laboratory, Imperial Cancer Research Fund, PO Box 123, Lincoln's Inn Fields, London, England WC2A 3PX (Mr Altman).

METHODS

Influential statistical articles published after 1950 were identified from two books that reprinted important statistical articles,[1,2] from a list of the most cited articles in medical journals, and from personal knowledge (Table 1). Several articles relate to survival analysis[6,9,11,13,14] or meta-analysis,[5,7] two of the strongest growth areas (in both medicine and medical statistics) in recent years. Unfortunately, in some important areas of statistical methods there was no key article that could be widely cited by a large proportion of users, such as logistic regression and sample size calculations for clinical trials. We have included some articles that were published in medical journals (notably, cancer journals) when these seemed to be the primary source of the new method, and also one book.

For each article, the time pattern of citations in the medical literature was ascertained. Citations prior to 1971 were obtained by hand searching of printed volumes of the *Science Citation Index*,[23] as were citations for a few of the later articles with relatively few citations. Citations from 1971 to 1992 were obtained using computer searches of the SciSearch database (Institute of Scientific Information, Philadelphia, Pa). These searches were carried out in July and August 1993, by which time citations for 1992 should have been virtually complete. We did not search for articles that had incorrect citations of the articles of interest. It is our impression that the rate of incorrect citations of these articles was about 10% (excluding errors in titles). Some minor inconsistency between the two methods of searching may have arisen through problems in identifying what constitutes a medical journal. For comparison, similar citation analyses were performed for two heavily cited expository statistical articles published in medical journals.[21,22]

We also sought evidence from longitudinal studies of the statistical content of articles in medical journals to examine changes in the methods used over time.

RESULTS

Figure 1 shows cumulative numbers of citations for the articles listed in Table 1 divided into four decades—the 1950s, 1960s, 1970s, and 1980s. The article by Cox[14] was excluded because it has been cited much more often than the other articles. It is shown in Fig 2, together with the article by Kaplan and Meier.[6] These two articles are frequently cited together in articles reporting the results of survival analyses. They were published 14 years apart, and Fig 2 shows that the citations for the earlier article have risen in parallel with those for the Cox article, but about 14 years later in relation to the year of publication. These are now two of the most heavily cited articles in medical journals. The rise in citations for the article by Kaplan and Meier[6] is especially marked given that it received only six citations in medical journals in the first 10 years after publication.

Annual citations for the articles published in the four decades do show some evidence of decreasing lag times between the introduction and widespread use of new statistical methods. Newer technical innovations still typically take 4 to 6

Table 1.—Statistical Articles Included in This Study

Source, y	Topic
Methodological articles	
Cornfield,[3] 1951	Odds ratio
Cochran,[4] 1954	χ^2 Trend test
Woolf,[5] 1955	Combining 2×2 tables
Kaplan and Meier,[6] 1958	Survival curve
Mantel and Haenszel,[7] 1958	Stratified 2×2 table
Cohen,[8] 1960	κ Statistic
Mantel,[9] 1963	Survival analysis
Box and Cox,[10] 1964	Transformations
Mantel,[11] 1966	Survival analysis
Elston and Stewart,[12] 1971	Heredity
Peto and Peto,[13] 1972	Log rank test
Cox,[14] 1972	Proportional hazards regression
Dempster et al,[15] 1977	EM algorithm
Efron,[16] 1979	Bootstrap
Hanley and McNeil,[17] 1982	Receiver operating characteristic curve
Geman and Geman,[18] 1984	Gibbs sampling
Breiman et al,[19] 1984	Classification and regression trees
Zeger and Liang,[20] 1986	Longitudinal data
Expository articles	
Peto et al,[21] 1977	Log rank test
Bland and Altman,[22] 1986	Method comparison

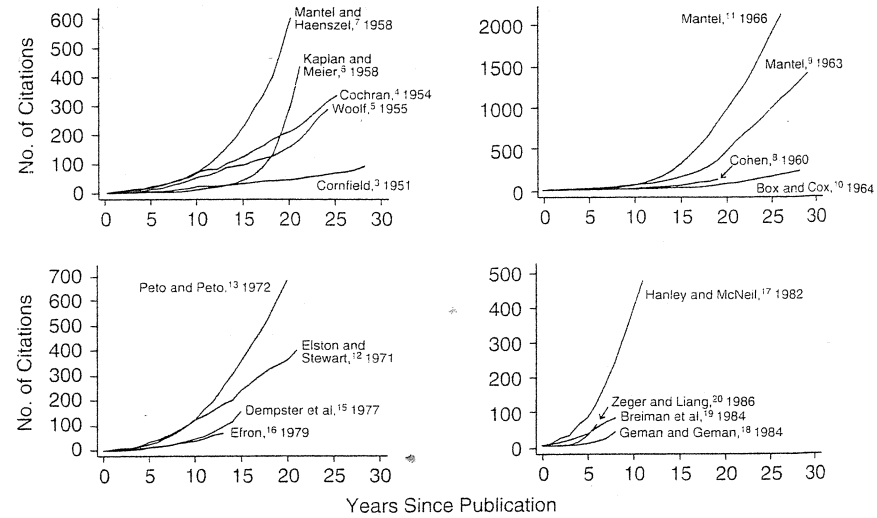


Fig 1.—Cumulative citations in medical journals for selected articles published in 1950 through 1959 (top left), 1960 through 1969 (top right), 1970 through 1979 (bottom left), and 1980 through 1989 (bottom right).

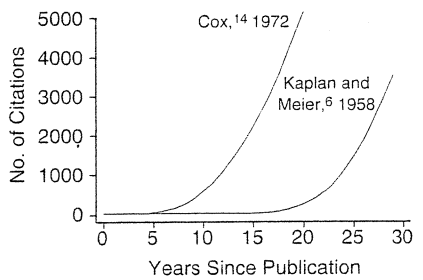


Fig 2.—Cumulative citations in medical journals for two heavily cited articles on survival analysis methods.

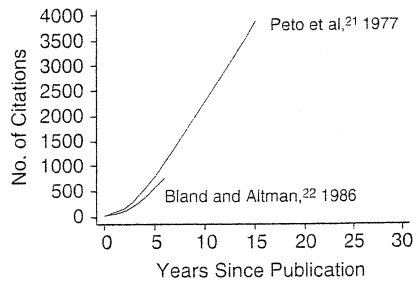


Fig 3.—Cumulative citations in medical journals for two expository articles.

years before they achieve 25 citations in the medical literature. Few methodological advances of the 1980s seem yet to have been widely cited in medical journals. By contrast, expository articles in medical journals can reach 500 citations within 4 to 5 years (Fig 3). Citations for one of the two expository articles[21] have leveled out, with a roughly constant number of citations each year. Most of the methodological articles (notably, the heavily cited articles) have increasing numbers of citations each year.

Few authors have studied changes over time in the use of statistical methods in one journal. Hayden[24] gave a brief summary of the rise in the use of simple statistical methods in *Pediatrics* from 1952 to 1982, while Felson et al[25] described similar changes in *Arthritis and Rheumatism* from 1967 to 1968 vs 1982. The most detailed information we are aware of relates to the *New England Journal of Medicine*. Articles published in 1978 and 1979,[26] 1989,[27] and 1990[28] have been reviewed using the same set of categories.[26] A large increase was noted during this period in the use of most statistical methods, especially the more complex methods (Table 2). It is notable that survival analysis and logistic regression were found in almost a third of original articles published in 1989 and 1990.

COMMENT

Citation studies are rightly criticized as a means of grading researchers,[29] but we think they provide a valuable measure of the impact of a new methodological development on medical research. Figure 1 suggests that technology diffusion may have speeded up during the last 30 to 40 years, although there is still usually a lag of several years before medical citations begin to accrue.

We used cumulative citations rather than annual citations, as we feel the total impact is more relevant in this context and that fluctuations in the annual counts obscure the trends. For the purposes of documenting technology transfer, it is not the actual number of citations but the shape of the citation curve that is most informative. This shape seems not to have changed greatly during four decades. Almost all of the curves for these classic articles have a dormant early phase followed by a somewhat dramatic takeoff. The general shape does not seem to vary in relation to how heavily cited an article is. There are, however, a few exceptions to this pattern, notably the article by Hanley and McNeil[17] (Fig 1). Developments that have probably contributed to the more rapid diffusion of statistical methods into the medical literature are the increasing number of statisticians working in medicine, the accessibility of powerful desktop computers to medical researchers, and the more rapid development and dissemination of software to implement new statistical methods.

Our analyses took no account of the large increase in the number of articles being published each year in medical journals (1730 journals published in 1950, increasing in 10-year intervals to 2800, 4420, 6780, and 9480) (Ulrich's International Serials database, Bowker Electronic Publishing). However, this increase has been almost linear since 1970, so adjustment for the increasing size of the literature would not greatly alter the shapes of the curves. Furthermore, such adjustment is not appropriate if, as seems likely, researchers today need to access many more articles in a greater number of journals than their predecessors. Huth[30] found a large increase between 1950 and 1985 in the number of

Table 2.—Statistical Methods Used in Original Articles in the *New England Journal of Medicine* in 1978 and 1979[26] and 1989 and 1990[27,28]

Topic	1978-1979, % (n=332)	1989-1990, % (n=215)
Any statistical analysis	73	88
t Tests	44	39
Contingency tables	27	33
Pearson correlation	12	18
Survival methods/ logistic regression	11	31
Nonparametric tests	11	23
Epidemiologic statistics	9	18
Analysis of variance	8	17
Simple linear regression	8	13
Transformation	7	7
Multiple regression	5	10
Multiway tables	4	8
Nonparametric correlation	4	5
Multiple comparisons	3	7
"Other methods" (not on original list)	3	14

Table 3.—Newer Statistical Methods That May Be Seen More Often in the Coming Years

Method	Description	Purpose
Bootstrap (also called resampling; related to the jackknife)[17]	Multiple new data sets are generated by random sampling "with replacement" from the original data	To calculate SEs or assess the stability of a statistical model, often when standard assumptions are unreliable or the sampling distribution is unknown
Gibbs sampling[18,34]	Random sampling from conditional distributions within a complex structure	Bayesian estimation of complex models
Generalized additive models[35]	Nonparametric smoothing of explanatory variables in regression	To replace regression when assumptions are not tenable
Classification and regression trees[19,36] (also known as recursive partitioning)	Division of a set of subjects by combinations of characteristics, to minimize the differences within groups and to maximize the differences between groups	To find combinations of variables of predictive importance
Models for longitudinal data ("general estimating equations")[20]	Modeling repeated measurements of an outcome variable while allowing for covariates	Regression for multiple assessments of outcome
Models for hierarchical data (also called multilevel models)[37]	Fitting mixed linear models to hierarchical data using iterative generalized least squares	Modeling data with more than one level of variation (eg, within and between patients)
Neural networks[38]	Nonparametric modeling of complex data	To provide nonlinear approximations to multivariable functions or for classification

different journals being cited in articles published in the *New England Journal of Medicine*.

Independent evidence for genuine changes in the use of statistics comes from studies that have looked at the same journals across time. The few such studies that we are aware of have shown large increases in the use of statistical methods and a tendency to use more complicated methods.[24-28] Thus, there is clearly a strong component of increased use and complexity of statistics independent of the total journal expansion. It is relevant that the number of original articles published per year by the *New England Journal of Medicine* decreased during the period of the studies summarized in Table 2.

Cumulative citations for the methodological articles considered generally curve upward, indicating that the annual number of citations keeps increasing. By contrast, the two expository articles considered show a much more rapid accrual of citations (starting in the year of publication) but near-linear cumulative citation curves, indicating a fairly steady annual citation rate. Expository statistical articles in medical journals can reach 500 citations within 4 to 5 years (Fig 3). Both articles we considered[21,22] described methods previously published in statistical journals[13,31] without achieving many citations in medical journals. These citation figures suggest that expository articles are valuable, especially for topics that are not usually included in medical statistics textbooks. Indeed, the International Committee of Medical Journal Editors guidelines state, "References for study design and statistical methods should be to standard works (with pages stated) when possible rather than to papers in which the designs or methods were originally reported."[32] Expository articles cowritten by a statistician and medical researcher may be especially helpful—a recent example considers receiver operating characteristic curves.[33] Unfortunately, such crossover articles require a considerable amount of work, and such activity (being a form of teaching) may not be helpful to the statistician's or researcher's career in comparison with either more methodological or medical articles.

Several complex statistical methods introduced in the 1980s are beginning to be seen more frequently. Although it is not possible to identify recent articles that will turn out to be major breakthroughs, most of the newer methods are sophisticated. Journals should expect to see growing numbers of articles using them. Methods likely to be seen more often are described briefly in Table 3. Software is available for all of these techniques, and some are beginning to be included in well-known statistics packages. It is worth noting that by the time a topic reaches medical journals there may be a large methodological literature. Ripley[38] notes that there are already more than a dozen journals and at least 15 texts devoted to neural networks.

The evidence of time trends within one major journal (Table 2) supports the idea that there is an ever-increasing variety of statistical methods appearing in medical articles. The speed with which new methods are introduced may pose problems for statistical referees, for the physicians who read the published work, and for the journals themselves. Referees may not be able to judge new methods that they have not yet learned. Physicians may feel that they have no chance of understanding the new methods (even if they are comfortable with more traditional methods) and will have to take the results of such studies on faith. The journals, in whom that faith is being entrusted, may bear an increasing burden to ensure that the methods are indeed valid, since most of their audience will be unable to assess that for themselves.

We think that the following developments are possible and may be desirable in the future:

- Authors using complex methods will be asked to supply additional supporting material for referees but not for publication. This might take the form of a formal appendix in the submitted manuscript, which is peer reviewed (and possibly modified) but not published. It should be supplied by authors to readers on request.
- Because statistical refereeing will be a more difficult process (because of both the novelty and the complexity of methods), medical journals may need to recruit panels of methodological reviewers who specialize in specific methods.
- Editors of medical journals should encourage or actively solicit more crossover (expository) articles on new methods, perhaps with both medical and statistical authors.
- More postgraduate training for medical researchers should be developed, with formal accreditation, both in basic statistical methods and also to help those who wish to keep abreast of newer methods.

It is likely that the statistical education of physicians, already poor,[39,40] will in the future lag even further behind the

methods that are used in medical journals. Already the standard methods taught in an introductory course would leave a reader unable to judge a high percentage of articles published in the *New England Journal of Medicine*, and that proportion is likely to increase with time.

We thank Scott Zeger, PhD, for suggesting this topic of investigation. We are grateful to Will Russell-Edu for carrying out the computer citation searches.

References

1. Greenland S, ed. *Evolution of Epidemiologic Ideas: Annotated Readings on Concepts and Methods*. Chestnut Hill, Mass: Epidemiology Resources Inc; 1987.
2. Kotz S, Johnson N, eds. *Breakthroughs in Statistics: Volume II: Methodology and Distribution*. New York, NY: Springer Publishing Co; 1992.
3. Cornfield J. A method of estimating comparative rates from clinical data: applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst*. 1951;11:1269-1275.
4. Cochran WG. Some methods for strengthening the common χ^2 tests. *Biometrics*. 1954;10:417-451.
5. Woolf B. On estimating the relation between blood group and disease. *Ann Hum Genet*. 1955; 11:251-253.
6. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53:457-481.
7. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 1958;22:719-748.
8. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37-46.
9. Mantel N. Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *J Am Stat Assoc*. 1963;58:690-700.
10. Box GEP, Cox DR. An analysis of transformations (with discussion). *J R Stat Soc B*. 1964;26: 211-252.
11. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*. 1966;50:163-170.
12. Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. *Hum Heredity*. 1971;21:523-542.
13. Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). *J R Stat Soc A*. 1972;135:185-207.
14. Cox DR. Regression models and life tables (with discussion). *J R Stat Soc B*. 1972;34:187-220.
15. Dempster AP, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B*. 1977;39:1-38.
16. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat*. 1979;7:1-26.
17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29-36.
18. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images. *IEEE Trans Pattern Anal Machine Intell*. 1984; 6:721-741.
19. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Belmont, Calif: Wadsworth; 1984.
20. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*. 1986;42:121-130.
21. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient, II: analysis and examples. *Br J Cancer*. 1977;35:1-39.
22. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307-310.
23. Institute of Scientific Information. *Science Citation Index*. Philadelphia, Pa: Institute of Scientific Information; 1951-1971.
24. Hayden GF. Biostatistical trends in *Pediatrics*: implications for the future. *Pediatrics*. 1983;72:84-87.
25. Felson DT, Cupples LA, Meenan RF. Misuse of statistical methods in *Arthritis and Rheumatism*: 1982 versus 1967-68. *Arthritis Rheum*. 1984;27:1018-1022.
26. Emerson JD, Colditz G. Use of statistical analysis in the *New England Journal of Medicine*. *N Engl J Med*. 1983;309:707-713.
27. Emerson JD, Colditz G. Use of statistical analysis in the *New England Journal of Medicine*. In: Bailar JC III, Mosteller F, eds. *Medical Uses of Statistics*. 2nd ed. Boston, Mass: NEJM Books; 1992: 45-57.
28. Altman DG. Statistics in medical journals: developments in the 1980s. *Stat Med*. 1991;10:1897-1913.
29. Seglen PO. Citation frequency and journal impact: valid indicators of scientific quality? *J Intern Med*. 1991;229:109-111.
30. Huth E. The information explosion. *Bull N Y Acad Med*. 1989;65:647-661.
31. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician*. 1983;32:307-317.
32. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *JAMA*. 1993;269:2282-2286.
33. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental tool in clinical medicine. *Clin Chem*. 1993;39:561-577.
34. Gilks WR, Clayton DG, Spiegelhalter DJ, et al. Modelling complexity: applications of Gibbs sampling in medicine. *J R Stat Soc B*. 1993;55:39-52.
35. Hastie T, Tibshirani R. Generalized additive models. *Stat Sci*. 1986;1:297-318.
36. Ciampi A, Lawless JF, McKinney SM, Singhal K. Regression and recursive partition strategies in the analysis of medical and survival data. *J Clin Epidemiol*. 1988;41:737-748.
37. Goldstein H. *Multilevel Models in Educational and Social Research*. London, England: Griffin; 1986.
38. Ripley BD. Statistical aspects of neural networks. In: Barndorff-Nielsen OE, Jensen JL, Kendall WS, eds. *Networks and Chaos: Statistical and Probabilistic Aspects*. London, England: Chapman and Hall; 1993:40-123.
39. Wulff HR, Andersen B, Brandenhoff P, Guttler F. What do doctors know about statistics? *Stat Med*. 1987;6:3-10.
40. Altman DG, Bland JM. Improving doctors' understanding of statistics (with discussion). *J R Stat Soc A*. 1991;154:223-267.

Journal of Applied Statistics
Vol. 32, No. 5, 461–474, July 2005

Routledge
Taylor & Francis Group

The Most-Cited Statistical Papers

THOMAS P. RYAN* & WILLIAM H. WOODALL**

*National Institute of Standards and Technology, Gaithersburg, Maryland, USA, **Department of Statistics, Virginia Tech, Blacksburg, Virginia, USA

ABSTRACT *We attempt to identify the 25 most-cited statistical papers, providing some brief commentary on each paper on our list. This list consists, to a great extent, of papers that are on non-parametric methods, have applications in the life sciences, or deal with the multiple comparisons problem. We also list the most-cited papers published in 1993 or later. In contrast to the overall most-cited papers, these are predominately papers on Bayesian methods and wavelets. We briefly discuss some of the issues involved in the use of citation counts.*

KEY WORDS: Citations, history of statistics

Citations in General

There has been much discussion of the uses of citation counts in the literature, although not with respect to the statistics literature with the exceptions of Stigler (1994), Altman & Goodman (1994), and Theoharakis & Skordia (2003). Austin (1993) assessed the reliability of citation counts in making tenure and promotion decisions in academia, while Gilbert (1977) and Edge (1979) have considered citation counts as measures of the influence of research. See also Cronin (1984).

Edge (1979) criticized citation counts as being overused to measure intellectual linkages. Others have made similar criticisms. Despite such criticisms, however, the use of citation counts seems to be increasing. The National Research Council, for example, uses citation rates as one measure to rank PhD programmes in statistics and other fields. In addition, citation counts appear to be increasingly used in promotion decisions in academia, in addition to ranking scientific journals. Using ISI *Journal Citation Reports*, for example, one can determine that among the 71 journals in the Statistics and Probability category, *Statistical Science* ranked fifth in citation impact factor and 16th in the total number of citations received in 2002 with 1,051.

In attempting to determine the causal factors for highly cited papers, Donoho (2002) gave a list of suggestions for writing papers that would receive a large number of citations. At the top of his list was 'Develop a method which can be applied on statistical data of a kind whose prevalence is growing rapidly'. For example, if someone could develop 'the' approach to data mining, the paper would undoubtedly garner a huge number of citations.

Correspondence Address: W. H. Woodall, Department of Statistics, Virginia Tech, Blacksburg, VA 24061-0439, USA. Email: bwoodall@vt.edu

0266-4763 Print/1360-0532 Online/05/050461–14
DOI: 10.1080/02664760500079373

There is generally a time lag of several years before new methodology is implemented in software, so it is not surprising that number 2 on Donoho's list was 'Implement the method in software, place examples of the software's use in the paper, make the software of broad functionality, and give the software away for free.'

Garfield (1998) reported that for the period 1945–1988 the majority of cited papers in science were cited only once. In a controversial citation analysis, it was shown that 55% of papers published during 1981–1985 received no citations within five years of their publication (Hamilton, 1990). From the same data, Hamilton (1991) broke down the 55% of uncited papers and indicated that there was a huge variation across various disciplines, ranging from 9.2% of papers uncited in atomic, molecular, and chemical physics, to 86.9% in engineering. Pendlebury (1991), however, disagreed with Hamilton's analysis and reported that only 22.4% of science articles published in 1984 remained uncited by the end of 1988.

Papers are cited at different rates in different fields. ScienceWatch (1999) reported that for the years 1981–1997 a paper in mathematics needed at least 291 citations to rank in the top 0.01%, while it took 1,823 citations for the corresponding ranking in molecular biology and genetics.

What should we make of these numbers? A sceptic might contend that these studies show that much research has little or no value. Indeed, it seems apparent that most published papers do not influence the work of other researchers, although there have of course been innumerable instances in which researchers have failed to acknowledge related work. Overall, however, it seems clear that the distribution of papers in regard to their impact has a huge amount of right skewness.

The 25 Most-Cited Papers

In this section we provide our list of the 25 most-cited statistical papers. We did not limit ourselves to the primary statistical journals. We considered for inclusion only papers in which the author(s) proposed a new statistical method, modified an existing statistical method, or used an existing statistical method in a novel way to address an important scientific problem. The application of this set of criteria is necessarily subjective to a large extent. As discussed by Straf (2003), there is no generally accepted definition of 'statistics'.

The citation counts are those given on the Institute for Scientific Information (ISI) Web of Science (as of 1 December 2003). Since the Web of Science does not include all scientific journals, the counts are all undercounts. In addition, we did not attempt to make adjustments for incorrect citation information, e.g., citations that had incorrect page or volume numbers. Taking into account these factors could lead to some reordering of the top 25 or even to some papers dropping off the list. A more significant issue, however, is the fact that the ISI Web of Science citation counts does not include citations before 1945. This is thus a problem for papers published well before then, and the problem is compounded by the fact that it was more difficult for the early papers to accumulate citations since the number of scientific journals was much smaller at the time they were published than is now the case. In addition, as a method becomes a generally accepted part of statistics, e.g., the one-sample *t*-test, the citation rate of the paper in which the method was initially proposed decreases. We also calculated a *current* annual citation rate for some papers (reported in parentheses when available). This is a conservative value since it was obtained by doubling the number of citations received during a period of less than six months in the last part of 2003.

Some very highly cited papers on fuzzy logic, such as the one by Zadeh (1965) with 5,022 citations (338 per year), were not considered to be statistical papers even though there is a connection between fuzzy logic and statistics, as discussed by Laviolette *et al.* (1995). Similarly, Hopfield (1982), on the topic of neural networks with 3,574 citations (156 per year), was not included. Reed & Muench (1938), with 10,974 citations (242 per year), was not included due to the simplicity of the proposed method. Wright (1931) with 2,218 citations (144 per year) was not included since the method and results were judged to be primarily probabilistic, not statistical. In addition, there are some highly cited papers by well-known statisticians that we did not consider to be statistical enough to warrant inclusion in our list, e.g., Cooley & Tukey (1965) with 2,872 citations (78 per year) and Nelder & Mead (1965) with 5,635 citations (426 per year). Some of these decisions are debatable since Cooley & Tukey (1965) was included by Kotz & Johnson (1997) with an introduction written by I. J. Good.

It would not be surprising if some papers with significant statistical content have been overlooked in our study. In addition, it might be argued that some of the papers on our list should have been excluded for one reason or another. We welcome input on these issues from the readers of our paper.

The following is our list with some brief commentary.

(1) With 25,869 citations (currently cited 1,984 times per year),

> Kaplan, E. L. & Meier, P. (1958) Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, 53, pp. 457–481.

Kaplan & Meier (1958) proposed a non-parametric method for estimating the proportion of items in a population whose lifetime exceeded some specified time t from censored survival data. This type of data is very common in medical studies. This paper not only has by far the highest number of citations of all statistics papers, but it has also been ranked among the top five most cited papers for the entire field of science. Based on data from *Journal Citation Reports*, the total number of citations received by this paper exceeds twice the number of citations received by *all Journal of the American Statistical Association* papers in 2002. This paper appeared in Kotz & Johnson (1992b, pp. 311–338) as a breakthrough paper in statistics with an introduction written by N. E. Breslow.

Kaplan (1983) reported that he and Meier had, in fact, each submitted separate manuscripts to the *Journal of the American Statistical Association.* Due to their similarity, the editor recommended that their papers be combined into one manuscript. It took them four years to resolve the differences between their approaches, during which time they were concerned that someone else might publish the idea.

Interestingly, Garfield (1989) gave this paper as an example of one that was slow to receive recognition. Indeed, Figure 3 in Garfield (1989) shows that the paper received very few citations per year through the early 1970s (i.e., for the first 15 years after it was published). It was cited only 25 times from 1958–1968. But, starting in 1975, the number of citations per year began to increase sharply and continued to increase monotonically through 1989, the last year covered by the graph. Meier is quoted in personal communication that year as stating that the needs of applied researchers were 'quite well met' by the existing methodology, and it was not until the advent of computers and the increasing mathematical sophistication of clinical researchers that the Kaplan–Meier method grew in importance and eventually was recognized as the standard.

Despite its popularity, the Kaplan–Meier method has not been without controversy. Miller (1983) wrote a paper entitled 'What price Kaplan–Meier?' in which he claimed

that the Kaplan–Meier estimator was inefficient and suggested that analysts should use some parametric assumptions whenever possible. Eighteen years later, Meier (2001) responded to the paper with a talk entitled 'The price of Kaplan–Meier.' Meier believed Miller's (1983) conclusions were incorrect and initially believed that references to it would taper off for that reason. His presentation was motivated in part by the number of citations of Miller's paper. Also, see Meier *et al.* (2004).

(2) With 18,193 citations (1,342 per year),

Cox, D. R. (1972) Regression models and life tables, *Journal of the Royal Statistical Society, Series B*, 34, pp. 187–220.

The topic of this paper is the regression analysis of censored failure time data, which has far-reaching applications in the biomedical sciences. Cox (1972) used a semiparametric model for the hazard function, which has significant advantages over using parametric models for the failure time.

This paper appeared in Kotz & Johnson (1992b, pp. 519–542) as a breakthrough paper in statistics with an introduction written by R. L. Prentice. See Reid (1994) for some interesting background on this paper from D. R. Cox. Interestingly, it is reported that a key insight into the statistical analysis method first came to Professor Cox when he was quite ill with the flu and was recalled later only with some difficulty. Cox (1986) also provided some background on the paper.

(3) With 13,108 citations (256 per year),

Duncan, D. B. (1955) Multiple range and multiple *F*-tests, *Biometrics*, 11, pp. 1–42.

David Duncan presented his now-famous multiple range test for comparing the means of several populations at the Joint Meetings of the Institute of Mathematical Statistics and the Eastern North American Region of the Biometric Society in March of 1954. Although Duncan also proposed multiple *F*-tests, and in fact this was his original emphasis, these tests have not enjoyed the popularity of his multiple range test because they were more cumbersome to use.

Duncan (1977) gave some historical background on this paper. He also recommended that the methods in Duncan (1975) be used in place of his multiple range test.

(4) With 9,504 citations (488 per year),

Marquardt, D. W. (1963) An algorithm for least squares estimation of non-linear parameters, *Journal of the Society for Industrial and Applied Mathematics*, 2, pp. 431–441.

The Marquardt algorithm proposed in this paper is used to estimate the parameters in a nonlinear model. See Hahn (1995) and Marquardt (1979) for some interesting background information on this paper.

(5) With 8,720 citations (114 per year),

Litchfield, J. T. & Wilcoxon, F. A. (1949) A simplified method of evaluating dose-effect experiments, *Journal of Pharmacological and Experimental Therapeutics*, 96, pp. 99–113.

The authors proposed a rapid graphical method for approximating the median effective dose and the slope of dose-percent effect curves. Litchfield (1977) credited Wilcoxon's

intense interest in collaboration for the development of the proposed method. When Litchfield joined the laboratories where Wilcoxon was working, the two were discussing the method at Wilcoxon's request even before Litchfield had seen his employer or checked in with the personnel department.

(6) With 8,151 citations (1,590 per year),

> Bland, J. M. & Altman, D. G. (1986) Statistical methods for assessing agreement between two clinical measurements, *Lancet*, 1 (8476), pp. 307–310.

The authors described simple statistical methods and graphs originally proposed by Altman & Bland (1983) for using paired data to assess the differences between measurements obtained by two different measurement systems. (The paper is available online at http://www.users.york.ac.uk/∼mb55/meas/ba.htm) See Bland & Altman (1992) and Bland & Altman (1995) for descriptions of the genesis and impact of this paper.

(7) With 6,788 citations (914 per year),

> Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap, *Evolution*, 39, pp. 783–791.

The context of evolutionary biology is phylogeny, the connections between all groups of organisms as understood by ancestor/descendant relationships. According to I. Hoeschele (personal communication), the human genome project and sequencing projects for other organisms provide an unprecedented amount of data to which the methods in this paper and those in Nei (1972), our Number 13 paper, can be applied. The resulting information is immensely valuable in understanding questions in evolution and in inferring the functions of genes. Phylogenetics is a very active area of research, in particular in the context of comparative genome analysis and genome-scale adaptation of methods. For more information on this topic, the reader is referred to Holmes (2003).

Felsenstein (1985) considered an application of the bootstrap method, whereas more fundamental statistical issues were addressed in the bootstrap paper in Efron (1979), which narrowly missed being in our list with 1,889 citations (156 per year). Efron (1979) appeared in Kotz & Johnson (1992b, pp. 519–542) as a breakthrough paper in statistics, with an introduction written by R. J. Beran.

(8) With 6,579 citations (126 per year),

> Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. & Smith, K. G. (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Part II. Analysis and examples, *British Journal of Cancer*, 35, pp. 1–39.

Sir Richard Peto, the first author, and Sir David Cox and Nathan Mantel, who appear in other places on this list, are among the distinguished group of co-authors of this paper. The paper is the second of a two-part report to the UK Medical Research Council's Leukemia Steering Committee. This report was focused on efficient methods of analysis of data from randomized clinical trials for which the duration of survival among different groups of patients is to be compared.

(9) With 6,006 citations (422 per year),

> Mantel, N. & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute*, 22, pp. 719–748.

These authors proposed a chi-square test with one degree of freedom for testing the association of disease incidence using 2×2 contingency tables.

(10) With 5,260 citations (300 per year),

> Mantel, N. (1966) Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports*, 50, pp. 163–170.

Mantel (1966) was also cited by Garfield (1989) as a paper that was slow to receive recognition. Mantel was apparently philosophical about this, stating in personal communication to Garfield in 1989, 'Actually, slow initial rise characterizes nearly everything', and also reasoned that his method was slow to gain recognition by statisticians and epidemiologists because it was published in a cancer journal.

(11) With 4,306 citations (492 per year),

> Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (C/R: pp. 22–37), *Journal of the Royal Statistical Society, Series B*, 39, pp. 1–22.

The Expectation Maximization (EM) algorithm is used for maximum likelihood estimation with data for which some variables are unobserved. Much has been written about the algorithm, which coupled with its various applications, including those involving censored data and truncated data, helps to explain the large number of citations. A well-regarded book by McLachlan & Krishnan (1997) has been written about the algorithm. The name 'EM' was coined by Dempster, Laird & Rubin in this paper, but the method was apparently used in some form much earlier by a few researchers, including McKendrick (1926) and Hartley (1958), who introduced the procedure for calculating maximum likelihood estimates for the general case of count data.

(12) With 3,819 citations (32 per year),

> Wilkinson, G. N. (1961) Statistical estimations in enzyme kinetics, *Biochemical Journal*, 80, pp. 324–336.

The author gave an account of the weighted linear and nonlinear regression methods applicable to general problems in enzyme kinetics. The Michaelis–Menten model, which is used frequently in enzyme kinetics, was used to illustrate aspects of nonlinear regression.

(13) With 3,672 citations (142 per year),

> Nei, M. (1972) Genetic distance between populations, *The American Naturalist*, 106, pp. 283–292.

Nei (1972) proposed a measure of genetic distance based on the identity of genes between populations. The measure can be applied to any pair of organisms.

(14) With 3,511 citations (118 per year),

> Dunnett, C. W. (1955) A multiple comparison procedure for comparing several treatments with a control, *Journal of the American Statistical Association*, 50, pp. 1096–1121.

A very interesting article about Professor Dunnett and how his work on multiple comparisons against a control evolved can be found at www.ssc.ca/main/about/history/dunnett_e.html. Additionally, Professor Dunnett has been kind enough to provide us with information relating to this paper and subsequent developments. His work with Bob Bechhofer and Milton Sobel on ranking and selection led to development of the multivariate-*t* distribution (Dunnett & Sobel, 1954, 1955), which fortuitously turned out to be the appropriate distribution for making multiple comparisons involving a control. Dunnett (1955) formulated the problem in terms of simultaneous confidence intervals, which was the same approach that John Tukey and Henry Scheffé had taken in their work.

Because of the great extent to which multiple comparison procedures are used by researchers outside the field of statistics, it is relevant to question the extent to which more recent papers that have defined the current state-of-the-art may have been overlooked. Indeed, Dunnett & Tamhane (1991, 1992) presented step-up and step-down methods (somewhat analogous to forward selection and backward elimination in linear regression) with these methods being superior to one-stage procedures in terms of maximizing power. Despite the superiority of these procedures, Dunnett & Tamhane (1991), for example, has only 31 citations.

(15) With 3,444 citations (280 per year),

> Akaike, H. (1974) A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19, pp. 716–723.

This is a paper in which Akaike proposed a criterion for estimating the dimensionality of a model using the criterion now known as Akaike's Information Criterion (AIC). This paper has over three times as many citations as Akaike (1973), which was included in Kotz & Johnson (1992a, pp. 599–624) as a breakthrough paper in statistics, with a discussion written by J. de Leeuw.

(16) With 2,837 citations (376 per year),

> Liang, K.-Y. & Zeger, S. (1986) Longitudinal data analysis using generalized linear models, *Biometrika*, 73, pp. 13–22.

This paper was reprinted by Kotz & Johnson (1997, pp. 463–482) as a breakthrough paper in statistics with a discussion by P. J. Diggle. Liang & Zeger (1986) dealt with longitudinal studies in which the response measurement was a count. They derived a generalized estimating equations (GEE) methodology, which is now widely used.

(17) With 2,810 citations (22 per year),

> Cutler, S. J. & Ederer, F. (1958) Maximum utilization of the life table method in analyzing survival, *Journal of Chronic Diseases*, 8, pp. 699–712.

The authors presented the rationale and computational details of the actuarial or life-table method for analysing data on patient survival. The method makes use of all survival information accumulated up to the closing date of a study. Cutler (1979) reported that he and Ederer were sharing a hotel room at a scientific meeting when the question leading to the paper came to him at 5 a.m. He promptly woke Ederer to discuss his idea. Cutler (1979) also stated that the paper did not represent a methodological breakthrough. The authors demonstrated that the life-table method could be used to extract the maximum amount

of information from the data being collected in the newly organized cancer reporting system.

(18) With 2,764 citations (240 per year),

> Geman, S. & Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, pp. 721–741.

This paper was included by Kotz & Johnson (1997, pp. 123–126) as a breakthrough paper with a discussion by P. J. Huber. Geman & Geman (1984) modified Markov chain Monte Carlo methods and applied them to Bayesian models for the computation of posterior probabilities.

(19) With 2,529 citations (120 per year),

> Box, G. E. P. & Cox, D. R. (1964) An analysis of transformations, *Journal of the Royal Statistical Society, Series B*, 26, pp. 211–243 (discussion pp. 244–252).

DeGroot (1987) provided some interesting background on this paper from an interview with Professor Box. Box recounted, for example, that he and Cox were on a committee of the Royal Statistical Society and several people suggested that they collaborate. Their motivation and the idea of the paper sprung, to some extent, from the similarities of their family names.

Box & Cox (1964) presented a very useful family of power transformations that have typically been used to transform the dependent variable in a regression model so as to try to meet the assumptions of homoscedasticity and normality of the error terms. The right side of the model can then be transformed in the same manner so as to retrieve the quality of the fit before the dependent variable was transformed.

(20) With 2,512 citations (76 per year),

> Mantel, N. (1963) Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure, *Journal of the American Statistical Association*, 58, pp. 690–700.

The author extended the methods in Mantel & Haenszel (1959), Number 9 on our list, in two ways, as it was recognized that the methods are not limited to retrospective studies and the number of levels of the study factor of interest was allowed to be greater than two.

(21) With 2,456 citations (46 per year),

> Dunnett, C. W. (1964) New tables for multiple comparisons with a control, *Biometrics*, 20, pp. 482–491.

In this paper, exact critical values are given for the method of Dunnett (1955), Number 14 on our list, when two-sided comparisons are made with a control.

(22) With 2,302 citations (42 per year),

> Kramer, C. Y. (1956) Extension of multiple range tests to group means with unequal numbers of replications, *Biometrics*, 12, pp. 307–310.

Kramer (1956) proposed an approximate method for extending multiple range tests to cases for which the sample sizes are unequal. Kramer's work was strongly related to the

methodology proposed by John Tukey in 1953, whose work was not published. Nevertheless, because of the close connection, Kramer's method for the unbalanced case is known as the Tukey–Kramer procedure. (See Benjamini & Braun, 2002, for a discussion of this issue.)

(23) With 2,248 citations (72 per year),

Fisher, R. A. (1953) Dispersion on a sphere, *Proceedings of the Royal Society of London, Series A*, 217, pp. 295–305.

Fisher (1953) presented a theory of errors that is believed to be appropriate for measurements on a sphere and derived a test of significance that was stated as being 'the analogue of "Student's test" in the Gaussian theory of errors'. The paper can be viewed online at http://www.library.adelaide.edu.au/digitised/fisher/249.pdf). According to Garfield (1977), this paper had only 277 citations between 1961 and 1975, but was Fisher's most frequently cited paper during that time period.

(24) With 2,219 citations (240 per year),

Schwarz, G. (1978) Estimating the dimension of a model, *Annals of Statistics*, 6, pp. 461–464.

Schwartz's Bayesian Information Criterion (BIC), introduced in this paper, is a criterion for model selection that is often mentioned with Akaike's AIC criterion.

(25) With 2,014 citations (382 per year),

Weir, B. S. & Cockerham, C. C. (1984) Estimating *F*-statistics for the analysis of population structure, *Evolution*, 38(6), pp. 1358–1370.

As Professor Weir informed us, the number of citations of this paper has risen every year since its publication as different groups of researchers have become interested in genetic population structure. These groups include ecologists, conservationists and, interestingly enough, forensic scientists.

Comments on the Top 25 List

The most-cited statistical papers fare well when compared to the most-cited papers in science. Garfield (1990) ranked the 100 most-cited papers in the 1945–1988 *Science Citation Index*. Duncan (1972), Litchfield & Wilcoxon (1949), Kaplan & Meier (1958), Marquardt (1963), and Cox (1972) ranked Numbers 24, 29, 55, 92 and 94, respectively. Kaplan & Meier (1958) and Cox (1972) had 'only' 4,756 and 3,392 citations, respectively, in Garfield's study.

All papers on our list were published prior to 1987. A dotplot of the publication years of the 25 papers on our list is shown in Figure 1.

There is no question that the field of a paper is related to the number of citations. This is evident from the number of papers in biostatistics on our list. Similarly, of the 27 'highly cited authors in mathematics and statistics' listed by Kruse (2002) in *AmStat News*, the

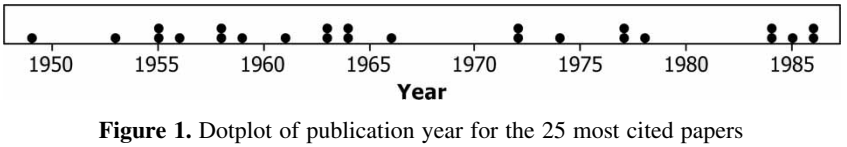


Figure 1. Dotplot of publication year for the 25 most cited papers

four who would rank the highest in terms of the number of citations per paper (for papers published during 1991–2001 and also cited during that period) are all biostatisticians.

If we traced the development of statistical methodology and theory (as Efron, 2001, did, concentrating on 1950–1980), we would certainly expect that there would be a strong correlation between the influence of a paper and its number of citations. The most influential papers in statistics have large citation counts, but only a few have enough citations to make our list. Efron (2001) listed non-parametric and robust methods first in impact, followed by the Kaplan–Meier method and Cox's method, with logistic regression and generalized linear models (GLM) mentioned third, while stating that logistic regression has had a huge effect on biostatistics. It has been known for some time that Kaplan & Meier (1958) and Cox (1972) were the two most-cited papers in statistics. See, for example, Stigler (1994).

Only a few of the most influential papers on the field of statistics are included on our list. Only five are included in Kotz & Johnson (1992a, 1992b, 1997) as representing 'breakthrough papers in statistics'. Four of our most cited papers, Duncan (1955), Kramer (1956), and Dunnett (1955, 1964) are on the topic of multiple comparisons. Multiple comparison methods are widely used in statistical practice, but without a major influence on the field of statistics itself. Tukey (1991), for example, downplayed the importance of the method of Duncan (1955) calling it a 'distraction'. To more effectively measure impact with respect to the field of statistics, it would be better to count only citations that appeared in statistical journals.

It is interesting to note that Nathan Mantel was author or co-author on four of our 25 papers. For more information on his contributions to statistics, the reader is referred to his obituary in *AmStat News* (July, 2002, pp. 35–36) or to http://members.aol.com/savilon/nmantel.html. He did much of his work at the National Cancer Institute, retiring from there in 1974. He was a very active consultant and undoubtedly many of his major research contributions had their origins in his consulting work. Sir D. R. Cox was author or co-author of three of the papers on our list. See Reid (1994) for information on his background.

Most-cited Papers Published Since 1993

For a perspective on the changing emphases of statistics over time, we also studied the most-cited statistical papers published in 1993 or later. Our list of the fifteen most-cited papers was obtained by first obtaining citation counts for all papers written during this time by the ten most-cited statisticians (for citations received for papers written and citations received between 1 January 1993 and 30 June 2003) listed in the November 2003 *AmStat News* (also see http://www.in-cites.com/top/2003/third03-math.html). These were the following: David L. Donoho (1,354 citations), Iain M. Johnstone (1,203 citations), Adrian E. Raftery (1,117 citations), Adrian F. M. Smith (866 citations), Peter Hall (827 citations), Donald B. Rubin (792 citations), Jianqing Fan (768 citations), Gareth O. Roberts (725 citations), Robert E. Kass (723 citations), and Siddhartha Chib (708 citations). We then checked the citation counts for all papers published in 1993 or later in the following statistical journals given by ISI *Journal Citation Reports* in the Statistics and Probability category as having the most citations in 2002 (number of citations is in parentheses): *Journal of the American Statistical Association* (11,318), *Econometrica* (9,458), *Biometrics* (7,469), *Biometrika* (6,742), *Annals of Statistics* (5,566), *Statistics in Medicine* (4,755), *Journal of the Royal Statistical Society, Series B* (4,755), and *Technometrics* (2,514). (Note that *Fuzzy Sets and Systems* with 3,626 citations was not included in our search.) It is possible that some papers were overlooked.

The following is our list:

1. Breslow, N. E. & Clayton, D. G. (1993) Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, 88, pp. 9–25. (558 citations)
2. Tierney, L. (1994) Markov-chains for exploring posterior distributions, *Annals of Statistics*, 22, pp. 1701–1728. (541 citations)
3. Kass, R. E. & Raftery, A. E. (1995) Bayes Factors, *Journal of the American Statistical Association*, 90, pp. 773–795. (533 citations)
4. Donoho, D. L. & Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, 81, pp. 425–455. (480 citations)
5. Smith, A. F. M. & Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov-chain Monte-Carlo methods, *Journal of the Royal Statistical Society, Series B*, 55, pp. 3–23. (444 citations)
6. Green, P. J. (1995) Reversible jump Markov-chain Monte Carlo computation and Bayesian model determination, *Biometrika*, 82, pp. 711–732. (479 citations)
7. Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B*, 57, pp. 289–300. (294 citations)
8. Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1995) Wavelet shrinkage – asymptopia, *Journal of the Royal Statistical Society, Series B*, 57, pp. 301–337. (293 citations)
9. Donoho, D. L. (1995) De-noising by soft thresholding, *IEEE Transactions on Information Theory*, 41, pp. 613–627. (292 citations)
10. Grambsch, P. M. & Therneau, T. M. (1994) Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika*, 81, pp. 515–526. (261 citations)
11. Donoho, D. L. & Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage, *Journal of the American Statistical Association*, 90, pp. 1200–1224. (257 citations)
12. Bound, J., Jaeger, D. A. & Baker, R. M. (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak, *Journal of the American Statistical Association*, 90, pp. 443–450. (252 citations)
13. Albert, J. H. & Chib, S. (1993) Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, 88, pp. 669–679. (246 citations)
14. Stock, J. H. & Watson, M. W. (1993) A simple estimator of cointegrating vectors in higher-order integrated systems, *Econometrica*, 61, pp. 783–820. (244 citations)
15. Chib, S. & Greenberg, E. (1995) Understanding the Metropolis–Hastings algorithm, *The American Statistician*, 49, pp. 327–335. (240 citations)

The most cited papers presented here tend to be on topics related to Bayesian methods and wavelets, although the topics of multiple testing and proportional hazards modelling are represented. It is interesting to note that it often takes quite a few years for the number of citations of a paper to reach its maximum rate. A number of the 25 overall most-cited papers are cited now at much higher rates than the most-cited papers of the last decade.

Conclusions

We find the study of citation counts to be very interesting. It is surprising that relatively little research has been done on citation counts, rates and patterns in the field of statistics.

Garfield (1979: 16) described early work in this area, including the Citation Index for Statistics and Probability, 'a cumulative one-time effort that covers the journal literature of the field from its inception, early in the twentieth century, through 1966'. This was compiled by John Tukey and published in 1973 as part of the 'Information Access Series' of R&D Press. It provided comprehensive coverage of 40 statistics journals and selective coverage of an additional 100 journals.

In our view it would be very interesting to examine a list of the most-cited papers in each of the top statistics journals (see Campbell & Julious, 1994) or in different application areas of statistics. Also, it would be useful to identify papers projected to enter the top 25 most-cited statistical papers and, more generally, 'hot papers' that have attained unusually high citation rates shortly after publication. For more on this latter topic, the reader is referred to Garfield (2000).

Acknowledgements

The authors gratefully acknowledge input from I. Hoeschele, C. W. Dunnett, J. Felsenstein, C. King, B. S. Weir, J. M. Bland, E. Garfield, I. J. Good and D. G. Altman.

References

Akaike, H. (1973) Information theory and the maximum likelihood principle, in: B. N. Petrov & F. Csàki (Eds) *Second International Symposium on Information Theory* (Budapest: Akademiai Kiàdo).

Akaike, H. (1974) A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19, pp. 716–723.

Altman, D. G. & Bland, J.M. (1983) Measurement in medicine: the analysis of method comparison studies, *The Statistician*, 32, pp. 307–317.

Altman, D. G. & Goodman, S. N. (1994) Transfer of technology from statistical journals to the biomedical literature: past trends and future predictions, *Journal of the American Medical Association*, 272, pp. 129–132. Available at http://www.ama-assn.org/public/peer/7_13_94/pv3108x.htm.

Austin, A. (1993) The reliability of citation counts in judgments on promotion, tenure, and status, *Arizona Law Review*, pp. 829–829.

Benjamini, Y. & Braun, H. (2002) John Tukey's contributions to multiple comparisons, *Annals of Statistics*, 30, pp. 1576–1594.

Bland, J. M. & Altman, D. G. (1986) Statistical methods for assessing agreement between two clinical measurements, *Lancet*, 1 (8476), pp. 307–310.

Bland, J. M. & Altman, D. G. (1992) Comparing methods of clinical measurement. [Citation Classic], *Current Contents*, 20, p. 8.

Bland, J. M. & Altman, D. G. (1995) Comparing two methods of clinical measurement: a personal history, *International Journal of Epidemiology*, 24 (Suppl. 1), pp. S7–S14.

Box, G. E. P. & Cox, D. R. (1964) An analysis of transformations, *Journal of the Royal Statistical Society, Series B*, 26, pp. 211–243 (discussion pp. 244–252).

Campbell, M. J. & Julious, S. A. (1994) *Statistics in Medicine*: citations of papers in the first ten years, *Statistics in Medicine*, 13, pp. 3–10.

Cooley, J. W. & Tukey, J. W. (1965) An algorithm for the machine calculation of complex Fourier series, *Mathematics of Computation*, 19, pp. 297–301.

Cox, D. R. (1972) Regression models and life tables, *Journal of the Royal Statistical Society, Series B*, 34, pp. 187–220.

Cox, D. R. (1986) This week's citation classic, *Current Contents*, CC/Number 42, p. 16.

Cronin, B. (1984) *The Citation Process, the Role and Significance of Citations in Scientific Communication* (Taylor Graham).

Cutler, S. J. (1979) This week's citation classic, *Current Contents*, Number 16, p. 356.

Cutler, S. J. & Ederer, F. (1958) Maximum utilization of the life table method in analyzing survival, *Journal of Chronic Diseases*, 8, pp. 699–712.

DeGroot, M. H. (1987) A conversation with George Box, *Statistical Science*, 2, pp. 239–258.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (C/R: pp. 22–37), *Journal of the Royal Statistical Society, Series B*, 39, pp. 1–22.

Donoho, D. (2002) How to be a widely cited author in the mathematical sciences, *in-cites*. Available at http://www.in-cites.com/scientists/DrDavidDonoho.html.

Duncan, D. B. (1955) Multiple range and multiple *F*–tests, *Biometrics*, 11, pp. 1–42.

Duncan, D. B. (1975) *t* tests and intervals for comparisons suggested by the data, *Biometrics*, 31, pp. 339–359.

Duncan. D. B. (1977) This week's citation classic, *Current Contents*, Number 4, p. 10.

Dunnett, C. W. (1955) A multiple comparison procedure for comparing several treatments with a control, *Journal of the American Statistical Association*, 50, 1096–1121.

Dunnett, C. W. (1964) New tables for multiple comparisons with a control, *Biometrics*, 20, pp. 482–491.

Dunnett, C.W. & Sobel, M. (1954) A bivariate generalization of Student's *t*–distribution with tables for certain special cases, *Biometrika*, 41, pp. 153–169.

Dunnett, C.W. & Sobel, M. (1955) Approximations to the probability integral and certain percentage points of a multivariate analogue of Student's *t*–distribution, *Biometrika*, 42, pp. 258–260.

Dunnett, C. W. & Tamhane, A. C. (1991) Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts, *Statistics in Medicine*, 10, pp. 939–947.

Dunnett, C. W. & Tamhane, A. C. (1992) A step-up multiple test procedure, *Journal of the American Statistical Association*, 87, pp. 162–170.

Edge, D. (1979) Quantitative measures of communication in science: a critical review, *History of Science*, 17, p. 102.

Efron, B. (1979) Bootstrap methods: another look at the jackknife, *The Annals of Statistics*, 7, pp. 1–26.

Efron, B. (2001) Statistics in the 20th century, and the 21st, Invited Lecture at the Statistiche Woche 2001 in Vienna, Austria, 16–19 October 2001. Published in Fetschrift 50 Jahre Österreichische Statistische Gesellschaft, Seiten 7–19. Available at http://www.statistik.tuwien.ac.at/oezstat/festschr02/papers/efron.pdf.

Engle, R. F. & Granger, C. W. J. (1987) Co-integration and error correction: representation, estimation and testing, *Econometrica*, 55, pp. 251–276.

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap, *Evolution*, 39, pp. 783–791.

Fisher, R. A. (1953) Dispersion on a sphere, *Proceedings of the Royal Society of London, Series A*, 217, pp. 295–305.

Garfield, E. (1977) The 250 most-cited primary authors, 1961–1975. Part III. Each author's most-cited publication, *Current Comments: Essays of an Information Scientist*, 3, pp. 348–363. Available at http://www.garfield.library.upenn.edu/essays/v3p348y1977–78.pdf.

Garfield, E. (1979) *Citation Indexing – its Theory and Application in Science, Technology, and Humanities* (New York: Wiley).

Garfield, E. (1989) Delayed recognition in scientific discovery: citation frequency analysis aids the search for case histories, in: *Essays of an Information Scientist: Creativity, Delayed Recognition, and other Essays*, 12, pp. 154–160, and in *Current Contents*, 23, pp. 3–9, June 5, 1989.

Garfield, E. (1990) The most-cited papers of all time, SCI 1945–1988. Part 1A. The SCI top 100 – will the Lowry method ever be obliterated?, *Current Comments*, 7, pp. 3–14. Available at http://www.garfield.library.upenn.edu/essays/v13p045y1990.pdf.

Garfield, E. (1998) The use of journal impact factors and citation analysis for evaluation of science, Presented at Cell Separation, Hematology, and Journal Citation Analysis, Mini–Symposium in tribute to Arne Bøyum, Rikshospitalet, Oslo.

Garfield, E. (2000) The evolution of 'Hot Papers', *The Scientist*, 14. Available at www.garfield.library.upenn.edu/commentaries/tsv14(14)p04y20000710.pdf.

Geman, S. & Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, pp. 721–741.

Gilbert, G. N. (1977) Referencing as persuasion, *Social Studies of Science* 7, pp. 113–122.

Hahn, G. J. (1995) A conversation with Donald Marquardt, *Statistical Science*, 10, pp. 377–393.

Hamilton, D. P. (1990) Publishing by and for ? the numbers, *Science*, 250, pp. 1331–1332.

Hamilton, D. P. (1991) Research papers: who's uncited now?, *Science*, 251, p. 25.

Hartley, H. O. (1958) Maximum likelihood estimation from incomplete data. *Biometrics*, 14, pp. 174–194.

Holmes, S. (2003) Bootstrapping phylogenetic trees: theory and methods, *Statistical Science*, 18, pp. 241–255.

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities, *Proceedings of the National Academy of Sciences*, 79, pp. 2554–2558.

Kaplan, E. L. (1983) This week's citation classic, *Current Contents*, CC/Number 24, p. 14.

Kaplan, E. L. & Meier, P. (1958) Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, 53, pp. 457–481.

Kotz, S. & Johnson, N. L. (1992a) *Breakthroughs in Statistics Volume I* (New York: Springer-Verlag).

Kotz, S. & Johnson, N. L. (1992b) *Breakthroughs in Statistics Volume II* (New York: Springer-Verlag).
Kotz, S. & Johnson, N. L. (1997) *Breakthroughs in Statistics Volume III* (New York: Springer-Verlag).
Kramer, C. Y. (1956) Extension of multiple range tests to group means with unequal numbers of replications, *Biometrics*, 12, pp. 307–310.
Kruse, M. (2002) Members, *JASA* top math citations, *AmStat News*, Issue 305, November, pp. 6–7.
Laviolette, M., Seaman Jr., J. W., Barrett, J. D. & Woodall, W. H. (1995) A probabilistic and statistical view of fuzzy methods, (with discussion) *Technometrics*, 37, pp. 249–292.
Liang, K.-Y. & Zeger, S. (1986) Longitudinal data analysis using generalized linear models, *Biometrika*, 73, pp. 13–22.
Litchfield, J. T. (1977) This week's citation classic, *Current Contents*, Number 7, p. 8.
Litchfield, J. T. & Wilcoxon, F. A. (1949) A simplified method of evaluating dose-effect experiments, *Journal of Pharmacological and Experimental Therapeutics*, 96, pp. 99–113.
Marquardt, D. W. (1979) This week's citation classic, *Current Contents*, CC/Number 27, p. 20.
Mantel, N. (1963) Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure, *Journal of the American Statistical Association*, 58, pp. 690–700.
Mantel, N. (1966) Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports*, 50, pp. 163–170.
Mantel, N. & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute*, 22, pp. 719–748.
McKendrick, A. G. (1926) Applications of mathematics to medical problems, *Proceedings of the Edinburgh Mathematics Society*, 44, pp. 98–130.
McLachlan, G. J. & Krishnan, T. (1997) *The EM Algorithm and Extensions* (New York: Wiley).
Meier, P. (2001) The price of Kaplan–Meier, paper presented at the 9th Merck–Temple Conference on Research Topics in Pharmaceutical Statistics. Available at http://www.sbm.temple.edu/~biostat/conf9.html.
Meier, P., Karrison, T., Chappell, R. & Xie, H. (2004) The price of Kaplan–Meier, *Journal of the American Statistical Association*, 99, pp. 890–896.
Miller Jr., R. G. (1983) What price Kaplan–Meier?, *Biometrics*, 39, pp. 1077–1081.
Nei, M. (1972) Genetic distance between populations, *The American Naturalist*, 106, pp. 283–292.
Nelder, J. A. & Mead, R. (1965) A simplex method for function minimization, *Computer Journal*, 8, pp. 308–313.
Pendlebury, D. A. (1991) Science, citation, and funding, *Science*, 251, pp. 1410–1411.
Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. & Smith, K. G. (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Part II. Analysis and examples, *British Journal of Cancer*, 35, pp. 1–39.
Reed, L. J. & Muench, H. (1938) A simple method of estimating 50 percent endpoints, *American Journal of Hygiene*, 27, pp. 493–497.
Reid, N. (1994) A conversation with Sir David Cox, *Statistical Science*, 9, pp. 439–455.
ScienceWatch (1999) Citations reveal concentrated influence: some fields have it, but what does it mean?, *ScienceWatch*, 10(1). Available at http://www.sciencewatch.com/jan-feb99/sw_jan–feb99_page2.htm.
Schwarz, G. (1978) Estimating the dimension of a model, *Annals of Statistics*, 6, pp. 461–464.
Stigler, S. M. (1994) Citation patterns in the journals of statistics and probability, *Statistical Science*, 9, pp. 94–108.
Straf, M. L. (2003) Statistics: the next generation, *Journal of the American Statistical Association*, 98, pp. 1–6.
Theoharakis, V. & Skordia, M. (2003) How do statisticians perceive statistics journals?, *The American Statistician*, 57, pp. 115–123.
Tukey, J. W. (1991) The philosophy of multiple comparisons, *Statistical Science*, 6, pp. 100–116.
Weir, B. S. & Cockerham, C. C. (1984) Estimating *F*–Statistics for the analysis of population structure, *Evolution*, 38(6), pp. 1358–1370.
Wilkinson, G. N. (1961) Statistical estimations in enzyme kinetics, *Biochemical Journal*, 80, pp. 324–336.
Wright, S. (1931) Evolution in Mendelian populations, *Genetics*, 16, pp. 97–159.
Zadeh, L. A. (1965) Fuzzy sets, *Information and Control*, 8, pp. 338–353.

General

Statistics and Ethics: Some Advice for Young Statisticians

Stephen B. VARDEMAN and Max D. MORRIS

We write to young statisticians about the nature of statistics and their responsibilities as members of the statistical profession. We observe that the practice of the discipline is inherently moral and that this fact has serious implications for their work. In light of this, we offer some advice about how they should resolve to think and act.

KEY WORDS: Graduate study; Integrity; Principle; Professional practice; Research; Teaching.

Dear Gentle Reader:

So, you are embarking on a career in statistics. Good. It is a genuinely noble pursuit, though this may be hard to see as you wrestle with new-to-you technical issues varying from "How do I get this SAS job to run?" to "How do I show this thing is UMVU?" and on occasion find yourself wondering "What is the point of all this?"

This last question about purpose is actually a very important and quite serious one. It has implications that run far beyond your present pain (and joy) of "getting started." How you answer it will affect not only you, but also the profession, and human society at large. We write to offer some advice and encouragement, and to say how we hope you frame your answer to this simultaneously practical and cosmic question.

What *are* this subject and this profession really all about? And why *are* you doing what you are doing? For sure, there are details to learn (and keep current on throughout a career). There is everything from the seemingly uncountable number of tricks of first year probability theory, to statistical computing, to nonlinear models. It initially looks like "soup to nuts." You know that statistics is about collecting and handling data. That is true, but incomplete; there is much more than that at work here.

The vital point is that this discipline provides tools, patterns of thought, and habits of heart that will allow you to deal with data *with integrity*. At its core statistics is not about cleverness and technique, but rather about *honesty*. Its real contribution to society is primarily *moral*, not technical. It is about doing *the right thing* when interpreting empirical information. Statisticians are *not* the world's best computer scientists, mathematicians, or scientific subject matter specialists. We *are* (potentially, at least) the best at the *principled* collection, summarization, and analysis of data. Our subject provides a framework for dealing transparently and consistently with empirical information from *all* fields; means of seeing and portraying what is true; ways of avoiding being fooled by both the ill intent (or ignorance) of others and our own incorrect predispositions. The mix of theory and methods that you are discovering is the best available for achieving these noble ends. The more you practice with it, the sharper will become your (fundamentally moral) judgments about what is appropriate in handling empirical information.

Others from areas ranging from philosophy to physics might well object that we have claimed too much, wrapping statistics in a cloak of virtue to the apparent exclusion of other disciplines. After all, thoughtful scientists and humanists from a variety of fields are engaged in the pursuit of truth. And any serious education has moral dimensions. Our point, however, is that the particular role that the profession plays in science and society should not be viewed as amoral, and that this fact constrains how we all must think and act as its members.

That society expects our profession to play this kind of role can be seen in the place statistics has as arbiter of what is sufficient evidence of efficacy and safety to grant FDA approval of a drug, or enough evidence to support an advertiser's claim for the effectiveness of a consumer product. And it can be seen in the fact that many disciplines have "statistical significance" requirements for results appearing in their journals.

Society also recognizes that when statistical arguments are abused, whether through malice or incompetence, genuine harm is done. How else could a book titled *How to Lie With Statistics* (Huff 1954) have ever been published and popular? The famous line (attributed by Mark Twain (1924) to Benjamin Disraeli) "There are three kinds of lies: lies, damned lies, and statistics" witnesses effectively to society's distaste for obfuscation or outright dishonesty cloaked in the garb of statistical technology. Society disdains hypocrisy. It hates crooked lawyers, shady corporate executives, and corrupt accountants, and it has contempt for statisticians and statistical work that lack integrity. But young statisticians sometimes find themselves being "encouraged" to offer questionable interpretations of data. This pressure can come even from well-meaning individuals who believe that their only interest is in ensuring that their position is treated "fairly." Maintaining an independent and principled point-of-

Stephen B. Vardeman is Professor, and Max D. Morris is Professor, Department of Statistics and Department of Industrial and Manufacturing Systems Engineering, Iowa State University Ames, IA 50011-1210 (E-mail: vardeman@iastate.edu). The authors gratefully acknowledge the generous input of a number of colleagues. Karen Kafadar, Bob Stephenson, Bill Meeker, and Dean Isaacson provided detailed comments on a first draft of this article. And the input of Ken Ryan, Tammy Brown, Mike Moon, Bill Notz, Tom Dubinin, Frank Peters, David Moore, Bill Duckworth, Bruce Held, Dennis Gilliland, Bobby Mee, Doug Bonett, Dan Nettleton, and Hal Stern is also gratefully acknowledged.

 DOI: 10.1198/0003130031072

view in such contexts is critical if a statistician hopes to avoid becoming a part of Disraeli's third "lie."

So, you are embarking upon a noble and serious business. We take as given that you have a basic moral sense and a strong desire to personally do good. We also take as self-evident that integrity is a pattern of life, not an incident. Principled people consistently do principled work, regardless of whether it serves their short-term personal interests. Integrity is not something that is turned on and off at one's convenience. It cannot be generally lacking and yet be counted on to appear in the nick of time when the greater good calls. This implies that what you choose to think and do now, early in your career, are very good predictors of what you will think and do throughout the whole of it. You are setting patterns that will endure over a professional lifetime and substantially influence the nature and value of what you can hope to accomplish.

A fair amount has been written about professional ethics in statistics and we do not propose to review it all or comment on every issue that has been raised. For example, Deming's (1986) article is fundamentally a discussion of ethics. Both the American Statistical Association (1999) and the International Statistical Institute (1985) have official statements on ethical guidelines for statisticians. And in a more general setting, the National Academy of Sciences (1995) has published a useful booklet that is primarily about ethics in science and has implications for statistical practice.

Our more specific goal here is to suggest some things that a high view of the discipline means for your present work and attitudes. Aiming to speak to both statistics graduate students and recent grads, we'll begin with some implications for life in graduate school, and then move on to implications for an early career in the discipline.

ADVICE FOR STATISTICS GRADUATE STUDENTS

"Graduate student ethics" (or for that matter "professional ethics") is really just "plain ethics" expressed in a graduate student (or professional) world. A discussion of it really boils down to consideration of circumstances and issues that arise in a particular graduate student (or professional) setting. So an obvious place to begin is with general student responsibilities. If you are still in graduate school, we urge you to be scrupulous about your conduct in the courses you take. Here are some specifics:

- Resolve to never accept credit for work that is not your own. It should make no difference to you whether an exam is proctored or unproctored. Whatever the homework policy of the course, make it your practice to clearly note on your papers places where you have gained from discussions with classmates or consulting old problem sets of others. It's simply right to give others credit where it is deserved and it's simply wrong to take credit where it is undeserved.

- If course policy is that everyone is "completely on their own," resolve in advance to politely refuse to discuss with peers topics that are off-limits, even if others violate the policy. It may seem a small thing at the time, but you are setting life trajectories that are bigger than the particular incidents.

- Determine to never take advantage of (or over) your peers. If you join a group study session, be ready to make your fair contribution, not just to benefit from the input of others. If you have legitimate access to old files or notes or textbooks that are helpful, let others know about them so that they can benefit as well.

What do these three points say? Simply that you should play by the rules set out and be clear and honest about all contributions made to the work you turn in. Why would anyone do otherwise? Honestly, only to gain an undeserved advantage in a course grade, or to avoid some effort. But a student willing to cut corners for an A or a free weekend will have serious difficulty not cutting corners in later professional responsibilities when the reward is a promotion or pay raise or a free weekend.

Some additional issues are related to the notion of "doing the hard thing." Everyone has things that come harder for them than others. It's human nature to want to avoid what is difficult and to even convince ourselves that really, the easy thing is what is important and the hard thing is worthless. But that is not only obviously silly, it has moral implications. Here is some advice for the student reader:

- Understand that acquiring an advanced education is a difficult enterprise, that there may be times when you *feel* like complaining about this, but that it doesn't really help to do so. Whining wastes energy and can poison the learning atmosphere for others. You are engaged in a noble, if difficult, pursuit. Give it your best shot without complaining. After all, most things worth doing *are* hard.

- Resolve to work on your weaknesses rather than excuse them. Doing good statistical work is important, and demands the best possible personal tool kit. The reasoning "I find methods (theory) easier than theory (methods), so I'll just do methods (theory)" implicitly and quite wrongly assumes that one can do good statistical work with half a tool kit.

- Decide not to denigrate the strengths of others. Give other people credit for what they can do that you cannot. Find your niche without minimizing the honest efforts and contributions of others.

- Determine to take the courses that will enable you to be the best-educated and most effective statistician you can be. These are often academically demanding, and may not form a particularly easy route to a high GPA. While difficulty, per se, is not necessarily a measure of how often you will find the material in a course useful, it *is* related to the mental discipline you will develop. If you choose a course that covers material you could easily pick up on your own or because it is taught by a professor who demands little in exchange for an A, you've cheated yourself. The choices you make about curriculum are moral choices, not just choices of convenience. You have a limited time in graduate school . . . use it wisely. How effective you will be as a professional depends on it. Besides, your choices say something nontrivial about the personal character that you are developing.

- Purpose to do what your thesis or dissertation advisor sets for you to do, as independently as you can. While it may seem that some assignments are arbitrary or unnecessary, remember

that you do not have your advisor's experience as a researcher *or* educator. This person knows what you know, what your abilities are, and the difficulty of your problem. He or she is trying to help you to develop as a responsible and independent member of the profession, one accustomed to consistently working up to your capabilities. Focusing your energy on the challenge of the problem and the opportunity it represents will take you much farther than wasting your energy in grumbling or in negotiating to be led through every detail of a solution.

It is worth adding a further note related to this last point. The advisor–advisee experience has the potential to be invigorating and rewarding (both professionally and personally) for both parties. Think of the efforts you put into it not only as a requirement for the degree, but as the beginning of what may be one of your most important and cherished long-term relationships. Find someone to work with who you like and respect, and put your energy into the enterprise.

Most statistics graduate students work as graduate assistants. Assistants should remember first that an assistantship is not a fellowship, but rather a job. And it is axiomatic that principled people return honest effort for their pay. If you are working on a faculty member's grant, that person must produce quality work in line with the interests of some outside entity. Do what you can to help him or her. If you are a teaching assistant, there are lectures to conscientiously prepare and deliver, papers to carefully grade, and students to help. If you are a consultant, people with real problems of data analysis will appear at your door seeking aid. They need your best effort and advice. Let us amplify a bit:

- If you are a research assistant it is understood that you have "your own" class work and thesis or dissertation to attend to. But some of your weekly hours are first committed to providing the help (programming, library work, report writing, etc.) your employer needs. There are important educational benefits that accrue as you practice at these duties. But the most fundamental reason to carry them out conscientiously and cheerfully is simply that it is the right thing to do. (And it is wrong to think that cutting corners now doesn't say anything about later behavior. Life will always be hectic and there is no reason to expect your work habits after finishing school to be better than the ones you are developing now.)

- If you are a teaching assistant, purpose to make the best of the fact that along with some conscientious, motivated, and pleasant students, you will deal with some unpleasant, intentionally ignorant, lazy, and dishonest students. It simply comes with the territory. For your part, make it a point to model integrity and purpose for all of them. Do your best to convey that what you are teaching them really does matter and how they do it matters as well. Resolve that whatever your "style"/personality (from animated to reserved) your body language will convey a genuine willingness to help. The job takes patience—plan on it. Resolve to treat all of your students well, whether or not their behavior in any sense merits that. And it should go without saying that although you want to be pleasant and approachable, propriety and impartiality dictate that you are their instructor or TA, not their pal.

- If your assignment is to help with statistical consulting, you are already wrestling (at a "trainee" level) with some of the serious issues faced by one segment of our profession. Carefully consider and handle these now, as you begin to see how the "human element" of statistical consulting requires thoughtful and principled discipline. You're going to have to argue with yourself in conversations like:

— What looks to me like the thing that *should* be done would take two hours to explain and several more hours of my time to implement, while this client would be happy with something less appropriate that I could explain in five minutes . . .

— This client *really* wants "A" to be true, but these data look inconclusive . . .

— This looks pretty much OK except for that oddity over there that the client doesn't really want to discuss . . .

Graduate Student Reader, keep your eyes open during this graduate student experience. Watch your faculty and emulate the ones who take seriously what they do. There are some fine role models in our university statistics departments, excellent members of the profession. Find them, and learn as much as you can about what they think and how they practice statistics.

ADVICE FOR YOUNG PROFESSIONAL STATISTICIANS

Many of the themes we've introduced in the context of graduate study have their logical extensions to early professional life. But there are also other matters that we've not yet raised. We proceed to discuss some of the less obvious extrapolations and further ethical issues faced by young statisticians, organizing our advice around the topics of (1) research/publication, (2) teaching, and (3) professional practice.

If you have finished a Ph.D., you have been introduced to the craft of research in statistical theory or methods. You are in a position to help develop the profession's supporting body of knowledge and to contribute to our journals. It's important to consider the corresponding responsibilities. These are tied closely to a proper view of the purpose of publication in statistics. Published statistical research should provide reliable and substantial new theory or methodology that has genuine potential to ultimately help statisticians in the practice of the discipline. Statistical publication should not be treated as a game. It is, and should be treated as, a serious and moral business. Here are some points of advice issuing from this high view of what the research and publication activity is all about:

- Resolve that if you choose to submit work for publication, it will be complete and represent your best effort. Submitting papers of little intrinsic value, half-done work, or work sliced into small pieces sent to multiple venues is an abuse of an important communication system and is not honorable scholarship. It is not the job of editors or referees to proofread or complete your papers, or to insist that you follow up on important issues that you know exist. See the "Let's just send it off and let the reviewers

sort it out" impulse for what it is, a temptation to off-load your work to someone else. And the "I'll just submit this half-done thing to an outlet that will print anything" strategy does nothing of real value for anyone. It wastes time and effort of those in the review system, and when "successful" it dilutes our literature. This makes important work harder to find, and in the end calls into question our very reason to exist as a profession.

• Purpose that when asked to do the job of a referee, you will do it thoroughly, impartially, and in as timely a manner as possible. There is no obvious short-term payoff to doing what is right here. But the integrity and currency of the scientific publication process depend on competent and principled referees taking the job seriously. Resolve never to do a shoddy/cursory review job, or worse yet to let calculations about personalities (and personal advantage) govern how you judge a piece of work. Even though many statistics journals use a "double-blind" system, the profession is small, and you will find it increasingly rare that you have no idea who authored a paper you receive for review. So remember that the *spirit* of the blind review policy is honorable, and that you have an obligation to conduct your review in this spirit even when you cannot be completely "blind." And do what you can as an individual to help fix the widely recognized problem that the review process in statistics is presently much slower than in many other disciplines.

• Decide to routinely take the advice of editors and referees regarding papers that you submit for publication. Occasions are rare where editors or referees have it all wrong or purposely treat an author unfairly. Most often, the advice they offer is constructive and when followed substantially improves an article. Until an editor signals clearly that he or she has no further interest in a piece you have submitted, you should almost always make good faith efforts to revise your paper in accord with his or her advice. Serial journal-shopping for a venue that will publish a submission with essentially no revision may minimize the total effort an author expends on a paper, but the practice wastes the overall energy of the profession and has a negative effect on the overall quality of what is published.

• Determine to be scrupulous about giving credit where it is due. If another has contributed substantially to the content of a paper, co-authorship is typically appropriate and should be offered. (On the other hand, *never* list a colleague as co-author of a paper until you have that person's explicit permission to do so.) And include acknowledgments of others deserving thanks for less extensive, but real, help with an article.

• Resolve to acknowledge priority and the derivative nature of your work with due humility. If after the fact of publication you find that some of your results can be found in earlier work, immediately send an acknowledgment to that effect to the journal where your paper appeared. In writing your papers in the first place, we encourage you to be forthright and helpful about what you know is already published on your subject, delineating carefully what others have already said and where your new contribution lies. (No one ever really "starts from scratch." Don't fall prey to the temptation to leave unsaid what you know is already known, thinking that to do so strengthens your own position.) And *never* borrow published/copyrighted words, even of your own authorship, without acknowledgment. To do so is plagiarism and is completely unacceptable. (This caution extends, by the way, to thesis and dissertation work, even if that work is never submitted to a journal for formal publication.)

A note related to this last point: Avoiding plagiarism places an extra burden on students whose writing skills are not strong, especially those struggling with English as a second language. But it is essential to find one's own words and not simply copy or even paraphrase those of another (even for parts of a paper that are background and obviously don't purport to provide new technical content). This is a very serious integrity issue.

Next, let's consider issues relevant to teaching of statistics as a professional. There are reasons to do this whether or not you have plans for a career at a college or university. Teaching/training is increasingly done "in house" by corporations and consultants, and it could be argued that most professional presentations are essentially teaching efforts. The logical extension of the advice offered above to graduate teaching assistants is, of course, relevant here. But there is an important extra dimension to discuss, related to the freedom and responsibility that a professional has in answering the question "What will govern what and how I teach?" Will it be "What's easy for me?" Or will it be "What will get the best short-term reaction from the students?" Or will it be "My best professional judgement as to what the students need for the long term and my best understanding of how to effectively convey that information?" This is a moral choice. Here is some amplification:

• Determine that you won't fall into the trap of organizing all courses around your technical specialty. This is an issue of fundamental humility and recognition that none of us has put all that is needed into our personal little package (to say nothing about the matter of "truth in advertising!"). But we suspect that you know what we are talking about, having seen people turn every course they teach into a platform to show off their own work.

• Purpose not to be governed by what is easy to do. This is not an entirely separate issue from the previous one. But we are also thinking about cases where the case is not so blatant or not tied directly to one's specialty. It's a lot of work to learn new methods and software to include in a course, to freshen examples, to develop new laboratories and assignments for students, to replace outdated topics and means of presentation. And it's sometimes possible to "get by" without investing that effort. But doing so is simply wrong. We urge you not to take that route.

• Resolve to do the best for your students, whether or not they appreciate your efforts in the short term. We live in a "consumer" society. There is huge pressure on teachers in all contexts to make students happy. But statistics is hard, and students DON'T know what they need. You will. We hope that you opt to do your best to provide that, not simply what will get the best crowd reaction. Lots of jokes, little in the way of course demands, and high grades can please many audiences. And leave students ignorant. Of course we should aim to be engaging in our presentation of our subject. But the point of teaching is to genuinely improve subject matter knowledge and the reasoning powers of students. It is not to produce feel-good experiences for them. (In this regard, we were recently dismayed to see an Iowa community college president quoted in the *Des Moines Register* (2001) as

proudly saying "We are really a service organization first and an educational institution second." While that may in fact be true, it is a terrible commentary on the state of the institution.)

Those of you beginning academic careers will face enormous demands for early success. Most universities require substantial accomplishments in both research and teaching during the first six years of employment, and some place the bar so high that seemingly superhuman effort is required. If numbers of refereed publications and instructor evaluations are the "keys to success," can you afford to have *real quality* as your primary goal? Is there enough time in six short years to accomplish all that is required if you take our advice seriously? These are real and hard questions. How you use your assistant professorship is critical to your long-term professional success, and it is obvious that you must take your institution's expectations into account. But, we urge you as you face these issues to remember that one who spends an assistant professorship cutting corners is at best prepared to be an associate professor who knows how to cut corners . . . not one who has learned how to make a difference.

Turning finally to the area of professional practice, we note that most of what has been written about ethical guidelines for statisticians concerns what is appropriate in public practice, in lending aid to others in the impartial and efficient collection and analysis of their data. This is understandable, as (1) the discipline's whole reason to exist is ultimately to provide such aid and (2) this activity is both subtle and full of pitfalls. Both the ethical guidelines and public skepticism typified in the "lies" quote of Disraeli point to the fact that statistics can be used to form highly technical and even technically correct support for statements which are in fact not true. We might hope this could happen only when nonstatisticians practice statistics without proper technical understanding of the subject. But statistical lies are by definition immoral uses of statistical arguments, whether technically correct or not, and stem from societal pressures that affect statisticians and nonstatisticians alike. What then must you do in society to preserve the discipline's (and your own) integrity?

First, recognize that *a professional statistician should never behave like a courtroom lawyer.* The practice of law is based on an adversarial model in which each lawyer represents an assigned point of view—that which will yield the most positive outcome for his or her client. While the use of lies and intentionally misleading statements is prohibited in legal proceedings, legal strategy certainly does involve the selective use of evidence so as to present the truth (or some part of it) in the light most favorable to a particular point of view. But a key aspect of this model of litigation is that decisions are made by an unbiased authority (a judge or jury) based not on the case presented by a single side, but only after arguments presented by all parties are heard.

Statisticians usually do not operate in such well-controlled adversarial systems. If you *do* work in this kind of arena you must keep absolutely clear the distinction between an objective analyst and an advocate, and never purport to be (or think yourself) the first when you are the second. If you are employed by an organization (whether on a permanent basis or as a consultant) you are by definition not disinterested in its well-being. And if you are working "pro bono" for a cause you support, you are not disinterested in furthering the cause. In either case, it is axiomatic that your professional judgment is potentially clouded by what you (quite naturally) want to be true. And you will be no fair judge of the extent to which this clouding has occurred. There is real danger here. There is little that is more damning to the discipline than for one of its professionals, implicitly claiming some degree of objectivity, to be publicly exposed as overstating a statistical case in favor of his or her employer or cause.

More commonly, statisticians function as consultants to those who must make decisions. We do this through careful and thoughtful design of data collection mechanisms and analysis of assembled data. But "careful and thoughtful" here are words that acknowledge a critical fact: *Statistical analysis of data can only be performed within the context of selected assumptions, models, and/or prior distributions.* A statistical analysis is actually the extraction of substantive information from data and assumptions. And herein lies the rub, understood well by Disraeli and others skeptical of our work: For given data, an analysis *can* usually be selected which will result in "information" more favorable to the owner of the analysis than is objectively warranted.

The only "cure" for this difficulty is statistical practice based on assumptions embodying an informed, balanced, and honest representation of what is known. "Known," not "wished for," "desired," "convenient," or even "other-than-worst-fears." This has implications for how statisticians must be and act if they are to be both effective and ethical.

- *Statisticians must be knowledgeable about the system under study.* They should not present themselves as competent to analyze data from systems about which they have no substantive understanding. Real data are not "context-free."

- *On the other hand, statisticians must recognize and acknowledge the limitations of their "subject matter" knowledge.* Data and variation are ubiquitous. Knowing how to handle them can give you important and even uncommon insights in a variety of contexts where you have limited subject matter credentials. But the fact that you can make contributions in league with experts in a variety of fields doesn't substitute for credentials in those fields. The credibility of the statistical profession depends upon its members being scrupulous about what they know and what they don't know. Never forget that you are not the context expert.

- *Statisticians must go out of their way to see that their analyses allow interpretations of the available data which are tenable but not popular in the statistician's organization.* This does not mean "be a troublemaker," but it does mean that you should carefully think through how available data would be interpreted by those with all possible rational points of view.

- *Statisticians must write complete reports stating the results of their entire informed thought processes—including what they know, what they have assumed, what they have decided cannot be assumed, and what conclusions tenable assumptions support.* Our reports should contain "complete and sufficient" analyses upon which any rational point of view can be argued. If you come to the conclusion that one of the spectrum of sensible interpretations is "best" in a particular application, make it your goal to be absolutely transparent about your reasoning. People

should be able to easily see your full set of model assumptions, understand what methodology you have used to make inferences in that model, and have access to diagnostic and robustness work you have done. (This advice is sound in general. But it is perhaps especially relevant to explicitly Bayesian analyses. A consumer of a posterior distribution has a moral right to know how strongly it depends upon the prior.) Honest statistical work has nothing to hide. It says what it says. It doesn't try to obscure points where alternative conclusions are possible if other assumptions are made or different analysis paths are followed, and admits where model fits are short of perfection or conclusions are highly model-dependent.

As a statistician, your allegiance must be to finding the conclusions which can be supported by data and careful assumptions. Does this make the business of assumption selection more difficult than it seemed in your statistics coursework? Does it seem as though you must take these issues more personally and seriously than our favorite semi-academic phrase "Let $X_1, X_2, \ldots, X_n$ be iid $F \ldots$?" Does it sound like your formulation of these assumptions may have more to do with nonmathematical *values* than has been discussed in your textbooks? Yes, this and more is true. Ethical statistical practice requires that you *take responsibility* for acquiring substantive understanding, knowing all rational points of view, and making decisions well beyond those based entirely in data.

- You must examine yourself to see that you are not even subconsciously leaning toward analyses which you believe will "please the boss" or yourself, or simplify the problem unjustifiably. This means that you cannot afford to think of yourself as a data technician or a hired gun. You must be secure enough to simultaneously separate any prior vested interest (yours or others') in the outcome from your analysis, and meld together seamlessly everything you know about the subject matter of your investigation with the structure of your statistical work. *You cannot do this unless you have strength of character and integrity.*

- You must not stop with the obvious or even the most likely explanation of data, but find ways to examine them so that all rational viewpoints can be informed. This means that you will work harder and longer than anyone who reads your reports will ever know. You will not rest until you *know* you understand all the information contained in the data, where "information" is defined by the context of your work across the spectrum of rational viewpoints. *You cannot do this unless you develop an ethic of self-reliance, thoroughness, and hard work.*

- You must understand fully what your assumptions say and what they imply. You must not claim that the "usual assumptions" are acceptable due to the robustness of your technique unless you really understand the implications and limits of this assertion in the context of your application. And you must absolutely never use any statistical method without realizing that you are implicitly making assumptions, and that the validity of your results can never be greater than that of the most questionable of these. *You cannot do this unless you remain dedicated to being the best technical statistician you can possibly be, understanding that this involves knowing and understanding the mathematical arguments as well as the computational techniques behind every tool you need.*

Well there it is, more than enough advice to keep a young statistician busy for a career. We hope we don't sound too much like myopic cranks, finding "serious ethical issues" to raise in even the most mundane contexts. Instead, we hope that we have argued effectively that ethical matters are central to our discipline and provided some insight into issues that this raises. We further hope that you determine to take the matter of principle most seriously.

Carry on, Gentle Reader.

[Received April 2002. Revised November 2002.]

REFERENCES

American Statistical Association (1999), "Ethical Guidelines for Statistical Practice," http://www.amstat.org/profession/ethicalstatistics.html

Deming, W. E. (1986), "Principles of Professional Statistical Practice," in *Encyclopedia of Statistical Science* (vol. 7), eds. S. Kotz and N. Johnson, New York: Wiley.

Des Moines Register (2001), "Western Iowa Tech Among Nation's Fastest Growing Schools," December 31, 2001, pp. B-1.

Huff, D. (1954), *How to Lie with Statistics*, New York: Norton.

International Statistical Institute (1985), "Declaration on Professional Ethics," http://www.cbs.nl/isi/ethics.htm.

National Academy of Sciences (1995), *On Being a Scientist: Responsible Conduct in Research*, Washington, DC: National Academy Press. Also available online at http://www.nap.edu/books/0309051967/html/.

Twain, M. (1924), *Mark Twain's Autobiography*, New York: Harper & Brothers.

Likelihood Theory

This appendix is just an overview of the likelihood theory used in this book. For greater detail or a more gentle introduction, the reader is advised to consult a book on theoretical statistics such as Cox and Hinkley (1974), Bickel and Doksum (1977) or Rice (1998).

A.1 Maximum Likelihood

Consider n independent discrete random variables, $Y_1, \ldots, Y_n$, with probability distribution function $f(y|\theta)$ where θ is the, possibly vector-valued, parameter. Suppose we observe $\mathbf{y} = (y_1, \ldots, y_n)^T$, then we define the likelihood as:

$$P(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{n} f(y_i|\theta) = L(\theta|y)$$

So the likelihood is a function of the parameter(s) given the data and is the probability of the observed data given a specified value of the parameter(s).

For continuous random variables, $Y_1, \ldots, Y_n$ with probability density function $f(y|\theta)$, we recognize that, in practice, we can only measure or observe data with limited precision. We may record y_i, but this effectively indicates an observation in the range $[y_i^l, y_i^u]$ so that:

$$P(Y_i = y_i) = P(y_i^l \leq y_i \leq y_i^u) = \int_{y_i^l}^{y_i^l} f(u|\theta)du \approx f(y_i|\theta)\delta_i$$

where $\delta_i = y_i^u - y_i^l$. We can now write the likelihood as:

$$L(\theta|y) \approx \prod_{i=1}^{n} f(y_i|\theta) \prod_{i=1}^{n} \delta_i$$

Now provided that δ_i is relatively small and does not depend on θ, we may ignore it and the likelihood is the same as in the discrete case.

As an example, suppose that Y is binomially distributed $B(n, p)$. The likelihood is:

$$L(p|y) = \binom{n}{y} p^y (1-p)^{n-y}$$

The *maximum likelihood* estimate (MLE) is the value of the parameter(s) that gives the largest probability to the observed data, or in other words, maximizes the likelihood function. The value at which the maximum occurs, $\hat{\theta}$, is the maximum likelihood estimate. In most cases, it is easier to maximize the log of likelihood function,

$l(\theta|y) = \log L(\theta|y)$. Since log is a monotone increasing function, the maximum occurs at the same $\hat{\theta}$.

In a few cases, we can find an exact analytical solution for $\hat{\theta}$. For the binomial, we have the log-likelihood:

$$l(p|y) = \log \binom{n}{y} + y \log p + (n-y) \log(1-p)$$

The *score function*, $u(\theta)$, is the derivative of the log-likelihood with respect to the parameters. For this example, we have:

$$u(p) = \frac{dl(p|y)}{dp} = \frac{y}{p} - \frac{n-y}{1-p}$$

We can find the maximum likelihood estimate $\hat{p}$ by solving $u(p) = 0$. We get $\hat{p} = y/n$. We should also verify that this stationary point actually represents a maximum.

Usually we want more than an estimate; some measure of the uncertainty in the estimate is valuable. This can be obtained via the Fisher information which is:

$$I(\theta) = \text{var}\, u(\theta) = -E \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T}$$

If there is more than one parameter, $I(\theta)$ will be a matrix. The information at $\hat{\theta}$ is the second derivative at the maximum. Large values indicate high curvature so that the maximum is well defined and even close alternatives will have much lower likelihood. This would indicate a high level of confidence in the estimate. One can show that the variance of $\hat{\theta}$ can be estimated by:

$$\hat{\text{var}}\,(\hat{\theta}) = I^{-1}(\hat{\theta})$$

under mild conditions. Sometimes it is difficult to compute the expected value of the matrix of second derivatives. As an alternative, the observed, rather than expected, value at $\hat{\theta}$ may be used instead. For the binomial example this gives:

$$\hat{\text{var}}\,\hat{p} = \hat{p}(1-\hat{p})/n$$

We illustrate these concepts by plotting the log-likelihood for two binomial datasets: one where $n = 25, y = 10$ and another where $n = 50, y = 20$. We construct the log-likelihood function:

```
> loglik <- function(p,y,n) lchoose(n,y) + y*log(p) + (n-y)*log(1-p)
```

For ease of presentation, we normalize by subtracting the log-likelihood at the maximum likelihood estimate:

```
> nloglik <- function(p,y,n) loglik(p,y,n) - loglik(y/n,y,n)
```

Now plot the two log-likelihoods, as seen in Figure A.1:

```
> pr <- seq(0.05,0.95,by=0.01)
> matplot(pr,cbind(nloglik(pr,10,25),nloglik(pr,20,50)),type="l",
  xlab="p",ylab="log-likelihood")
```

We see that the maximum occurs at $p = 0.4$ in each case at a value of zero because of the normalization. For the larger sample, we see greater curvature and hence more information.

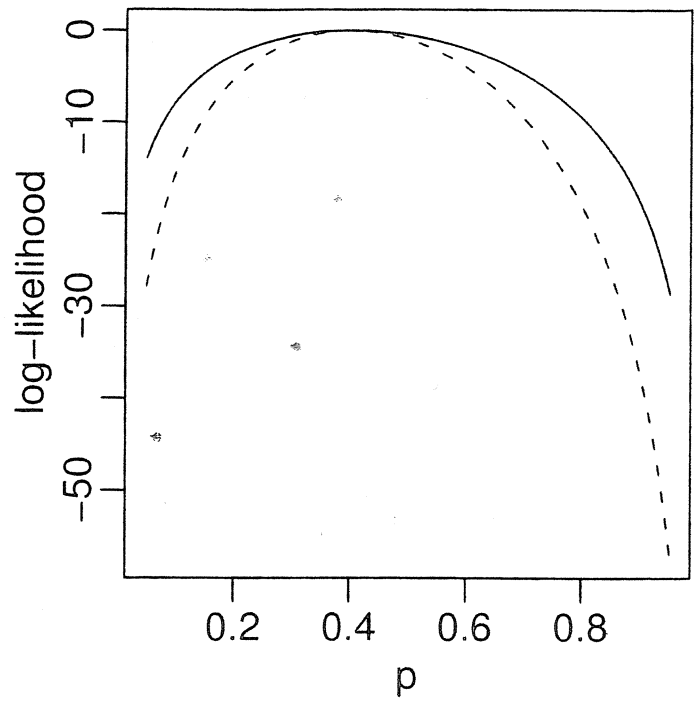


Figure A.1 *Normalized binomial log-likelihood for* $n = 25, y = 10$ *shown with a solid line and* $n = 50, y = 20$ *shown with a dotted line.*

Examples where likelihood can be maximized explicitly are confined to simple cases. Typically, numerical optimization is necessary. The *Newton–Raphson* method is the most well-known technique. Let θ_0 be an initial guess at θ, then we update using:

$$\theta_1 = \theta_0 - H^{-1}(\theta_0)u(\theta_0)$$

where H is the Hessian matrix of second derivatives:

$$H(\theta) = \frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}$$

We iterate this method, putting θ_1 in place of θ_0 and so on, until the procedure (hopefully) converges. This method works well provided the log-likelihood is smooth and convex around the maximum and that the initial value is reasonably close. In less well-behaved cases, several things can go wrong:

- The likelihood has multiple maxima. The maximum that Newton–Raphson finds will depend on the choice of initial estimate. If you are aware that multiple maxima may exist, it is advisable to try multiple starting values to search for the overall maximum. The number and choice of these starting values is problematic. Such problems are common in fitting neural networks, but rare for generalized linear models.
- The maximum likelihood may occur at the boundary of the parameter space. This means that perhaps $u(\hat{\theta}) \neq 0$, which will confuse the Newton–Raphson method.

Mixed effect models have several variance parameters. In some cases, these are maximized at zero, which causes difficulties in the numerical optimization.

- The likelihood has a large number of parameters and is quite flat in the neighborhood of the maximum. The Newton–Raphson method may take a long time to converge.

The Fisher scoring method replaces H with $-I$ and sometimes gives superior results. This method is used in fitting GLMs and is equivalent to iteratively reweighted least squares.

A minimization function that uses a Newton-type method is available in R. We demonstrate its use for likelihood maximization. Note that we need to minimize $-l$ because `nlm` minimizes, not maximizes:

```
> f <- function(x) -loglik(x,10,25)
> mm <- nlm(f,0.5,hessian=T)
```

We use a starting value of 0.5 and find the optimum at:

```
> mm$estimate
[1] 0.4
```

The inverse of the Hessian at the optimum is equal to the standard estimate of the variance:

```
> c(1/mm$hessian,0.4*(1-0.4)/25)
[1] 0.0096016 0.0096000
```

Of course, this calculation is not necessary for the binomial, but it is useful for cases where exact calculation is not possible.

A.2 Hypothesis Testing

Consider two nested models, a larger model Ω and a smaller model ω. Let $\hat{\theta}_\Omega$ be the maximum likelihood estimate under the larger model, while $\hat{\theta}_\omega$ be the corresponding value when θ is restricted to the range proscribed by the smaller model. The *likelihood ratio test statistic* is:

$$2\log(L(\hat{\theta}_\omega)/L(\hat{\theta}_\Omega)) = 2(l(\hat{\theta}_\omega) - l(\hat{\theta}_\Omega))$$

Under some regularity conditions, this statistic is asymptotically distributed χ^2 with degrees of freedom equal to the difference in the number of identifiable parameters in the two models. The approximation may not be good for small samples and may fail entirely if the regularity conditions are broken. For example, if the smaller model places some parameters on the boundary of the parameter space, the χ^2 may not be valid. This can happen in mixed effects models when testing whether a particular variance component is zero.

The *Wald test* may be used to test hypotheses of the form $H_0 : \theta = \theta_0$ and the test statistic takes the form:

$$(\hat{\theta} - \theta_0)^T I(\hat{\theta})(\hat{\theta} - \theta_0)$$

Under the null, the test statistic has approximately a χ^2 distribution with degrees of freedom equal to the number of parameters being tested. Quite often, one does not wish to test all the parameters and the Wald test is confined to a subset. In particular,

if we test only one parameter, $H_0 : \theta_i = \theta_{i0}$, the square root of the Wald test statistic is simply:

$$z = \frac{\hat{\theta}_i - \theta_{i0}}{se(\hat{\theta}_i)}$$

This is asymptotically normal. For a Gaussian linear model, these are the t-statistics and have an exact t-distribution, but for generalized linear and other models, the normal approximation must suffice.

The *score test* of the hypothesis $H_0 : \theta = \theta_0$ uses the statistic:

$$u(\theta_0)^T I^{-1}(\theta_0) u(\theta_0)$$

and is asymptotically χ^2 distributed with degrees of freedom equal to the number of parameters being tested.

There is no uniform advantage to any of these three tests. The score test does not require finding the maximum likelihood estimate, while the likelihood ratio test needs this computation to be done for both models. The Wald test needs just one maximum likelihood estimate. However, although the likelihood ratio test requires more information, the extra work is often rewarded. Although the likelihood ratio test is not always the best, it has been shown to be superior in a wide range of situations. Unless one has indications to the contrary or the computation is too burdensome, the likelihood ratio test is the recommended choice.

These test methods can be inverted to produce confidence intervals. To compute a $100(1-\alpha)\%$ confidence interval for θ, we calculate the range of hypothesized θ_0 such that $H_0 : \theta_0 = 0$ would not be rejected at the α level. The computation is simple for the single-parameter Wald test where the confidence interval for θ_i is:

$$\hat{\theta}_i \pm z^{1-\alpha/2} se(\hat{\theta}_i)$$

where z is the appropriate quantile of the normal distribution. The computation is trickier for the likelihood ratio test. If we are interested in a confidence interval for a single parameter θ_i, we will need to compute the log-likelihood for a range of θ_i with the other θ set to the maximizing values. This is known as the *profile likelihood* for θ_i. Once this is computed as $l_i(\theta_i|y)$, the confidence interval is:

$$\{\theta_i : 2(l(\hat{\theta}_i|y) - l(\theta_i|y)) < \chi_1^{1-\alpha}\}$$

As an example, this type of calculation is used in the computation of the confidence interval for the transformation parameter used in the Box–Cox method.

We can illustrate this by considering a binomial dataset where $n = 100$ and $y = 40$. We plot the normalized log-likelihood in Figure A.2 where we have drawn a horizontal line at half the distance of the 0.95 quantile of χ_1^2 below the maximum:

```
> pr <- seq(0.25,0.55,by=0.01)
> plot(pr,nloglik(pr,40,100),type="l",xlab="p",ylab="log-likelihood")
> abline(h=-qchisq(0.95,1)/2)
```

All p that have a likelihood above the line are contained within a 95% confidence interval for p. We can compute the range by solving for the points of intersection:

```
> g <- function(x) nloglik(x,40,100)+qchisq(0.95,1)/2
> uniroot(g,c(0.45,0.55))$root
```

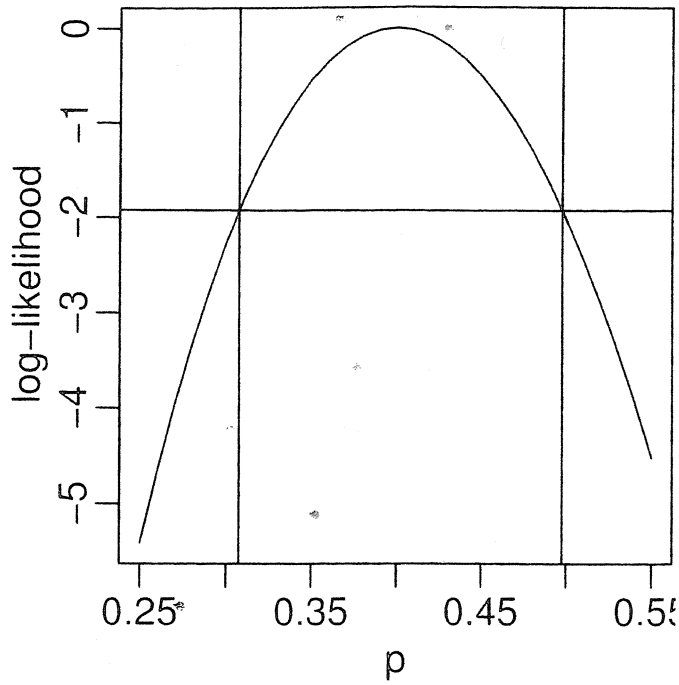


Figure A.2 *Likelihood ratio test-based confidence intervals for binomial p.*

```
[1] 0.49765
> uniroot(g,c(0.25,0.35))$root
[1] 0.30743
> abline(v=c(0.49765,0.30743))
```

The confidence interval is $(0.307, 0.498)$ as is indicated by the vertical lines on the plot. We can compute the Wald test-based interval as:

```
> se <- sqrt(0.4*(1-0.4)/100)
> cv <- qnorm(0.975)
> c(0.4-cv*se,0.4+cv*se)
[1] 0.30398 0.49602
```

which is very similar, but not identical, to the LRT-based intervals.

Suppose we are interested in the hypothesis, $H_0 : p = 0.5$. The LRT and p-value are:

```
> (lrstat <- 2*(loglik(0.4,40,100)-loglik(0.5,40,100)))
[1] 4.0271
> pchisq(lrstat,1,lower=F)
[1] 0.044775
```

So the null is barely rejected at the 5% level. The Wald test gives:

```
> (z <- (0.5-0.4)/se)
[1] 2.0412
> 2*pnorm(z,lower=F)
[1] 0.041227
```

Again, not very different from the LRT. The score test takes more effort to compute. The observed information is:

$$\frac{-d^2 l(p|y)}{dp^2} = \frac{y}{p^2} + \frac{n-y}{(1-p)^2}$$

We compute the score and information at $p = 0.5$ and then form the test and get the p-value:

```
> (sc <- 40/0.5-(100-40)/(1-0.5))
[1] -40
> (obsinf <- 40/0.5^2+(100-40)/(1-0.5)^2)
[1] 400
> (score.test <- 40*40/400)
[1] 4
> pchisq(4,1,lower=F)
[1] 0.0455
```

The outcome is again slightly different from the previous two tests. Asymptotically, the three tests agree. We have a moderate size sample in the example, so there is little difference. More substantial differences could be expected for smaller sample sizes.

APPENDIX A

Delta Method, Maximum Likelihood Theory, and Information

A.1 Delta Method

One simple technique commonly used for deriving variance estimators for functions of random variables is called the *delta method*. Suppose that a random variable X has mean μ and variance σ^2. Suppose further that we construct a new random variable Y by transforming X, $Y = g(X)$, for a continuously differentiable function $g(\cdot)$. Then, by Taylor's theorem, $g(X) = g(\mu) + g'(\mu)(X - \mu) + O((X - \mu)^2)$. Ignoring the higher order terms, we have that

$$E[Y] = E[g(X)] \approx g(\mu)$$

and

$$\begin{aligned} \text{Var}[Y] &= E[g(X) - g(\mu)]^2 \\ &\approx g'(\mu)^2 E[X - \mu]^2 \\ &= \sigma^2 g'(\mu)^2. \end{aligned}$$

This simple approximation to the variance of Y is often referred to as the *delta method*.

Example A.1. Suppose $Y = \log(X)$, then $g'(x) = 1/x$, so $\text{Var}(Y) \approx \sigma^2/\mu^2$. □

A.2 Asymptotic Theory for Likelihood Based Inference

Suppose that Y is a random variable with the p.d.f. $f_Y(y;\theta)$ where θ is an unknown parameter of dimension p, and $f_Y(y;\theta)$ is twice differentiable in a neighborhood of the true value of θ. The *likelihood function*, $L_{\mathbf{Y}}(\theta)$, is the density function for the observed values of Y viewed as a function of the parameter θ. If $Y_i, Y_2, \ldots, Y_n$ are an i.i.d. sample, then, letting $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$,

$$L_{\mathbf{Y}}(\theta) = \prod_{j=1}^{n} f_Y(y_j;\theta). \tag{A.1}$$

The *maximum likelihood estimate* (MLE) of θ is the value of θ, $\hat{\theta}$, that maximizes $L_{\mathbf{Y}}(\theta)$, or equivalently, the value of θ for which

$$U_{\mathbf{Y}}(\theta) = \frac{\partial \log L_{\mathbf{Y}}(\theta)}{\partial \theta} = 0.$$

For the cases we will consider, this equation will have a unique solution. The function $U_{\mathbf{Y}}(\theta)$ is known as the *efficient score* for θ. It can be shown that $E_\theta U_{\mathbf{Y}}(\theta) = 0$ where E_θ denotes expectation when Y has distribution $f_Y(y;\theta)$.

An important special case is the one in which $f_Y(y;\theta)$ is part of an *exponential family*. That is, $f_Y(y;\theta) = \exp(\eta(x,\theta)T(y) - A(\eta(x,\theta)))$ where x is a subject level covariate (possibly vector valued), $\eta(x,\theta)$ is a function of the covariate and the parameter θ, and $A(\eta)$ is a function of η which forces $f_Y(y;\theta)$ to integrate to one. In this case, $A'(\eta) = ET(y)$ and the score function has the form $U_{\mathbf{Y}}(\theta) = \sum_i B(x,\theta)(T(y) - A'(\eta))$, where $B(x,\theta) = \partial\eta(x,\theta)/\partial\theta$. Hence, the score function is a linear combination of the centered, transformed observations $T(y) - ET(x)$ and the solution to the score equations (A.1) satisfies $\sum_i B(x,\theta)T(y) = \sum_i B(x,\hat{\theta})ET(y)$.

Example A.2. Suppose we have the Gaussian linear model $y_i = x_i^T\beta + \epsilon_i$ where x_i and β are p-dimensional vectors and the ϵ_i are i.i.d. $N(0,\sigma^2)$. For simplicity we will assume that σ^2 is known. The log-likelihood is

$$\begin{aligned} \log L_{\mathbf{Y}}(\beta) &= -n\log(2\sigma^2)/2 - \sum_i (y_i - x_i^T\beta)^2/2\sigma^2 \\ &= -\frac{1}{2\sigma^2}\sum_i y_i x_i^T\beta - (x_i^T\beta)^2 + y_i^2 - n\log(2\sigma^2)/2. \quad \text{(A.2)} \end{aligned}$$

The score equations are

$$U_{\mathbf{Y}}(\beta) = -\sum x_i(y_i - x_i^T\beta)/\sigma^2 = 0.$$

Since, in general, x_i is a vector, the score function is vector valued, so the score equations are a linear system of p equations in p unknowns. If the model has an intercept term, β_1, so that $x_i = (1, x_{2i}, x_{3i}, \ldots, x_{pi})^T$, then we have that the sum of all the observations equals the sum of their expected values, $\sum_i y_i = \sum_i x_i^T\beta$. If we have two treatment groups and one of the covariates is the indicator of treatment, letting R_j be the set of indices of subjects in treatment group $j = 1, 2$, we have

$$\sum_{i\in R_j} y_i = \sum_{i\in R_j} x_i^T\beta.$$

In some settings it may be computationally simpler to fit the model by computing the expected values directly by forcing the required marginal totals to equal the expected totals, rather than by direct estimation of model parameters. □

The expected value of the derivative of $-U_{\mathbf{Y}}(\theta)$ with respect to θ is known

as the *Fisher information*, $\mathcal{I}(\theta)$. In the one dimensional case it can be shown that

$$\begin{aligned}\mathcal{I}(\theta) &= -E_\theta[\frac{\partial^2 \log L_{\mathbf{Y}}(\theta)}{\partial\theta^2}] \\ &= E_\theta[U_{\mathbf{Y}}(\theta)^2] \\ &= \mathrm{Var}_\theta[U_{\mathbf{Y}}(\theta)].\end{aligned}$$

In many situations, we cannot compute $\mathcal{I}(\theta)$ directly, but will need to use an estimate obtained from the data. It can be shown that under modest regularity conditions, $\hat{\theta} \overset{a}{\sim} N(\theta, \mathcal{I}^{-1}(\theta))$ where $\overset{a}{\sim}$ indicates the asymptotic distribution, e.g., asymptotically $\hat{\theta}$ has a normal distribution with mean θ and variance $\mathcal{I}^{-1}(\theta)$. Note that $\mathcal{I}(\theta)$ is of the expected curvature of the log-likelihood function at the true value of θ. Larger values of $\mathcal{I}(\theta)$ indicate that the likelihood function is more sharply peaked, and therefore, estimates of θ are more precise.

These results can be generalized to the case where θ is a p-dimensional vector with no difficulty. In this case the score, $U_{\mathbf{Y}}(\theta)$ is a p-dimensional vector of partial derivatives, and the Fisher information, $\mathcal{I}(\theta)$, is a matrix of partial derivatives. The asymptotic covariance matrix of $\hat{\theta}$ is the matrix inverse $\mathcal{I}^{-1}(\theta)$.

A.3 Hypothesis Testing

Three commonly used approaches for testing $H_0\colon \theta = \theta_0$ are as the *likelihood ratio test*, the *score test*, and the *Wald test*. Here we assume that θ has dimension p. In general there may be other unknown parameters, known as *nuisance parameters*, that are not of interest but need to be taken into account.

Example A.3. Suppose that we have two binomial samples, $y_i \sim \mathrm{Bin}(n_i, \pi_i)$, $i = 1, 2$, with y_i the number of successes, n_i the number of trials and π_i the success probability in each sample. If the null hypothesis is $H_0\colon \pi_1 = \pi_2$, we can let $\Delta = \pi_2 - \pi_2$ so that H_0 is equivalent to $H_0\colon \Delta = 0$. We may write the joint distribution of y_1 and y_2 in terms of Δ and either π_1 or π_2. We may say that Δ is the parameter of interest and π_1 is the nuisance parameter. □

In general, if we let ν be the (possibly vector valued) nuisance parameter and $L_{\mathbf{Y}}(\theta, \nu)$ be the likelihood function, then the score function has two components, $U_{\mathbf{Y}}(\theta, \nu) = (U_{\theta,\mathbf{Y}}(\theta, \nu), U_{\nu,\mathbf{Y}}(\theta, \nu))$, where $U_{\theta,\mathbf{Y}}(\theta, \nu) = \partial \log L_{\mathbf{Y}}(\theta, \nu)/\partial\theta$ and $U_{\nu,\mathbf{Y}}(\theta, \nu) = \partial \log L_{\mathbf{Y}}(\theta, \nu)/\partial\nu$. Under $H_1\colon \theta \neq \theta_0$ let $(\hat{\theta}, \hat{\nu})$ be the solution to $U_{\mathbf{Y}}(\theta, \nu) = 0$ while under H_0, let $\tilde{\nu}$ be the solution to $U_{\theta,\mathbf{Y}}(\theta_0, \nu) = 0$.

The Fisher information can be written in partitioned matrix form as

$$\mathcal{I}(\theta, \nu) = -\begin{bmatrix} U_{\theta,\theta,\mathbf{Y}}(\theta, \nu) & U_{\theta,\nu,\mathbf{Y}}(\theta, \nu) \\ U_{\nu,\theta,\mathbf{Y}}(\theta, \nu) & U_{\nu,\nu,\mathbf{Y}}(\nu, \nu) \end{bmatrix} = \begin{bmatrix} \mathcal{I}_{\theta,\theta} & \mathcal{I}_{\theta,\nu} \\ \mathcal{I}_{\nu,\theta} & \mathcal{I}_{\nu,\nu} \end{bmatrix}$$

where

$$U_{s,t,\mathbf{Y}}(\theta, \nu) = \frac{\partial \log L_{\mathbf{Y}}(\theta, \nu)}{\partial s \partial t}.$$

The covariance matrix of the vector $(\hat{\theta}, \hat{\nu})$ can be written

$$\begin{aligned}\mathrm{Var}(\hat{\theta}, \hat{\nu}) &= \mathcal{I}(\theta, \nu)^{-1} \\ &= \begin{bmatrix} (\mathcal{I}_{\theta,\theta} - \mathcal{I}_{\theta,\nu}\mathcal{I}_{\nu,\nu}^{-1}\mathcal{I}_{\nu,\theta})^{-1} & -(\mathcal{I}_{\theta,\theta} - \mathcal{I}_{\theta,\nu}\mathcal{I}_{\nu,\nu}^{-1}\mathcal{I}_{\theta,\nu})^{-1}\mathcal{I}_{\theta,\nu}\mathcal{I}_{\nu,\nu}^{-1} \\ \mathcal{I}_{\nu,\nu}^{-1}\mathcal{I}_{\nu,\theta}(\mathcal{I}_{\theta,\theta} - \mathcal{I}_{\theta,\nu}\mathcal{I}_{\nu,\nu}^{-1}\mathcal{I}_{\theta,\nu})^{-1} & (\mathcal{I}_{\nu,\nu} - \mathcal{I}_{\nu,\theta}\mathcal{I}_{\theta,\theta}^{-1}\mathcal{I}_{\nu,\theta}) \end{bmatrix}.\end{aligned}$$

The three tests can now be described.

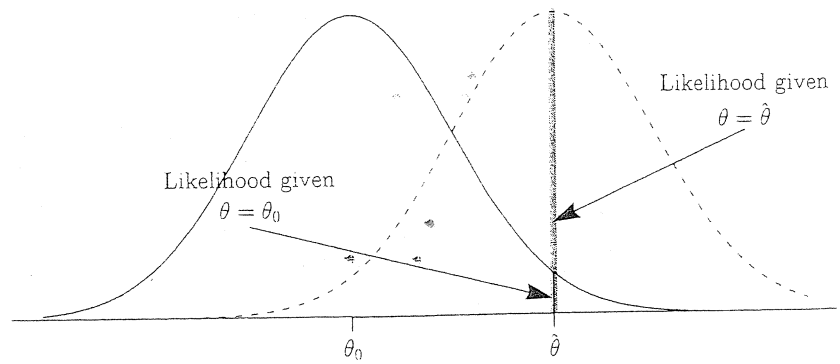


Figure A.1 *Graphical illustration of likelihood ratio test.*

1. Likelihood Ratio Test (LRT). The test statistic is twice the log-likelihood ratio:

$$2 \log \frac{L_{\boldsymbol{Y}}(\hat{\theta}, \hat{\nu})}{L_{\boldsymbol{Y}}(\theta_0, \tilde{\nu})} \overset{\text{a}}{\sim} \chi_p^2 \text{ under } H_0$$

   where χ_p^2 is the χ^2 distribution with p degrees of freedom. Figure A.1 illustrates the principle underlying the LRT when there are no nuisance parameters. Since $(\hat{\theta}, \hat{\nu})$ is the MLE, the likelihood evaluated at $(\hat{\theta}, \hat{\nu})$ is at least as large as the likelihood evaluated at $(\theta_0, \hat{\nu})$. If the likelihood ratio is large enough, there is strong evidence from the data that the observations do not arise from the distribution $f_Y(y; \theta_0)$.

2. Wald test. Since the upper left block of the matrix $\mathcal{I}(\theta)^{-1}$ is the asymptotic covariance matrix of $\hat{\theta}$, we have that

$$(\hat{\theta} - \theta_0)^T(\mathcal{I}_{\theta,\theta} - \mathcal{I}_{\theta,\nu}\mathcal{I}_{\nu,\nu}^{-1}\mathcal{I}_{\nu,\theta})(\hat{\theta} - \theta_0) \overset{\text{a}}{\sim} \chi_p^2 \text{ under } H_0,$$

   where the $\mathcal{I}_{s,t}$ are evaluated at $(\hat{\theta}, \tilde{\nu})$.

3. Score (Rao) test. We have under H_0, $U_{\boldsymbol{Y}}(\theta_0, \hat{\nu}) \overset{a}{\sim} N(0, \mathcal{I}(\theta_0, \nu))$. Hence the test statistic is

$$U_{\boldsymbol{Y}}(\theta_0, \tilde{\nu})\mathcal{I}(\theta_0, \tilde{\nu})^{-1}U_{\boldsymbol{Y}}(\theta_0, \tilde{\nu}) \overset{\text{a}}{\sim} \chi_p^2 \text{ under } H_0.$$

   In the exponential family case, the score test assesses the difference between the (possibly transformed) observed data, and its expectation. Since the

second component of $U_{\boldsymbol{Y}}(\theta_0, \tilde{\nu})$ is zero, this can also be written

$$U_{\theta,\boldsymbol{Y}}(\theta_0,\nu)^T(\mathcal{I}_{\theta,\theta} - \mathcal{I}_{\theta,\nu}\mathcal{I}_{\nu,\nu}^{-1}\mathcal{I}_{\nu,\theta})^{-1}U_{\theta,\boldsymbol{Y}}(\theta_0,\nu) \overset{a}{\sim} \chi_p^2 \text{ under } H_0,$$

where the $\mathcal{I}_{s,t}$ are evaluated at $(\theta_0, \tilde{\nu})$. Note that unlike the LRT and the Wald test, there is no need to compute the MLE $\hat{\theta}$. The score test is often used to assess the association between the outcome and large numbers of covariates without the need to fit many different models.

Under modest assumptions, these tests are asymptotically equivalent. For ordinary linear regression models with i.i.d. Gaussian errors and known variance, these tests are in fact identical. For other models, the Wald and score tests are quadratic approximations to the likelihood ratio test. The Wald test is based on a quadratic approximation to the LRT at the MLE, while the score test is an approximation at θ_0. These approximations are illustrated by figure A.2.

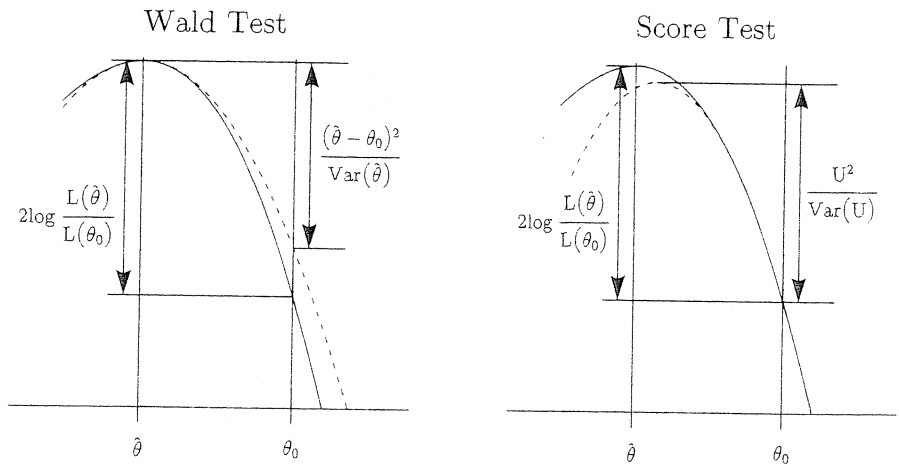


Figure A.2 *Illustrations showing Wald and score tests as quadratic approximations to the likelihood ratio test. The solid line represents the log-likelihood function and the dashed line represents the quadratic approximation.*

Example A.4. Returning to the binomial example, we have

$$L_{\boldsymbol{Y}}(\Delta,\pi_1) = K(y_1,n_1,y_2,n_2)\pi_1^{y_1}(1-\pi_1)^{n_1-y_1}(\pi_1+\Delta)^{y_2}(1-\pi_1-\Delta)^{n_2-y_2}$$

for a function $K(\cdot)$ which does not depend on the unknown parameters. The components of the score function are

$$\begin{aligned} U_{\Delta,\boldsymbol{Y}}(\Delta,\pi_1) &= \frac{y_2}{\pi_1+\Delta} - \frac{n_2-y_2}{1-\pi_1-\Delta} \\ U_{\pi_1,\boldsymbol{Y}}(\Delta,\pi_1) &= \frac{y_1}{\pi_1} - \frac{n_1-y_1}{1-\pi_1} + \frac{y_2}{\pi_1+\Delta} - \frac{n_2-y_2}{1-\pi_1-\Delta}. \end{aligned}$$

It is easy to show that, under $H_0\colon \Delta = 0$, the MLE for π_1 is the overall

mean $\tilde{\pi}_1 = (y_1 + y_2)/(n_1 + n_2)$, and, under H_1: $\Delta \neq 0$, $\hat{\pi}_1 = y_1/n_1$ and $\widehat{\Delta} = y_2/n_2 - y_1/n_1$.

The log-likelihood ratio is:

$$\begin{aligned}
\log\left(\frac{L_{\boldsymbol{Y}}(\widehat{\Delta}, \hat{\pi}_1)}{L_{\boldsymbol{Y}}(0, \hat{\pi}_1)}\right) &= y_1 \log \hat{\pi}_1 + (n_1 - y_1)\log(1 - \hat{\pi}_1) \\
&\quad + y_2 \log(\hat{\pi} + \widehat{\Delta}) + (n_2 - y_2)\log(1 - \hat{\pi}_1 - \Delta) \\
&\quad - y_1 \log \tilde{\pi}_1 - (n_1 - y_1)\log(1 - \tilde{\pi}_1) \\
&\quad - y_2 \log \tilde{\pi}_1 - (n_2 - y_2)\log(1 - \tilde{\pi}_1) \\
&= y_1 \log \frac{\hat{\pi}_1}{\tilde{\pi}_1} + (n_1 - y_1)\log(\frac{1 - \hat{\pi}_1}{1 - \tilde{\pi}_1}) \\
&\quad + y_2 \log(\frac{\hat{\pi}_1 + \widehat{\Delta}}{\tilde{\pi}_1}) + (n_2 - y_2)\log(\frac{1 - \hat{\pi}_1 - \Delta}{1 - \tilde{\pi}_1}) \\
&= \sum O \log \frac{O}{E} \qquad \text{(A.3)}
\end{aligned}$$

where the sum in equation (A.3) is over the four cells in the two-by-two table of survival status by treatment, O represents the observed value in a given cell and E represents its expected value under H_0. Asymptotically, the statistic $2\sum O \log \frac{O}{E}$ has a χ^2_1 distribution under H_0.

To perform the Wald and score tests, we need the Fisher information matrix. The elements of the Fisher information matrix are

$$\begin{aligned}
\mathcal{I}_{\Delta,\Delta} = \mathcal{I}_{\Delta,\pi_1} = \mathcal{I}_{\pi_1,\Delta} &= E[\frac{y_2}{(\pi_1 + \Delta)^2} + \frac{(n_2 - y_2)}{(1 - \pi_1 - \Delta)^2}] \\
&= \frac{n_2}{(\pi_1 + \Delta)} + \frac{n_2}{(1 - \pi_1 - \Delta)} \\
&= \frac{n_2}{(\pi_1 + \Delta)(1 - \pi_1 - \Delta)}
\end{aligned}$$

where the expectation is taken under H_1.

Similarly,

$$\mathcal{I}_{\pi_1,\pi_1} = \frac{n_1}{(\pi_1)(1 - \pi_1)} + \frac{n_2}{(\pi_1 + \Delta)(1 - \pi_1 - \Delta)}.$$

The variance of $\widehat{\Delta}$ is, therefore,

$$\begin{aligned}
\mathrm{Var}\widehat{\Delta} &= (\mathcal{I}_{\Delta,\Delta} - \mathcal{I}_{\Delta,\pi_1}\mathcal{I}^{-1}_{\pi_1,\pi_1}\mathcal{I}_{\pi_1,\Delta})^{-1} \\
&= \frac{\pi_1(1 - \pi_1)}{n_1} + \frac{(\pi_1 + \Delta)(1 - \pi_1 - \Delta)}{n_2}.
\end{aligned}$$

Note that this is identical to the variance obtained directly using the binomial variance, Var $y_i/n_i = \pi_i(1 - \pi_i)/n_i$. The variance estimate is obtained by replacing π_1 and Δ by their estimates.

Therefore, the Wald test statistic for H_0 has form

$$\frac{(y_2/n_2 - y_1/n_1)^2}{\hat{\pi}_1(1 - \hat{\pi}_1)/n_1 + (\hat{\pi}_1 + \widehat{\Delta})(1 - \hat{\pi}_1 - \widehat{\Delta})/n_2}.$$

The score test is similar. Under H_0, we have

$$\begin{aligned} U_{\Delta,\boldsymbol{Y}}(0,\tilde{\pi}_1) &= \frac{y_2}{\tilde{\pi}_1} - \frac{n_2 - y_2}{1-\tilde{\pi}_1} \\ &= \frac{1}{\tilde{\pi}_1(1-\tilde{\pi}_1)}(y_2 - n_2\tilde{\pi}_1) \end{aligned}$$

which is proportional to the observed number of events in group 2 minus the expected number under H_0. Under H_0, the variance of $U_{\Delta,\boldsymbol{Y}}(0,\tilde{\pi}_1)$ is $\mathcal{I}_{\Delta,\Delta} - \mathcal{I}_{\Delta,\pi_1}\mathcal{I}^{-1}_{\pi_1,\pi_1}\mathcal{I}_{\pi_1,\Delta} = (1/n_1 + 1/n_2)^{-1}/\pi_1(1-\pi_1)$, so the score test statistic is

$$\begin{aligned} &\left(\frac{y_2 - n_2\tilde{\pi}_1}{\tilde{\pi}_1(1-\tilde{\pi}_1)}\right)^2 \tilde{\pi}_1(1-\tilde{\pi}_1)(\frac{1}{n_1} + \frac{1}{n_2}) \\ &= \frac{(y_2 - n_2(y_1+y+2)/(n_1+n_2))^2(n_1+n_2)^3}{n_1 n_2 (y_1+y_2)(n_1+n_2-y_1-y_2)} \end{aligned}$$

which is the (uncorrected) Pearson χ^2 test statistic. It can be shown that the score statistic can also be written $\sum(O-E)^2/E$, where O and E are as in equation (A.3).

Note that the Wald test and the score test have similar form; the difference between them is that the Wald test uses the variance computed under H_1, while the score test uses the variance computed under H_0. □

A.4 Computing the MLE

In most situations there is no closed form solution for the MLE, $\hat{\theta}$, but finding the solution requires an iterative procedure. One commonly used method for finding $\hat{\theta}$ is the *Newton-Raphson* procedure.

We begin with an initial guess $\hat{\theta}_0$, from which we generate a sequence of estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots$ until convergence is achieved. For $i = 1, 2, 3, \ldots$, $\hat{\theta}_i$ is derived from $\hat{\theta}_{i-1}$ by first applying Taylor's theorem. We have that

$$U_{\boldsymbol{Y}}(\theta) = U_{\boldsymbol{Y}}(\hat{\theta}_{i-1}) + U_{\theta,\boldsymbol{Y}}(\hat{\theta}_{i-1})(\theta - \hat{\theta}_{i-1}) + \text{ higher order terms.}$$

Ignoring the higher order terms, we set the above equal to zero and solve for θ yielding

$$\hat{\theta}_i = \hat{\theta}_{i-1} - U_{\theta,\boldsymbol{Y}}(\hat{\theta}_{i-1})^{-1} U_{\boldsymbol{Y}}(\hat{\theta}_{i-1}). \tag{A.4}$$

Iteration stops once consecutive values of $\hat{\theta}_i$ differ by a sufficiently small amount.

Note that if θ is a vector of length p, the score function $U_{\boldsymbol{Y}}(\theta)$ will be a vector of partial derivatives of length p, and $U_{\theta,\boldsymbol{Y}}(\hat{\theta}_{i-1})$ will be the matrix of second order partial derivatives. Equation (A.4) will be an equation involving vectors and matrices.

Statistical Inference by Casella & Berger (2002)

a. $Y_n X_n \to aX$ *in distribution.*

b. $X_n + Y_n \to X + a$ *in distribution.*

The proof of Slutsky's Theorem is omitted, since it relies on a characterization of convergence in distribution that we have not discussed. A typical application is illustrated by the following example.

Example 5.5.18 (Normal approximation with estimated variance) Suppose that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to \text{n}(0,1),$$

but the value of σ is unknown. We have seen in Example 5.5.3 that, if $\lim_{n\to\infty} \text{Var } S_n^2 = 0$, then $S_n^2 \to \sigma^2$ in probability. By Exercise 5.32, $\sigma/S_n \to 1$ in probability. Hence, Slutsky's Theorem tells us

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sigma}{S_n} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to \text{n}(0,1).$$

5.5.4 The Delta Method

The previous section gives conditions under which a standardized random variable has a limit normal distribution. There are many times, however, when we are not specifically interested in the distribution of the random variable itself, but rather some function of the random variable.

Example 5.5.19 (Estimating the odds) Suppose we observe $X_1, X_2, \ldots, X_n$ independent Bernoulli(p) random variables. The typical parameter of interest is p, the success probability, but another popular parameter is $\frac{p}{1-p}$, the *odds*. For example, if the data represent the outcomes of a medical treatment with $p = 2/3$, then a person has odds 2 : 1 of getting better. Moreover, if there were another treatment with success probability r, biostatisticians often estimate the *odds ratio* $\frac{p}{1-p}/\frac{r}{1-r}$, giving the relative odds of one treatment over another.

As we would typically estimate the success probability p with the observed success probability $\hat{p} = \sum_i X_i/n$, we might consider using $\frac{\hat{p}}{1-\hat{p}}$ as an estimate of $\frac{p}{1-p}$. But what are the properties of this estimator? How might we estimate the variance of $\frac{\hat{p}}{1-\hat{p}}$? Moreover, how can we approximate its sampling distribution?

Intuition abandons us, and exact calculation is relatively hopeless, so we have to rely on an approximation. The Delta Method will allow us to obtain reasonable approximate answers to our questions.

One method of proceeding is based on using a Taylor series approximation, which allows us to approximate the mean and variance of a function of a random variable. We will also see that these rather straightforward approximations are good enough to obtain a CLT. We begin with a short review of Taylor series.

Definition 5.5.20 If a function $g(x)$ has derivatives of order r, that is, $g^{(r)}(x) = \frac{d^r}{dx^r}g(x)$ exists, then for any constant a, the *Taylor polynomial of order r about a* is

$$T_r(x) = \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i.$$

Taylor's major theorem, which we will not prove here, is that the *remainder* from the approximation, $g(x) - T_r(x)$, always tends to 0 faster than the highest-order explicit term.

Theorem 5.5.21 (Taylor) *If* $g^{(r)}(a) = \frac{d^r}{dx^r}g(x)\big|_{x=a}$ *exists, then*

$$\lim_{x\to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0.$$

In general, we will not be concerned with the explicit form of the remainder. Since we are interested in approximations, we are just going to ignore the remainder. There are, however, many explicit forms, one useful one being

$$g(x) - T_r(x) = \int_a^x \frac{g^{(r+1)}(t)}{r!}(x-t)^r dt.$$

For the statistical application of Taylor's Theorem, we are most concerned with the *first-order* Taylor series, that is, an approximation using just the first derivative (taking $r = 1$ in the above formulas). Furthermore, we will also find use for a multivariate Taylor series. Since the above detail is univariate, some of the following will have to be accepted on faith.

Let $T_1, \ldots, T_k$ be random variables with means $\theta_1, \ldots, \theta_k$, and define $\mathbf{T} = (T_1, \ldots, T_k)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Suppose there is a differentiable function $g(\mathbf{T})$ (an estimator of some parameter) for which we want an approximate estimate of variance. Define

$$g_i'(\boldsymbol{\theta}) = \frac{\partial}{\partial t_i} g(\mathbf{t})\big|_{t_1=\theta_1,\ldots,t_k=\theta_k}.$$

The first-order Taylor series expansion of g about $\boldsymbol{\theta}$ is

$$g(\mathbf{t}) = g(\boldsymbol{\theta}) + \sum_{i=1}^{k} g_i'(\boldsymbol{\theta})(t_i - \theta_i) + \text{Remainder}.$$

For our statistical approximation we forget about the remainder and write

$$g(\mathbf{t}) \approx g(\boldsymbol{\theta}) + \sum_{i=1}^{k} g_i'(\boldsymbol{\theta})(t_i - \theta_i). \tag{5.5.7}$$

Now, take expectations on both sides of (5.5.7) to get

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\theta}} g(\mathbf{T}) &\approx g(\boldsymbol{\theta}) + \sum_{i=1}^{k} g_i'(\boldsymbol{\theta})\mathrm{E}_{\boldsymbol{\theta}}(T_i - \theta_i) \qquad (5.5.8) \\ &= g(\boldsymbol{\theta}). \qquad (T_i \text{ has mean } \theta_i) \end{aligned}$$

We can now approximate the variance of $g(\mathbf{T})$ by

$$\begin{aligned}\mathrm{Var}_\theta\, g(\mathbf{T}) &\approx \mathrm{E}_\theta\left([g(\mathbf{T}) - g(\boldsymbol{\theta})]^2\right) && \text{(using (5.5.8))}\\ &\approx \mathrm{E}_\theta\left(\left(\sum\nolimits_{i=1}^{k} g_i'(\boldsymbol{\theta})(T_i - \theta_i)\right)^2\right) && \text{(using (5.5.7))}\\ &= \sum_{i=1}^{k}[g_i'(\boldsymbol{\theta})]^2\mathrm{Var}_\theta\, T_i + 2\sum_{i>j} g_i'(\boldsymbol{\theta})g_j'(\boldsymbol{\theta})\mathrm{Cov}_\theta(T_i, T_j), \qquad (5.5.9)\end{aligned}$$

where the last equality comes from expanding the square and using the definition of variance and covariance (similar to Exercise 4.44). Approximation (5.5.9) is very useful because it gives us a variance formula for a general function, using only simple variances and covariances. Here are two examples.

Example 5.5.22 (Continuation of Example 5.5.19) Recall that we are interested in the properties of $\frac{\hat{p}}{1-\hat{p}}$ as an estimate of $\frac{p}{1-p}$, where p is a binomial success probability. In our above notation, take $g(p) = \frac{p}{1-p}$ so $g'(p) = \frac{1}{(1-p)^2}$ and

$$\begin{aligned}\mathrm{Var}\left(\frac{\hat{p}}{1-\hat{p}}\right) &\approx [g'(p)]^2\,\mathrm{Var}(\hat{p})\\ &= \left[\frac{1}{(1-p)^2}\right]^2 \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3},\end{aligned}$$

giving us an approximation for the variance of our estimator.

Example 5.5.23 (Approximate mean and variance) Suppose X is a random variable with $\mathrm{E}_\mu X = \mu \neq 0$. If we want to estimate a function $g(\mu)$, a first-order approximation would give us

$$g(X) = g(\mu) + g'(\mu)(X - \mu).$$

If we use $g(X)$ as an estimator of $g(\mu)$, we can say that approximately

$$\begin{aligned}\mathrm{E}_\mu g(X) &\approx g(\mu),\\ \mathrm{Var}_\mu\, g(X) &\approx [g'(\mu)]^2\mathrm{Var}_\mu\, X.\end{aligned}$$

For a specific example, take $g(\mu) = 1/\mu$. We estimate $1/\mu$ with $1/X$, and we can say

$$\begin{aligned}\mathrm{E}_\mu\left(\frac{1}{X}\right) &\approx \frac{1}{\mu},\\ \mathrm{Var}_\mu\left(\frac{1}{X}\right) &\approx \left(\frac{1}{\mu}\right)^4 \mathrm{Var}_\mu X.\end{aligned}$$

Using these Taylor series approximations for the mean and variance, we get the following useful generalization of the Central Limit Theorem, known as the *Delta Method*.

Theorem 5.5.24 (Delta Method) *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathrm{n}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then*

$$\sqrt{n}[g(Y_n) - g(\theta)] \to \mathrm{n}(0, \sigma^2[g'(\theta)]^2) \text{ in distribution.} \tag{5.5.10}$$

Proof: The Taylor expansion of $g(Y_n)$ around $Y_n = \theta$ is

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \text{ Remainder}, \tag{5.5.11}$$

where the remainder $\to 0$ as $Y_n \to \theta$. Since $Y_n \to \theta$ in probability it follows that the remainder $\to 0$ in probability. By applying Slutsky's Theorem (Theorem 5.5.17) to

$$\sqrt{n}[g(Y_n) - g(\theta)] = g'(\theta)\sqrt{n}(Y_n - \theta),$$

the result now follows. See Exercise 5.43 for details. □

Example 5.5.25 (Continuation of Example 5.5.23) Suppose now that we have the mean of a random sample $\bar{X}$. For $\mu \neq 0$, we have

$$\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right) \to \mathrm{n}\left(0, \left(\frac{1}{\mu}\right)^4 \mathrm{Var}_\mu X_1\right)$$

in distribution.

If we do not know the variance of X_1, to use the above approximation requires an estimate, say S^2. Moreover, there is the question of what to do with the $1/\mu$ term, as we also do not know μ. We can estimate everything, which gives us the approximate variance

$$\widehat{\mathrm{Var}}\left(\frac{1}{\bar{X}}\right) \approx \left(\frac{1}{\bar{X}}\right)^4 S^2.$$

Furthermore, as both $\bar{X}$ and S^2 are consistent estimators, we can again apply Slutsky's Theorem to conclude that for $\mu \neq 0$,

$$\frac{\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right)}{\left(\frac{1}{\bar{X}}\right)^2 S} \to \mathrm{n}(0, 1)$$

in distribution.

Note how we wrote this latter quantity, dividing through by the estimated standard deviation and making the limiting distribution a standard normal. This is the only way that makes sense if we need to estimate any parameters in the limiting distribution. We also note that there is an alternative approach when there are parameters to estimate, and here we can actually avoid using an estimate for μ in the variance (see the score test in Section 10.3.2). ‖

There are two extensions of the basic Delta Method that we need to deal with to complete our treatment. The first concerns the possibility that $g'(\mu) = 0$. This could

happen, for example, if we were interested in estimating the variance of a binomial variance (see Exercise 5.44).

If $g'(\theta) = 0$, we take one more term in the Taylor expansion to get

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \frac{g''(\theta)}{2}(Y_n - \theta)^2 + \text{ Remainder.}$$

If we do some rearranging (setting $g' = 0$), we have

$$g(Y_n) - g(\theta) = \frac{g''(\theta)}{2}(Y_n - \theta)^2 + \text{ Remainder.} \tag{5.5.12}$$

Now recall that the square of a $\mathrm{n}(0,1)$ is a χ^2_1 (Example 2.1.9), which implies that

$$\frac{n(Y_n - \theta)^2}{\sigma^2} \to \chi^2_1$$

in distribution. Therefore, an argument similar to that used in Theorem 5.5.24 will establish the following theorem.

Theorem 5.5.26 (Second-order Delta Method) *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathrm{n}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not 0. Then*

$$n[g(Y_n) - g(\theta)] \to \sigma^2 \frac{g''(\theta)}{2}\chi^2_1 \text{ in distribution.} \tag{5.5.13}$$

Approximation techniques are very useful when more than one parameter makes up the function to be estimated and more than one random variable is used in the estimator. One common example is in growth studies, where a ratio of weight/height is a variable of interest. (Recall that in Chapter 3 we saw that a ratio of two *normal* random variables has a Cauchy distribution. The ratio problem, while being important to experimenters, is nasty in theory.)

This brings us to the second extension of the Delta Method, to the multivariate case. As we already have Taylor's Theorem for the multivariate case, this extension contains no surprises.

Example 5.5.27 (Moments of a ratio estimator) Suppose X and Y are random variables with nonzero means μ_X and μ_Y, respectively. The parametric function to be estimated is $g(\mu_X, \mu_Y) = \mu_X/\mu_Y$. It is straightforward to calculate

$$\frac{\partial}{\partial \mu_X} g(\mu_X, \mu_Y) = \frac{1}{\mu_Y}$$

and

$$\frac{\partial}{\partial \mu_Y} g(\mu_X, \mu_Y) = \frac{-\mu_X}{\mu_Y^2}.$$

The first-order Taylor approximations (5.5.8) and (5.5.9) give

$$\mathrm{E}\left(\frac{X}{Y}\right) \approx \frac{\mu_X}{\mu_Y}$$

and

$$\begin{aligned}\mathrm{Var}\left(\frac{X}{Y}\right) &\approx \frac{1}{\mu_Y^2}\mathrm{Var}\,X + \frac{\mu_X^2}{\mu_Y^4}\mathrm{Var}\,Y - 2\frac{\mu_X}{\mu_Y^3}\mathrm{Cov}(X,Y)\\ &= \left(\frac{\mu_X}{\mu_Y}\right)^2\left(\frac{\mathrm{Var}\,X}{\mu_X^2} + \frac{\mathrm{Var}\,Y}{\mu_Y^2} - 2\frac{\mathrm{Cov}(X,Y)}{\mu_X\mu_Y}\right).\end{aligned}$$

Thus, we have an approximation for the mean and variance of the ratio estimator, and the approximations use only the means, variances, and covariance of $\bar{X}$ and Y. Exact calculations would be quite hopeless, with closed-form expressions being unattainable. ‖

We next present a CLT to cover an estimator such as the ratio estimator. Note that we must deal with multiple random variables although the ultimate CLT is a univariate one. Suppose the vector-valued random variable $\mathbf{X} = (X_1, \ldots, X_p)$ has mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ and covariances $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$, and we observe an independent random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and calculate the means $\bar{X}_i = \sum_{k=1}^n X_{ik}$, $i = 1, \ldots, p$. For a function $g(\mathbf{x}) = g(x_1, \ldots, x_p)$ we can use the development after (5.5.7) to write

$$g(\bar{x}_1, \ldots, \bar{x}_p) = g(\mu_1, \ldots, \mu_p) + \sum_{k=1}^p g_k'(\mathbf{x})(\bar{x}_k - \mu_k),$$

and we then have the following theorem.

Theorem 5.5.28 (Multivariate Delta Method) *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample with $\mathrm{E}(X_{ij}) = \mu_i$ and $\mathrm{Cov}(X_{ik}, X_{jk}) = \sigma_{ij}$. For a given function g with continuous first partial derivatives and a specific value of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ for which $\tau^2 = \Sigma\Sigma\sigma_{ij}\frac{\partial g(\boldsymbol{\mu})}{\partial \mu_i}\cdot\frac{\partial g(\boldsymbol{\mu})}{\partial \mu_j} > 0$,*

$$\sqrt{n}[g(\bar{X}_1, \ldots, \bar{X}_s) - g(\mu_1, \ldots, \mu_p)] \to \mathrm{n}(0, \tau^2) \text{ in distribution}.$$

The proof necessitates dealing with the convergence of multivariate random variables, and we will not deal with such multivariate intricacies here, but will take Theorem 5.5.28 on faith. The interested reader can find more details in Lehmann and Casella (1998, Section 1.8).

5.6 Generating a Random Sample

Thus far we have been concerned with the many methods of describing the behavior of random variables—transformations, distributions, moment calculations, limit theorems. In practice, these random variables are used to describe and model real phenomena, and observations on these random variables are the data that we collect.

Biometrics 60, 807–811
September 2004

Confidence Interval Estimation of the Intraclass Correlation Coefficient for Binary Outcome Data

Guangyong Zou[1,2,*] and Allan Donner[2]

[1]Robarts Clinical Trials, Robarts Research Institute, London, Ontario N6A 5K8, Canada
[2]Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario N6A 5C1, Canada
[*]*email:* gzou@robarts.ca

Summary. We obtain closed-form asymptotic variance formulae for three point estimators of the intraclass correlation coefficient that may be applied to binary outcome data arising in clusters of variable size. Our results include as special cases those that have previously appeared in the literature (Fleiss and Cuzick, 1979, *Applied Psychological Measurement* **3,** 537–542; Bloch and Kraemer, 1989, *Biometrics* **45,** 269–287; Altaye, Donner, and Klar, 2001, *Biometrics* **57,** 584–588). Simulation results indicate that confidence intervals based on the estimator proposed by Fleiss and Cuzick provide coverage levels close to nominal over a wide range of parameter combinations. Two examples are presented.

Key words: Agreement; Cluster-randomization trials; Common correlation model; Delta method; Exchangeability; Kappa; Reliability; Variance.

1. Introduction

The intraclass correlation coefficient (ICC), a quantitative measure of the resemblance among observations within classes (clusters), is one of the most widely applied and versatile indices in applied research. For example, it is frequently used to quantify the familial aggregation of disease in genetic epidemiological studies (Cohen, 1980; Liang, Qaqish, and Zeger, 1992). In reliability studies, the ICC is an index measuring the level of interobserver agreement (Barto, 1966; Fleiss and Cuzick, 1979; Kraemer, Periyakoil, and Noda, 2002), while in health care delivery research, the ICC has been used to measure the efficiency of hospital staff (Gange et al., 1996). This parameter is also critical for estimating the required size of a cluster randomization trial (Cornfield, 1978).

Inference procedures for the ICC are well developed for the case of continuous data under the assumption of multivariate normality, as summarized by Donner (1986). In contrast, techniques for binary data have been less well developed, with the emphasis primarily on point estimation (see Ridout, Demétrio, and Firth [1999] for an excellent review). Aside from some computationally intensive procedures (e.g., Feng and Grizzle, 1992), two popular approaches for such inferences are based on generalized estimating equations (GEE) and the beta-binomial (BB) distribution (Lui, Cumberland, and Kuo, 1996). However, recent research has shown that the GEE approach, which was not designed for inference concerning the ICC, may result in confidence interval coverage which is substantially below nominal (Evans, Feng, and Peterson, 2001). A disadvantage of the BB model is that it is too restrictive to be relied on for inferences concerning the ICC when the class sizes are variable (Feng and Grizzle, 1992). This approach is further limited by the assumption that "the binary observations within a cluster are assumed to be a finite subset of an infinite exchangeable sequence of random variables" (Bowman, 2001).

We also note that Mak (1988) has derived a formula for the variance of an ICC estimator. However, the resulting expression is equivalent to that obtained using the GEE approach in that the expectations of the third and fourth moments are replaced by observed values (Shoukri and Martin, 1992).

In Section 2, we adopt the common correlation model (Madsen, 1993) to derive explicit variance formulae for three estimators of the ICC previously found to perform well in terms of mean square error and bias by Ridout et al. (1999). Confidence interval methods based on these formulae are described in Section 3. In Section 4, we evaluate the performance of these methods using Monte Carlo simulation. In Section 5, we provide examples using data from two previously published studies, one addressing familial aggregation of a respiratory condition, and the other focusing on interrater agreement. The article concludes with some final remarks in Section 6.

2. The Large Sample Variance of the ICC Estimators

2.1 *Assumptions and Point Estimators*

Consider a random sample of k clusters of size n_i ($i = 1, 2, \ldots, k$), where the responses X_{ij} ($j = 1, 2, \ldots, n_i$) in the ith class are dichotomous with success and failure coded as 1 and 0, respectively. The probability of success π is assumed to be identical for all individuals, i.e., $\Pr(X_{ij} = 1) = \pi$ for all i, j, an assumption usually referred to as "exchangeability." (Note that when this assumption is in doubt, it may be tested; see

Stefanescu and Turnbull [2003].) A second assumption is that the observations from different clusters are independent, while each pair of observations within the same cluster have a common correlation given by $\rho = \text{corr}(X_{ij}, X_{il})$, $j \neq l$.

Under these assumptions, sample estimates for π and ρ are readily available. An intuitive and simple estimator for π is given by

$$\hat{\pi} = \frac{\sum_{i=1}^{k} Y_i}{N},$$

where $Y_i = \sum_j X_{ij}$ is the total number of successes in class i, and $N = \sum n_i$ is the total number of observations in the study. At least 20 different estimators for ρ have been proposed in various areas of research, as reviewed by Ridout et al. (1999). The simulation results reported by these authors identified three of these as most accurate in terms of bias and mean square error. The first two estimators are obtained by applying formulae for continuous data directly to binary data while the third is commonly used in the context of reliability studies by Fleiss and Cuzick (1979).

The analysis of variance (ANOVA) estimator is given by

$$\hat{\rho}_{\text{A}} = \frac{\text{MSB} - \text{MSW}}{\text{MSB} + (n_{\text{A}} - 1)\text{MSW}}, \tag{1}$$

where

$$\text{MSB} = \frac{1}{k-1}\left\{\sum \frac{Y_i^2}{n_i} - \frac{\left(\sum Y_i\right)^2}{N}\right\},$$

$$\text{MSW} = \frac{1}{N-k}\left\{\sum Y_i - \sum \frac{Y_i^2}{n_i}\right\},$$

and

$$n_{\text{A}} = \frac{1}{k-1}\left(N - \frac{\sum n_i^2}{N}\right).$$

The Pearson pairwise estimator with constant weights is given by

$$\hat{\rho}_{\text{P}} = \frac{1}{\hat{\mu}(1-\hat{\mu})}\left[\frac{\sum Y_i(Y_i-1)}{\sum n_i(n_i-1)} - \hat{\mu}^2\right], \tag{2}$$

where

$$\hat{\mu} = \frac{\sum Y_i(n_i-1)}{\sum n_i(n_i-1)}.$$

Finally, the kappa-type estimator proposed by Fleiss and Cuzick (1979) is given by

$$\hat{\rho}_{\text{FC}} = 1 - \frac{\sum Y_i(n_i - Y_i)/n_i}{(N-k)\hat{\pi}(1-\hat{\pi})}. \tag{3}$$

Note that the Pearson estimator and the Fleiss–Cuzick estimator are identical in the case of constant cluster size (Ridout et al., 1999).

2.2 *Variance Derivation*

Under the common correlation assumption the exchangeable model can be written (Madsen, 1993) as

$$\Pr(Y=y) = \begin{cases} \rho(1-\pi) + (1-\rho)(1-\pi)^n, & y = 0, \\ \binom{n}{y}(1-\rho)\pi^y(1-\pi)^{n-y}, & 1 \le y \le n-1, \\ \rho\pi + (1-\rho)\pi^n, & y = n. \end{cases} \tag{4}$$

Note that for (4) to be a probability mass function, ρ must satisfy

$$\max\left[-\frac{(1-\pi)^n}{(1-\pi)-(1-\pi)^n}, -\frac{\pi^n}{\pi-\pi^n}\right] \le \rho \le 1. \tag{5}$$

A straightforward calculation yields the moment generating function of Y as

$$M_Y(t) = \rho\left[1 - \pi\left\{1 - \exp(tn)\right\}\right] + (1-\rho)\left[1 - \pi + \pi\exp(t)\right]^n,$$

which yields the lth moment as

$$EY^l = \left.\frac{d^l}{dt^l} M_Y(t)\right|_{t=0}.$$

Noting that $\hat{\rho}_A$ and $\hat{\rho}_{\text{FC}}$ are functions of $S_1 = \sum Y_i$ and $S_2 = \sum Y_i^2/n_i$, it can be shown that (S_1, S_2) is distributed asymptotically as bivariate normal with variance–covariance matrix

$$\begin{aligned}\boldsymbol{\Sigma} &= \begin{pmatrix} \text{var}(S_1) & \text{cov}(S_1,S_2) \\ \text{cov}(S_1,S_2) & \text{var}(S_2) \end{pmatrix} \\ &= \sum_{i=1}^{k} \begin{pmatrix} \text{var}(Y_i) & \text{cov}\left(Y_i, Y_i^2/n_i\right) \\ \text{cov}\left(Y_i, Y_i^2/n_i\right) & \text{var}\left(Y_i^2/n_i\right) \end{pmatrix}. \end{aligned} \tag{6}$$

Application of the delta method (Agresti, 2002, p. 579) yields the asymptotic distribution for $\hat{\rho}$ as

$$\sqrt{k}\left(\hat{\rho} - \rho\right) \to N\left(0, \Phi^{\text{T}}\boldsymbol{\Sigma}\Phi\right),$$

where

$$\Phi = \begin{pmatrix} \dfrac{\partial\hat{\rho}}{\partial S_1} \\ \dfrac{\partial\hat{\rho}}{\partial S_2} \end{pmatrix}$$

is evaluated at $ES_1 = N\pi$ and $ES_2 = k\pi(1-\pi) + \pi(1-\pi)(N-k)\rho + N\pi^2$ for $\hat{\rho}_{\text{A}}$ and $\hat{\rho}_{\text{FC}}$. After some straightforward but tedious calculation, a consistent variance estimator for $\hat{\rho}_{\text{A}}$ is obtained as

$$\begin{aligned}\text{var}(\hat{\rho}_{\text{A}}) = {} & \left[(k-1)n_{\text{A}}N(N-k)\right]^2 / \lambda^4 \\ & \times \left\{2k + \left(\frac{1}{\pi(1-\pi)} - 6\right)\sum n_i^{-1} \right. \\ & + \left[\left(\frac{1}{\pi(1-\pi)} - 6\right)\sum n_i^{-1} - 2N + 7k - 8k^2/N \right. \\ & \left.\left. - \frac{2k(1-k/N)}{\pi(1-\pi)} + \left(\frac{1}{\pi(1-\pi)} - 6\right)\sum n_i^2\right]\rho \right.\end{aligned}$$

$$+\left[\frac{N^2-k^2}{\pi(1-\pi)}-2N-k+4k^2/N+\left(7-8k/N-\frac{2(1-k/N)}{\pi(1-\pi)}\right)\sum n_i^2\right]\rho^2+\left(\frac{1}{\pi(1-\pi)}-4\right)\left(\frac{N-k}{N}\right)^2\left(\sum n_i^2-N\right)\rho^3\Bigg\}, \tag{7}$$

where

$$\lambda=(N-k)\left[N-1-n_{\mathrm{A}}(k-1)\right]\rho+N(k-1)(n_{\mathrm{A}}-1).$$

Similar steps yield the estimated variance of $\hat{\rho}_{\mathrm{FC}}$, given by

$$\begin{aligned}\mathrm{var}(\hat{\rho}_{\mathrm{FC}})=(1-\rho)&\times\left\{\left[\frac{1}{\pi(1-\pi)}-6\right]\frac{\sum n_i^{-1}}{(N-k)^2}+\left[2N+4k-\frac{k}{\pi(1-\pi)}\right]\frac{k}{N(N-k)^2}\right.\\&+\left[\frac{\sum n_i^2}{N^2\pi(1-\pi)}-\frac{(3N-2k)(N-2k)\sum n_i^2}{N^2(N-k)^2}-\frac{2N-k}{(N-k)^2}\right]\rho\\&\left.+\left[4-\frac{1}{\pi(1-\pi)}\right]\frac{\sum n_i^2-N}{N^2}\rho^2\right\},\end{aligned} \tag{8}$$

which reduces to the null variance derived by Fleiss and Cuzick (1979) when $\rho=0$.

Since $\hat{\rho}_{\mathrm{P}}$ is a function of S_1 and $S_3=\sum Y_i^2$, a similar derivation yields the variance of $\hat{\rho}_{\mathrm{P}}$ as

$$\begin{aligned}\mathrm{var}(\hat{\rho}_{\mathrm{P}})=&\frac{(1-\rho)}{\left[\sum n_i(n_i-1)\right]^2}\\&\times\left\{2\sum n_i(n_i-1)+\rho\left[\frac{1}{\pi(1-\pi)}-3\right]\right.\\&\times\sum n_i^2(n_i-1)^2+\rho^2\left[4-\frac{1}{\pi(1-\pi)}\right]\\&\left.\times\sum n_i(n_i-1)^3\right\}.\end{aligned} \tag{9}$$

In the case of constant cluster size $n_i=n$ for all i, expressions (8) and (9) simplify to

$$\mathrm{var}(\hat{\rho})=\frac{1-\rho}{k}\left[\frac{2}{n(n-1)}-\left\{3-\frac{1}{\pi(1-\pi)}\right\}\rho+\frac{n-1}{n}\left\{4-\frac{1}{\pi(1-\pi)}\right\}\rho^2\right], \tag{10}$$

which is identical to the variance formula derived by Bloch and Kraemer (1989) for $n=2$ and to that derived by Altaye, Donner, and Klar (2001) for $n=3$.

3. Confidence Interval Construction

An obvious approach to constructing a confidence interval for ρ is to obtain the large sample limits given by

$$\hat{\rho}\pm z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\hat{\rho})},$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution. However, several simulation studies have shown that this procedure does not perform well with extreme values of π and ρ or when k is small (e.g., Donner and Eliasziw, 1992; Altaye et al., 2001). Alternatively, one may attempt to use Fisher's z transformation to improve the normality of the sampling distribution of $\hat{\rho}$. Unfortunately it has been shown that this transformation is of limited use when applied to nonnormal data (Berry and Mielke, 2000).

We propose here to invert a modified Wald test, an approach which has been shown to provide accurate results when computing confidence limits for the difference between two intraclass kappa coefficients (Donner and Zou, 2002). This approach is also conceptually straightforward since the above variance formulae can be regarded as cubic functions of ρ. Therefore, we may write

$$(\hat{\rho}-\rho)^2=z_{\alpha/2}^2\widetilde{\mathrm{var}}(\hat{\rho}), \tag{11}$$

where $\widetilde{\mathrm{var}}(\hat{\rho})$ is the appropriate variance expression with $\hat{\pi}$ substituted for π. The confidence limits for ρ are then given by the two admissible roots of this equation, which may be found explicitly. For the ANOVA method, we replace $\hat{\rho}$ with $\hat{\rho}_{\mathrm{A}}$ and $\widetilde{\mathrm{var}}(\hat{\rho})$ with $\widetilde{\mathrm{var}}(\hat{\rho}_{\mathrm{A}})$ in equation (11). In a similar fashion we may also obtain confidence limits for ρ using either the Pearson estimator or the Fleiss–Cuzick estimator, which we refer to as the Pearson and FC methods, respectively.

4. Simulation Study

A simulation study was performed to evaluate the coverage levels of the three methods described above. For this purpose, we generated variable cluster sizes from a truncated negative binomial distribution with mean and variance given, respectively, by 3.12 and 4.52, which correspond to the U.S. sibship size distribution in 1950 (Brass, 1958; Donner and Koval, 1987). Other parameter values considered were $\rho=$ 0.1, 0.2, 0.3, 0.5, 0.8; $\pi=0.1$, 0.3, 0.5; and $k=25$, 50, 100, 200. For each of the 60 parameter combinations, 1000 data sets were generated from the common correlation model given by (4), followed by construction of a two-sided 95% confidence interval using the methods described above.

The performance of these methods was evaluated in terms of the observed percent coverage. Since negative values of ρ are usually considered implausible in most application areas, we truncated $\hat{\rho}$ at 0 for any calculated negative value. All programming was implemented using `SAS IML`.

Results in Table 1 show that the confidence interval based on the Fleiss–Cuzick estimator for ρ is slightly conservative when $\pi=0.1$ and $k=25$, but maintains 95% nominal coverage level very well provided $k\geq 50$. On the other hand, the performance of the confidence interval based on the ANOVA estimator is somewhat erratic, even though it is

Table 1
Empirical coverage percent based on 1000 *runs for three methods of constructing a* 95% *two-sided confidence interval for the ICC with binary data*

		π											
		$k = 25$			$k = 50$			$k = 100$			$k = 200$		
ρ	Method	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
0.1	ANOVA	60.0	97.1	97.2	94.4	97.0	96.7	98.7	98.2	96.2	99.9	98.2	93.1
	Pearson	97.4	96.7	97.3	96.6	96.6	97.5	96.4	97.1	96.5	98.1	96.5	95.6
	FC	97.6	96.8	98.3	96.5	96.1	97.6	96.3	97.6	97.0	96.8	96.9	93.6
0.2	ANOVA	50.9	97.4	97.1	88.9	97.9	95.1	97.1	97.4	92.6	99.3	97.1	91.2
	Pearson	98.6	97.1	98.1	97.4	97.2	97.9	97.6	96.7	96.5	98.1	95.9	95.1
	FC	97.7	97.0	97.8	96.9	97.3	96.3	96.4	96.0	93.9	97.1	95.0	94.4
0.3	ANOVA	40.4	97.1	94.4	81.1	97.8	93.5	98.7	97.9	91.7	99.7	96.9	90.7
	Pearson	99.2	98.7	97.9	98.8	98.1	96.5	98.6	97.2	95.6	97.8	96.3	95.8
	FC	98.1	97.5	96.8	97.5	95.0	96.3	97.4	96.0	94.5	95.8	94.5	94.3
0.5	ANOVA	20.2	89.8	92.0	50.1	95.8	89.3	84.2	96.7	90.2	99.8	97.1	86.5
	Pearson	100	97.8	96.5	99.5	95.8	95.7	99.1	95.9	95.5	97.0	95.7	95.1
	FC	99.5	95.5	94.6	97.2	94.5	94.7	94.5	95.1	94.6	94.9	95.6	94.3
0.8	ANOVA	2.0*	56.2	54.4	3.0	77.0	49.3	7.1	91.6	39.5	22.2	96.3	35.1
	Pearson	93.3*	92.8	93.8	91.1	95.1	94.6	91.0	95.5	93.8	93.4	94.9	94.6
	FC	93.7*	95.8	95.2	94.0	95.6	95.4	93.0	95.9	93.7	94.5	95.4	93.7

*Since in 3% of data sets generated in this case $\hat{\rho}$ was undefined ($\hat{\pi} = 0$), the coverage was calculated over the remaining 970 runs. Cluster sizes are generated according to a truncated negative binomial model with mean 3.12 and variance 4.52.

commonly recommended for point estimation. We also note that the method based on the Pearson estimator performs better than that based on the ANOVA estimator, but not as well as that based on the Fleiss–Cuzick estimator. In particular, confidence interval construction based on the Pearson estimator tends to yield conservative limits unless ρ is high ($\geq$0.8).

5. Examples

As a first example, we analyze the data presented in Example 3 of Liang et al. (1992), where the familial aggregation of chronic obstructive pulmonary disease (COPD) is used as a measure of how genetic and environmental factors may contribute to disease etiology. The data involve 203 siblings from 100 families with size ranging from 1 to 6, with the binary response of interest indicating whether a given sibling of a COPD patient has impaired pulmonary function. We obtain $\hat{\pi} = 0.296$, with the values of $\hat{\rho}_{\mathrm{A}}$, $\hat{\rho}_{\mathrm{P}}$, and $\hat{\rho}_{\mathrm{FC}}$ (standard errors) given by 0.186 (0.129), 0.260 (0.131), and 0.180 (0.107), respectively. The likelihood estimates (standard error) obtained using a BB and a saturated exchangeable model as proposed by Stefanescu and Turnbull (2003) are given by 0.270 (0.145) and 0.200 (0.089), respectively. The corresponding 95% confidence intervals using the modified Wald method are given by (0, 0.441), (0.068, 0.523), and (0.008, 0.402).

As a second example, we consider the data presented by Lipsitz, Laird, and Brennan (1994). In this data set, 26 patients with psychiatric disorders are classified by at least three and at most six psychiatrists into two categories (neurosis versus other disorder). The value of $\hat{\pi}$ is given by 0.401, and the values (standard error) of $\hat{\rho}_{\mathrm{A}}, \hat{\rho}_{P}$, and $\hat{\rho}_{\mathrm{FC}}$ by 0.422 (0.116), 0.408 (0.117), and 0.409 (0.114). Therefore the resulting 95% confidence intervals are given by (0.217, 0.633), (0.205, 0.625), and (0.210, 0.621), respectively. Note that the estimated standard error for $\hat{\rho}_{\mathrm{FC}}$ is in close agreement with simulation results presented by Lipsitz et al. (1994, Table 4).

6. Final Remarks

As many as 20 different point estimators of the ICC have been proposed across a diverse number of application areas. Ridout et al. (1999) provided a systematic review and evaluation of these estimators, focusing on bias and mean square error. We have extended their results by providing closed-form variance expressions for three of these point estimators, filling a long-standing gap in the literature (e.g., see Fleiss et al., 1979; Kraemer et al., 2002). The results of a simulation study lead us to recommend a modified Wald method as having excellent properties for confidence interval construction. Simulation results indicate that this method, used in conjunction with the Fleiss–Cuzick estimator, performs very well in sample sizes of 50 or more.

The results in this article apply to studies involving a reasonably large number of clusters, each of relatively small size. Thus they are most applicable to family studies and other application areas where these conditions apply. Our results also depend on the assumption of a common correlation among observations within the same cluster. Future research that extends this model to accommodate more complex correlation structures, such as arise in many genetic epidemiology studies, would clearly be worthwhile.

The implementation of the recommended confidence interval procedure using `SAS IML` and `S_plus` is available from the *Biometrics* website under the link "Data Sets/Computer Code."

Acknowledgements

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

Résumé

Nous obtenons une formule de la variance asymptotique pour trois estimateurs du coefficient de corrélation intra-classe qui peut être utilisé à des données binaires issues de groupes de taille variable. Nos résultats incluent les cas développés dans la littérature (Fleiss et Cuzick, 1979, *Applied Psychological Measurement* **3,** 537–554; Bloch et Kramer, 1989, *Biometrics* **45,** 269–287; Altaye, Donner et Klar, 2001, *Biometrics* **57,** 584–588). Les résultats des simulations montrent que les intervalles de confiance basés sur l'estimateur proposé par Fleiss et Cuzick (1979) conduisent des niveaux de couverture proche de la valeur nominale pour une large échelle de combinaisons des paramètres. Deux exemples sont présentés.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition. New York: Wiley.

Altaye, M., Donner, A., and Klar, N. (2001). Inference procedures for assessing interobserver agreement among multiple raters. *Biometrics* **57,** 584–588.

Barto, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* **19,** 3–11.

Berry, K. J. and Mielke, P. W. (2000). A Monte Carlo investigation of the Fisher z transformation for normal and nonnormal distributions. *Psychological Reports* **87,** 1101–1114.

Bloch, D. A. and Kraemer, H. C. (1989). 2 × 2 Kappa coefficients: Measures of agreement and association." *Biometrics* **45,** 269–287.

Bowman, D. (2001). Effects of correlation in modeling clustered binary data. *Journal of Statistical Computing and Simulation* **69,** 369–389.

Brass, W. (1958). Models of birth distribution in human populations. *Bulletin of the International Statistical Institute* **36**(2), 165–167.

Cohen, B. H. (1980). Chronic obstructive pulmonary disease: A challenge in genetic epidemiology. *American Journal of Epidemiology* **112,** 274–288.

Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology* **108,** 100–102.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review* **54,** 67–82.

Donner, A. and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: Confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine* **11,** 1511–1519.

Donner, A. and Koval, J. J. (1987). A procedure for generating group sizes from a one-way classification with a specified degree of imbalance. *Biometrical Journal* **29,** 181–187.

Donner, A. and Zou, G. (2002). Interval estimation for a difference between intraclass kappa statistics. *Biometrics* **58,** 209–215.

Evans, B. A., Feng, Z., and Peterson, A. V. (2001). A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance parameter estimation in longitudinal studies and group randomized trials. *Statistics in Medicine* **20,** 3353–3373.

Feng, Z. and Grizzle, J. E. (1992). Correlated binomial variates: Properties of estimator of intraclass correlation and its effect on sample size calculation. *Statistics in Medicine* **11,** 1607–1614.

Fleiss, J. L. and Cuzick, J. (1979). The reliability of dichotomous judgments: Unequal numbers of judges per subject. *Applied Psychological Measurement* **3,** 537–542.

Fleiss, J. L., Nee, J. C. M., and Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin* **86,** 974–977.

Gange, S. J., Munoz, A., Saez, M., and Alonso, J. (1996). Use of the beta-binomial distribution to model the effect of policy changes on appropriateness of hospital stays. *Applied Statistics* **45,** 371–382.

Kraemer, H. C., Periyakoil, V. S., and Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine* **21,** 2109–2129.

Liang, K. Y., Qaqish, B., and Zeger, S. L. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society B* **54,** 3–40.

Lipsitz, S. R., Laird, N. M., and Brennan, T. A. (1994). Simple moment estimates of the κ-coefficient and its variance. *Applied Statistics* **43,** 309–323.

Lui, K.-J., Cumberland, W. G., and Kuo, L. (1996). An interval estimate for the intraclass correlation in beta-binomial sampling. *Biometrics* **52,** 412–425.

Madsen, R. W. (1993). Generalized binomial distributions. *Communications in Statistics, Part A—Theory and Methods* **22,** 3065–3086.

Mak, T. K. (1988). Analysing intraclass correlation for dichotomous variables. *Applied Statistics* **37,** 344–352.

Ridout, M. S., Demétrio, C. G. B., and Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics* **55,** 137–148.

Shoukri, M. M. and Martin, S. W. (1992). Estimating the number of clusters for the analysis of correlated binary response variables from unbalanced data. *Statistics in Medicine* **11,** 751–760.

Stefanescu, C. and Turnbull, B. W. (2003). Likelihood inference for exchangeable binary data with varying cluster sizes. *Biometrics* **59,** 18–24.

Received August 2003. *Revised January* 2004.
Accepted February 2004.

ELSEVIER

Journal of Statistical Planning and Inference 97 (2001) 3–7

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Two score and 10 years of score tests

C. Radhakrishna Rao

Department of Statistics, 325 Joab L. Thomas Bldg., Pennsylvania State University, University Park, PA 16802, USA

S.J. Poti one of my co-workers at the Indian Statistical Institute in the early 1940s was working on a practical problem where he had to test a null hypothesis on a single parameter, H_0: $\theta = \theta_0$ (θ has the given value θ_0), when it was known a priori that the alternative $\theta > \theta_0$. He asked me whether an efficient test could be constructed for this purpose. I told him that Neyman and Pearson constructed what is called a locally unbiased most powerful (LUMP) test for two-sided alternatives and a similar method could be used to construct a locally most powerful one-sided (LMPOS) test. We need only to find a test (critical region of a given size) for which the power function has the maximum slope at θ_0 on one side. If $P(X,\theta)$ is the density at the value X in the sample space, then an application of Neyman–Pearson lemma gives the critical region as

$$w\colon P'(X,\theta_0) \geqslant \lambda P(X,\theta_0) \tag{1}$$

or $S(X,\theta_0) \geqslant \lambda$ where $S(X,\theta) = \mathrm{d}\log P(X,\theta)/\mathrm{d}\theta$ is Fisher's score function. The constant λ is determined such that the size of the critical region has a given value α. The result was published as a short note (Rao and Poti, 1946). We also suggested that a two-sided test such as $|S(X,\theta_0)| > \lambda$ would be a good competitor to LUMP test of Neyman and Pearson. Of course, a critical region of the form

$$w\colon \{S(X,\theta_0) > \lambda_1\} \bigcup \{S(X,\theta_0) < \lambda_2\} \tag{2}$$

would provide a better test.

In 1946, I was deputed to work on an anthropometric project in the Museum of Anthropology and Ethnology at the Cambridge University, UK. I took this opportunity of contacting R.A. Fisher who was then Balfour Professor of Genetics at the Cambridge University and register for a Ph.D. degree in statistics under his guidance. Fisher agreed but insisted that I should spend some time in his Genetics Laboratory where he was breeding mice to map the chromosomes (i.e., locating the positions of various genes on different chromosomes). I thought that this would be a good experience and agreed to work a few hours in his genetics laboratory every day in addition to my regular work at

E-mail address: ccrl@psu.edu (C. Radhakrishna Rao).

0378-3758/01/$ - see front matter
PII: S0378-3758(00)00342-6

the Museum. Fisher assigned to me the problem of mapping four genes on one of the six chromosomes of mice, by estimating the linkages or recombination probabilities of segments between genes. (Later I learnt that all his students were geneticists working for a Ph.D. degree in genetics, and I was the only one who wrote a thesis in statistics under his guidance.)

I started mating mice of different genotypes to collect the necessary data. At the same time, I started to develop the appropriate statistical methods for the analysis of experimental data. Each experiment provided data containing independent information on the same set of parameters (recombination probabilities in the various segments of the chromosomes). The problem was one of meta analysis, i.e., of combining the information from different experiments for the estimation of parameters. In such cases, it is often necessary to examine whether the parameters involved in different experiments are the same or not. My solution was as follows.

Let X_i be the observed sample, $l_i(X_i, \theta_i)$, the log likelihood, $\Phi_i(X_i, \theta_i) = \partial l_i(X_i, \theta_i)/\partial\theta_i$, the p-vector of scores for the vector parameter θ_i and $I(\theta_i)$, Fisher information matrix for the ith experiment, $i = 1, \ldots, k$. To test the hypothesis

$$H_0: \theta_1 = \cdots = \theta_k. \tag{3}$$

I suggested the statistic

$$\sum_{i=1}^{k} [\Phi_i(X_i, \hat{\theta})]'[I_i(\hat{\theta})]^{-1}[\Phi_i(X_i, \hat{\theta})], \tag{4}$$

where $\hat{\theta}$ is the maximum likelihood estimate, under the assumption $\theta_1 = \cdots = \theta_k$, which is obtained as a solution of the equation

$$\sum_{i=1}^{k} \Phi_i(X_i, \theta) = 0. \tag{5}$$

Statistic (4) is shown to be distributed as chi-square on $p(k-1)$ degrees of freedom in large samples. $\hat{\theta}$ would be the appropriate estimate of the common parameter obtained from the whole data if the value of (4) is not large. Otherwise, we have to examine the nature of the differences between parameters in different experiments.

I wrote a paper setting out the detailed steps for analyzing the data involving the segregation of several factors in matings of different genotypes. In an appendix to the paper, I discussed the general theory of asymptotic tests based on scores from which statistic (4) was derived. I showed the paper to Fisher. He thumbed through it and said, "The paper is probably good but I would like to see numerical results". He also suggested that I should write a separate paper on the theoretical results.

After I acquired the data, I did the necessary computations and give him the revised paper. He was pleased and asked me to submit the paper to the *Journal of Heredity*, a new journal for reporting research work in genetics. The paper was accepted and published in 1950 (see Rao, 1950). The theoretical portion of the paper appeared as a separate publication in the *Proceedings of the Cambridge Philosophical Society* (Rao, 1948). In this paper, I considered the general problem of testing simple and composite hypotheses concerning a vector parameter θ based on the vector score function $\Phi(X, \theta)$

and information matrix $I(\theta)$, where X is the observed sample. To test the simple hypothesis H_0: $\theta = \theta_0$, the test statistic was defined as

$$\max_a \frac{[a'\Phi(X,\theta_0)]^2}{a'I(\theta_0)a} = [\Phi(X,\theta_0)]'[I(\theta_0)]^{-1}[\Phi(X,\theta_0)], \tag{6}$$

where a is a p-vector of constants. It was shown that statistic (6) is asymptotically distributed as χ^2 on p degrees of freedom. Note that when $p = 1$, a takes only two values ± 1, and test reduces to $|\Phi(X,\theta_0)| > \lambda$ as discussed in Rao and Poti (1946). For testing a composite hypothesis, the suggested statistic was

$$[\Phi(X,\hat{\theta})]'[I(\hat{\theta})]^{-1}[\Phi(X,\hat{\theta})], \tag{7}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ under the restriction of the composite hypothesis. Statistic (4) used in the analysis of genetic data is a special case of test (7).

I knew at the time I developed the score test, now referred to as RS (Rao's score), there were two alternative tests, the likelihood ratio L of Neyman and Pearson (1928) whose asymptotic distribution was derived by Wilks (1938) and Wald's (1943) W. These three test statistics, L, W and RS which are referred as the "holy trinity" are asymptotically equivalent under null as well as Pitman alternatives (see Serfling, 1980, p. 156). The RS seemed to be attractive as it involved less computations and is invariant for transformation of parameters. Further several well-known large sample tests like Pearson Chisquare could be identified as score tests. I mentioned as a conjecture in the first edition of my book *Linear Statistical Inference and its Applications* (Rao, 1965, Section 6.2) that RS is likely to be locally more powerful than L and W. Peers (1971) showed that this conjecture is not true as stated. On the basis of his result which I did not examine carefully, I omitted my conjecture in the second edition of my book (Rao, 1973).

The score test went unnoticed for a number of years after it was introduced. It was resurrected by the Indian School of Statisticians in the eighties, who studied its optimum properties and showed that my conjecture holds with some modifications, and in some respects it has more attractive features than L and W, contrary to what Peers thought. For details of these developments, reference may be made to papers by Ghosh (1991), Li (1999) and Mukherjee (1993).

I may also mention that I used the score function in deriving sequential tests of null hypotheses (Rao, 1951) which has applications in quality control (Box and Ramirez, 1992) and clinical trials (Bradley, 1953).

It may be noted that a few years after my 1948 paper appeared, the same score test was introduced by Silvey (1954) under the name Lagrangian multiplier test, and this terminology appears in econometrics literature (see Byron (1968) who was probably the first to introduce the RS statistic in econometrics). Neyman (1954, 1959) introduced what is called $C(\alpha)$ test which is similar to the score test when there is one main parameter and several nuisance parameters. Neyman used $\sqrt{n}$ consistent estimates of the nuisance parameters under the null hypothesis for the main parameter, instead of ml estimates used by me (Rao, 1948) in the computation the score statistic. Hall and Mathiason (1990) extended Neyman's $C(\alpha)$ test to more than one main parameter using

an argument similar to the one used in the derivation of the score test. The resulting test was named by them as Neyman–Rao test.

I am glad to see that score type tests are beginning to be used in different areas of study and research. They have found applications in econometrics (see Amemiya, 1985, pp. 142–146, 206–207, 469; Bera and Bilias, 1999; Bera, 1999; Bera and Mukherjee, 1999 for general surveys) and in survival analysis (Klein and Moeschberger, 1997, pp. 407–410, 429–433). Score tests are discussed in books on asymptotic statistical inference (Lehman, 1999, pp. 451, 529, 532, 534, 539, 570; Serfling, 1980, pp. 155–160). Score type tests based on estimating equations have also been introduced (Boos, 1992; Sen, 1982). Some other modulli on score tests are given in Rao (1961).

I would like to mention that I would not have thought of score tests if I had not worked on a particular practical problem in genetics which Fisher asked me to investigate. I realized the importance of the *score function* in combining information from different independent sources and tried to develop a theory of inference *based primarily on scores*. Score tests have some attractive features and I am glad to see that my 1948 paper has been included in *Breakthroughs in Statistics*: 1890–1990, Vol. 3, edited by Kotz and Johnson. A foreword to this paper by P.K. Sen contains a discussion of the merits and demerits of the score test. A special invited paper session on 50 years of Rao's score test was organized at the Joint Meetings of ASA, IMS, SSC and BS held at Anaheim, August 10–14, 1997. I am glad to see the current interest in score tests and their modifications and generalizations, and I wish to thank A.K. Bera and R. Mukherjee for putting together papers on the current state of the art for publication in a special issue of the *Journal of Statistical Planning and Inference*. This would encourage further research into several unresolved problems in the application of score tests. Of course, no known method in statistics is universally applicable and it is important to know in which situations, a particular method is efficient.

I may also mention that mathematical genetics I learnt by attending Fisher's lectures and talking to his students when I was in Cambridge had another benefit. It enabled me to guide students for the Ph.D. degree in mathematical genetics on my return to the Indian Statistical Institute. At least two of my students who received the Ph.D. degree are leading figures in statistical genetics today.

References

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press, Cambridge, MA.

Bera, A.K., 1999. Hypotheses testing in 20th century with special reference to testing with misspecified models. J. Statist. Plann. Inference, in press.

Bera, A.K., Bilias, Y., 1999. Rao's score, Neyman's $C(\alpha)$ and Silvey's LM tests: an essay on historical developments and some new results. J. Statist. Plann. Inference 97 (2001) 9.

Bera, A.K., Mukherjee, R., 1999. 50 years of Rao's score test (a special issue). J. Statist. Planning Inference 97 (2001) 3.

Boos, D.D., 1992. On generalized score tests. Am. Stat. 46, 327–333.

Box, G., Ramirez, J., 1992. Cumulative score charts. Qual. Reliab. Eng. Int. 8, 17–27.

Bradley, R.A., 1953. Some statistical methods in taste testing and quality evaluation. Biometrics 9, 22–38.

Byron, R.P., 1968. Methods for estimating demand equations using prior information: a series of experiments with Australian data. Austr. Econ. Papers 7, 227–248.

Ghosh, J.K., 1991. Higher order asymptotics for the likelihood ratio, Rao's and Wald's tests. Statist. Probab. Lett. 12, 505–509.

Hall, W.J., Mathiason, D.J., 1990. On large-sample estimation and testing in parametric models. Internat. Statist. Rev. 58, 77–97.

Klein, J.P., Moeschberger, H.L., 1997. Survival Analysis. Springer, New York.

Lehman, E.L., 1999. Elements of Large Sample Theory. Springer, New York.

Li, B., 1999. Sensitivity of Rao's test, the Wald test and likelihood ratio test to nuisance parameters. J. Statist. Plann. Inference 97 (2001) 57.

Mukherjee, R., 1993. Rao's score test: recent asymptotic results. In: Maddala, G.S., Rao, C.R., Vinod, H. (Eds.), Handbook of Statistics, Vol. 11. North-Holland, Amsterdam, pp. 363–379.

Neyman, J., 1954. Sur une famille de tests asymptotiques des hypotheses statistiques compasses. Trab. Estadist. 5, 161–168.

Neyman, J., 1959. Optimal asymptotic tests of composite statistical hypotheses. In: Grenander, U. (Ed.), Probability and Statistics (Harald Cramér Volume). Wiley, New York, pp. 212–234.

Neyman, J., Pearson, E.S., 1928. On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 20, 175–240.

Peers, H.W., 1971. Likelihood ratio and associated test criteria. Biometrika 58, 577–587.

Rao, C.R., 1948. Large sample tests of statistical hypotheses concerning several parameters with application to problems of estimation. Proc. Cambridge Philos. Soc. 44, 50–57.

Rao, C.R., 1950. Methods of scoring linkage data giving the simultaneous segregation of three factors. Heredity 4, 37–59.

Rao, C.R., 1951. Sequential tests of null hypotheses. Sankhyā 10, 361–370.

Rao, C.R., 1961. A study of large sample criteria through properties of efficient estimates. Sankhyā A 23, 25–40.

Rao, C.R., 1965. Linear Statistical Inference and its Applications, 1st Edition, 1965; 2nd Edition, 1973. Wiley, New York.

Rao, C.R., 1973. Linear Statistical Inference and its Applications, 2nd Edition. Wiley, New York.

Rao, C.R., Poti, S.J., 1946. On locally most powerful tests when alternatives are one-sided. Sankhyā 7, 439–440.

Sen, P.K., 1982. On *M* tests in linear models. Biometrika 69, 93–101.

Serfling, R.J., 1980. Approximation Theorems of Mathematical Statistics. Wiley, New York.

Silvey, S.D., 1954. The Lagrangian multiplier test. Ann. Math. Statist. 30, 389–407.

Wald, A., 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Amer. Math. Soc. 54, 426–482.

Wilks, S.S., 1938. The large sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Statist. 9, 60–62.

Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability
Author(s): J. Neyman
Source: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, Vol. 236, No. 767 (Aug. 30, 1937), pp. 333-380
Published by: The Royal Society
Stable URL: http://www.jstor.org/stable/91337
Accessed: 09/07/2009 10:30

X—Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability

By J. NEYMAN

Reader in Statistics, University College, London

(Communicated by H. JEFFREYS, F.R.S.—Received 20 November, 1936—Read 17 June, 1937)

CONTENTS

I—INTRODUCTORY

(*a*) *General Remarks, Notation, and Definitions*

We shall distinguish two aspects of the problems of estimation : (i) the practical and (ii) the theoretical. The practical aspect may be described as follows :

(i*a*) The statistician is concerned with a population, π, which for some reason or other cannot be studied exhaustively. It is only possible to draw a sample from this population which may be studied in detail and used to form an opinion as to the values of certain constants describing the properties of the population π. For example, it may be desired to calculate approximately the mean of a certain character possessed by the individuals forming the population π, etc.

(i*b*) Alternatively, the statistician may be concerned with certain experiments which, if repeated under apparently identical conditions, yield varying results. Such experiments are called random experiments, (*see* p. 338). To explain or describe

the machinery of the varying results of random experiments certain mathematical schemes are drawn up involving one or more parameters, the values of which are not fixed. The statistician is then asked to provide numerical values of these parameters, to be calculated from experimental data and upon the assumption that the mathematical model of the experiments is correct.

The situation may be exemplified by the counts of α-particles ejected by some radioactive matter. The physicists have here elaborated a mathematical model of the phenomenon involving only one numerical parameter, namely, the average duration of life of an atom, and the statistician is asked to use the results of the available observations to deduce the numerical value of this parameter.

In both cases described, the problem with which the statistician is faced is the problem of estimation. This problem consists in determining what arithmetical operations should be performed on the observational data in order to obtain a result, to be called an estimate, which presumably does not differ very much from the true value of the numerical character, either of the population π, as in (i*a*), or of the random experiments, as in (i*b*).

(ii) The theoretical aspect of the problem of statistical estimation consists primarily in putting in a precise form certain vague notions mentioned in (i). It will be noticed that the problem in its practical aspect is not a mathematical problem, and before attempting any mathematical solution we must substitute for (i) another problem, (ii), having a mathematical sense and such that, for practical purposes, it may be considered as equivalent to (i).

The vague non-mathematical elements in (i) are connected with the sentence describing the meaning of the word estimate. What exactly is meant by the statement that the value of the estimate " presumably " should not differ very much from the estimated number ? The only established branch of mathematics dealing with conceptions bearing on the word " presumably " is the calculus of probability. It therefore seems natural to base the precise definition of an estimate on conceptions of probability. It is easy to see that the connexion of the problem considered with the theory of probability does not stop here and that the conditions of the problem themselves are, mathematically, clear only if they are expressed in the same terms of probability.

In (i*a*) we speak of a statistician drawing a sample from the population studied. It is known that if the sample is systematically selected and not drawn " at random " the conclusions concerning the population π formed on its basis are, as a rule, false and at the present state of our knowledge impossible to justify. On the other hand, we know that justifiable and frequently correct conclusions are possible only when the process of drawing the sample is " random ", though the randomness may be at times more or less restricted. I have put the word " random " in inverted commas because it is very difficult to define what is meant by it in practice.* We try to achieve randomness by more or less complicated devices, using roulette,

* This point requires a longer discussion, which I hope to be able to publish in a separate paper.

dice, etc. Theoretically, however, the situation is clear : when we speak of a random sample we mean that it is drawn so that (1) the probability of each individual of the population being included in the sample is the same, and (2) separate drawings are mutually independent, except in the case of dependence resulting from the population being finite, when the individual drawn is not returned to the population before the next drawing.

Leaving apart on one side the practical difficulty of achieving randomness and the meaning of this word when applied to actual experiments, I want to call attention to the fact that the conditions of the problem in (i*a*) may be mathematically described as follows.

Denote X, Y, . . ., Z, the characters of the individuals of the population π, in which we are interested and by $x, y, \ldots . z$, respectively the values of these characters corresponding to some particular individual. For example, if the population π consists of certain plants, X may mean the weight of the roots, Y the colour of the flowers, Z the weight of the seeds, etc. The method of random sampling adopted, together with the properties of the population π, some of which may be known and others doubtful, determine the probability,* say P {E}, of the occurrence of any possible system, E, of values of X, Y, ..., Z in the individuals which may be drawn to form the sample. Denote by θ_1 the numerical character of the population π which it is desired to estimate : this, for example, may be the mean value of X, the regression coefficient of Z on X, the mean square contingency of Z and Y, etc. The probability P {E} will depend on the value of θ_1 and in most cases on the values of certain other parameters, say, θ_2, θ_3, . . ., etc.

We see, therefore, that the problem with which the theoretical statistician is faced is as follows :

Sampling randomly from the population π, it is possible to obtain samples, say

$$E_1, E_2, \ldots, E_N, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

where each sample is described by means of values of the characters X, Y, . . ., Z, corresponding to each of the individuals forming the sample. The probability of any sample E_i, say $P\{E_i | \theta_1, \theta_2, \ldots \theta_l\}$, depends on a certain number, l, of parameters θ_i, the values of which are unknown, describing the properties of the population π. The problem consists in determining how to use the sample which may be actually obtained in order to estimate θ_1.

We see that the conditions of the problem in (i*a*) are expressed in terms of probability. The same holds good with regard to the problem in (i*b*), which shows that the distinction between (i*a*) and (i*b*) is only superficial. In fact, random experiments differ from those which are not considered as random only by the circumstance that the mathematical model devised for their description involves

* If the population π is finite. Otherwise the method of sampling and the properties of the population will determine the elementary probability law of X, Y, ..., Z considered as random variables. For the definitions of random variables and their probability laws, *see* p. 340 below.

probabilities. Each model of this kind determines the range of the possible results of random experiments and also the probability of each such result, depending upon one parameter or more, the numerical value of which is unknown.

We come to the conclusion that both the conditions of the problem of estimation and the satisfactory solution sought, if expressed accurately, are expressed in terms of probability. Before we proceed to the final formulation of the problem, it will be useful to give a short review of the forms of some solutions which have been advanced in the past. For this we shall need to define the terms probability, random variable, and probability law. These definitions are needed not because I introduce some new conceptions to be described by the above terms, but because the theory which is developed below refers only to some particular systems of the theory of probability which at the present time exist,* and it is essential to avoid misunderstandings.

I find it convenient to use the word probability in the following connexion : " the probability of an object, A, having a property B ". This may include as particular cases : " probability of a result, A, of a certain experiment having the property B of actually occurring " (= probability of the result A — for short) and " the probability of a proposition, A, of having the property, B, of being true ". All these ways of speaking could be shortened in obvious ways.

I want to emphasize at the outset that the definition of probability as given below is applicable only to certain objects A and to certain of their properties B—not to all possible. In order to specify the conditions of the applicability of the definition of the probability, denote by (A) the set of all objects which we agree to denote by A. (A) will be called the fundamental probability set. Further, let (B) denote the set of these objects A which possess some distinctive property B and finally, ((B)), a certain class of subsets (B') , (B''), . . ., corresponding to some class of properties B', B'', etc.

It will be assumed†

(1) that the class ((B)) includes (A), so that (A) ε ((B)) and

* It may be useful to point out that although we are frequently witnessing controversies in which authors try to defend one or another system of the theory of probability as the only legitimate, I am of the opinion that several such theories may be and actually are legitimate, in spite of their occasionally contradicting one another. Each of these theories is based on some system of postulates, and so long as the postulates forming one particular system do not contradict each other and are sufficient to construct a theory, this is as legitimate as any other. In this, of course, the theories of probability are not in any sort exceptional.

Both Euclidean and non-Euclidean geometries are equally legitimate, but, *e.g.*, the statement " the sum of angles in a linear triangle is always equal to π " is correct only in the former. In theoretical work the choice between several equally legitimate theories is a matter of personal taste only. In problems of application the personal taste is again the decisive moment, but it is certainly influenced by considerations of the relative convenience and the empirical facts.

† The problem of the definition of measure in relation to the theory of probability has been recently discussed by Łomnicki and Ulam (1934), who quote an extensive literature. A systematic outline of the theory of probability based on that of measure is given by Kolmogoroff (1933). *See also* Borel (1925–26) ; Lévy (1925) ; Fréchet (1937).

(2) that for the class ((B)) it was possible to define a single-valued function, m (B), of (B) which will be called the measure of (B). The sets (B) belonging to the class ((B)) will be called measurable. The assumed properties of the measure are as follows :

(*a*) Whatever (B) of the class ((B)), m (B) ≥ 0.

(*b*) If (B) is empty (does not contain any single element), then it is measurable and m (B) $= 0$.

(*c*) The measure of (A) is greater than zero.

(*d*) If (B_1), (B_2) . . . (B_n) . . . is any at most denumerable set of measurable subsets, then their sum, (ΣB_i), is also measurable. If the subsets of neither pair (B_i) and (B_j) (where $i \neq j$) have common elements, then m (ΣB_i) $= \sum_{i=1}^{\infty} m$ (B_i).

(*e*) If (B) is measurable, then the set ($\overline{B}$) of objects A non-possessing the property B is also measurable and consequently, owing to (*d*), m (B) + m ($\overline{B}$) = m (A).

Under the above conditions the probability, P{B|A}, of an object A having the property B will be defined as the ratio $P\{B|A\} = \frac{m\ (B)}{m\ (A)}$. The probability P{$B_1$|A}, or P{$B_1$} for short, may be called the absolute probability of the property B_1. Denote by B_1 B_2 the property of A consisting in the presence of both B_1 and B_2. It is easy to show that if (B_1) and (B_2) are both measurable then (B_1 B_2) will be measurable also. If m (B_2) > 0, then the ratio, say P{$B_1|B_2$} = m (B_1B_2)/m (B_2), will be called the relative probability of B_1 given B_2. This definition of the relative probability applies when the measure m (B_2) as defined for the fundamental probability set (A) is not equal to zero. If, however, m (B_2) $= 0$ and we are able to define some other measure, say m', applicable to (B_2) and to a class of its subsets including (B_1 B_2) such that m' (B_2) > 0, then the relative probability of B_1 given B_2 will be defined as $P\{B_1|B_2\} = m'\ (B_1B_2)/m'\ (B_2)$. Whatever may be the case, we shall have $P\{B_1B_2\} = P\{B_1\}P\{B_2|B_1\} = P\{B_2\}P\{B_1|B_2\}$.

It is easy to see that if the fundamental probability set is finite, then the number of elements in any of its subsets will satisfy the definition of the measure. On the other hand, if (A) is the set of points filling up a certain region in n-dimensioned space, then the measure of Lebesgue will satisfy the definition used here. These two definitions will be used wherever applicable.

If (A) is infinite but the objects A are not actually points (*e.g.*, if they are certain lines, etc.), the above definition of probability may be again applied, provided it is possible to establish a one to one correspondence between the objects A and other objects A', forming a class of sets where the measure has already been defined. If (B) is any subset of (A) and (B') the corresponding subset of (A'), then the measure of (B) may be defined as being equal to that of (B'). It is known that a similar

definition of measure of subsets of (A) could be done in more than one way. Such is, for instance, the historical example considered by BERTRAND, POINCARÉ, and BOREL when the objects A are the chords in a circle C of radius r and the property B consists of their length, l, exceeding some specified value, B. It may be useful to consider two of the possible ways of treating this problem.

1. Denote by x the angle between the radius perpendicular to any given chord A and any fixed direction. Further, let y be the distance of the chord A from the centre of the circle C. If A′ denotes a point on the plane with coordinates x and y, then there will be a one to one correspondence between the chords A of length $0 \leqq l < 2r$ and the points of a rectangle, say (A′), defined by the inequalities $0 < x \leqq 2\pi$ and $0 < y \leqq r$. The measure of the set of chords A with lengths exceeding B could be defined as being equal to the area of that part of (A′) where $0 < y \leqq \sqrt{r^2 - (\frac{1}{2}\mathrm{B})^2}$. It follows that the probability in which we are interested is $\mathrm{P}\{l > \mathrm{B}\} = (r^2 - (\frac{1}{2}\mathrm{B})^2)^{\frac{1}{2}} r^{-1}$.

2. Denote by x and y the angles between a fixed direction and the radii connecting the ends of any given chord A. If A″ denotes a point on a plane with coordinates x and y, then there will be a one to one correspondence between the chords of the system (A) and the points A″ within the parallelogram (A″) determined by the inequalities $0 < x \leqq 2\pi$, $x \leqq y \leqq x + \pi$. The measure of the set of chords A with their lengths exceeding B may be defined as being equal to the area of that part of (A″) where $2r \sin \frac{1}{2} y > \mathrm{B}$.

Starting with this definition $\mathrm{P}\{l > \mathrm{B}\} = 1 - 2 \arcsin (\mathrm{B}/2r)\, \pi^{-1}$.

It is seen that the two solutions differ, and it may be asked which of them is correct. The answer is that both are correct but they correspond to different conditions of the problem. In fact, the question " what is the probability of a chord having its length larger than B " does not specify the problem entirely. This is only determined when we define the measure appropriate to the set (A) and its subsets to be considered. We may describe this also differently, using the terms " random experiments " and " their results ". We may say that to have the problem of probability determined, it is necessary to define the method by which the randomness of an experiment is attained. Describing the conditions of the problem concerning the length of a chord leading to the solution (1), we could say that when selecting at random a chord A, we first pick up at random the direction of a radius, all of them being equally probable, and then, equally at random, we select the distance between the centre of the circle and the chord, all values between zero and r being equally probable. It is easy to see what would be the description in the same language of the random experiment leading to the solution (2). We shall use sometimes this way of speaking, but it is necessary to remember that behind such words, as *e.g.*, " picking up at random a direction, all of them being equally probable ", there is a definition of the measure appropriate to the fundamental probability set and its subsets. I want to emphasize that in this paper the sentence like the one taken in inverted commas is no more than a way of describing the fundamental probability set and the appropriate measure. The conception of " equally probable " is not in any way involved in the

definition of probability adopted here, and it is a pure convention that the statement

" In picking up at random a chord, we first select a direction of radius, all of them being equally probable and then we choose a distance between the centre of the circle and the chord, all values of the distance between zero and r being equally probable."	means no more and no less than	" For the purpose of calculating the probabilities concerning chords in a circle, the measure of any set (A_1) of chords is defined as that of the set (A'_1) of points with coordinates x and y such that for any chord A_1 in (A_1), x is the direction of the radius perpendicular to A_1 and y the distance of A_1 from the centre of the circle. (A_1) is measurable only if (A'_1) is so."

However free we are in mathematical work in using wordings we find convenient, as long as they are clearly defined, our choice must be justified in one way or another. The justification of the way of speaking about the definition of the measure within the fundamental probability set in terms of imaginary random experiments lies in the empirical fact, which BORTKIEWICZ insisted on calling the law of big numbers. This is that, given a purely mathematical definition of a probability set including the appropriate measure, we are able to construct a real experiment, possible to carry out in any laboratory, with a certain range of possible results and such that if it is repeated many times, the relative frequencies of these results and their different combinations in small series approach closely the values of probabilities as calculated from the definition of the fundamental probability set. Examples of such real random experiments are provided by the experience of roulette (BORTKIEWICZ, 1917), by the experiment with throwing a needle* so as to obtain an analogy to the problem of Buffon, and by various sampling experiments based on TIPPETT's Tables of random numbers (1927).

These examples show that the random experiments corresponding in the sense described to mathematically defined probability sets are possible. However, frequently they are technically difficult, *e.g.*, if we take any coin and toss it many times, it is very probable that the frequency of heads will not approach $\frac{1}{2}$. To get this result, we must select what could be called a well-balanced coin and we have to work out an appropriate method of tossing. Whenever we succeed in arranging the technique of a random experiment, say E, such that the relative frequencies of its different results in long series sufficiently approach, in our opinion, the probabilities calculated from a fundamental probability set (A), we shall say that the set (A) adequately represents the method of carrying out the experiment E. The theory developed below is entirely independent of whether the law of big numbers holds

* This is mentioned by BOREL (1910). I could not find the name of the performer of the experiment.

good or not. But the applications of the theory do depend on the assumption that it is valid. The questions dealt with in the present section are of fundamental importance. However, they do not constitute the main part of the paper and therefore are necessarily treated very briefly. The readers who may find the present exposition not sufficiently clear may be referred for further details to the work of KOLMOGOROFF (1933, *see* particularly p. 3 *et seq.*). I should state also that an excellent theoretical explanation of the experimental phenomena mentioned, connected with the previous work of POINCARÉ and SMOLUCHOWSKI, has been recently advanced by HOPF (1934).

We shall now draw a few obvious but important conclusions from the definition of the probability adopted.

(1) If the fundamental probability set consists of only one element, any probability calculated with regard to this set must have the value either zero or unity.

(2) If all the elements of the fundamental probability set (A) possess a certain property B_0, then the absolute probability of B_0 and also its relative probability given any other property B_1, must be equal to unity, so that $P\{B_0\} = P\{B_0|B_1\} = 1$. On the other hand, if it is known only that $P\{B_0\} = 1$, then it does not necessarily follow that $P\{B_0|B_1\}$ must be equal to unity.

We may now proceed to the definition of a random variable. We shall say that x is a random variable if it is a single-valued measurable function (not a constant) defined within the fundamental probability set (A), with the exception perhaps of a set of elements of measure zero. We shall consider only cases where x is a real numerical function. If x is a random variable, then its value corresponding to any given element A of (A) may be considered as a property of A, and whatever the real numbers $a < b$, the definition of (A) will allow the calculation of the probability, say $P\{a \leqq x < b\}$ of x having a value such that $a \leqq x < b$.

We notice also that as x is not constant in (A), it is possible to find at least one pair of elements, A_1 and A_2, of (A) such that the corresponding values of x, say $x_1 < x_2$, are different. If we denote by B the property distinguishing both A_1 and A_2 from all other elements of (A) and if $a < b$ are two numbers such that $a < x_1 < b < x_2$ then $P\{a \leqq x < b|B\} = \frac{1}{2}$. It follows that if x is a random variable in the sense of the above definition, then there must exist such properties B and such numbers $a < b$ that $0 < P\{a \leqq x < b|B\} < 1$.

It is obvious that the above two properties are equivalent to the definition of a random variable. In fact, if x has the properties (*a*) that whatever $a < b$ the definition of the fundamental probability set (A) allows the calculation of the probability $P\{a \leqq x < b\}$, and (*b*) that there are such properties B and such numbers $a < b$ that $0 < P\{a \leqq x < b|B\} < 1$, then x is a random variable in the sense of the above definition.

The probability $P\{a \leqq x < b\}$ considered as a function of a and b will be called the integral probability law of x.

A random variable is here contrasted with a constant, say θ, which will be defined as a magnitude, the numerical values of which corresponding to all elements of the

set (A) are all equal. If θ is a constant, then whatever $a < b$, and B, the probability $P\{a \leq \theta < b|B\}$ may have only values unity or zero according to whether θ falls in between a and b or not.

Keeping in mind the above definitions of the variables, in discussing them we shall often use the way of speaking in terms of random experiments. In the sense of the convention adopted above, we may say that x is a random variable when its values are determined by the results of a random experiment.

It is important to keep a clear distinction between random variables and unknown constants. The 1000th decimal, X_{1000}, in the expansion of $\pi = 3{\cdot}14159\ldots$ is a quantity unknown to me, but it is not a random variable since its value is perfectly fixed, whatever fundamental probability set we choose to consider. We could say alternatively that the value that X_{1000} may have does not depend upon the result of any random experiment.

Similarly, if we consider a specified population, say the population π_{1935} of persons residing permanently in London during the year 1935, any character of this population will be a constant. In the sense of the terms used here, there will be no practical meaning in a question concerning the probability that the average income, say I_{1935}, of the individuals of this population is, say, between £100 and £300. As the fundamental probability set consists of only one element, namely I_{1935}, the value of this probability is zero or unity, and to ascertain it we must discover for certain whether $£100 \leq I_{1935} < £300$ or not. This is, of course, possible, though it might involve great practical difficulty, just as it is possible to find the actual value of X_{1000}, the 1000th figure in the expansion of π. Any calculations showing that $P\{100 \leq I_{1935} < 300\}$ has a greater value than zero and smaller than unity must be either wrong or based on some theory of probability other than the one considered here.

This is the point where the difference between the theory of probability adopted here and that developed by JEFFREYS (1931) comes to the front. According to the latter, previous economic knowledge may be used to calculate the probability $P\{a \leq I_{1935} < b|B\}$ where $a < b$ are any numbers and the result of the calculations may be represented by any fraction, not necessarily by zero or unity.

The above examples must be contrasted with the following ones. We may consider the probability of a figure X, in the expansion of π falling between any specified limits $a < b$ and find it to be equal, *e.g.*, to $\frac{1}{2}$. This is possible when we first define a random method of drawing a figure out of those which serve to represent the expansion of π. If this is done, then X is a random variable and the X_{1000} previously defined will be one of its particular values.

Similarly, it is probably not impossible to construct a more or less adequate mathematical model of fluctuations in the size of income, in which the yearly average income, I, of the permanent population of London will be a random variable. The I_{1935} previously defined will be a particular value of I, observed at the end of the year 1935.

It is true that any constant, ξ, might be formally considered as a random variable

with the integral probability law $P\{a \leq \xi < b\}$ having only values unity or zero according to whether ξ falls between a and b or not. If we pass from letters to figures this will lead to formulae like $P\{1 \leq 2 < 3\} = 1$, or $P\{3 \leq 2 < 4\} = 0$.

Of course, in practice we shall have generally some unknown number ξ instead of 2 in the above formulae and accordingly we shall not know what are the actual values of the probabilities. In order to find these values, it would be necessary to obtain some precise information as to the value of ξ. It follows that the consideration of such probabilities is entirely useless, since whatever we are able to express in using them, we can say more simply by means of equations or inequalities.

For this reason, when defining a random variable, we require its probability law to be able to have values other than zero and unity. The other case may be set aside as trivial.

In the following development we shall have to consider at once several random variables

$$X_1, X_2, \ldots X_n. \quad \ldots\ldots\ldots\ldots \quad (2)$$

It will be convenient to denote by E any combination of their particular values and to interpret each such combination E as a point (the sample point) in an n-dimensional space W (the sample space), having its coordinates equal to the particular values of the variables (2). If w denotes any region in W, then the probability, say $P\{E \varepsilon w\}$, of the sample point falling within w considered as a function of w will be described as the integral probability law of the variables (2).

We shall consider only cases where there exists a non-negative function $p(E) \equiv p(x_1, \ldots, x_n)$ determined and integrable in the whole sample space W, such that for any region w

$$P\{E \varepsilon w\} = \int \ldots \int_w p(E)\, dx_1 \ldots dx_n. \quad \ldots\ldots\ldots \quad (3)$$

The function $p(E)$ will be called the elementary probability law of the X's in (2). It is easy to show that when $p(x_1, \ldots, x_{n-1}, x_n)$ is known, then $p(x_1, \ldots, x_{n-1})$ may be calculated by integrating $p(x_1, \ldots x_n)$ with regard to x_n from $-\infty$ to $+\infty$.

When dealing with several probability laws calculated in relation to probability sets depending on some variables, say $y_1 \ldots y_m$, in order to avoid misunderstandings, we shall use the notation $p(x_1 \ldots x_n | y_1 \ldots y_m)$ or $p(E | y_1 \ldots y_m)$. If $p(x_1, \ldots x_k, x_{k+1}, \ldots x_n)$ is the probability law of $x_1, x_2 \ldots x_k, x_{k+1}, \ldots x_n$ and if for a given system of the x's, $p(x_{k+1}, \ldots x_n) > 0$ then, for that system, the relative probability law of $x_1, x_2, \ldots x_k$ given $x_{k+1}, \ldots x_n$, denoted by $p(x_1, \ldots x_k | x_{k+1}, \ldots x_n)$, will be defined by the relation $p(x_1, x_2, \ldots x_k, \ldots x_n) = p(x_{k+1}, \ldots x_n)\, p(x_1, \ldots x_k | x_{k+1}, \ldots x_n)$.

With the above definitions and notation we may now formulate the problem of estimation as follows :

Let

$$X_1, X_2, \ldots X_n \quad \ldots\ldots\ldots\ldots \quad (4)$$

be a system of n random variables, the particular values of which may be given by observation. The elementary probability law of these variables

$$p(x_1 \ldots x_n|\theta_1, \theta_2, \ldots \theta_l) \quad \ldots\ldots\ldots\ldots \quad (5)$$

depends in a known manner upon l parameters $\theta_1 \ldots \theta_l$, the values of which are not known. It is required to estimate one (or more) of these parameters, using the observed values of the variables (4), say

$$x'_1, x'_2, \ldots x'_n \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

(b) Review of the Solutions of the Problem of Estimation Advanced Hereto*

The first attempt to solve the problem of estimation is connected with the theorem of Bayes and is applicable when the parameters $\theta_1, \theta_2, \ldots \theta_l$ in (5) are themselves random variables. The theorem of Bayes leads to the formula

$$p(\theta_1, \theta_2, \ldots \theta_l|x'_1, x'_2, \ldots x'_n) = \frac{p(\theta_1, \theta_2, \ldots \theta_l)\, p(x'_1, x'_2, \ldots x'_n|\theta_1 \ldots \theta_l)}{\int \ldots \int p(\theta_1, \theta_2, \ldots \theta_l)\, p(x'_1, x'_2, \ldots x'_n|\theta_1, \ldots \theta_l)\, d\theta_1 \ldots d\theta_l}, \quad (7).$$

representing the probability law of $\theta_1, \theta_2, \ldots \theta_l$, calculated under the assumption that the observations have provided the values (6) of the variables (4). Here $p(\theta_1, \ldots \theta_l)$ denotes the probability law of the θ's, called *a priori*, and the integral in the denominator extends over all systems of values of the θ's. The function $p(\theta_1, \theta_2, \ldots \theta_l|x'_1, x'_2 \ldots x'_n)$ is called the *a posteriori* probability law of θ's. In cases where the *a priori* probability law $p(\theta_1, \theta_2, \ldots \theta_l)$ is known, the formula (7) permits the calculation of the most probable values of any of the θ's and also of the probability that θ_i, say, will fall in any given interval, say, $a \leq \theta_i < b$. The most probable value of θ_i, say $\overset{v}{\theta}_i$, may be considered as the estimate of θ_i and then the probability, say

$$P\{\overset{v}{\theta}_i - \Delta < \theta_i < \overset{v}{\theta}_i + \Delta|E'\}, \quad \ldots\ldots\ldots \quad (8)$$

will describe the accuracy of the estimate $\overset{v}{\theta}_i$, where Δ is any fixed positive number and E′ denotes the set (6) of observations.

It is known that, as far as we work with the conception of probability as adopted in this paper, the above theoretically perfect solution may be applied in practice only in quite exceptional cases, and this for two reasons :

(*a*) It is only very rarely that the parameters $\theta_1, \theta_2, \ldots \theta_l$ are random variables. They are generally unknown constants and therefore their probability law *a priori* has no meaning.

* This review is not in any sense complete. Its purpose is to exemplify the attempts to solve the problem of estimation.

(*b*) Even if the parameters to be estimated, $\theta_1, \theta_2, \ldots \theta_l$, could be considered as random variables, the elementary probability law *a priori*, $p(\theta_1, \theta_2, \ldots \theta_l)$, is usually unknown, and hence the formula (7) cannot be used because of the lack of the necessary data.

When these difficulties were noticed, attempts were made to avoid them by introducing some new principle lying essentially outside the domain of the objective theory of probability.

The first of the principles advanced involved the assumption that when we have no information as to the values of the θ's, it is admissible to substitute in formula (7) some function of the θ's selected on intuitive grounds, *e.g.*,

$$p(\theta_1, \theta_2, \ldots \theta_l) = \text{const.} \quad \ldots\ldots\ldots\ldots \quad (9)$$

and use the result, say

$$p_1(\theta_1, \ldots \theta_l | \mathrm{E}') = \frac{p(x'_1, x'_2, \ldots x'_n | \theta_1, \ldots \theta_l)}{\int \ldots \int p(x'_1, x'_2 \ldots x'_n | \theta_1, \ldots \theta_l)\, d\theta_1 \ldots d\theta_l}, \quad \ldots\ldots \quad (10)$$

as if this were the *a posteriori* probability law of the θ's.

This procedure is perfectly justifiable on the ground of certain theories of probability, *e.g.*, as developed by HAROLD JEFFREYS, but it is not justifiable on the ground of the theory of probability adopted in this paper. In fact, the function $p_1(\theta_1 \ldots \theta_l | \mathrm{E}')$ as defined by (10) will not generally have the property serving as a definition of the elementary probability law of the θ's. Its integral over any region w in the space of the θ's will not be necessarily equal to the ratio of the measures of two sets of elements belonging to the fundamental probability set, which we call the probability. Consequently, if the experiment leading to the set of values of the x's is repeated many times and if we select such experiments (many of them) in which the observed values were the same, $x'_1, x'_2 \ldots x'_n$, the assumed validity of the law of big numbers (in the sense of BORTKIEWICZ) will not guarantee that the frequency of cases where the true value of θ_i falls within $\overset{v}{\theta}_i - \Delta < \theta_i < \overset{v}{\theta}_i + \Delta$ will approach the value of (8), if this is calculated from (10). Moreover, if the θ's are constant, this frequency will be permanently zero or unity, thus essentially differing from (8).

The next principle I shall mention is that advocating the use of the so-called unbiassed estimates and leading to the method of least squares. Partly following MARKOFF (1923), I shall formulate it as follows :

In order to estimate a parameter θ_i involved in the probability law (5), we should use an unbiassed estimate or, preferably, the best unbiassed estimate.

A function, F_i, of the variables (4) is called an unbiassed estimate of θ_i if its mathematical expectation is identically equal to θ_i, whatever the actual values of $\theta_1, \theta_2, \ldots \theta_l$. Thus,

$$\mathscr{E}(\mathrm{F}_i) \equiv \theta_i. \quad \ldots\ldots\ldots\ldots \quad (11)$$

An unbiassed estimate F_i is called the best if its variance, say

$$V_{F_i} = \mathscr{E}(F_i - \theta_i)^2, \quad \ldots\ldots\ldots\ldots \quad (12)$$

does not exceed that of any other unbiassed estimate of θ_i.

It is known that MARKOFF provided a remarkable theorem leading, in certain cases, to the calculation of the best of the unbiassed estimates which are linear functions of the variables (4). The advantage of the unbiassed estimates and the justification of their use lies in the fact that in cases frequently met the probability of their differing very much from the estimated parameters is small.

The other principle, which is to a certain extent in rivalry with that of the unbiassed estimate, is the principle of maximum likelihood. This consists in considering $L = \text{const.} \times p(x'_1, x'_2 \ldots x'_n | \theta_1 \ldots \theta_l)$, where x'_i denotes the observed value of X_i, as a function of the parameters θ_i, called the likelihood. It is advocated that the values of L may serve as a measure of our uncertainty or confidence in the corresponding values of the θ's. Accordingly, we should have the greatest confidence in the values, say, $\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_l$, for which L is a maximum. $\hat{\theta}_i$ obviously is a function of $x'_1 \ldots x'_n$; it is called the maximum likelihood estimate of θ_i.

As far as I am aware, the idea of the maximum likelihood estimates is due to KARL PEARSON, who applied the principle in 1895 (*see* particularly pp. 262–265), among others to deduce the now familiar formula for estimating the coefficient of correlation. However, he did not insist much on the general applicability of the principle. This was done with great emphasis by R. A. FISHER, who invented the term likelihood, and in a series of papers (FISHER, 1925) stated several important properties of the maximum likelihood estimates, to the general effect that it is improbable that their values will differ very much from those of the parameters estimated. In fact, the maximum likelihood estimates appear to be what could be called the best " almost unbiassed " estimates. Many of FISHER's statements, partly in a modified form, were subsequently proved by HOTELLING (1932), DOOB (1934), and DUGUÉ (1936). An excellent account of the present state of the theory is given by DARMOIS (1936).

In certain cases the unbiassed estimates are identical with those of maximum likelihood; in others we know only the maximum likelihood estimate, and then there is no " competition " between the two principles. But it sometimes happens that both principles may be applied and lead to different results. Such is, for instance, the case when it is known that the variables (4) are all independent and each of them follows the same normal law, so that

$$p(E|\xi, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right) e^{-\frac{\Sigma(x_i-\xi)^2}{2\sigma^2}}. \quad \ldots\ldots\ldots \quad (13)$$

The maximum likelihood estimate of the variance, σ^2, is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \ldots\ldots\ldots \quad (14)$$

while the unbiassed estimate is, say,

$$\bar{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 , \quad \ldots\ldots\ldots \quad (15)$$

and the question arises which of them to use. Obviously this is a question of principle, and the arguments, like " you must use (15) because the expectation of $\bar{\sigma}^2$ is equal to σ^2 ", do not prove much by themselves. It is perhaps remarkable that some of the authors who, when discussing theory, advocate the use of the maximum likelihood estimate, use in practice the estimate (15).

The formulae (14) and (15) may be used to illustrate the meaning of the expression " almost unbiassed " estimate, used above. Familiar formulae show that the expectation of $\hat{\sigma}^2$ is

$$\mathscr{E}(\hat{\sigma}^2) = \left(1 - \frac{1}{n}\right)\sigma^2 , \quad \ldots\ldots\ldots \quad (16)$$

thus showing a " negative bias," $n^{-1}\sigma^2$. If we increase the number of observations, n, the bias tends to zero, which justifies the terms " almost unbiassed " or " consistent " estimate attached to (14).

(c) *Estimation by Unique Estimate and by Interval*

In the preceding pages we have described briefly three of the several important principles advanced for the calculation of estimates. All of them represent attempts to solve the problem which might be called the problem of a unique estimate of an unknown parameter which reduces to determining a function of the observations, the value of which presumably does not differ very much from that of the estimated parameter.

We shall now call attention to the fact that apart from the problem of a unique estimate, the requirements of practical statistical work brought to the front another problem which we shall call the problem of estimation by interval.

Denote generally by θ the parameter to be estimated and by T its estimate, deduced from some principle or another. Whatever the principle, it is obviously impossible to assume that in any particular case T is exactly equal to θ. Therefore, the practical statistician required some measure of the accuracy of the estimate T. The generally accepted method of describing this accuracy consists in calculating the estimate, say S_T^2, of the variance V_T of T and in writing the result of all the calculations in the form $T \pm S_T$.

Behind this method of presenting the results of estimating θ, there is the idea that the true value of θ will frequently lie between the value of T minus a certain multiple of S_T and T plus perhaps some other multiple of S_T. Therefore, the smaller S_T the more accurate is the estimate T of θ.

If we look through a number of recent statistical publications, we shall find that it is exceedingly rare that the values of unique estimates are given without the $\pm S_T$.

We shall find also that the comments on the values of T are largely dependent on those of S_T. This shows that what the statisticians have really in mind in problems of estimation is not the idea of a unique estimate but that of two estimates having the form, say

$$\underline{\theta} = T - k_1 S_T \quad \text{and} \quad \overline{\theta} = T + k_2 S_T, \quad \ldots\ldots\ldots \quad (17)$$

where k_1 and k_2 are certain constants, indicating the limits between which the true value of θ presumably falls.

In this way the practical work, which is frequently in advance of the theory, brings us to consider the theoretical problem of estimating the parameter θ by means of the interval $(\underline{\theta}, \overline{\theta})$, extending from $\underline{\theta}$ to $\overline{\theta}$. These limits will be called the lower and upper estimates of θ respectively. It is obvious that if the values of k_1 and k_2 in (17) are not specified, then the real nature of the two estimates is not determined.

In what follows, we shall consider in full detail the problem of estimation by interval. We shall show that it can be solved entirely on the ground of the theory of probability as adopted in this paper, without appealing to any new principles or measures of uncertainty in our judgements. In so doing, we shall try to determine the lower and upper estimates, $\underline{\theta}$ and $\overline{\theta}$, which assure the greatest possible accuracy of the result, without assuming that they must necessarily have the commonly adopted form (17).

II—Confidence Intervals

(*a*) *Statement of the Problem*

After these somewhat long preliminaries, we may proceed to the statement of the problem in its full generality.

Consider the variables (4) and assume that the form of their probability law (5) is known, that it involves the parameters $\theta_1, \theta_2, \ldots, \theta_l$, which are constant (not random variables), and that the numerical values of these parameters are unknown. It is desired to estimate one of these parameters, say θ_1. By this I shall mean that it is desired to define two functions $\overline{\theta}(E)$ and $\underline{\theta}(E) \leq \overline{\theta}(E)$, determined and single valued at any point E of the sample space, such that if E′ is the sample point determined by observation, we can (1) calculate the corresponding values of $\underline{\theta}(E')$ and $\overline{\theta}(E')$, and (2) state that the true value of θ_1, say θ_1^0, is contained within the limits

$$\underline{\theta}(E') \leq \theta_1^0 \leq \overline{\theta}(E'), \quad \ldots\ldots\ldots\ldots \quad (18)$$

this statement having some intelligible justification on the ground of the theory of probability.

This point requires to be made more precise. Following the routine of thought established under the influence of the Bayes Theorem, we could ask that, given the sample point E′, the probability of θ_1^0 *falling within the limits* (18) should be large, say, $\alpha = 0{\cdot}99$, etc. If we express this condition by the formula

$$P\{\underline{\theta}(E') < \theta_1^0 < \overline{\theta}(E') | E'\} = \alpha, \quad \ldots\ldots\ldots \quad (19)$$

we see at once that it contradicts the assumption that θ_1^0 is constant. In fact, on this assumption, whatever the fixed point E′ and the values $\underline{\theta}$ (E′) and $\bar{\theta}$ (E′), the only values the probability (19) may possess are zero and unity. For this reason we shall drop the specification of the problem as given by the condition (19).

Returning to the inequalities (18), we notice that while the central part, θ_1^0, is a constant, the extreme parts $\underline{\theta}$ (E′) and $\bar{\theta}$ (E′) are particular values of random variables. In fact, the coordinates of the sample point E are the random variables (4), and if $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are single-valued functions of E, they must be random variables themselves.

Therefore, whenever the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are defined in one way or another, but the sample point E is not yet fixed by observation, we may legitimately discuss the probability of $\underline{\theta}$ (E) and $\bar{\theta}$ (E) fulfilling any given inequality and in particular the inequalities analogous to (18), in which, however, we must drop the dashes specifying a particular fixed sample point E′. We may also try to select $\underline{\theta}$ (E) and $\bar{\theta}$ (E) so that the probability of $\underline{\theta}$ (E) falling short of θ_1^0 and at the same time of $\bar{\theta}$ (E) exceeding θ_1^0, is equal to any number α between zero and unity, fixed in advance. If θ_1^0 denotes the true value of θ_1, then of course this probability must be calculated under the assumption that θ_1^0 *is* the true value of θ_1. Thus we can look for two function $\underline{\theta}$ (E) and $\bar{\theta}$ (E), such that

$$P\{\underline{\theta}\,(E) \leq \theta_1^0 \leq \bar{\theta}\,(E) \mid \theta_1^0\} = \alpha. \quad \ldots\ldots\ldots \quad (20)$$

and require that the equation (20) holds good *whatever* the value θ_1^0 of θ_1 and *whatever* the values of the other parameters $\theta_2, \theta_3, \ldots, \theta_l$, involved in the probability law of the X's may be.

The functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) satisfying the above conditions will be called the lower and the upper confidence limits of θ_1. The value α of the probability (20) will be called the confidence coefficient, and the interval, say δ (E), from $\underline{\theta}$ (E) to $\bar{\theta}$ (E), the confidence interval corresponding to the confidence coefficient α.

It is obvious that the form of the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) must depend upon the probability law $p\,(E|\theta_1 \ldots \theta_l)$.

It will be seen that the solution of the mathematical problem of determining the confidence limits $\underline{\theta}$ (E) and $\bar{\theta}$ (E) provides the solution of the practical problem of estimation by interval. For suppose that the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are determined so that the equation (20) does hold good whatever the values of all the parameters $\theta_1, \theta_2, \ldots \theta_l$ may be, and α is some fraction close to unity, say $\alpha = 0{\cdot}99$. We can then tell the practical statistician that whenever he is certain that the form of the probability law of the X's is given by the function $p\,(E|\theta_1, \theta_2, \ldots \theta_l)$ which served to determine $\underline{\theta}$ (E) and $\bar{\theta}$ (E), he may estimate θ_1 by making the following three steps : (*a*) he must perform the random experiment and observe the particular values $x_1, x_2, \ldots x_n$ of the X's ; (*b*) he must use these values to calculate the corresponding values of $\underline{\theta}$ (E) and $\bar{\theta}$(E) ; and (*c*) he must state that $\underline{\theta}\,(E) < \theta_1^0 < \bar{\theta}\,(E)$, where θ_1^0 denotes the true value of θ_1. How can this recommendation be justified ?

The justification lies in the character of probabilities as used here, and in the law of great numbers. According to this empirical law, which has been confirmed by numerous experiments, whenever we frequently and independently repeat a random experiment with a constant probability, α, of a certain result, A, then the relative frequency of the occurrence of this result approaches α. Now the three steps (*a*), (*b*), and (*c*) recommended to the practical statistician represent a random experiment which may result in a correct statement concerning the value of θ_1. This result may be denoted by A, and if the calculations leading to the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are correct, the probability of A will be constantly equal to α. In fact, the statement (*c*) concerning the value of θ_1 is only correct when $\underline{\theta}$ (E) falls below θ_1^0 and $\bar{\theta}$ (E), above θ_1^0, and the probability of this is equal to α whenever θ_1^0 is the true value of θ_1. It follows that if the practical statistician applies permanently the rules (*a*), (*b*) and (*c*) for purposes of estimating the value of the parameter θ_1, in the long run he will be correct in about 99 per cent. of all cases.

It is important to notice that for this conclusion to be true, it is not necessary that the problem of estimation should be the same in all the cases. For instance, during a period of time the statistician may deal with a thousand problems of estimation and in each the parameter θ_1 to be estimated and the probability law of the X's may be different. As far as in each case the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are properly calculated and correspond to the same value of α, his steps (*a*), (*b*), and (*c*), though different in details of sampling and arithmetic, will have this in common—the probability of their resulting in a correct statement will be the same, α. Hence the frequency of actually correct statements will approach α.

It will be noticed that in the above description the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results *will* tend to α.* Consider now the case when a sample, E′, is already drawn and the calculations have given, say, $\underline{\theta}$ (E′) = 1 and $\bar{\theta}$ (E′) = 2. Can we say that in this particular case the probability of the true value of θ_1 falling between 1 and 2 is equal to α ?

The answer is obviously in the negative. The parameter θ_1 is an unknown constant and no probability statement concerning its value may be made, that is except for the hypothetical and trivial ones

$$P\{1 \leq \theta_1^0 \leq 2\} = \left\{ \begin{array}{ll} 1 & \text{if} \quad 1 \leq \theta_1^0 \leq 2 \\ 0 & \text{if either } \theta_1^0 < 1 \quad \text{or} \quad 2 < \theta_1^0, \end{array} \right\} \quad . \quad . \quad . \quad (21)$$

which we have decided not to consider.

The theoretical statistician constructing the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E), having the above property (20), may be compared with the organizer of a game of chance in which the gambler has a certain range of possibilities to choose from while, whatever

* This, of course, is subject to restriction that the X's considered *will* follow the probability law assumed.

he actually chooses, the probability of his winning and thus the probability of the bank losing has permanently the same value, $1 - \alpha$.

The choice of the gambler on what to bet, which is beyond the control of the bank, corresponds to the uncontrolled possibilities of θ_1 having this or that value. The case in which the bank wins the game corresponds to the correct statement of the actual value of θ_1. In both cases the frequency of " successes " in a long series of future " games " is approximately known. On the other hand, if the owner of the bank, say, in the case of roulette, knows that in a particular game the ball has stopped at the sector No. 1, this information does not help him in any way to guess how the gamblers have betted. Similarly, once the sample E′ is drawn and the values of $\underline{\theta}$ (E′) $\bar{\theta}$ and (E′) determined, the calculus of probability adopted here is helpless to provide answer to the question of what is the true value of θ_1.

(*b*) *Solution of the Problem of Confidence Intervals*

In order to find the solution of the problem of confidence intervals, let us suppose that it is already solved and that $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are functions determined and single valued in the whole sample space, W, and such that the equality (20) holds good whatever the true values of the parameters $\theta_1, \theta_2, \ldots \theta_l$. It will be convenient to interpret the situation geometrically. For this purpose we shall need to consider the space, G, of $n + 1$ dimensions which we shall call the general space. The points in this space will be determined by $n + 1$ coordinates $x_1, x_2, \ldots x_n, \theta_1$, the first n of which are the particular values of the random variables (4) and thus determine the position of a sample point, E, in the n-dimensional space W, and the last coordinate θ_1 is one of the possible values of the parameter θ_1 in the probability law p (E$|\theta_1 \ldots \theta_l$) which we desire to estimate.

Consequently, if we consider any hyperplane, G (θ_1) in G corresponding to the equation $\theta_1 =$ const., this may be interpreted as an image of the sample space W. We notice also that to any point E in the sample space W there will correspond in G a straight line, say L (E), parallel to the axis Oθ_1. If $x_1', x_2' \ldots x'_n$ are the coordinates of E′, then the line L (E′) will correspond to the equations $x_i = x_i'$ for $i = 1, 2, \ldots n$.

Consider now the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E). On each line L (E), they will determine two points, say B (E) and C (E) with coordinates

$$x_1, x_2, \ldots x_n, \underline{\theta}\ (\mathrm{E}) \quad \ldots\ldots\ldots\ldots \quad (22)$$

and

$$x_1, x_2, \ldots x_n, \bar{\theta}\ (\mathrm{E}) \quad \ldots\ldots\ldots\ldots \quad (23)$$

respectively, where $x_1, x_2, \ldots x_n$ are the coordinates of the sample point E. The interval between B (E) and C (E) will be the image of the confidence interval δ (E) corresponding to the sample point E. If we fix a value of $\theta_1 = \theta_1'$ and a sample point E′, then the hyperplane G (θ_1') may cut or may not cut the confidence interval δ (E′). If G (θ_1') does cut δ (E′), let a (θ_1', E′) denote the point of intersection.

The position is illustrated in fig. 1, in which, however, only three axes of coordinates are drawn, Ox_1, Ox_n, and $O\theta_1$. The line L (E′) is represented by a dotted vertical line and the confidence interval δ (E′) by a continuous section of this line, which is thicker above and thinner below the point a (θ'_1, E′) of its intersection with the hyperplane G (θ_1'). The confidence interval δ (E″) corresponding to another sample point, E″, is not cut by G (θ_1') and is situated entirely above this hyperplane.

Now denote by A (θ_1') the set of all points a (θ'_1, E) in G (θ'_1) in which this hyperplane cuts one or the other of the confidence intervals δ (E), corresponding to any sample point. It is easily seen that the coordinate θ_1 of any point belonging to A (θ'_1) is equal to θ'_1 and that the remaining coordinates $x_1, x_2, \ldots x_n$ satisfy the inequalities

$$\underline{\theta}\,(E) \leq \theta'_1 \leq \bar{\theta}\,(E). \quad . \; . \quad (24)$$

In many particular problems it is found that the set of points A (θ_1) thus defined is filling up a region. Because of this A (θ'_1) will be called the region of acceptance corresponding to the fixed value of $\theta_1 = \theta'_1$.

It may not seem obvious that the region of acceptance A (θ_1) as defined above must exist (contain points) for any value of θ_1. In fact, it may seem possible that for certain values of θ_1 the hyperplane G (θ_1) may not cut any of the intervals δ (E). It will, however, be seen below that this is impossible.

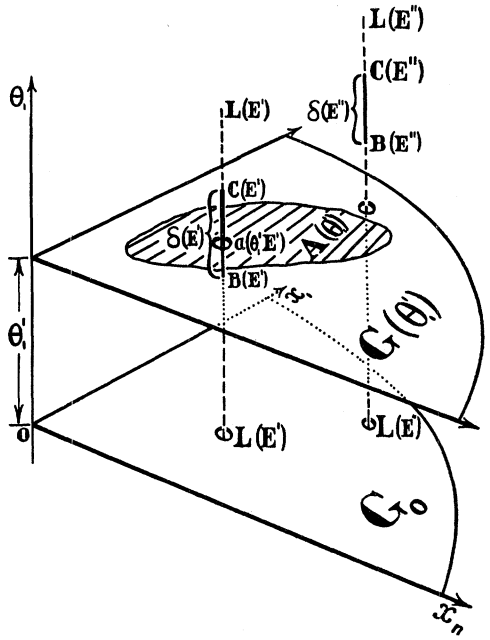


FIG. 1—The general space G.

As mentioned above, the coordinates x_1, $x_2, \ldots x_n$ of any sample point E determine in the space G the straight line L (E) parallel to the axis of θ_1. If this line crosses the hyperplane G (θ_1) in a point belonging to A (θ_1) it will be convenient to say that E falls within A (θ_1).

If for a given sample point E the lower and the upper estimates satisfy the inequalities $\underline{\theta}\,(E) \leq \theta'_1 \leq \bar{\theta}\,(E)$, where θ'_1 is any value of θ_1, then it will be convenient to describe the situation by saying that the confidence interval δ (E) covers θ'_1. This will be denoted by δ (E) Cθ'_1.

The conception and properties of the regions of acceptance are exceedingly important from the point of view of the theory given below. We shall therefore discuss them in detail proving separately a few propositions, however simple they may seem to be.

Proposition I—Whenever the sample point E falls within the region of acceptance A (θ'_1), corresponding to any fixed value θ'_1 of θ_1, then the corresponding confidence interval δ (E) must cover θ'_1.

Proof—This proposition is a direct consequence of the definition of the region of

acceptance. Suppose it is not true. Then there must be at least one sample point, say E′, which falls within A (θ'_1) and such that either $\underline{\theta}\,(E') \leqq \bar{\theta}\,(E') < \theta'_1$ or $\theta'_1 < \underline{\theta}\,(E') \leqq \bar{\theta}\,(E')$. Comparing these inequalities with (24) which serve as a definition of the region of acceptance A (θ'_1), we see that E′ could not fall within A (θ'_1), which proves the Proposition I.

Proposition II—If a confidence interval $\delta\,(E'')$ corresponding to a sample point E″ covers a value θ'_1 of θ_1, then the sample point E″ must fall within A (θ'_1).

Proof—If $\delta\,(E'')$ covers θ'_1, then it follows that $\underline{\theta}\,(E'') \leqq \theta'_1 \leqq \bar{\theta}\,(E'')$. Comparing these inequalities with (24) defining the region A (θ'_1), we see that E″ must fall within A (θ'_1).

If we agree to denote generally by $\{B \varepsilon A\}$ the words " B belongs to A " or " B is an element of A ", then we may sum up the above two propositions by writing the identity

$$\{E \varepsilon A\,(\theta'_1)\} \equiv \{\delta\,(E)\; C\theta'_1\} \equiv \{\underline{\theta}\,(E) \leqq \theta'_1 \leqq \bar{\theta}\,(E)\}, \quad \ldots \quad (25)$$

meaning that the event consisting in the sample point E falling within the region of acceptance A (θ'_1) is equivalent to the other event which consists in θ'_1 being covered by $\delta\,(E)$.

Corollary I—It follows from the Proposition I and II that whatever may be the true values $\theta'_1, \theta'_2 \ldots \theta'_l$ of the θ's, the probability of any fixed value θ''_1 of θ_1 being covered by $\delta\,(E)$ is identical with the probability of the sample point E falling within A (θ''_1).

$$P\{\delta\,(E)\; C\theta''_1 | \theta'_1, \ldots \theta'_l\} = P\{\underline{\theta}\,(E) \leqq \theta''_1 \leqq \bar{\theta}\,(E) | \theta'_1, \theta'_2, \ldots \theta'_l\}$$
$$= P\{E \varepsilon A\,(\theta''_1) | \theta'_1, \theta'_2, \ldots \theta'_l\}. \quad (26)$$

Proposition III—If the functions $\underline{\theta}\,(E)$ and $\bar{\theta}\,(E)$ are so determined that whatever may be the true values of $\theta_1, \theta_2, \ldots \theta_l$, the probability, P, of the true value of θ_1 being covered by the interval $\delta\,(E)$ extending from $\underline{\theta}\,(E)$ to $\bar{\theta}\,(E)$ is always equal to a fixed number α, then the region of acceptance A (θ'_1) corresponding to any fixed value θ'_1 of θ_1 must have the property that the probability

$$P\{E \varepsilon A\,(\theta'_1) | \theta'_1, \theta_2, \ldots \theta_l\} = \alpha, \quad \ldots \ldots \quad (27)$$

whatever may be the values of the parameters $\theta_2, \theta_3, \ldots \theta_l$.

Proof—Assume that θ'_1 happens to be the true value of θ_1 and denote generally by θ'_i the true value of θ_i, for $i = 2, 3, \ldots l$. The probability P, as defined in conditions of the Proposition III, may be expressed by means of the formula

$$P = P\{\underline{\theta}\,(E) \leqq \theta'_1 \leqq \bar{\theta}\,(E) | \theta'_1, \theta'_2, \ldots \theta'_l\}. \quad \ldots \ldots \quad (28)$$

Owing to (26), which holds good for any $\theta'_1, \theta'_2, \ldots \theta'_l$, we may write also

$$P = P\{E \varepsilon A\,(\theta'_1) | \theta'_1, \theta'_2, \ldots \theta'_l\}. \quad \ldots \ldots \quad (29)$$

If P is permanently equal to α, then $P\{E \varepsilon A(\theta'_1) | \theta'_1, \theta'_2, \ldots \theta'_l\}$ must be also equal to α, whatever $\theta'_1, \theta'_2, \ldots \theta'_l$, which proves the proposition.

Corollary II—It follows from the Proposition III that whatever the value θ'_1 of θ_1, the region of acceptance $A(\theta'_1)$ could not be empty. In fact, if for any value θ'_1 the region $A(\theta_1')$ as defined above did not contain any points at all, then the probability $P\{E \varepsilon A(\theta'_1) | \theta'_1, \ldots \theta'_l\}$ would be zero, which would contradict the Proposition III.

Proposition III describes the fundamental property of any single region of acceptance $A(\theta_1)$ taken separately. We shall now prove three propositions concerning the whole set of the regions $A(\theta_1)$ corresponding to all possible values of θ_1.

Proposition IV—If the functions $\underline{\theta}(E)$ and $\bar{\theta}(E) \geqq \underline{\theta}(E)$ are single valued and determined for any sample point E, then whatever the sample point E', there will exist at least one value of θ_1, say θ'_1, such that the point E' will fall within $A(\theta'_1)$.

Proof—Consider the values of $\underline{\theta}(E)$ and $\bar{\theta}(E)$ corresponding to the point E' and let θ'_1 be any value of θ_1 satisfying the condition $\underline{\theta}(E') \leqq \theta'_1 \leqq \bar{\theta}(E')$. Comparing these inequalities with (24), we see that E' must fall within $A(\theta'_1)$, which proves the proposition.

Proposition V—If a sample point E' falls within the regions of acceptance $A(\theta'_1)$ and $A(\theta''_1)$ corresponding to θ'_1 and $\theta''_1 > \theta'_1$ respectively, then it will also fall within the region of acceptance $A(\theta'''_1)$ corresponding to any θ'''_1 such that $\theta'_1 < \theta'''_1 < \theta''_1$.

Proof—If the sample point E' falls within $A(\theta'_1)$ and $A(\theta''_1)$ then, owing to (24), it follows that

$$\underline{\theta}(E') \leqq \theta'_1 < \theta''_1 \leqq \bar{\theta}(E'). \quad \ldots\ldots\ldots \quad (30)$$

Accordingly, whatever θ'''_1 such that $\theta'_1 < \theta'''_1 < \theta''_1$, it follows that

$$\underline{\theta}(E') < \theta'''_1 < \bar{\theta}(E'), \quad \ldots\ldots\ldots \quad (31)$$

which shows that E' falls within $A(\theta''')$.

Proposition VI—If a sample point E' falls within any of the regions $A(\theta_1)$ for $\theta'_1 < \theta_1 < \theta''_1$ where θ'_1 and θ''_1 are fixed numbers, then it must also fall within $A(\theta'_1)$ and $A(\theta''_1)$.

Proof—Suppose that the proposition is not true and that, for example, E' does not fall within $A(\theta'_1)$. Then it follows that

$$\theta'_1 < \underline{\theta}(E'). \quad \ldots\ldots\ldots\ldots\ldots \quad (32)$$

Let θ'''_1 be a number exceeding θ'_1 but smaller than either $\underline{\theta}(E')$ and θ'', so that $\theta' < \theta''' < \theta''$ and $\theta''' < \underline{\theta}(E')$. It follows from the definition (24) of $A(\theta_1)$ that E' does not fall within $A(\theta''')$, contrary to the assumption that for any θ_1 such that $\theta'_1 < \theta_1 < \theta''_1$ the point E' falls within $A(\theta_1)$. Similarly it is possible to prove that E' must fall within $A(\theta''_1)$.

The four propositions III, IV, V, and VI describe the necessary conditions which must be satisfied by the regions of acceptance A (θ_1), either separately by each of them or collectively, if the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are determined and single valued in the whole sample space W and if the equation (20) holds good for any value of θ_1; that is to say when they determine the confidence intervals required.

We shall now prove the reciprocal proposition, showing that if it is possible to select on each hyperplane G (θ_1) a region A (θ_1) having the properties as described in the conclusions of the propositions III to VI, then the system of these regions may be used to define the functions $\underline{\theta}$ (E) $\leq \bar{\theta}$ (E) which will be determined and single valued at any sample point E; further, their system will have the property that for any value ${\theta_1}^0$ of θ_1 the equality (20) will hold good, whatever the values of the other parameters $\theta_2, \theta_3, \ldots \theta_l$.

Suppose, therefore, that on each hyperplane G (θ_1) there is defined a region, A′ (θ_1), such that

(i) P $\{E \varepsilon A' (\theta_1) | \theta_1\} = \alpha$, whatever the values of $\theta_2, \theta_3, \ldots \theta_l$.

(ii) Whatever the sample point E, there exists at least one value ${\theta'}_1$ of θ_1 such that E falls within A′ $({\theta'}_1)$.

(iii) If a sample point E falls within A′ $({\theta'}_1)$ and A′ $({\theta''}_1)$ where ${\theta'}_1 < {\theta''}$, then, whatever θ''', such that ${\theta'}_1 < {\theta'''}_1 < {\theta''}_1$, the point E falls within A′ $({\theta'''}_1)$.

(iv) If a sample point E falls within A′ (θ_1) for any θ_1 satisfying the inequalities ${\theta'}_1 < \theta_1 < {\theta''}_1$, then it falls also within A′ $({\theta'}_1)$ and A′ $({\theta''}_1)$.

Denote by $\underline{\theta}'$ (E) the lower and by $\bar{\theta}'$ (E) the upper bound of values of θ_1 for which a fixed sample point E falls within A′ (θ_1).

Proposition VII—If the regions A′ (θ_1) satisfy the conditions (i), (ii), (iii), and (iv), then the functions $\underline{\theta}'$ (E) and $\bar{\theta}'$ (E) are the lower and the upper confidence limits of θ_1, *i.e.*, such that

(*a*) they are determined and single valued at any point E and $\underline{\theta}'$ (E) $\leq \bar{\theta}'$ (E),

(*b*) whatever the true value ${\theta_1}^0$ of θ_1, the probability

$$P \{\underline{\theta}' (E) \leq {\theta_1}^0 \leq \bar{\theta}' (E) | {\theta_1}^0\} = \alpha, \quad \ldots\ldots\ldots \quad (33)$$

independently of the values of the other parameters $\theta_2, \theta_3, \ldots \theta_l$.

Proof—The property (*a*) of functions $\underline{\theta}'$ (E) and $\bar{\theta}'$ (E) follows directly from the condition (ii) and the definition of $\underline{\theta}'$ (E) and $\bar{\theta}'$ (E). We may notice, however, that $\underline{\theta}'$ (E) and $\bar{\theta}'$ (E) are not necessarily finite.

To prove the property (*b*), it will be sufficient to show that whatever ${\theta_1}^0$

$$P \{\underline{\theta}' (E) \leq {\theta_1}^0 \leq \bar{\theta}' (E) | {\theta_1}^0\} = P \{E \varepsilon A' ({\theta_1}^0) | {\theta_1}^0\}, \quad \ldots\ldots \quad (34)$$

and then refer to the condition (i).

For this purpose we notice first that owing to the definition of $\underline{\theta}'$ (E) and $\bar{\theta}'$ (E), whenever E falls within A′ $({\theta_1}^0)$, then it must follow that $\underline{\theta}'$ (E) $\leq {\theta_1}^0 \leq \bar{\theta}'$ (E).

It remains to show that inversely, if for any point E, $\underline{\theta}'(E) \leq \theta_1^0 \leq \bar{\theta}'(E)$, then this point must fall within $A'(\theta_1^0)$.

Suppose for a moment that this is not true and that there is a sample point E′ not falling within $A'(\theta_1^0)$ and such that $\underline{\theta}'(E') \leq \theta_1^0 \leq \bar{\theta}'(E')$.

It is easily seen that in such a case, either $\underline{\theta}'(E') = \theta_1^0$ or $\theta_1^0 = \bar{\theta}'(E')$ or both, if $\underline{\theta}'(E') = \bar{\theta}'(E')$. In fact, if $\underline{\theta}'(E') < \theta_1^0 < \bar{\theta}'(E')$, then $\underline{\theta}'(E')$ and $\bar{\theta}'(E')$, being the lower and the upper bounds of the values of θ_1 for which E′ falls within $A'(\theta_1)$, there would exist two values of θ_1, say θ'_1 and θ''_1, such that E′ is falling within $A'(\theta'_1)$ and $A'(\theta''_1)$ and

$$\underline{\theta}'(E') \leq \theta'_1 < \theta_1^0 < \theta''_1 \leq \bar{\theta}'(E'). \quad \ldots\ldots \quad (35)$$

It would then follow from the condition (iii) that E′ falls within $A'(\theta_1^0)$, contrary to the assumption. Therefore, we cannot assume that $\underline{\theta}'(E') < \bar{\theta}_1^0 < \theta'(E')$.

Now it is easy to see that if

$$\underline{\theta}'(E') = \theta_1^0 = \bar{\theta}'(E') \quad \ldots\ldots\ldots\ldots \quad (36)$$

then E′ must fall within $A'(\theta_1^0)$. In fact, $\underline{\theta}'(E')$ and $\bar{\theta}'(E')$ are respectively the lower and the upper bounds of the values of θ_1 for which E′ falls within $A'(\theta_1)$. If they are both equal to θ_1^0, then θ_1^0 must be the only value of θ_1 for which E′ falls within $A'(\theta_1)$.

It remains to consider only such cases where either $\underline{\theta}'(E') = \theta_1^0 < \bar{\theta}'(E')$ or $\underline{\theta}'(E') < \theta_1^0 = \bar{\theta}'(E')$. In both cases $\underline{\theta}'(E') < \bar{\theta}'(E')$. We notice first that, whatever θ_1, within the limits

$$\underline{\theta}'(E') < \theta_1 < \bar{\theta}'(E') \quad \ldots\ldots\ldots\ldots \quad (37)$$

the sample point E′ must fall within $A'(\theta_1)$. Otherwise either $\underline{\theta}'(E')$ and $\bar{\theta}'(E')$ would not be respectively the lower and the upper bounds of values of θ_1 for which E′ falls within $A'(\theta_1)$, or else the condition (iii) would not be satisfied. Now it follows from (iv) that E′ must fall both within $A'(\theta'_1)$ and $A'(\theta''_1)$ where $\theta'_1 = \underline{\theta}'(E')$ and $\theta''_1 = \bar{\theta}'(E')$ and therefore within $A'(\theta_1^0)$, which completes the proof of the Proposition VII.

Thus the problem of constructing the system of confidence intervals is equivalent to that of selecting on each hyperplane, $G(\theta_1)$, regions $A(\theta_1)$ satisfying the conditions (i)–(iv). Obviously, these regions will have the property of being regions of acceptance.

Before going any further with the theory and discussing the problem of how to choose the most appropriate system of the regions of acceptance, we shall illustrate the results already reached on two examples. These have been selected so as to reduce to a minimum the technical difficulties in carrying out the necessary calculations which might easily conceal the essential points of the theory to be illustrated. It is obvious that under the circumstances the examples could hardly fail to be somewhat artificial. However, at the end of the paper the reader will find examples having direct practical importance.

(c) *Example I*

Consider first the case where the probability law of the random variables considered depends only upon one unknown parameter θ, which it is desired to estimate. Assume further, for simplicity, that the number of random variables, the particular values of which may be given by observation is $n = 2$ and that their elementary probability law $p(x_1, x_2|\theta)$ is known to be

$$\left.\begin{aligned} &p(x_1, x_2|\theta) = \theta^{-2} \quad \text{for} \quad 0 < x_1, x_2 < \theta \\ \text{and} \quad &p(x_1, x_2|\theta) = 0 \quad \text{for any other system of values of } x_1 \text{ and } x_2 \end{aligned}\right\}. \tag{38}$$

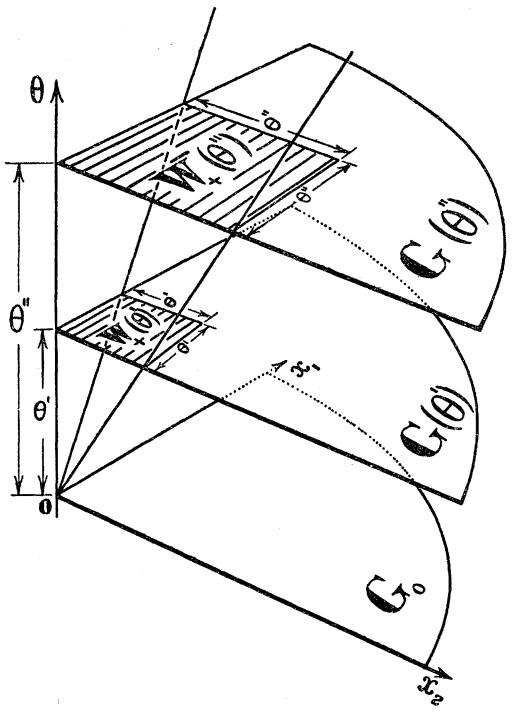


FIG. 2.

The value of θ is unknown and it is desired to construct a system of confidence intervals for its estimation.

The sample space W is now of two dimensions, *i.e.*, a plane. As the coordinates of the sample point must be positive, we may consider that W is limited by the conditions $0 < x_1$ and $0 < x_2$. Denote by $W_+(\theta)$ the part of W in which $p(x_1, x_2|\theta)$ is positive. Obviously $W_+(\theta)$ is a square with its side equal to θ.

Fig. 2 represents the general space G with two planes $G(\theta)$ on which the corresponding squares $W_+(\theta)$ are marked.

According to Proposition VII, the construction of the system of confidence intervals will be completed if we manage to select on each of the planes $G(\theta)$ a region of acceptance $A(\theta)$ satisfying (i)–(iv). Now it is easily seen that it is possible to suggest many systems of regions, some of which will and some of which will not satisfy all these conditions. We shall consider three systems, which will be denoted by S_1, S_2, and S_3, and the particular regions forming these systems by $A_1(\theta)$, $A_2(\theta)$, and $A_3(\theta)$ respectively.

(1) Fix any value of θ and let the region of acceptance $A_1(\theta)$ be defined by the inequalities

$$\beta\theta < x_i < \theta \quad \text{for} \quad i = 1, 2, \ldots\ldots\ldots \tag{39}$$

where β is a positive number less than unity and so selected as to satisfy the condition

$$P\{E_\varepsilon A_1(\theta)|\theta\} = \alpha. \ldots\ldots\ldots\ldots \tag{40}$$

Obviously

$$P\{E_\varepsilon A_1(\theta)|\theta\} = (1 - \beta)^2, \ldots\ldots\ldots \tag{41}$$

and it follows that

$$\beta = 1 - \alpha^{\frac{1}{2}}. \quad \ldots\ldots\ldots\ldots \quad (42)$$

The regions $A_1(\theta)$ defined by (39) will form the system S_1. If β is selected as indicated in (42), they will satisfy the condition (i). Now it is easy to see that they will not satisfy the condition (ii) and that therefore the system S_1 does not present a suitable choice of regions of acceptance which would determine the confidence intervals.

To see this, take any sample point E′ with coordinates x'_1, x'_2, and see whether it is always possible to find a value of $\theta = \theta'$, such that E′ will fall within $A_1(\theta')$. Owing to (39), such a value θ' should satisfy the inequalities

$$\beta\theta' < x'_1, x'_2 < \theta', \quad \ldots\ldots\ldots\ldots \quad (43)$$

or, if l and L denote respectively the smaller and the greater of the numbers x'_1 and x'_2, then

$$L < \theta' < l\beta^{-1}. \quad \ldots\ldots\ldots\ldots \quad (44)$$

This shows that the value θ' such that E′ falls within $A_1(\theta'_1)$ can be found only if $L < l\beta^{-1}$, or $\beta L < l$. Now if $l = x'_1 \leqq x'_2 = L$, then these inequalities lead to the condition $\beta x'_2 < x'_1$. If, on the contrary, $l = x'_2 \leqq x'_1 = L$, then $\beta x'_1 < x'_2$. Accordingly, none of the sample points E″ with coordinates x''_1 and x''_2 such that either

$$0 < x''_2 < \beta x''_1 \quad \text{or} \quad 0 < x''_1 < \beta x''_2 \quad \ldots\ldots \quad (45)$$

will fall within any of the regions $A_1(\theta)$ forming the system S_1, and it follows that they could not serve as regions of acceptance. Fig. 3 (i) illustrates the situation. Here cross-hatched areas correspond to (45).

(2) The second system S_2 of regions $A_2(\theta)$ we shall consider might be suggested by intuition. It follows from the definition of the probability law $p(x_1, x_2|(\theta)$ that x_1 and x_2 are mutually independent and that they vary from zero to θ. Under these circumstances, the mean $\bar{x} = \frac{1}{2}(x_1 + x_2)$ will vary symmetrically about $\frac{1}{2}\theta$ and therefore $2\bar{x} = x_1 + x_2 = T$ could be considered as an estimate of θ itself.

Denote by $A_2(\theta)$ a region in $G(\theta)$ defined by the inequalities

$$\theta - \Delta \leqq x_1 + x_2 \leqq \theta + \Delta, \quad \ldots\ldots\ldots \quad (46)$$

where Δ is so selected as to have $P\{E \varepsilon A_2(\theta)|\theta\} = \alpha$. Simple calculations give

$$P\{E \varepsilon A_2(\theta)|\theta\} = 1 - \left(\frac{\Delta}{\theta}\right)^2 = \alpha, \quad \ldots\ldots\ldots \quad (47)$$

and it follows that $\Delta = \theta(1 - \alpha)^{\frac{1}{2}}$. Substituting this value in (46), we get

$$\theta(1 - (1 - \alpha)^{\frac{1}{2}}) \leqq x_1 + x_2 \leqq \theta(1 + (1 - \alpha)^{\frac{1}{2}}) \quad \ldots\ldots \quad (48)$$

as the final definition of the region $A_2(\theta)$. Fig. 3 (ii) shows the form of the region.

It is easily seen that the system S_2 of regions thus defined satisfies all the conditions (i)–(iv).

For example, in order to check the condition (ii), we may notice that whatever the positive numbers x'_1 and x'_2, the value

$$\theta' = \frac{x'_1 + x'_2}{1 - (1 + \alpha)^{\frac{1}{2}}} \quad \ldots \ldots \ldots \ldots \quad (49)$$

satisfies the inequalities (48) which means that the sample point E′ with coordinates x'_1 and x'_2 falls within $A_2(\theta')$.

The other conditions (iii) and (iv) are checked equally easily. Thus the regions $A_2(\theta)$ may be considered as regions of acceptance. Let us now see how they determine the lower and the upper confidence limits of θ, say $\underline{\theta}_2(E)$ and $\bar{\theta}_2(E)$. According to the definition, $\underline{\theta}_2(E)$ is the lower bound of the values θ' of θ for which the sample point E falls within $A_2(\theta')$. If x_1 and x_2 are the coordinates of E, then it follows from (48) that θ' could not be smaller than, but may be as small as, $(x_1 + x_2)(1 + (1 - \alpha)^{\frac{1}{2}})^{-1}$, which means that

$$\underline{\theta}_2(E) = \frac{x_1 + x_2}{1 + (1 - \alpha)^{\frac{1}{2}}}. \quad \ldots \ldots \ldots \quad (50)$$

Similarly we get from (48) that θ' may be as large as, but could not exceed, $(x_1 + x_2)(1 - (1 - \alpha)^{\frac{1}{2}})^{-1}$, which shows that

$$\bar{\theta}_2(E) = \frac{x_1 + x_2}{1 - (1 - \alpha)^{\frac{1}{2}}}. \quad \ldots \ldots \ldots \quad (51)$$

Formerly we used the symbol $\delta(E)$ to denote the confidence interval extending from $\underline{\theta}(E)$ to $\bar{\theta}(E)$. Now we shall use the same symbol to denote the *length* of the confidence interval. We shall have from (50) and (51), say

$$\delta_2(E) = \bar{\theta}_2(E) - \underline{\theta}_2(E) = 2(x_1 + x_2)\frac{\sqrt{1 - \alpha}}{\alpha}. \quad \ldots \quad (52)$$

Now we may use (50) and (51) for estimating θ. If the observations provided the values of x_1 and x_2, say x'_1 and x'_2, we should state that

$$\frac{x'_1 + x'_2}{1 + (1 - \alpha)^{\frac{1}{2}}} \leq \theta \leq \frac{x'_1 + x'_2}{1 - (1 - \alpha)^{\frac{1}{2}}}. \quad \ldots \ldots \quad (53)$$

Whatever value of α may be fixed in advance, such that $0 < \alpha < 1$, we may be certain that the frequency of the statement in the form (53) being correct will, in the long run, approach α.

The accuracy of estimation corresponding to a fixed value of α may be measured by the lengths of the confidence intervals (52).

(3) The regions $A_3(\theta)$ forming the third set, S_3, will be defined by the inequalities

$$q\theta \leq L < \theta \quad \ldots\ldots\ldots\ldots \quad (54)$$

where L denotes again the larger of the two numbers x_1 and x_2, and q a number between zero and unity to be determined so as to satisfy the condition (i)

$$P\{E \varepsilon A_3(\theta) | \theta\} = P\{q\theta \leq L < \theta | \theta\} = \alpha. \quad \ldots\ldots\ldots \quad (55)$$

Fig. 3 (iii) shows the relationship between $W_+(\theta)$ and $A_3(\theta)$ which lies outside the square adjoining the origin of coordinates with its side equal to $q\theta$.

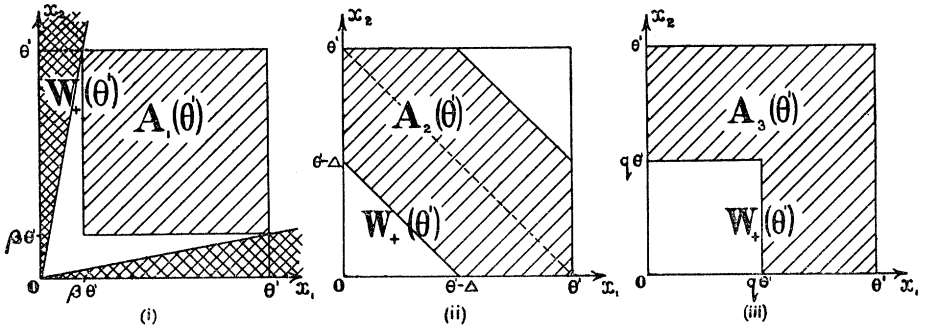


FIG. 3.

It may be useful to deduce the probability law of L for a more general case, when the number n of the x's considered is arbitrary, all of them being independent and following the same probability law

$$\left.\begin{aligned} p(x_i) &= 1/\theta \qquad \text{for } 0 < x_i < \theta \\ p(x_i) &= 0 \qquad \text{elsewhere} \end{aligned}\right\} \quad \ldots\ldots\ldots \quad (56)$$

For this purpose we notice that for any positive constant $L' \leq \theta$

$$P\{L < L' | \theta\} = \prod_{i=1}^{n} \int_0^{L'} p(x_i)\, dx_i = \left(\frac{L'}{\theta}\right)^n \quad \ldots\ldots\ldots \quad (57)$$

Differentiating this expression with regard to L', we may obtain the elementary probability law of L. The probability in the left-hand side of (55) may be obtained directly from (57) and we have, for $n = 2$,

$$P\{q\theta \leq L < \theta | \theta\} = 1 - q^2 = \alpha. \quad \ldots\ldots\ldots \quad (58)$$

Hence

$$q = (1 - \alpha)^{\frac{1}{2}}. \quad \ldots\ldots\ldots\ldots \quad (59)$$

Thus the inequality (54) defining the region $A_3(\theta)$ becomes

$$\theta(1 - \alpha)^{\frac{1}{2}} \leq L < \theta. \quad \ldots\ldots\ldots\ldots \quad (60)$$

It is easily seen that the system S_3 satisfies the conditions (i)–(iv) and therefore may be considered as a system of regions of acceptance defining the lower and the upper confidence limits of θ and hence the confidence intervals. In order to obtain the lower limit, $\underline{\theta}$ (E), fix any sample point E and consider (54). It is easily seen that if L is the larger of the coordinates of E, then the lower bound of the θ's for which E falls within A_3 (θ) is given by

$$\underline{\theta}_3 (E) = L. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (61)$$

On the other hand, it is seen also from (54) that the upper bound of the same θ's is obtained from $q\bar{\theta}$ (E) = L, thus

$$\bar{\theta}_3 (E) = L (1 - \alpha)^{-\frac{1}{2}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (62)$$

It follows that the length of the confidence interval is, say,

$$\delta_3 (E) = \frac{1 - (1 - \alpha)^{\frac{1}{2}}}{(1 - \alpha)^{\frac{1}{2}}} L. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (63)$$

The formulae (61) and (62) could be used to estimate θ, and in applying them we shall be correct, in the long run, in 100α per cent. of all cases.

It is interesting to compare the two systems of confidence intervals (50) and (51), (61) and (62). For this purpose let us choose $\alpha = \frac{3}{4}$. The statements concerning the value of θ using the two confidence intervals will be

$$\tfrac{4}{3}\bar{x} \leq \theta \leq 4\bar{x}, \quad \delta_2 (E) = \tfrac{8}{3}\bar{x}, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (64)$$

and

$$L \leq \theta \leq 2L, \quad \delta_3 (E) = L, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (65)$$

where $\bar{x}$ is the arithmetic mean of x_1 and x_2. Assume that in two different cases, A and B, the observations gave $x'_1 = x'_2 = 1$ and $x''_1 = 0{\cdot}1$, $x''_2 = 1{\cdot}9$ respectively. Then using (64) we shall get, in both cases,

$$\tfrac{4}{3} \leq \theta \leq 4, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (66)$$

while using (65)

$$1 \leq \theta \leq 2 \quad \text{and} \quad 1{\cdot}9 \leq \theta \leq 3{\cdot}8 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (67)$$

in cases A and B respectively.

The two pairs of inequalities do not agree and a superficial examination may lead to the conclusion that there is some contradiction in the theory.

It is perhaps not so bad with the sample A, for which the two confidence intervals (66) and (67) partly overlap but do not cover each other. But in the case of the sample B the interval (67) is entirely included within (66). Are these intervals equally reliable ?

Before this question could be answered, it must be made more precise. What is exactly meant by the words " equally reliable ", and do they refer to the numerically defined intervals, viz., (4/3, 4) and (1·9, 3·8), or to the whole systems of intervals as given by (64) and (65) ?

The theory of confidence intervals as explained in preceding pages does give reasons for considering the systems (64) and (65) as " equally reliable ". By this is meant that (1) if a random experiment determining the values of x_1 and x_2 is performed many times and (2) if the random variables x_1 and x_2 follow the probability law (38) where the value of $\theta > 0$ in each experiment may be the same or different—without any limitation whatsoever—then the frequency of cases where the intervals (64) and (65) calculated for each experiment would actually cover the true value of θ will be, in the long run, the same, namely, $\alpha = 3/4$.

On the other hand, if the words " equally reliable " in the above question refer to the numerical intervals (4/3, 4) and (1·9, 3·8), then the theory of confidence intervals does not give any reasons for judging them equally reliable or not.

It may be useful to illustrate the above statements with a simple sampling experiment which the reader may wish to perform.

Imagine that in a period of time the statistician is faced 400 times with the problem of estimating θ. The true value of θ may be in all those 400 cases the same, or it may vary from case to case in an absolutely arbitrary manner. Assume, for instance, that in a set of 400 random experiments the distribution of θ is as set up in the following table (or any other) :

True θ	Frequency
1	155
2	37
10	8
20	10
30	190

Next take TIPPETT's random sample tables (1927) and consider each of the numbers composed of four digits as a decimal fraction. Write down from the table 400 couples of figures. The figures of the first 155 couples consider as particular values of x_1 and x_2 determined by 155 experiments with true $\theta = 1$. The figures in the next 37 couples multiply by 2 and consider the products as forming the results of 37 further experiments where $\theta = 2$. The figures in the next 8 couples should be multiplied by 10, those in the next 10 couples by 20, and finally those in the remaining 190 couples by 30.

Substitute the obtained results in formulae (64) and (65) and see in each case whether the calculated interval covers the true value of θ, *i.e.*, 1, 2, 10, 20, or 30, whichever the case may be. It will be seen that the relative frequency of cases where the confidence intervals either calculated from (64) or from (65) will actually cover the true θ will be approximately equal to $\alpha = 0{\cdot}75$. Of course, there will be no perfect agreement with this figure, but it would be extremely surprising if the observed frequency fell outside the limits of 0·69 and 0·81. This result is entirely independent of the distribution of true θ's, and the above table may be altered as desired, without any limitation.

If there is little to choose between the two systems of confidence intervals (50) and

(51), and (62) and (63) from the point of view of probability of correct statements, there are other aspects which easily determine the choice. In problems of estimation by interval, it is natural to try to get as narrow confidence intervals as possible. Comparing again (66) and (67), we find that the latter interval is considerably shorter than the former. It is easy to see that this is a general rule. In fact, whatever the mean, $\bar{x}$, if both x_1 and x_2 are necessarily positive, then

$$\bar{x} \leqq L < 2\bar{x}, \quad \ldots\ldots\ldots\ldots\ldots\ldots (68)$$

and it follows from (64) and (65) that

$$(3/8)\,\delta_2\,(E) < \delta_3\,(E) < (3/4)\,\delta_2\,(E), \quad \ldots\ldots\ldots (69)$$

showing that the length of the confidence interval determined by (62) and (63) is always less than 3/4 of that determined by (50) and (51). It is obvious, therefore, that the confidence intervals defined by (62) and (63) compared to the other system have definite advantages. These advantages, however, are independent of the conception of probability.

Using again the analogy with the games of chance, we may say that while the rules of the two kinds of game, as described by the two pairs of inequalities (50) and (51), (62) and (63), assure the same probability of winning, the sums which could be won in each case are different, and this is the reason why we prefer the " game " (62) and (63).*

(*d*) *Example II*

Let us now consider an example in which the probability law of the random variables considered depends upon two parameters θ_1 and θ_2, our problem being to estimate the value of θ_1. In order to remove all technical difficulties which might screen the essential points of the theory, we shall again consider a simple case where the number of the random variables is $n = 2$. Suppose that it is known for certain that

$$\left.\begin{aligned} &p\,(x_1, x_2|\theta_1, \theta_2) = \frac{2}{\theta_1^{\,2}} - \theta_1\,\theta_2 + 3\,\theta_2\,x_1 \text{ for } 0 < x_1, x_2 \text{ and } x_1 + x_2 \leqq \theta_1 \\ &p\,(x_1, x_2|\theta_1, \theta_2) = 0 \text{ for any other system of values of the } x\text{'s.} \end{aligned}\right\}. \quad (70)$$

As to the parameters θ_1 and θ_2, it is known only that $\theta_1 > 0$ and $-1 < \theta_1^{\,3}\theta_2 \leqq 2$. The sample space W is limited to the first quadrant of the plane of the x's, and its positive part, $W_+(\theta_1)$, corresponding to any fixed value of θ_1, is formed by a triangle as suggested in fig. 4.

In order to see at once the difficulties introduced by the fact that the probability law (70) depends upon *two* parameters, while we are interested in one only, let us try to solve the problem of confidence intervals by a guess. In Example I, the more

* This point will be discussed later. *See* pp. 370 *et seq.*

satisfactory confidence intervals were determined by regions of acceptance belonging to S_3, having their internal boundary similar to that of the external, the latter being also the external boundary of $W_+(\theta)$.

As the conditions of the problem in Example II present many features similar to those in Example I, let us try to use as regions of acceptance the regions $A_1(\theta_1)$, constructed in the same manner as the more successful regions of acceptance in Example I.

The region $A_1(\theta_1)$ will be limited by the axes of coordinates, by the straight line $x_1 + x_2 = \theta_1$ and by a parallel to that line, corresponding to the equation $x_1 + x_2 = a\,\theta_1$, where $a < 1$ will be a constant which we shall try to determine so as to satisfy the condition (i).

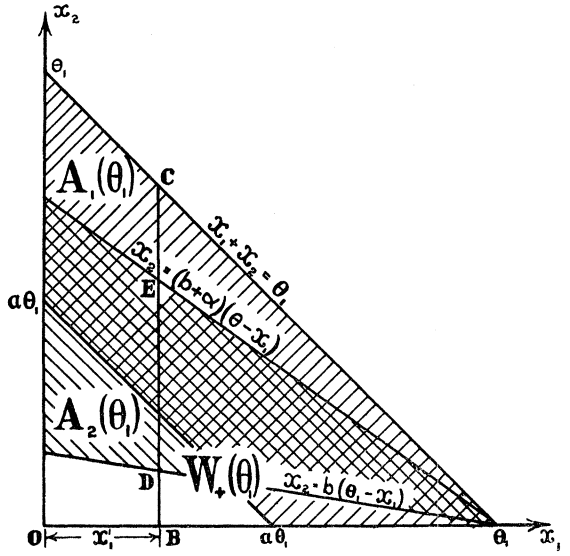


FIG. 4—B C represents $W_+(x'_1)$; D E represents $w(x'_1)$

We have

$$P\{E\varepsilon A_1(\theta_1)|\theta_1, \theta_2\} = P\{a\theta_1 \leq x_1 + x_2 \leq \theta_1|\theta_1, \theta_2\}$$

$$= 1 - \int_0^{a\theta_1} dx_1 \int_0^{a\theta_1 - x_1} p(x_1, x_2|\theta_1, \theta_2)\, dx_2$$

$$= 1 - a^2 + \tfrac{1}{2}a^2(1 - a)\,\theta_1{}^3\theta_2 \quad \ldots\ldots \quad (71)$$

Now it is easy to see that the regions $A_1(\theta_1)$ cannot be used as regions of acceptance.

In fact, it follows from the proposition III that the regions $A_1(\theta_1)$ could only be used as regions of acceptance if, for any fixed value of $\theta_1 = \theta'_1$, the probability $P\{E\varepsilon A_1(\theta'_1)|\theta'_1, \theta_2\}$ were equal to α *irrespective of what is the true value of* θ_2. Looking at the last line of (71), we see that if a and $\theta_1 = \theta'_1$ are fixed, the probability $P\{E\varepsilon A_1(\theta'_1)|\theta'_1, \theta_2\}$ still depends on θ_2 and that, according to the value of this parameter, it may be smaller or larger than the prescribed α.

We see, therefore, that in cases where the probability law of the x's depends upon some more parameters, say $\theta_2, \theta_3, \ldots \theta_l$ besides θ_1, which it is desired to estimate,

the choice of the regions of acceptance must be limited to those, A (θ_1), for which the value of the probability $P\{E \varepsilon A(\theta_1) | \theta_1, \theta_2, \ldots \theta_l\} = \alpha$ and is independent of the values of the parameters $\theta_2 \ldots . \theta_l$.

Regions of this type which have been considered elsewhere (NEYMAN and PEARSON, 1933) are called similar to the sample space with regard to the parameters θ_2, $\theta_3, \ldots \theta_l$, and of size α. If certain limiting conditions are satisfied by the elementary probability law of the X's, it is known also how to construct the most general similar region. Therefore, under these conditions, we are able to select the regions of acceptance, not only satisfying the condition (i) but also some other conditions concerning the relative width of the confidence intervals which will be discussed below.

The conditions under which we are able to construct the most general region similar to the sample space with regard to the parameter θ_2 are not satisfied by the probability law (70). Therefore, we are not able to construct any region similar to W with regard to θ_2. However, a few theoretical remarks which follow allow the construction of a rather broad family, say F, of such regions. It is just possible that an advance of our knowledge on the subject will show that the only regions similar to W with regard to θ_2 are those belonging to F.

(e) *Family of Similar Regions Based on a Sufficient System of Statistics*

Denote by $p(E|\theta_1, \theta_2, \ldots \theta_l)$ the probability law of random variables $X_1, X_2, \ldots X_n$ depending on l parameters $\theta_1, \theta_2, \ldots \theta_l$, by $W(T_1, T_2, \ldots T_s)$, or W (T) for short, the locus of points in the sample space W where some statistics* $T_1, T_2, \ldots T_s$ have certain constant values and finally by $w(T_1, T_2, \ldots T_s)$, or w (T), a part of W (T) which may be defined in one way or another. We shall assume that the T's possess continuous partial derivatives with regard to the X's. We may now prove the following proposition.

Proposition VIII—If the statistics $T_1, T_2, \ldots T_s$ form a sufficient set with regard to the parameters $\theta_2, \theta_3, \ldots \theta_l$, then the probability of the sample point E falling within w (T) calculated under the assumption that it has fallen within W (T) or

$$P\{E \varepsilon w(T) | E \varepsilon W(T)\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (72)$$

is independent of $\theta_2, \theta_3, \ldots \theta_l$ and is a function of θ_1 only.

In proving this proposition, we shall start by expressing its conditions analytically. The condition that the statistics $T_1, T_2, \ldots T_s$ form a sufficient system with regard to $\theta_2, \theta_3, \ldots \theta_l$ is equivalent to (i) that $T_1, T_2, \ldots T_s$ are algebraically independent and (ii) that the elementary probability law of the X's can be presented in the form of the product

$$p(E|\theta_1, \theta_2, \ldots \theta_l) \equiv p(T_1, T_2, \ldots T_s|\theta_1, \theta_2, \ldots \theta_l) f(E|\theta_1), \quad . \quad . \quad (73)$$

* For the definitions of the terms used in this section, *see* NEYMAN and PEARSON (1936, *b*).

where $p\,(T_1, T_2, \ldots T_s \mid \theta_1, \theta_2, \ldots \theta_l)$ means the elementary probability law of the T's and $f\,(E|\theta_1)$ is a function of the x's and possibly of θ_1, but quite independent of $\theta_2, \theta_3, \ldots \theta_l$.* The word " equivalent " means that whenever $T_1, \ldots T_s$ form a sufficient set then both (i) and (ii) must hold good and that, inversely, whenever (i) and (ii) are true, then the statistics $T_1 \ldots T_s$ must form a sufficient set.

Introduce a new system of n-variables $T_1, T_2, \ldots T_s, t_{s+1}, \ldots t_n$, including the statistics T_i, which form the sufficient set, and transforming the original space W of the x's into another n-dimensional space W'. As the T's are algebraically independent, it is always possible to arrange so as to have a one to one correspondence between W and W', except perhaps for a set of points of measure zero. Denoting by E' the point in W' and using (73), we may write the probability law of the new variables in the form

$$p\,(E'|\theta_1, \theta_2, \ldots \theta_l) = p\,(T_1, T_2, \ldots T_s|\theta_1, \ldots \theta_l) f_1\,(E'|\theta_1), \quad . \quad (74)$$

where again $f_1\,(E'|\theta_1)$ does not depend upon $\theta_2, \theta_3 \ldots \theta_l$. Dividing both sides of (74) by $p\,(T_1, \ldots T_s|\theta_1, \theta_2, \ldots \theta_l)$, we shall obtain the relative probability law of $t_{s+1}, t_{s+2}, \ldots t_n$, given $T_1, T_2, \ldots T_s$,

$$p\,(t_{s+1}, t_{s+2}, \ldots t_n|\theta_1, \ldots \theta_l, T_1, \ldots T_s) = f_1\,(E'|\theta_1). \quad . \quad . \quad . \quad (75)$$

Now (72) represents the probability of E falling within w (T), calculated on the assumption that it fell on the hypersurface W (T). The image of W (T) in W' will be a prime, say W'(T), defined by $T_i = \text{const.}$, $i = 1, 2, \ldots s$, and the image of w (T) a part of W' (T), which we shall denote by w' (T). The position of the point E' on W' (T) corresponding to any fixed system of values of $T_1, T_2, \ldots T_s$ is determined by the coordinates $t_{s+1}, t_{s+2}, \ldots t_n$, and it follows that the probability in (72) is equal to the integral of (75) with regard to $t_{s+1}, t_{s+2}, \ldots t_n$ extending over the region w' (T).

As (75) is independent of $\theta_2, \theta_3, \ldots \theta_l$, so must be its integral taken over w' (T),

$$\begin{aligned} P\,\{E \varepsilon w\,(T)|E \varepsilon W\,(T)\} &= P\,\{E' \varepsilon w'\,(T)|E' \varepsilon W'\,(T)\} \\ &= \int \ldots \int_{w'(T)} p\,(t_{s+1}, \ldots t_n|T_1, T_2, \ldots T_s)\, dt_{s+1} \ldots dt_n \\ &= \int \ldots \int_{w'(T)} f_1\,(E'|\theta_1)\, dt_{s+1}\, dt_{s+2} \ldots dt_n \quad . \quad . \quad . \quad . \quad (76) \end{aligned}$$

This completes the proof of the proposition VIII. We may remark that for any fixed value of θ_1 and a fixed system of $T_1, T_2, \ldots T_s$ for which $p\,(T_1, \ldots T_s) > 0$ the region w (T) may be so selected as to ascribe to (76) any value between zero and unity which may be given in advance. It is also obvious that this could be done in an infinity of ways.

* This proposition has been stated without proof by NEYMAN and PEARSON (1936, *b*), p. 121. It may be easily proved following the lines indicated by NEYMAN (1935, *a*).

Proposition IX—If $T_1, T_2, \ldots T_s$ form a sufficient set of statistics with regard to $\theta_2, \theta_3, \ldots \theta_l$ and if for any system of values of the T's the region w (T) is so selected that, for a fixed value of $\theta_1 = \theta'_1$,

$$P\{E\varepsilon w(T)|E\varepsilon W(T)\} = \alpha, \quad \ldots\ldots\ldots \quad (77)$$

where $0 < \alpha < 1$, then, for that value $\theta_1 = \theta'_1$ the n-dimensional region w which would be obtained by combining together the regions w (T) corresponding to all possible systems of values of $T_1, T_2, \ldots T_s$, will be similar to the sample space W with regard to $\theta_2, \theta_3, \ldots \theta_l$ and will have its size equal to α, so that

$$P\{E\varepsilon w|\theta'_1\} = \alpha, \quad \ldots\ldots\ldots\ldots \quad (78)$$

whatever the values of $\theta_2, \theta_3, \ldots \theta_l$.

In order to prove Proposition IX, denote by w' the image of w in W'. Obviously w' will be a combination of the regions w' (T) and also

$$P\{E\varepsilon w|\theta'_1\} = P\{E'\varepsilon w'|\theta'_1\}, \quad \ldots\ldots\ldots\ldots \quad (79)$$

and therefore

$$P\{E\varepsilon w|\theta'_1\} = \int \ldots \int_{w'} p(E'|\theta'_1, \theta_2, \ldots \theta_l)\, dT_1\, dT_2 \ldots dt_n. \quad (80)$$

Using (74) and denoting by W'' the set of all possible systems of values of $T_1, T_2, \ldots T_s$, we obtain further

$$P\{E\varepsilon w|\theta'_1\} = \int \ldots \int_{W''} \Big\{ p(T_1, T_2, \ldots T_s|\theta_1, \theta_2, \ldots \theta_l) \int \ldots \int_{w'(T)} f_1(E'|\theta'_1)\, dt_{s+1} \ldots dt_n \Big\}\, dT_1 \ldots dT_s . \quad (81)$$

Owing to (77), this equation reduces to

$$P\{E\varepsilon w|\theta'_1\} = \alpha \int \ldots \int_{W''} p(T_1, \ldots T_s|\theta'_1, \theta_2, \ldots \theta_l)\, dT_1 \ldots dT_s = \alpha, \quad (82)$$

since the integral of $p(T_1, \ldots T_s|\theta'_1, \ldots \theta_l)$, taken over the set W'' of all possible systems of values of the T's, must be equal to unity, whatever the values of $\theta_1, \theta_2, \ldots \theta_l$. This proves the Proposition IX.

It follows that, whenever a system of statistics $T_1, T_2, \ldots T_s$ sufficient with regard to the parameters $\theta_2, \ldots \theta_l$ exists, we may construct an infinity of regions w, all of which will be similar to the sample space W and will have the same size α. To do so it is sufficient

(*a*) To select on any hypersurface W (T) a region w (T) satisfying the condition (77). Owing to Proposition VIII, this *is* always possible and in an infinity of ways.

(*b*) To combine all the regions w (T) corresponding to all possible systems of values of the T's.

The family of the regions similar to the sample space with regard to $\theta_2, \ldots \theta_l$ which may be thus obtained may be called the family based on the sufficient system of statistics $T_1, T_2, \ldots T_s$. It is possible that in certain cases similar regions will exist which do not enter into such families based on sufficient systems of statistics.

We may now go back to our Example II and see how the problem of confidence intervals could be solved.

(*f*) *Example IIa.*

Turning back to the probability law of x_1 and x_2 as defined in (70), it is easy to see that x_1 is a specific sufficient statistic with regard to θ_2. As a specific sufficient statistic with regard to one parameter is a particular case of a sufficient system of statistics, this fact, together with the Proposition IX, could be used in order to construct regions similar with regard to θ_2, which we require to serve us as regions of acceptance.

In order to see that x_1 is a specific sufficient statistic with regard to θ_2, let us calculate its elementary probability law. Integrating (70) with regard to x_2 between limits zero and $\theta_1 - x_1$, we easily obtain

$$\left.\begin{aligned} &p(x_1) = p(x_1, x_2|\theta_1, \theta_2)(\theta_1 - x_1) \quad \text{for} \quad 0 < x_1 \leqq \theta_1, \\ &p(x_1) = 0 \quad \text{for any other value of } x_1. \end{aligned}\right\} \quad \ldots \quad (83)$$

It is seen that $p(x_1)$ depends both on θ_1 and θ_2 and therefore we shall denote it by $p(x_1|\theta_1\theta_2)$. Now we can write

$$p(x_1, x_2|\theta_1, \theta_2) = p(x_1|\theta_1, \theta_2) f(\mathrm{E}|\theta_1), \quad \ldots \quad (84)$$

with $f(\mathrm{E}|\theta_1)$ defined as follows. For $0 < x_1, x_2$ and $x_1 + x_2 \leqq \theta_1$

$$f(\mathrm{E}|\theta_1) = (\theta_1 - x_1)^{-1}, \quad \ldots \quad (85)$$

and at any other point $f(\mathrm{E}|\theta_1) = 0$. As $f(\mathrm{E}|\theta_1)$ is independent of θ_2, it follows that x_1 is a specific sufficient statistic of θ_2.

Using Proposition IX, we may now construct regions which, for a fixed value of θ_1, will be similar to W with regard to θ_2. For this purpose we have to fix $\theta_1 = \theta'_1$ (say) and also the value of the sufficient statistic $x_1 = x'_1$. Next we consider the locus W (x'_1) where $x = x'$ and select any part of it $w(x')$ satisfying (77).

The combination of $w(x')$ corresponding to all values of x' between limits $0 < x' \leqq \theta'_1$ will give us a region similar to the sample space with regard to θ_2.

Now W (x'_1) is a straight line parallel to the axis Ox_2. In order to select its part $w(x')$, which may be represented by an interval, satisfying (77), we require the relative probability law of x_2, given x_1. Using the familiar relation

$$p(x_1, x_2) = p(x_1)\, p(x_2|x_1), \quad \ldots \quad (86)$$

and comparing it with (84) and (85), we find that for $0 < x_1 \leqq \theta_1$

$$\left.\begin{aligned} p\,(x_2|\theta_1, x_1) &= (\theta_1 - x_1)^{-1} \quad \text{for} \quad 0 < x_2 \leqq \theta_1 - x_1 \\ p\,(x_2|\theta_1, x_1) &= 0 \quad \text{for other values of } x_2. \end{aligned}\right\} \quad . \; . \quad (87)$$

It follows that the relative probability law of x_2, given x_1, is positive and constant for $0 < x_2 \leqq \theta_1 - x_1$ and is zero elsewhere on the line $W(x_1)$. Therefore the condition (77) concerning the interval $w(x'_1)$* to be one of the elements of the similar region w reduces to the requirement that the length of $w(x')$ should be in a constant proportion α to the length of the interval, say $W_+(x'_1)$, on $W(x'_1)$, where $p\,(x_2|\theta_1, x'_1)$ is positive.

We see that a number of regions similar to the sample space with regard to θ_2 could be obtained as follows. (*a*) Fix a value of $x = x' < \theta_1$ and select on the line $W_+(x'_1)$ corresponding to

$$x_1 = x'_1 \quad \text{and} \quad 0 < x_2 \leqq \theta_1 - x'_1, \quad . \; . \; . \; . \; . \; . \; . \quad (88)$$

any interval $w(x'_1)$, the length of which is equal to $\alpha\,(\theta_1 - x'_1)$. (*b*) Combine all such intervals together to form w.

We shall select as the regions of acceptance, $A_2(\theta_1)$, the regions constructed as described in (*a*) and (*b*) with an additional limitation, that the intervals $w(x_1)$ corresponding to different values of x_1 should be similarly situated on $W_+(x_1)$. Thus, for any $0 < x_1 < \theta_1$ we shall define the interval $w(x_1)$ by the inequalities

$$b\,(\theta_1 - x_1) < x_2 \leqq (b + \alpha)\,(\theta_1 - x_1), \quad . \; . \; . \; . \; . \; . \quad (89)$$

where b is any positive number not exceeding $1 - \alpha$. Combining all such intervals, which obviously satisfy (*a*), we shall obtain the region $A_2(\theta_1)$ which we shall use as a region of acceptance in estimating θ_1. As shown in fig. 4, the region $A_2(\theta_1)$ is limited by the axis Ox_2, and by two straight lines $x_2 = b\,(\theta_1 - x_1)$ and $x_2 = (b + \alpha)\,(\theta_1 - x_1)$. It is easy to check that $P\{E \varepsilon A_2(\theta_1)|\theta_1\} = \alpha$ whatever the value of θ_2, so that the condition (i) required for $A_2(\theta_1)$ to be a region of acceptance is satisfied. It is easily seen that the remaining conditions (ii)–(v) are also satisfied.

Now we may determine the confidence intervals for θ_1 resulting from the regions of acceptance $A_2(\theta_1)$. If x'_1 and x'_2 are the coordinates of any sample point E′ determined by observation, we see from (89) that the lower bound of values θ'_1 of θ_1 for which $E' \varepsilon A_2(\theta'_1)$ is

$$\underline{\theta}_1(E') = x'_1 + \frac{x'_2}{b + \alpha} \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (90)$$

* It is obvious that it is not necessary that $w(x')$ should be one single interval on $W_+(x')$. It could be formed by several such intervals subject to the condition that the sum of their lengths is equal to $\alpha\,(\theta_1 - x')$, *etc.*

The upper bound of θ'_1 is found from the same inequalities (89), namely,

$$\bar{\theta}_1 (E') = x'_1 + \frac{x'_2}{b}. \quad \ldots\ldots\ldots\ldots \quad (91)$$

These are two estimates of θ_1 determining the confidence interval δ (E'). The length of this interval for any given sample point, E, is

$$\delta (E) = \frac{\alpha x_2}{b (b + \alpha)}, \quad \ldots\ldots\ldots\ldots \quad (92)$$

and depends upon the value of b chosen. The larger b, the smaller δ (E) and therefore the more accurate estimation of θ_1. The confidence intervals giving the greatest accuracy correspond to $b = 1 - \alpha$.

We see again that after having assured that the probability of our being correct in statements concerning the estimated parameter is equal to α, we can proceed further and satisfy some requirements concerning the accuracy of these statements as measured by the length of the confidence intervals.

The above two examples are simple not only because they do not present any technical difficulties in calculating probability laws, etc., but also because the choice between the systems of confidence intervals suggested is easy, *e.g.*, if we use alternatively $b' = 1 - \alpha$ and $b'' < 1 - \alpha$, all the confidence intervals as determined by (90) and (91) corresponding to b' will be shorter than those corresponding to b''. There is therefore no doubt as to what value of b should be chosen.

This, however, is not always the case, and in general there are two or more systems of confidence intervals possible corresponding to the same confidence coefficient α, such that for certain sample points, E', the intervals in one system are shorter than those in the other, while for some other sample points, E'', the reverse is true.

This point is of some importance and I advise the reader, as a useful exercise, to consider a system of regions of acceptance, $A_3 (\theta_1)$, defined as follows :

(1) for $0 < x_1 \leq 1/2\ \theta_1$, $A_3 (\theta_1)$ contains all points in which

$$(1 - \alpha) (\theta_1 - x_1) \leq x_2 \leq \theta_1 - x_1, \quad \ldots\ldots\ldots \quad (93)$$

(2) for $1/2\ \theta_1 < x_1 < \theta_1$, $A_3 (\theta_1)$ contains all points in which

$$0 < x_2 < \alpha (\theta_1 - x_1). \quad \ldots\ldots\ldots\ldots \quad (94)$$

It is easy to see that the regions $A_3 (\theta_1)$ thus defined may serve as regions of acceptance. The reader will also easily find that for all sample points of the line $x_2 = \alpha x_1$ the confidence intervals as defined by regions $A_3 (\theta_1)$ will be shorter than those defined by (90) and (91) with $b = 1 - \alpha$. On the contrary, the confidence intervals for all sample points lying on the line $x_2 = q x_1$ with

$$0 < q < \frac{\alpha (1 - \alpha)}{1 - \alpha + \alpha^2}, \quad \ldots\ldots\ldots\ldots \quad (95)$$

will be greater than those defined by (90) and (91). The position is illustrated in fig. 5. Here it is not so clear which of the two systems of confidence intervals to choose. The analysis of the situation is given in the next section.

III—Accuracy of Confidence Intervals

(*a*) *Shortest Systems of Confidence Intervals*

If there are possible the systems of confidence intervals, say C_1 and C_2, such that for some sample points the intervals in C_1 are shorter than those in C_2, while for some other sample points the reverse is true, the choice between C_1 and C_2 may be based on the relative frequency or on the probability of having an interval of a given length. If using C_1 we have short confidence intervals more frequently than when using C_2, then the system C_1 will be probably considered as more satisfactory.

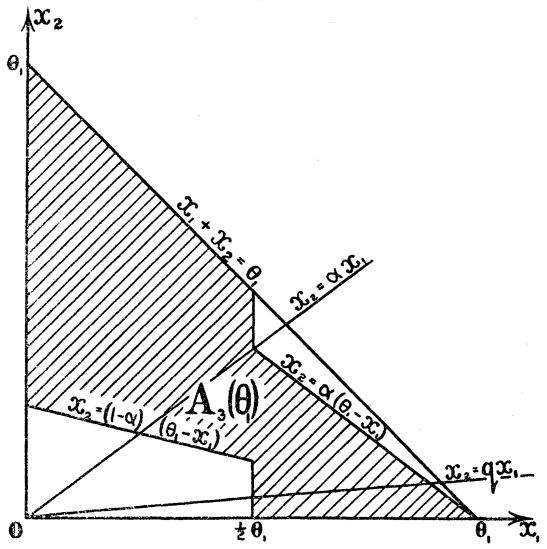


Fig. 5—Shaded area represents $A_3(\theta_1)$

The above statement may appeal to intuition, but it is obviously too vague to be used in practice.

Consider the general problem when the number n of the variables X which we may observe is arbitrary and the probability law of the X's, $p(E|\theta_1, \ldots \theta_l)$ depends on l parameters $\theta_1, \ldots \theta_l$, the first of which, θ_1, we desire to estimate. Denote by θ_1^0 the unknown true value and by θ'_1 any other value of the estimated parameter. Denote further by $\delta_i(E)$ the confidence interval for θ_1 corresponding to the sample point E and belonging to a particular system C_i, $(i = 1, 2 \ldots)$ of the confidence intervals established at a fixed confidence coefficient α. Thus we assume that, as in the above examples, we have several systems of confidence intervals $C_1, C_2, \ldots$ If all of them correspond to the same confidence coefficient α, then all of them satisfy the condition

$$P\{\delta(E)\, C\theta_1^0 | \theta_1^0\} \equiv \alpha, \quad \ldots\ldots\ldots \quad (96)$$

stating that, whatever θ_1^0 and whatever the values of other parameters $\theta_2, \ldots \theta_l$, the probability that the interval should cover the true value θ_1^0, is equal to α.

This is the common property of the systems of confidence intervals considered.

Now it is obvious that whilst it is desirable that the true value of $\theta_1 = \theta_1^0$ should be covered by the confidence interval $\delta(E)$ determined by an observed sample point E, it is not so with any other value of $\theta_1 = \theta'_1 \neq \theta_1^0$. In fact, the presence of the value $\theta'_1 \neq \theta_1^0$ within an interval $\delta(E)$ containing θ_1^0 is unnecessary and may be

interpreted as an indication that this interval is " too broad ". It is clearly impossible to avoid altogether covering the values of θ_1 which are not true. But we may try to diminish the frequency of δ (E) covering any value $\theta'_1 \neq \theta_1^0$ to a minimum. This leads us to the following definition of the shortest system of confidence intervals.

If a system, C_0, of confidence intervals δ_0 (E) has the property that whatever any other system C of intervals δ (E) corresponding to the same confidence coefficient α, whatever the true value of $\theta_1 = \theta_1^0$ and whatever any other value $\theta'_1 \neq \theta_1^0$

$$P\{\delta_0 (E)\, C\theta'_1 | \theta_1^0\} \leq P\{\delta (E)\, C\theta'_1 | \theta_1^0\}, \quad \ldots\ldots \quad (97)$$

then the system C_0 will be called the shortest system of confidence intervals.

The justification of this terminology is clear. When using C_0, the true value of $\theta_1 = \theta_1^0$ will be covered with the prescribed frequency α and any other value $\theta'_1 \neq \theta_1^0$, with a frequency not exceeding that corresponding to any other system, C corresponding to the same confidence coefficient α. This could be described by saying that the intervals δ_0 (E) are not *unnecessarily* broad.

The problem of determining the shortest system of confidence intervals is immediately reduced to that of finding appropriate regions of acceptance. In fact, using the Proposition I and II or the Corollary I expressed by (26), we may rewrite the condition (97) as follows :

$$P\{E \varepsilon A_0 (\theta'_1) | \theta_1^0\} \leq P\{E \varepsilon A (\theta'_1) | \theta_1^0\}, \quad \ldots\ldots \quad (98)$$

where $A_0 (\theta_1)$ and $A (\theta_1)$ denote the regions of acceptance leading to the systems of confidence intervals C_0 and C respectively.

If C_0 is the shortest system, then (98) should hold whatever θ_1^0 and θ'_1 and whatever the regions of acceptance $A (\theta_1)$, provided they correspond to the fixed confidence coefficient α. The condition (98) concerns the region of acceptance $A_0 (\theta'_1)$, and it must be combined with that expressed by the Proposition III, namely that

$$P\{E \varepsilon A_0 (\theta'_1) | \theta'_1\} = P\{E \varepsilon A (\theta'_1) | \theta'_1\} = \alpha, \quad \ldots\ldots \quad (99)$$

which must also hold for any θ'_1 and any values of the other parameters $\theta_2, \ldots \theta_l$.

We see that the problem of the shortest systems of confidence intervals corresponding to a confidence coefficient α is reduced to the following :

(1) Fix any value of $\theta_1 = \theta'_1$ and determine on the hyperplane $G (\theta'_1)$ a region $A (\theta'_1)$ similar to the sample space with regard to $\theta_2, \ldots \theta_l$ and of the size α.

(2) Out of all such regions $A (\theta'_1)$ choose the one, $A_0 (\theta'_1)$, for which the probability $P\{E \varepsilon A (\theta'_1) | \theta_1^0\}$, where θ_1^0 is any value of θ_1 different from θ'_1, is minimum.

(3) If the region $A_0 (\theta'_1)$ so found does not lose its property of minimizing $P\{E \varepsilon A (\theta'_1) | \theta_1^0\}$ when the value θ_1^0 is changed, and if the whole system of the regions $A_0 (\theta'_1)$ corresponding to all possible values of θ_1 satisfies the conditions (i)–(iv) of p. 354, then it may be used as the system of regions of acceptance and will

determine the shortest system of confidence intervals. The problem as described in (1) and (2) has already been considered in connexion with the theory of testing statistical hypotheses (NEYMAN and PEARSON, 1933) and its solution is known. However, it is also known that the region, $A_0(\theta'_1)$, satisfying the conditions (1) and (2) for a particular θ_1^0 does not always do so when that value of θ_1^0 is changed. It follows that the shortest systems of confidence intervals do not always exist. Still, they do exist occasionally. The reader acquainted with the joint paper mentioned will have no difficulty in checking that the confidence intervals determined by (61) and (62) in the case of the above Example I form the shortest system of confidence intervals. Applying the theory of the same paper, it is also easy to see that the confidence intervals defined by (90) and (91) with $b = 1 - \alpha$ form a system which is shortest of all those which could be constructed, using regions of acceptance belonging to the family based on the specific sufficient statistic x_1.

These, however, are rather rare cases. In order to emphasize this rareness, we shall prove the following proposition.

Proposition X

(1) If the probability law $p(E|\theta)$ of the X's, depending upon one parameter θ, is continuous in the whole sample space W and if at any point of this space it admits a continuous derivative with regard to θ not identically equal to zero, and admitting differentiation under the sign of the integral taken over W ;

(2) If $A(\theta')$ is a region in the sample space W and θ' and θ'' are two particular values of θ, such that

$$P\{E \varepsilon A(\theta')|\theta'\} = \alpha \ , \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (100)$$

and

$$P\{E \varepsilon A(\theta')|\theta''\} \leq P\{E \varepsilon A|\theta''\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (101)$$

where A is any other region in W such that $P\{E \varepsilon A|\theta'\} = \alpha$;

(3) If on the boundary of $A(\theta')$ there exists at least one point where $p(E|\theta')$ is not zero, then there must exist a third value of $\theta = \theta'''$, and a region B in W, such that

$$P\{E \varepsilon B|\theta'\} = \alpha \ . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (102)$$

$$P\{E \varepsilon A(\theta')|\theta'''\} > P\{E \varepsilon B|\theta'''\}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (103)$$

It will be noticed that the Proposition X means that if the probability law of the X's satisfies the condition (1), then the shortest system of confidence intervals generally do not exist. It follows also that in such cases the uniformly most powerful tests of hypotheses specifying the value of θ cannot exist.

We shall prove the Proposition X, starting with the assumption that it is not correct and that whatever the value θ''', either smaller or larger than θ', and whatever the region B satisfying (102) it follows that

$$P\{E \varepsilon A(\theta')|\theta'''\} \leq P\{E \varepsilon B|\theta'''\}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (104)$$

It is known (NEYMAN and PEARSON, 1933) that in such a case, whatever the sample point E′ within the region A (θ'), then for any θ,

$$p\,(E'|\theta) \leq k\,(\theta)\,p\,(E'|\theta'), \quad \ldots\ldots\ldots \quad (105)$$

where $k\,(\theta)$ depends only on θ and not on the x's. At any point, E″, outside A (θ') we should have

$$p\,(E''|(\theta) \geq k\,(\theta)\,p\,(E''|\theta'). \quad \ldots\ldots\ldots \quad (106)$$

Owing to the continuity of the probability law $p\,(E|\theta)$ we shall have at any point E‴ on the boundary of A (θ'_1)

$$p\,(E'''|\theta) = k\,(\theta)\,p\,(E'''|\theta'). \quad \ldots\ldots\ldots \quad (107)$$

We shall assume that $p\,(E'''|\theta') > 0$. As $p\,(E'''|\theta)$ admits a derivative with regard to θ, it follows that $k\,(\theta)$ must admit one. It follows also from (107) that if $\theta \longrightarrow \theta'$ then $k\,(\theta) \longrightarrow 1$. Differentiating (107) with regard to θ, and putting $\theta - \theta' = \Delta\theta$, we can write the following expansion of $k\,(\theta)$

$$\begin{aligned} k\,(\theta) &= 1 + \Delta\theta\, k'\,(\theta' + q\,\Delta\theta) \\ &= 1 + \Delta\theta\, p'\,(E'''|\theta' + q\,\Delta\theta)\, p^{-1}\,(E'''|\theta')\,, 0 < q < 1, \end{aligned} \quad (108)$$

where the dashes indicate differentiation with regard to θ. On the other hand, we can write also

$$p\,(E'|\theta) = p\,(E'|\theta'_1) + \Delta\theta\, p'\,(E'|\theta' + r\Delta\theta),\ 0 < r < 1. \ . \ . \quad (109)$$

Substituting (108) and (109) in (105) and rearranging, we get

$$\Delta\theta \left((p'\,(E'|\theta' + r\Delta\theta) - \frac{p'\,(E'''|\theta' + q\Delta\theta)\,p\,(E'|\theta')}{p\,(E'''|\theta')} \right) \leq 0, \quad . \quad (110)$$

and this inequality must hold good at any point E′ within A (θ') and for any value of $\Delta\theta$. It follows that

$$p'\,(E'|\theta') - \frac{p'\,(E'''|\theta')\,p\,(E'|\theta')}{p\,(E'''|\theta')} = 0 \quad \ldots\ldots \quad (111)$$

at any point E′ within A (θ'). In fact, if the expression in the left-hand side of (111) were not zero, then, owing to the continuity of $p'\,(E|\theta)$, for sufficiently small values of $\Delta\theta$, the expression in brackets in (110) would not be zero and would have a constant sign. As $\Delta\theta$ may be both positive and negative, the inequality (110) would not be satisfied. Using the inequality (106) holding good at any point outside A (θ') and repeating the above argument, we could easily find that (111) must hold good also outside A (θ') and therefore in the whole sample space W. Now it is easy to see that $p'\,(E|\theta')$ must be identically equal to zero, which contradicts the hypothesis (1) of the proposition X.

To show this we consider the integral

$$\int \dots \int_W p\,(\mathrm{E}|\theta)\,dx_1 \dots dx_n = 1. \quad \dots\dots\dots \quad (112)$$

Differentiating it with regard to θ and putting $\theta = \theta'$, we get

$$\int \dots \int_W p'\,(\mathrm{E}|\theta')\,dx_1 \dots dx_n = 0. \quad \dots\dots\dots \quad (113)$$

We can calculate $p'\,(\mathrm{E}|\theta')$ from (111) and substitute into (113). Using again (112) we find

$$\frac{p'\,(\mathrm{E}'''|\theta')}{p\,(\mathrm{E}'''|\theta')} = 0. \quad \dots\dots\dots\dots \quad (114)$$

Substituting this again in (111) we find $p'\,(\mathrm{E}|\theta'_1) = 0$, whatever the point E in W. This proves the Proposition X.

As the majority of probability laws with which we deal in practice, *e.g.*, the normal law, satisfy the conditions of Proposition X, it is seen that, for practical purposes, some other type of systems of confidence intervals is required, as the shortest systems generally do not exist.

(*b*) *One-sided Estimation*

The proof of the above proposition is based upon the circumstance that the left-hand side of the inequality (110) must not change its sign, while $\Delta\theta$ is both positive and negative.

It is therefore obvious that if it were for some reasons required to determine regions of acceptance $\mathrm{A}_0\,(\theta)$ satisfying the conditions

$$\mathrm{P}\,\{\mathrm{E}\varepsilon\mathrm{A}_0\,(\theta_1)|\theta_1\} = \alpha, \quad \dots\dots\dots \quad (112)$$

whatever the value of θ_1 and whatever the values of other unknown parameters involved in the probability law of the X's, and also the condition

$$\mathrm{P}\,\{\mathrm{E}\varepsilon\mathrm{A}_0\,(\theta'_1)|\theta''_1\} \leqq \mathrm{P}\,\{\mathrm{E}\varepsilon\mathrm{A}\,(\theta'_1)|\theta''_1\}, \quad \dots\dots \quad (113)$$

whatever any other region $\mathrm{A}\,(\theta'_1)$ satisfying (112) and whatever θ'_1 and θ''_1, *provided, however, the difference between them* $\theta'_1 - \theta''_1$, *is either always positive or always negative*, then the solution of this problem would exist more frequently than that of the problem of the shortest systems of confidence intervals.

The application of the regions of acceptance having the above properties is found useful in problems which may be called those of one-sided estimation. In frequent practical cases we are interested only in one limit which the value of the estimated parameter cannot exceed in one or in the other direction. When analysing seeds,

we ask for the minimum per cent. of germinating grains which it is possible to guarantee. When testing a new variety of cereals we are again interested in the minimum of gain in yield over the established standard which it is likely to give. In sampling manufactured products, the consumer will be interested to know the upper limit of the percentage defective which a given batch contains. Finally, in certain actuarial problems, we may be interested in the upper limit of mortality rate of a certain society group for which only a limited body of data is available.

In all these cases we are interested in the value of one parameter, say, θ_1, and it is desired to determine only one estimate of the same, either $\underline{\underline{\theta}}\,(E)$ or $\overline{\overline{\theta}}\,(E)$, which we shall call the unique lower and the unique upper estimate respectively. If θ_1 is the percentage of germinating seeds, we are interested in its lower estimate $\underline{\underline{\theta}}\,(E)$ so as to be able to state that $\underline{\underline{\theta}}\,(E) \leq \theta_1$, while the estimation of the upper bound $\overline{\overline{\theta}}\,(E)$ is of much less importance. On the other hand, if it is the question of the upper limit of mortality rate, θ_2, then we desire to make statements as to its value in the form $\theta_2 \leq \overline{\overline{\theta}}\,(E)$, etc.

These are the problems of one-sided estimation, and it is easy to see that their most satisfactory solution depends upon the possibility of constructing regions of acceptance satisfying (1) and (2), the latter with the restriction that the sign of the difference $\theta'_1 - \theta''_1$ is constant.

The two problems of the unique lower and the unique upper estimates are very similar, so that it will be sufficient to treat only one of them, *e.g.*, the first. Suppose, then, that we are interested in the unique lower estimate $\underline{\underline{\theta}}\,(E)$ of a parameter θ_1. Treating the problem from the point of view of confidence intervals, we desire to define a function $\underline{\underline{\theta}}\,(E)$ of the sample point E such that whatever may be the true value θ_1^0, of θ_1, the probability

$$P\{\underline{\underline{\theta}}\,(E) \leq \theta_1^0 | \theta_1^0\} = \alpha \quad \ldots\ldots\ldots \quad (114)$$

where α is the chosen confidence coefficient. Repeating the reasonings of the preceding sections, we find that this problem is equivalent with that of choosing appropriate regions of acceptance and that there is an infinity of solutions. Let us now specify the properties of a solution which would make it more desirable than any other.

For that purpose denote by θ_1^0 the unknown true value of θ_1 and by θ'_1 and θ''_1 any two other values such that

$$\theta'_1 < \theta_1^0 < \theta''_1. \quad \ldots\ldots\ldots\ldots \quad (115)$$

It is obvious that if we are interested only in the unique lower estimate of θ_1 and want the probability of $\underline{\underline{\theta}}\,(E)$ falling short of the true value θ_1^0 to be equal to α, we should not mind $\underline{\underline{\theta}}\,(E)$ being smaller than θ''_1. Therefore, when choosing the function $\underline{\underline{\theta}}\,(E)$, we should not formulate any restriction concerning its satisfying the inequality $\underline{\underline{\theta}}\,(E) < \theta''_1$, provided the equation (114) is satisfied. The position with regard to θ'_1 is different. If $\underline{\underline{\theta}}\,(E)$ happens to be smaller than θ'_1, then it will also be

smaller than θ_1^0 and our statement concerning the value of θ_1 based on $\underline{\underline{\theta}}(E)$ will be correct. However, it would also be correct if, say,

$$\underline{\underline{\theta}}(E) = \tfrac{1}{2}(\theta'_1 + \theta_1^0) > \theta'_1, \quad \ldots\ldots\ldots \quad (116)$$

and in such a case it would be more accurate and would undoubtedly be judged more desirable. Generalizing the above conclusion, we could say that whenever we are interested in the unique lower estimate $\underline{\underline{\theta}}(E)$ of a parameter θ_1, we should require it to have the property that whatever $\theta'_1 < \theta_1^0$, the chance of $\underline{\underline{\theta}}(E)$ falling short of θ'_1 should be as small as possible, thus

$$P\{\underline{\underline{\theta}}(E) < \theta'_1 | \theta_1^0\} = \text{minimum} \quad \ldots\ldots\ldots \quad (117)$$

for all values of θ'_1 and θ_1^0 such that $\theta'_1 < \theta_1^0$. This condition implies that the region of acceptance $A_0(\theta'_1)$ corresponding to any value of $\theta_1 = \theta'_1$ should have the property

$$P\{E \varepsilon A_0(\theta'_1) | \theta_1^0\} \leqq P\{E \varepsilon A | \theta_1^0\}, \quad \ldots\ldots\ldots \quad (118)$$

whatever $\theta_1^0 > \theta'_1$ and whatever any other region A such that

$$P\{E \varepsilon A | \theta'_1\} = P\{E \varepsilon A_0(\theta'_1) | \theta'_1\} = \alpha. \quad \ldots\ldots\ldots \quad (119)$$

Similarly, if it were desired to find the unique upper estimate $\overline{\overline{\theta}}(E)$ of θ_1, the most desirable solution would be determined by the regions of acceptance, $A^0(\theta_1)$ such that

$$P\{E \varepsilon A^0(\theta'_1) | \theta_1^0\} \leqq P\{E \varepsilon A | \theta_1^0\}, \quad \ldots\ldots\ldots \quad (120)$$

whatever $\theta_1^0 < \theta'_1$ and whatever the region A satisfying (119).

If unique estimates determined by (118) and (119) or (120) and (119) exist, they will be called the best one-sided estimates of θ_1.

Following the recent results (NEYMAN and PEARSON, 1933, 1936, *a*) concerning the theory of testing hypotheses, it is easy to establish formulae giving the best one-sided estimates in many important problems. Of these I shall mention one.

(*c*) *Example III*

Consider the case where the probability law of the X's is normal

$$p(E|\xi\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\Sigma(x_i-\xi)^2}{2\sigma^2}} \quad \ldots\ldots\ldots \quad (121)$$

with unknown ξ and σ and where it is desired to estimate ξ. Following the lines indicated, it is easily found that the best one-sided estimates of ξ are given by

$$\left.\begin{aligned} \overline{\overline{\xi}}(E) &= \bar{x} + ts \\ \underline{\underline{\xi}}(E) &= \bar{x} - ts \end{aligned}\right\}, \quad \ldots\ldots\ldots\ldots \quad (122)$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n(n-1)} \quad \ldots\ldots\ldots \quad (123)$$

and t may be taken from Fisher's tables corresponding to $P = 2(1 - \alpha)$.*

(*d*) *Short Unbiassed Systems of Confidence Intervals*

We must now consider the important case where we are interested in the two-sided estimation in which the probability law of the X's is highly regular so that, owing to Proposition X, the shortest systems of confidence intervals do not exist. We must formulate the properties of confidence intervals which could be considered as particularly satisfactory.

We shall start with the obvious remark that, if possible, the value of the estimated parameter which in a particular case happens to be true, should be covered by the confidence interval more frequently than any other value.

Alternatively, we may express this by saying that for any given value of $\theta_1 = \theta_1^0$ the probability of its being covered by the confidence interval δ (E) should be greatest when θ_1^0 happens to be the true value of θ_1. Therefore, whatever $\theta'_1 \neq \theta_1^0$, it must be

$$P\{\delta(E)\, C\theta_1^0 | \theta_1^0\} = \alpha \geq P\{\delta(E)\, C\theta_1^0 | \theta'_1\}. \quad \ldots\ldots \quad (124)$$

We may express this in still another manner, considering the probability of θ_1^0 being covered by the confidence interval δ (E) as a function of that value of θ_1 which happens to be true,

$$P\{\delta(E)\, C\theta_1^0 | \theta_1\} = f(\theta_1). \quad \ldots\ldots\ldots\ldots \quad (125)$$

The formula (124) requires that the function (125) should be maximum for $\theta_1 = \theta_1^0$ and that that maximum should be equal to α.

It seems to be obvious that if there are many systems of confidence intervals in which, whatever θ_1^0, the probability (125) considered as a function of θ_1, is maximum for $\theta_1 = \theta_1^0$, we should choose the system by which this maximum is the steepest, so that, while the true value of θ_1 is being shifted away from θ_1^0, the chance of θ_1^0 being covered by δ (E) diminishes in the quickest way.

These conditions may now be expressed in terms of equivalent conditions concerning the regions of acceptance.

* The properties of the formulae (122) giving the best one-sided estimates of ξ were found by the author in about 1930. Subsequently, these properties, together with an outline of the theory of estimation, were included in his lectures first given at the University of Warsaw, then, from 1934, at the University College, London, and also in a course of lectures at the University of Paris in January, 1936. References to these formulae may be found both in Polish and English statistical literature. *See* for instance: (1) W. PYTKOWSKI: "The Dependence of the Income of Small Farms upon their Area, the Outlay and the Capital Invested in Cows". Warsaw, 1932. *See* particularly pp. 28–29; (2) CLOPPER and PEARSON (1934).

Let $A(\theta_1^0)$ be a region of acceptance corresponding to some value θ_1^0 of θ_1, so that

$$P\{E \varepsilon A(\theta_1^0) | \theta_1^0\} = \alpha \quad \ldots \quad (126)$$

whatever θ_1^0. We have

$$f(\theta_1) = P\{\delta(E) C\theta_1^0 | \theta_1\} = P\{E \varepsilon A(\theta_1^0) | \theta_1\} \quad \ldots \quad (127)$$

and the above conditions concerning the confidence intervals appear to be equivalent with the condition that the right-hand side of (127), considered as a function of θ_1, should be a maximum for $\theta_1 = \theta_1^0$ and that this maximum should be as sharp as possible.

In cases where the elementary probability law of the X's, integrated over any region, admits two differentiations with regard to θ_1 under the integral sign, this leads to the following :

Whatever θ_1^0, and *whatever the values of other unknown parameters*, $\theta_2, \theta_3, \ldots \theta_l$,

$$\left.\frac{\partial P\{E \varepsilon A(\theta_1^0) | \theta_1\}}{\partial \theta_1}\right|_{\theta_1 = \theta_1^0} \equiv \int \ldots \int_{A(\theta_1^0)} p'(E | \theta_1^0, \ldots \theta_l)\, dx_1 \ldots dx_n \equiv 0 \quad \ldots \quad (128)$$

$$\left.\frac{\partial^2 P\{E \varepsilon A(\theta_1^0) | \theta_1\}}{\partial \theta_1^2}\right|_{\theta_1 = \theta_1^0} \equiv \int \ldots \int_{A(\theta_1^0)} p''(E | \theta_1^0, \ldots \theta_l)\, dx_1 \ldots dx_n = \text{minimum}, \quad (129)$$

where p' and p'' denote the derivatives with regard to θ_1.

The system of confidence intervals having the above properties will be called the short unbiassed system. The possibility of determining such systems depends on the possibility of determining the regions of acceptance satisfying (126), (128), and (129). This problem has been recently dealt with in the case where the number of the unknown parameters involved in the probability law of the X's is equal to one (NEYMAN and PEARSON, 1936, *a*) and to two (NEYMAN, 1935, *b*).

In such cases as treated in the papers referred to, the construction of the short unbiassed systems of confidence intervals does not present any difficulties.

In particular, if the probability law of the X's is as in (121), then the short unbiassed system of the confidence intervals for ξ is given by the formula

$$\bar{x} - ts \leq \xi \leq \bar{x} + ts \quad \ldots \quad (130)$$

where t should be taken from Fisher's tables for $P = 1 - \alpha$.

IV—SUMMARY

The main problem treated in this paper is that of confidence limits and of confidence intervals and may be briefly described as follows. Let $p(x_1, \ldots x_n | \theta_1, \theta_2, \ldots \theta_l) = p(E | \theta_1, \ldots \theta_l)$ be the elementary probability law of n random variables $x_1 \ldots x_n$ depending on l constant parameters $\theta_1, \theta_2, \ldots \theta_l$. The letter E stands here for $x_1, \ldots x_n$. Suppose that the analytical nature of $p(x_1, \ldots x_n | \theta_1, \ldots \theta_l)$ is known but the values of the parameters $\theta_1, \ldots \theta_l$ are unknown. It is required to determine two

single-valued functions of the x's, $\underline{\theta}$ (E) and $\bar{\theta}$ (E) having the property that, whatever the values of the θ's, say $\theta'_1, \theta'_2, \ldots \theta'_l$, the probability of $\underline{\theta}$ (E) falling short of θ'_1 and at the same time of $\bar{\theta}$ (E) exceeding θ'_1 is equal to a number α fixed in advance so that $0 < \alpha < 1$,

$$P\{\underline{\theta}\,(E) \leq \theta'_1 \leq \bar{\theta}\,(E) | \theta'_1, \theta'_2, \ldots \theta'_l\} = \alpha. \quad . \quad . \quad . \quad . \quad . \quad (131)$$

It is essential to notice that in this problem the probability refers to the values of $\underline{\theta}$ (E) and $\bar{\theta}$(E) which, being single-valued functions of the x's, are random variables. θ'_1 being a constant, the left-hand side of (131) *does not* represent the probability of θ'_1 falling within some fixed limits.

The functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are called the confidence limits for θ_1 and the interval ($\underline{\theta}$ (E) , $\bar{\theta}$ (E)) the confidence interval corresponding to the confidence coefficient α.

The problem thus stated has been completely solved for the case where $l = 1$, and it is found to possess an infinity of solutions. If $l \geq 2$ the solution obtained is limited to the case where there exists a sufficient set of statistics for $\theta_2, \theta_3, \ldots \theta_l$ and then again there is an infinity of solutions.

Methods were indicated by which it is possible to find among all possible solutions of the problem the one giving the confidence intervals which are shorter (in a sense defined in the text) than those corresponding to any other solution.

The confidence limits $\underline{\theta}$ (E) and $\bar{\theta}$ (E) may be looked upon as giving a solution of the statistical problem of estimating θ_1 independent of any knowledge of probabilities *a priori*. Once $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are determined corresponding to a value of α close to unity, say $\alpha = 0{\cdot}99$, the statistician desiring to estimate θ_1 may be recommended (1) to observe the values of the random variables $x_1, \ldots x_n$, (2) to calculate the corresponding values of $\underline{\theta}$ (E) and $\bar{\theta}$ (E), and (3) to state that the value of the parameter θ_1 is within the limits $\underline{\theta}\,(E) \leq \theta_1 \leq \bar{\theta}\,(E)$.

The justification of this recommendation lies in the fact that the three steps described are equivalent to a random experiment which may result either in a correct or in an erroneous statement concerning the value of θ_1, the probability of a correct statement being equal to $\alpha = 0{\cdot}99$. Thus the statistician following the above recommendation is in a position comparable with that of a game of chance with the probability of winning being equal to $\alpha = 0{\cdot}99$.

The method followed to determine the confidence limits for a single parameter permits an obvious generalization for the case where the number of parameters to be estimated simultaneously is greater than one.

Three previous publications concerning the confidence intervals for which I am either partly or wholly responsible (NEYMAN, 1934, MATUSZEWSKI, NEYMAN, and SUPIŃSKA, 1935, NEYMAN, 1935, *c*) refer to the simplest case where the number of random variables and that of the parameters to be estimated are equal to unity. The problem considered here is therefore far more general and also it is treated differently. Previously, the parameters to be estimated were considered as random variables following an arbitrary probability law which could be continuous or not and, even,

could reduce to unity just for one particular value of the parameter, being zero elsewhere. This arbitrariness of the probability law of the parameters served as an excuse, but the very assumption of its existence seemed to be an artificiality from which the present method of approach is entirely free.

Subsidiary results obtained include a method of constructing similar regions which is more general than the one known previously and the Proposition X bearing on the theory of testing hypotheses. It emphasizes the rareness of cases where there exists a uniformly most powerful test.

V—References

Borel, É. 1910 " Eléments de la Theorie des Probabilités," Paris.

—— 1925 " Principes et formules classiques," 1 fasc. du tome I du " Traité du Calcule des Probabilités et des ses Applications," Paris.

—— 1926 " Applications à l'Arithmétique," Ibidem fasc. 1, tome II.

Bortkiewicz, L. v. 1917 " Die Iterationen." Berlin.

Clopper, C. J., and Pearson, E. S. 1934 " Biometrika," **26**, 404–413.

Darmois, G. 1936 " Méthodes d'estimation," Paris.

Doob, J. L. 1934 ' Trans. Amer. Math. Soc.,' **36**, 759.

Dugué, D. 1936 ' C.R. Acad. Sci. Paris,' **193**, 452, 1732.

Fisher, R. A. 1925 ' Proc. Camb. Phil. Soc.,' **22**, 700–725.

Fréchet, M. 1937 " Recherches théoriques modernes sur le calcul des probabilités. Traité du Calcul des Probabilités et de ses Applications," Fasc. 3, tome 1, Paris.

Hopf, E. 1934 ' J. Math, Phys. Mass,' **13**.

Hotelling, H. 1932 ' Trans. Amer. Math. Soc.,' **32**, 847–859.

Jeffreys, H. 1931 " Scientific Inference," 1935.

Kolmogoroff, A. 1933 " Grundbegriffe der Wahrscheinlichkeitsrechnung," Berlin.

Lévy, P. 1925 " Calcul des Probabilités," Paris.

Łomnicki, Z., and Ulam, S. 1934 ' Fund. Math.,' **23**, 237–238.

Markoff, A. A. 1923 " Calculus of Probability " (Russian ed. iv), Moscow.

Matuszewski, T., Neyman, J., and Supińska, J. 1935 ' Supplement to J. Roy. Stat. Soc.,' **1**, 63–82.

Neyman, J., and Pearson, E. S. 1933 ' Phil. Trans.,' A, **231**.

—— 1936, *a* ' Stat. Res. Mem.,' **1**, 1–37.

—— 1936, *b* ' Stat. Res. Mem.,' **1**, 113–137.

Neyman, J. 1935, *a* ' G. Inst. Ital. Attuari,' **6**, 320.

—— 1935, *b* ' Bull. Soc. Math. Fr.,' **63**, 248–266.

Neyman, J. 1935, *c* ' Ann. Math. Stat.,' **6**, 111–116.

—— 1934 ' J. Roy. Stat. Soc.,' **97**, 589–593.

Pearson, K. 1895 ' Phil. Trans.,' **187**, 253–318.

Tippett, L. H. C. 1927 " Tracts for Computers," No. 15, Camb. Univ. Press.

Psychological Bulletin
1995, Vol. 118, No. 2, 193–198

0033-2909/95/$3.00

Tips for Writing (and Reading) Methodological Articles

Scott E. Maxwell and David A. Cole
University of Notre Dame

One reason many methodological articles are not very intelligible to their readers is because the content is often inherently difficult. However, a contributing factor in some cases is the tacit assumption that rules of good writing cease to apply when writing about statistics. The authors of this article argue that good writing becomes even more important as the content of the article becomes more complex. Furthermore, they believe that additional rules pertain to writing methodological articles and highlight various ways that methodological article authors can make their work more accessible (and less painful) to researchers who are not methodological specialists. The authors also suggest how nonspecialists can most effectively approach the task of reading a quantitative article.

For some psychologists, writing a methodological article is a fine art of obfuscating needlessly tedious and complex trivia. For others, reading a methodological article ranks right up there with a visit to the dentist's office. Many methodological articles, however, are not accessible to their intended readers, not necessarily because the material is so sophisticated but because the presentation of the material is so obtuse. Our goal in this article is to provide a few suggestions for writing methodological articles. Excellent articles are available on the writing of general psychology articles (e.g., Bem, 1987; Sternberg, 1988, 1992). Hence, we try to avoid repeating these points, except to say that all the rules for good nontechnical writing are at least as important for good technical writing if only because the material is often more complex. Our specific focus is on writing methodological articles for nonspecialists, although some of our comments may also pertain to authors who target specialists.

Quantitative methods articles in psychology take many different forms. Some articles are similar to substantive *Psychological Bulletin* articles insofar as they are literature reviews. The authors of these articles typically synthesize relevant methodological literature or present new statistical methods in a format that is appropriate to a nonstatistical audience. Other authors present the results of original research. The topics range from evaluations and comparisons of current statistical technologies to developments and introductions of qualitatively new research methodologies. Such articles may include highly technical mathematics or extensive computer simulation. Because of the diversity of these articles, we attempt to make points that are useful to as wide a range as possible of current and future methodological article authors.

Preparation

Defining Your Audience

"Perhaps the most important principle of good writing is to keep the reader uppermost in mind" (Knuth, Larrabee, & Roberts, 1989, p. 3). This principle is especially important in technical writing, where your audience may be remarkably diverse, ranging from methodologists who specialize precisely in the topic under investigation to researchers in very different fields who hope to apply a specific new technique in their next study.

Authors often overlook the fact that they wield considerable control over their readership by carefully choosing the journals to which they submit their work. At least three questions should be considered when selecting a journal in which to publish a methodological article. First, how technical is your presentation? The perfect article for a highly technical outlet such as *Psychometrika* may be almost unintelligible to the majority of *Psychological Bulletin* readers. Many journals (*Psychological Bulletin* included) explicitly proscribe the use of complex mathematics, such as calculus or matrix algebra. If not, the editor either requests the author to find a more accessible way to make the points or suggests to the author to submit the work to a more technical journal. Second, how specific is your methodological point? Among methodological journals, some (e.g., *Psychological Bulletin*) target a readership that uses a wide variety of methodologies. In general, articles in which highly specific points about a particular statistical technique are made belong in more specialized methodological journals (e.g., *Structural Equation Modeling*). If the point is more general or pertains to a wider variety of research paradigms, then broader methodological outlets may be more appropriate. Third, how specific are the implications of your article for a particular subdiscipline of psychology? Articles submitted to journals with broad readerships should have implications for researchers almost irrespective of their content area. Even when the technical level of the presentation is low, authors must still face the question of whether the practical implications of the article are broad enough to warrant publication in a journal such as *Psychological Bulletin* or whether a more specialized substantive journal might be more appropriate. Many area journals publish occasional methodological articles (e.g., *Journal of Applied Psychology*), have special sections on methodological advances (e.g., *Journal of Consulting and Clinical Psychology*), or even publish special issues on methodology (e.g., *Journal of Counseling Psychology* and *Journal of Family Psychology*). Consequently, an article on a specific topic, such as reaction times in cognitive tasks, would probably fit well in a cognitive journal,

Scott E. Maxwell and David A. Cole, Department of Psychology, University of Notre Dame.

Correspondence concerning this article should be addressed to Scott E. Maxwell, Department of Psychology, University of Notre Dame, Notre Dame, Indiana 46556.

whereas an article on reaction time research in general might cut across disciplines and thus be more appropriate for a journal with a broader readership.

After selecting a journal, continue to strive to write for as broad an audience as possible. Failure to relate specific methodological points to the variety of situations to which they might pertain unnecessarily limits the impact of the article. Use examples from diverse research areas; refer the reader to wide-ranging applications of your procedure; and elaborate on the implications of your methodology for diverse research paradigms. Pitching your article to too narrow an audience may not get it the attention it deserves.

Most articles have multiple audiences. A hierarchical structure permits an article to be read for its general ideas by some readers and for its specific details by others. Presenting a general overview of the problem and the solution early in the article enables all readers to walk away with the overall gist of the message. Then, increasing the amount of detail as the article progresses allows readers to go as far as they want (or need) into the intricacies of the methodology. At the same time, authors and readers alike need to be sensitive to the dangers of stopping too soon. Authors might motivate readers to persevere by issuing periodic cautionary notes that describe potential hazards of implementing this new technique (among other things) before reading the next section.

Obtain feedback on the article from a variety of sources. For example, sharing a draft of the article with other authors who have written articles in the same general area may provide valuable expert feedback. It may be especially useful to seek the opinions of individuals whose expertise and perspective differ from your own. For example, some authors may benefit from involving a methodological expert who can ensure the technical accuracy of the article. All authors may benefit from the input of a knowledgeable nonspecialist, who can endow the work with a healthy respect for some of the readers' primary concerns, paraphrase statistical jargon, enrich the article with substantive examples from nonquantitative journals, and maintain a focus on the article's practical implications.

Motivating the Reader

Most psychologists are content to continue plying the traditional statistics and methodologies learned in graduate school. A pretty serious wake-up call is needed to alert psychology authors to new alternatives. Before proving anything with numbers and formulas, prove to the reader that what you propose can make a real difference. A specialist who encounters your article may immediately appreciate the relevance and potential importance of your article simply by reading the title and the abstract. The nonspecialist, however, is likely to need more guidance. Consequently, be as explicit as possible about the purpose of the article. Furthermore, make the point as early as possible in the article; otherwise, many readers may not struggle beyond the first paragraph or even the abstract.

To some extent, the point is the same as Sternberg's (1992) advice that all psychology authors should "tell readers why they should be interested" (p. 12). This point is even more important when writing a methodological article, however, if only because there is likely to be a larger gap between the author's background and the reader's. The author may be drawn to the topic because of its theoretical elegance or mathematical challenge, whereas readers are more likely to be interested in knowing whether this article means that they should design their studies differently or analyze their data with a new technique.

As Knuth et al. (1989) stated, "present the reader with something straightforward to start off with" (p. 76). Hand the readers a statement that explains what the article is about and why they should read it. Most *Psychological Bulletin* articles have one of the following points at their center:

1. Methodological advances allow interesting questions to be answered that previously were not amenable to a solution.
2. Here is a way to increase your statistical power.
3. You may not be testing the hypothesis you thought you were.
4. If you have data that depart from standard assumptions, there may be better ways to analyze your data.
5. A new statistic is better than the standard statistic.

Remember, presenting a new solution is of little value if the reader does not understand the problem yet. A voluminous review of every nuance of a methodological conundrum is unlikely to hold anyone's interest unless one is working on the particular problem. If the problem is truly important, an author should be able to state in a few sentences at the beginning of the article what the problem is, why it is important for psychologists, and why it has been difficult to solve.

Reviewing the Literature

Stipulating that prospective authors conduct a thorough literature search prior to formulating a methodological article is hardly an earthshattering notion. Less obvious, however, is that searching the relevant literature for methodological articles is often quite different from reviewing the literature for substantive articles. The multidisciplinary nature of methodology requires that the researcher be familiar with previous work in a variety of other disciplines. What appears to be a new statistical technique in psychology may have already been proposed in the statistics literature. The *Current Index to Statistics* (American Statistical Association, 1994), an annual keyword index, is extremely useful for identifying relevant statistical literature on a particular topic.

Quantitative psychologists must also be aware of the methodological literatures in other social sciences. For example, authors on structural equation modeling often must be familiar with recent advances that have appeared in sociology literature (such as *Sociological Methods and Research* and *Sociological Methodology*). Finally, methodologists must be cognizant of the ideas transmitted to the next wave of researchers through recent methodology textbooks. Articles that critique methodologies from texts published a decade ago are not of much value if those presentations no longer appear in more recent books. Similarly, articles that constitute pedagogical reviews of already published methodologies must differ substantively from modern textbook presentations of the same material. Synthesizing literature that has heretofore appeared exclusively in specialized methodology journals may be quite valuable. Once new methodologies appear in textbooks, however, they are likely to be inappropriate journal topics even if previous literature reviews have not appeared in journal format.

Occasionally, relevant literature lurks in unexpected places.

In statistics, problems can sometimes be transformed in such a way that they take on an entirely different appearance (even though they are technically unchanged). Under the alternative guise, new literature, if not new insights, may be hiding.

Communicating Technical Material

Many psychologists' worst adult memories are from their first graduate statistics class. With a few well-chosen mathematical proofs and equations, you have the power to dredge up nightmares of endless take-home exams and to rekindle feelings of deep-seated insecurity—not exactly the recipe for tempting the reader past the first few opening paragraphs of your article. You might rationalize that these simply are not the people who will read your article anyway, but that is precisely the (unfortunate) point.

Some authors appear to operate from the assumption that clarity and rigor represent opposite ends of the same dimension: These authors argue that if everyone can understand their arguments, then their points must not have much insight. Certainly some arguments require a great deal of prior knowledge without which even the clearest prose fails to be comprehensible. Nevertheless, it does not follow that clarity and rigor are enemies of one another. The author must adopt a different attitude, such as by wondering how he or she can make this inherently difficult (and potentially tedious) material as accessible as possible.

Clarity is especially critical in technical writing where the presentation of ideas is usually cumulative. If the author does not communicate the first points clearly, readers will probably be lost and therefore be unable to appreciate the remainder of the article. Be aware of what the reader knows because either the material has already been presented in the article or some background knowledge can be safely assumed (Knuth et al., 1989). If your article is closely related to an earlier article, it is usually necessary to summarize the major points of the previous article in considerable detail. Do not expect readers to be familiar with recent articles, and do not require them to read the articles before they can comprehend yours. Good advice is generally to start at a lower technical level than you would think. Even more difficult, however, is to anticipate what the reader expects next. Prepare the reader for the relations between different sections of the article so that individual pieces become a coherent whole.

Presume that many readers will skim (or altogether skip) anything that even slightly resembles an equation. Why fight it? Too much mathematical material in an article written for nonspecialists may effectively reduce actual readership to zero. The most obvious solution is to relegate technical details to an appendix. This is frequently a useful strategy; however, authors must take care that the main message of the article is clear even to those who do not read the appendix.

At times, equations are necessary for the main message of the article, in which case they should not be placed in an appendix. Indeed, a statistics article in a specialized journal may (and perhaps should) contain as many equations as words. When it comes time for the unavoidable mathematical argument, consider a few simple steps:

1. Tell the reader what you are going to show and why it is important.
2. Define your terms clearly when you first introduce them (and do not be afraid to remind the reader of key terms along the way).
3. Within the mathematics section, do not forget that you can use words too. Phrases such as "substituting Equation 3 into Equation 4 produces the following" are far superior to phrases such as "it follows that" or insults such as "obviously." Remember, too, that symbolic expressions are parts of sentences and should be punctuated as such as well.
4. Pause periodically to explain particular equations and comment on how they fit into the big picture.
5. At the end of the mathematics section, provide a verbal summary of the main points and why they are important.

Formulas can often be made more comprehensible by the presentation of "special cases." For example, some formulas may become simpler when sample size becomes extremely large. Simplifications may also arise when certain terms are assumed to be equal to one another or to zero. Yet another simplification sometimes emerges when a formula is written for the special case of two groups or in its univariate form instead of the more general multivariate form. Even if the rest of the article uses the more complex form of the formula, readers will usually find this presentation to be more meaningful if they have been able to grasp the essential meaning of the formula through special cases.

Of course, authors must also exercise good judgment about how much verbal explanation surrounding the mathematical presentation will be useful to readers. Unnecessary verbiage simply slows readers down and can make concentrating on the major points more difficult. On a related point, although word variety can reduce repetition and subsequent boredom, technical terms should generally not be interchanged even when they have the same precise meaning because many readers may not know whether the change in working reflects a change in meaning.

Notation

The wise use of symbols in a quantitative article provides a clear and parsimonious form of communication. It is much simpler for both the reader and the author to write σ_j instead of "population standard deviation within group j." Whereas the advantage is most obvious in equations, the careful use of symbols in text can also prevent awkward and excessive verbiage. Careless or thoughtless notation, however, may frustrate the most dedicated reader even when the expository text of the article is exemplary. A few straightforward rules go a long way to ensure that symbols help rather than hinder the reader. For example, providing an explicit definition of each symbol when it is first introduced is essential. Even something as seemingly straightforward as n may need to be defined. Although the American Psychological Association's *Publication Manual* (1994) stipulates that n be used to denote sample size within a group and N be used to denote total sample size, some readers may not be aware of this notation. Even when the initial meaning is explicit and clear, readers may benefit from an occasional reminder of what a symbol represents, especially if it has not been used for several pages. Also helpful is to take advantage of mnemonic coding wherever possible. Standard notation should be used if it has been established. The *Publication Manual* (1994) provides an extensive list of common statistical abbrevi-

ations and symbols. Even when standard notation does not exist, it is still important to follow general conventions, such as using Greek letters to represent population parameters and Latin letters for sample statistics. Needless to say, the same symbol should never be used to represent two different concepts, nor should two different symbols be used to represent the same concept. Finally, authors must be aware of the need to balance the parsimony obtained from symbols with the added burden placed on readers to remember what each symbol represents. In general, the best advice is to use as few symbols as possible.

Examples and Figures

A mathematician's natural tendency is to derive the most general form of an expression first and only then consider special cases. This strategy can be effective in articles written for nonspecialists if the author explains the general problem thoroughly and builds a compelling case for needing the general form in the first place. Nonspecialists, however, often crave a few special cases as appetizers, which then whet their appetite for the most general case. Although this sequence is typically less elegant mathematically, beginning with concrete examples may allow nonspecialists to follow the underlying logic more easily. This approach is similar to the *particular–general–particular* teaching technique recommended by Rourke (as cited in Mosteller, 1980). To explain an abstract idea, begin with a specific example that motivates the need to develop a solution to the problem. A general approach to the problem can then be considered along with a general solution. A sense of closure and full understanding may be absent, however, unless the general principles are followed by their application to a specific problem.

Using the particular–general–particular strategy is often consistent with using appropriate examples. Numerical examples are especially helpful in methodological articles. Authors can fulfill the first step of Rourke's (cited in Mosteller, 1980) strategy by providing an initial discussion of a problem in need of a solution. Once the author has presented the general solution, the initial problem can be revisited through a numerical example. A dilemma facing the author is to make the example complicated enough to be realistic and yet simple enough to illustrate the general methodological principle clearly. At times, the best resolution of this dilemma may involve a succession of increasingly complicated examples (see Cole, 1987). Ideally, examples also provide sufficient information to allow readers to work through computations or programming themselves, so they can check the accuracy of their understanding as well as their ability to apply procedures to actual data. Sometimes providing a numerical example on the basis of a small number of cases is either so unrealistic as to be misleading or it is simply infeasible. However, authors should be aware that useful alternatives may exist in these cases. For example, Willett and Sayer (1994) provided complete longitudinal data on a subsample of cases and effectively integrated their presentation of the subsample with their discussion of the actual total sample. For some types of problems, presenting the sample covariance matrix (or other summary statistics) may be sufficient to allow readers to duplicate the authors' results (see MacCallum & Browne, 1993, for an example). Willett and Sayer's inclusion of the LISREL program code in an appendix also illustrates an additional approach for helping readers to check their understanding of the proposed method and to use it appropriately for their own data. Although examples are often essential for clear communication, both authors and readers must understand that examples in and of themselves do not establish desirable properties of a proposed method.

The juxtaposition of specific and general issues may be ideally suited for methodological articles that demonstrate how advances in computer software can offer new methodological opportunities. The impact of such a presentation can usually be greatly increased by couching the presentation in terms of more general methodological issues. Try to use software examples to illustrate fundamental methodological principles. Good examples are O'Brien and Kaiser's (1985) demonstration of how syntax choices in SPSS multivariate analysis of variance yield different analyses in repeated measures designs and Bryk and Raudenbush's (1987) discussion of how hierarchical linear modeling addresses basic questions in the analysis of change. The combination of computer software, a broad consideration of more general quantitative issues, and specific numerical examples enables readers to not just use the statistical program but also better understand the advantages and disadvantages of various data analytic strategies.

Another useful tool for communicating technical material is the use of figures. Figures may be useful for showing results from numerical examples or for displaying the results of simulation studies. An often overlooked advantage of figures, however, is their use for depicting mathematical relationships. Plotting mathematical functions often illuminates the meaning underlying an abstract mathematical expression. For example, some of our own work (Maxwell, 1994; Maxwell, Cole, Arvey, & Salas, 1991) illustrated how contour plots can show the meaning and practical implications of mathematical derivations. Recent advances in graphics software open the door to a multitude of possibilities for visual representations of multivariate data and relationships. Methodologists should be at the forefront of advances in graphics (see Cleveland, 1985, 1993; Tufte, 1983, 1990).

Simulation Studies

Much of the methodological work submitted to psychology journals involves simulation studies. Simulations can be extraordinarily valuable because they allow the author to describe properties of statistics under suboptimal conditions where underlying assumptions have not been met. As a consequence, mathematical derivations of properties may be cumbersome if not impossible. Effective communication of simulation studies involves special considerations beyond those of other methodological articles; simulation studies are experiments and must be described and interpreted in this light.

For example, careful thought must be given to the selection of specific parameter values to manipulate. An infinite number of ways exist for distributions to depart from homoscedasticity. How does the author select a realistic sample of distributions to examine? Although there is no simple answer, some sources can supply useful evidence of the types of distributions obtained in actual empirical work in the behavioral sciences (e.g., Micceri, 1989). Sawilowsky and Blair (1992) provided an example of how this type of information can be incorporated into the de-

sign of simulation studies. Of course, previous simulation studies in related areas can also provide a useful framework for selecting conditions to simulate.

As in all experiments, the author should be prepared to interpret the results obtained from the specific parameter values in the context of a broader theoretical framework. For example, the specific results obtained with exactly 20 or 50 participants per group in the simulation are valuable only to the extent that the author can establish a case for generalizing the findings to other sample sizes (even if these specific values were not included in the simulation). The author must also plan an appropriate number of simulation replications so that obtained results are sufficiently precise. Obtaining 8 significant results out of 100 simulated replications at an alpha level of .05 does not necessarily indicate that the test under consideration is liberal. The excessive error rate might simply reflect sampling error. Many replications are quite appropriate when a high degree of precision is required.

Simulation studies typically produce an enormous amount of data. After doing all of the work to generate the data, the author may be tempted to show the reader all the results of this massive effort. Authors, however, must distill this mass of information down to its essence, especially for a nonspecialist readership. Most important, the author must decide what conclusions emerge from systematic patterns in the data and organize the presentation of results accordingly. In addition to typical reports of proportions, means, and standard errors, Maxwell (1980) illustrated how correlates of the primary statistics can provide an even broader context for interpreting results obtained for the selected parameter values. Other approaches for establishing a broad framework include making an approximate argument (see the appendix of Hedges & Olkin, 1984, for an example) and using exact theory for simplified cases and developing large sample theory (see Hedges, Cooper, & Bushman, 1992, for both of these approaches). In addition, Harwell (1992) discussed methods for integrating results from simulation studies, which are valuable ideas for the prospective simulation researcher.

A final (or "first") concern for simulation studies is that they are sometimes completely unnecessary. Authors occasionally fail to appreciate the value of the analytic proof. If properties of a statistic can be derived mathematically under specified conditions, then there is no need to study the statistic through simulations under these same conditions. Such simulations add no information whatsoever to what is already known mathematically. Such simulations only serve to validate the algorithms used in the simulation itself. Thus, including such conditions in a simulation may be useful to verify that the simulation is correct under baseline conditions. Authors should not, however, make the mistake of inferring that these results are informative in and of themselves.

Once Burned, Twice Shy

Identify the limits of your findings early in the article. Imagine a reader's frustration at having plowed through a statistical treatise on distribution-free alternatives to maximum likelihood structural equation modeling only to discover at the end of the article that the sample size requirements are 10 times what the reader usually has available. Trudging through a second methodological masterpiece may not end up very high on this reader's list of things to do.

One frequent way in which limitations manifest themselves is through assumptions. Unfortunately, authors sometimes fail to state assumptions explicitly. Without a clear statement of assumptions, the reader has no starting point for statistical claims made in the article. Although a detailed statement of assumptions might best appear in an appendix, most articles would benefit from a general overview of the assumptions near the initial statement of the problem and proposed solution.

In a related vein, authors should avoid the temptation to present a new methodology as a panacea. In all likelihood, any new method carries with it some disadvantages as well as advantages. Authors do readers a disservice when their presentation is one sided. Although a certain degree of enthusiasm is understandable and even desirable, balance is also important.

Tips for Reading Methodological Articles

Not surprisingly, many of the tips for writing methodological articles apply equally well to reading quantitative articles. Ideally, the goals of the author and the reader are virtually identical. In many cases, the advice for authors can be generalized to readers simply by substituting *reader* for *author*.

Just as authors should often strive for a hierarchical structure, readers may also benefit from approaching a methodological article hierarchically. For many readers, attempting to read a technical article word for word from beginning to end is a guaranteed prescription for frustration. Instead, it is often far better to skim the article initially to develop a broad understanding of the article. A second reading might involve close reading of the introduction and the conclusion, again simply skimming the details of the justifications for the conclusions. Only on the third reading might there be any serious attempt to begin to understand the details of the actual argument. In any case, readers should frequently expect that they will need to reread methodological articles before they feel comfortable with their understanding of the material. Throughout this process, it is often helpful to take notes on the key points of the introduction and conclusions as well as on the basis for the conclusions. Similarly, readers can benefit from making a list of symbols and brief descriptions of what they represent.

Just as authors can improve clarity of technical points by presenting special cases, readers can also check their understanding of such points by considering special cases even if the author does not provide them for the reader's convenience. Along these lines, readers can also attempt to reproduce the results from a numerical example. Finally, when all else fails, readers can ask for help. Just as authors usually benefit from the advice of someone with a different perspective, readers may also discover that sharing an article with a colleague allows both individuals to reach a higher level of understanding.

Summary

Attention to fundamental rules for good writing is especially important when writing articles on methodology or statistics. Such basic rules are insufficient, however. Additional concerns arise as the content of psychological articles becomes increasingly technical or mathematical. With an eye toward improving

the written presentation of methodological material, we outline a number of tips for technical writing:

1. Keep the reader uppermost in mind.
2. Select the journal for your article carefully.
3. Write for as broad an audience as possible.
4. Obtain feedback from someone whose expertise and perspective is different from your own.
5. Be especially clear at the onset because methodological presentations are often cumulative.
6. Convince your reader that it is important to read this article.
7. Be aware of the unusually diverse literature that is relevant to methodological articles.
8. Be sensitive to what your readers do and do not know.
9. Strategically define your symbols.
10. Encapsulate and clearly summarize technical material.
11. Consider using the particular–general–particular approach to technical presentations.
12. A figure is worth a thousand equations.
13. Keep the work relevant to real-world situations.
14. Be mindful of the value of mathematical proofs.
15. Confess the limitations and shortcomings of even the best new methodologies.

Needless to say, following these guidelines and 100 others will not guarantee publication. The packaging will make the product pretty, it will get the article read, and it will help the material to be understood, but the bottom line will always be the quality of the authors' ideas and their ultimate relevance to psychological research.

References

American Psychological Association. (1994). *Publications manual* (4th ed.). Washington, DC: Author.

American Statistical Association. (1994). *Current index to statistics.* Alexandria, VA: Author.

Bem, D. J. (1987). Writing the empirical journal article. In M. P. Zanna & J. M. Darley (Eds.), *The complete academic* (pp. 171–201). New York: Random House.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–158.

Cleveland, W. S. (1985). *The elements of graphing data.* Monterey, CA: Wadsworth.

Cleveland, W. S. (1993). *Visualizing data.* Summit, NJ: Hobart Press.

Cole, D. A. (1987). The utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55,* 584–594.

Harwell, M. R. (1992). Summarizing Monte Carlo results in methodological research. *Journal of Educational Statistics, 17,* 297–313.

Hedges, L. V., Cooper, H., & Bushman, B. J. (1992). Testing the null hypothesis in meta-analysis: A comparison of combined probability and confidence interval procedures. *Psychological Bulletin, 111,* 188–194.

Hedges, L. V., & Olkin, I. (1984). Nonparametric estimators of effect size in meta-analysis. *Psychological Bulletin, 96,* 573–580.

Knuth, D. E., Larrabee, T., & Roberts, P. M. (1989). *Mathematical writing.* Washington, DC: Mathematical Association of America.

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin, 114,* 533–541.

Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational Statistics, 5,* 269–287.

Maxwell, S. E. (1994). Optimal allocation of assessment time in randomized pretest–posttest designs. *Psychological Bulletin, 115,* 142–152.

Maxwell, S. E., Cole, D. A., Arvey, R. D., & Salas, E. (1991). A comparison of methods for increasing power in randomized between-subject designs. *Psychological Bulletin, 110,* 328–337.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

Mosteller, F. (1980). Classroom and platform performance. *The American Statistician, 34,* 11–17.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin, 97,* 316–333.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111,* 352–360.

Sternberg, R. J. (1988). *The psychologist's companion* (2nd ed.). New York: Cambridge University Press.

Sternberg, R. J. (1992). How to win acceptances by psychology journals: 21 tips for better writing. *APS Observer, 5,* 12–13, 18.

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Tufte, E. R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116,* 363–381.

Received May 18, 1994
Revision received October 26, 1994
Accepted October 26, 1994 ■

Teacher's Corner

How to Read the Statistical Methods Literature: A Guide for Students

James R. MURPHY

Statistical methods papers are densely written. The writers assume that the readers already have sophisticated knowledge of the topic. In addition, a standard statistical notation has not been developed. Students who learn a technique in one notation may be confused when reading articles written with a different notation. This paper contains suggestions for making the student's task easier and more productive.

KEY WORDS: Pedagogy; Reading statistical methods; Teaching statistics.

1. INTRODUCTION

Several guides tell the nonstatistician how to read and interpret applied statistical results. Huff (1954) gives five points for the skeptical reader to keep in mind. Sackett (1991) and Colton (1979) each provide check lists to determine whether statistical methods are appropriately used in medical articles. There are also guides for reading papers containing complex mathematics, including Cowen (1991), Phanstiel (1990), Parke (1958), and Pemberton (1969). Schechtman (1987) suggests teaching biostatistics through reading the medical literature. All of the references given above provide useful information for reading technical material. However, there are specific problems in reading the statistical methods literature that these articles do not address. In this paper I present the outline of a general method for organizing such reading.

The statistical literature presents several challenges. First, various skills are needed in reading statistical methods: basic language skills, knowledge of statistical notation, algebraic skills, and, increasingly, some recognition of how computers function. Second, technical articles are rarely models of expository style. They tend to rely heavily on technical jargon and on an interplay between the written word and written notation. A concept that is difficult to explain in language is sometimes easily explained in notation, but the resulting dissonance between the two can make the article difficult to understand. Third, advanced articles assume that basic concepts do not need to be explained. Fourth, we are most comfortable with notation we learned in class. A different notation can be confusing. Finally, many papers discuss both statistical theory and the computational techniques necessary to implement the theory. The theory and computation are not always clearly separated.

Every statistician has to come to grips with these problems in reading the literature, but our individual solutions do not get passed on to students in any formal way. This means that each new group of students has to develop a way of dealing with these problems. In this paper I outline strategies that I use to read statistical methods papers. These suggestions were developed from trial and error, discussion with colleagues, and suggestions from Polya (1945). The outline was written for the student, and is intended for any class that requires reading papers from the literature. The first time that I give this outline to a class I ask students to use the outline while reading two articles that I select from journals such as *Statistics in Medicine* or the *Journal of the American Statistical Association*. The articles selected deal with the topic of the class, and are intended to be slightly above the knowledge level of the average student. Students write a summary of each article, answering the questions given in the outline. I have made no formal evaluation of the outline, but informal discussions with students suggest that it does help them to read the articles.

2. HOW TO READ THE STATISTICAL METHODS LITERATURE

2.1 Right Attitude and Environment (It is a Long Process; Be Comfortable)

I applaud everyone who finds it easy to read statistical methods papers. For the rest of us it is best to start with the right attitude: "This is going to take some time." I would set aside 4 hours at a minimum for a simple paper, and considerably longer for more complicated work. This does not have to be in a single large block of time, but it gives you an idea of the total amount of time that it might take. It helps to have a comfortable environment in which to work. A comfortable chair, good light, pencil, paper, and possibly a computer are useful accessories.

2.2 Focus on Why You are Reading the Article

The adage that "You can't see the forest for the trees" often applies when you read a complicated article. Before you begin to read, you should determine why you are reading this article. Focus on that main point.

James R. Murphy is Professor of Biostatistics, Department of Preventive Medicine and Biostatistics, University of Colorado Health Sciences Center, Denver, CO 80262.

Table 1. Sample Entry for a Bibliographic Database

Reference: Andrews, D. F. (1971), "Sequentially Designed Experiments for Screening Out Bad Models with *F*-tests," *Biometrika*, 58(3), 427.
Statistical theories involved: Linear models, sequential designs, sequential *F* tests.
Computational techniques used: Generation of random normal deviations, setting up a design space and choosing sample points based upon accumulating data, simulating data.
Distributional assumptions: Normal distribution; requires replicate measures.
Other relevant assumptions: The models discussed here exist in a hierarchy of polynomial models, and you are trying to choose the best order for the polynomial.
Dataset used: Simulated data.
Relevant cross-references: None dealing with designing a space for the experiment.
Notes: Possible use in Phase I trials or selecting models for decline rates in repeated measures. Basic proposal is to select design spaces that will allow you to determine which of a set of possible models is invalid, and then run your experiment on the most valid model.

A statistician has three basic reasons to read an article: general interest, relevance to a particular application, or broader knowledge of a specific statistical method. Gleser's (1986) suggestions for a fourth purpose of refereeing a paper are consistent with the advice given in this paper.

These reasons are not mutually exclusive. However, it helps to focus on one reason for reading a particular article. If the paper is of general interest, you would focus on the introduction and background. If it has information about an application, you would focus on the results and data section. Reading to improve your knowledge about statistical methodology requires the most comprehensive examination of the paper.

Whatever your reason for reading the article, it helps to start with Huff's first point: "*Who said it and where?*." Do the authors already have an established reputation in this methodology? Is the article published in a journal where these methods are likely to have had rigorous editorial scrutiny?

2.3 State the Problem in Your Terms

Read enough of the *abstract, introduction,* and *discussion* so that you can state the problem in a sentence or two. State the problem in notational terms with which you are familiar. Sketch a possible way to solve the problem (or several if they occur to you) in terms of your present knowledge. It may help to skim the article, reading only the topic sentences in each paragraph to make sure that you understand all aspects of the problem. If, after doing this you, cannot state the problem clearly in your terms, look at the references. Is there an earlier attempt to solve this problem? This earlier article may state the problem in more familiar terms or may be by someone you know to be a good researcher and writer. A general text covering the problem discussed in the article may give you related material for solving this problem. You may need several iterations to grasp the problem.

When you can state the problem in your terms, read the *introduction again along with the methods section.* Pay particular attention to the assumptions being made and the limitations that these place on the solution being offered. Compare the methods to the sketch of a solution that you made. How does it differ? What points did you miss that this method considers? What assumptions did you make compared to the ones made here? If you are reading for general interest, this may be as far as you want to go. I recommend making a brief outline of what you have just done for future reference (see Section 2.8).

2.4 Find a Similar Problem with which You are Familiar and Work Through the Technical Details of the New Problem by Relating it to the Familiar One

From this point on assume that you want to use the results in this paper either in an application or to understand and develop new theory. Simply reading an article does not give you a complete understanding of its contents. Using or teaching the methods in the article provides a more complete understanding. If you have an opportunity through a journal club or a class to teach someone else about the article, you should do so. Even if you cannot teach someone else, begin to use the methods in the article.

Start by relating this problem to one with which you are already familiar. Follow the arguments and manipulations of the familiar problem, and broaden them to include your new problem. For example, to understand a paper on estimating parameters for linear models with stochastic parameters, you could relate the problem to one with fixed parameters, and examine the differences in the matrix structures, the effect on the Gauss–Markov solutions, the variability of the estimates, etc. Starting with a familiar problem gives you a firm base for pushing into unknown territory. There may be several different ways to approach your new problem. Different starting points should get you to the same place. The solution to your new problem will fit with your expanding knowledge base, and can be used in other problems.

2.5 Apply the Problem to Data

This is similar to point 2.4, but emphasizes using your new knowledge in a concrete way. Work through the techniques using a dataset that you know. Think about what would happen if these data had a different distribution or structure. What happens if the assumptions are violated? Many applied papers supply data that demonstrate the use of the techniques discussed. Such a dataset may demonstrate the technique to best advantage.

Simple numerical examples may also be helpful. Try applying that new matrix manipulation on a 2×2 matrix, and see what happens. If appropriate, program the techniques and examine the statistics as they are generated.

2.6 Separate Theory from Technical Details of Execution

To understand and use a new technique with facility you will need to understand both its theory and method of execution. When you are starting to read, however, it is a good idea to separate theory from execution. An estimate derived from mixed model theory may require the EM algo-

rithm for calculations. When reading the article keep clear which part of the discussion concerns the EM algorithm and which concerns the theory. In a particularly complex paper you may want to go through the points in this outline once for the theory arguments and once for the execution of the theory.

2.7 Read the Article at Least Three Times Emphasizing Different Sections Each Time

You have now read the paper through once, examined all sections in some detail, and obtained a good general understanding of the paper. The second time through the paper examine the *internal consistency* of the arguments, concentrating on the methods and results sections. Are the assumptions necessary and sufficient? Are the logic and notation straightforward and understandable? Do you understand the statistics, the probability theory, and the mathematics? Could you explain and defend this technique to statisticians at your level of experience and understanding?

As you ask these questions also consider what the authors could have done to make the task easier for you. Everything you think of here should be a candidate for inclusion in your own papers. This is a good time to examine the references again, and possibly examine companion papers that will shed light on your remaining questions. Begin to talk to colleagues, and consider unsolved problems that await your solutions. You may want to present some of your thoughts and get feedback from a group at this point. You should feel comfortable doing this because you now have a firm grasp on parts of the problem and can explain what you are still confused about. Boen (1982) has good suggestions about making presentations and answering questions in front of an audience.

Finally, read the paper for *external consistency* or generalizibility. Scrutinize the introduction, results or examples, and discussion sections. Find out how to use this technique, what kind of data it is useful for, where it fits into a range of solutions for problems of this type, and whether tested, stable, well-supported computer programs are available. Look at other applications of this or similar techniques that are in the references.

2.8 Consider Setting Up an Annotated Database

An annotated list of references will make it easier to review a technique. You are not outlining the paper; you just need to put enough down to make the paper easier to read next time. This database could be as simple as notecards or as sophisticated as using a computerized reference manager. Table 1 gives an example from my database, but be creative. You do not want all of your effort in reading the paper the first time to be lost when you do not use the procedure for a period of time. Also, note the good writers, theorists, and applications people as you find them. Not all statisticians are equally good in all areas, but you can pick the best in each area to emulate. You might even consider adding notes on the best presenters at meetings and what makes them good.

3. FINAL COMMENTS

Mark your progress by what you have done, not by what there is to do. The amount of literature is increasing exponentially, and you will never be able to read it all. You may, however, be able to read all of the good articles on a given topic. If you keep track of articles with a database, you will be surprised at how much of the literature you do read.

[Received March 1995. Revised August 1996.]

REFERENCES

Boen, J. R., and Zahn, D. A. (1982), *The Human Side of Statistical Consulting*, Belmont, CA: Wadsworth.

Colton, T. (1979), *Statistics in Medicine*, Boston: Little, Brown.

Cowen, A. C. (1991), "Teaching and Testing Mathematical Reading," *American Mathematical Monthly*, 98, 50–53.

Gleser, L. J. (1986), "Some Notes on Refereeing," *The American Statistician*, 40, 310–315.

Huff, D. (1954), *How to Lie with Statistics*, New York: W. W. Norton.

Parke, N. (1958), *Guide to the Literature of Mathematics and Physics* (2nd ed.), New York: Dover.

Pemberton, J. (1969), *How to Find Out in Mathematics: A Guide to Sources of Information*, Oxford: Pergamon Press.

Phanstiel, O. (1990), "How to Read Chemistry," *Journal of Chemical Education*, 67, 57–59.

Polya, G. (1945), *How to Solve It: A New Aspect of Mathematical Method*, Princeton: Princeton University Press.

Sackett, D. L., Haynes, R., B., Guyatt, G. H., and Tugwell, P. (1991), *Clinical Epidemiology: A Basic Science for Clinical Medicine*, Boston: Little, Brown.

Schechtman, K. B., and Spitznagel, E. L. (1987), "Teaching Biostatistics with an Emphasis on Reading the Medical Literature," in *ASA Proceedings of Statistical Education*, pp. 111–115.

Probable Inference, the Law of Succession, and Statistical Inference

Edwin B. Wilson

Journal of the American Statistical Association, Vol. 22, No. 158. (Jun., 1927), pp. 209-212.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28192706%2922%3A158%3C209%3APITLOS%3E2.0.CO%3B2-%23

Journal of the American Statistical Association is currently published by American Statistical Association.

NOTES

PROBABLE INFERENCE, THE LAW OF SUCCESSION, AND STATISTICAL INFERENCE

By Edwin B. Wilson, *Harvard School of Public Health*

Probable Inference (Usual). If there be observed a certain frequency or rate p_0 in a population of n and if the corresponding standard deviation $(p_0q_0/n)^{1/2}=\sigma_0$ be computed, the common statement of probable inference is to say that: The probability that the true value of the rate p lies outside its limits $p_0-\lambda\sigma_0$ and $p_0+\lambda\sigma_0$ is less than or equal to P_λ. It is assumed that P_λ decreases with an increase of λ. If the criterion of Tchebycheff is used, P_λ is itself less than $1/\lambda^2$; but if the probability table is used, P_λ is the area under the probability curve beyond the ordinates $\pm\lambda\sigma_0$. The rule of Tchebycheff is exceedingly conservative in its estimate of P_λ, whereas the probability table gives a radical estimate.

Strictly speaking, the usual statement of probable inference as given above is elliptical. Really the chance that the true probability p lies outside a specified range is either 0 or 1; for p actually lies within that range or does not. It is the observed rate p_0 which has a greater or less chance of lying within a certain interval of the true rate p. If the observer has had the hard luck to have observed a relatively rare event and to have based his inference thereon, he may be fairly wide of the mark.

Probable Inference (Improved). A better way to proceed is to reason as follows: There is some rate p. Its standard deviation is $(pq/n)^{1/2}=\sigma$. The probability that an observation as bad as p_0 will occur, where p_0 lies outside the limits $p-\lambda\sigma$ and $p+\lambda\sigma$, is less than or equal to P_λ. This form of statement throws the emphasis upon the fallibility of a particular observation in respect to being typical of a general situation.

It is still possible to state the criterion in terms of the observed rate p_0 for the equation $(p_0-p)^2=\lambda^2pq/n$, where $q=1-p$, is quadratic in p and may be solved to find p. If $\lambda^2/n=t$, the solution is

$$p=\frac{p_0+t/2}{1+t}\pm\frac{\sqrt{p_0q_0t+t^2/4}}{1+t}.$$

The rule then may be stated as: If the true value of the probability p lies outside the range

$$\frac{p_0+t/2}{1+t}-\frac{\sqrt{p_0q_0t+t^2/4}}{1+t} \quad \text{and} \quad \frac{p_0+t/2}{1+t}+\frac{\sqrt{p_0q_0t+t^2/4}}{1+t},$$

the chance of having such hard luck as to have made an observation so bad as p_0 is less or equal to $P\lambda$. And this form of statement is not elliptical. It is the proper form of probable inference.

Concerning the range indicated, it may be remarked that it is not centered at the value p_0 but at the value $(p_0+t/2)/(1+t)$ which differs from p_0 by being displaced toward the value $1/2$ by the amount

$$\frac{p_0+t/2}{1+t}-p_0=\frac{t(1/2-p_0)}{1+t}=\frac{(q_0-p_0)t/2}{1+t}.$$

Moreover, the interval on either side of the mean is

$$R=\sqrt{p_0q_0/n+\lambda^2/4n^2}/(1+\lambda^2/n),$$

which is not identical with $\lambda\sigma_0$ computed from p_0 nor with that value $\lambda\sigma_c$ which might be computed from the central value $(p_0+t/2)/(1+t)$ of the range indicated. In fact $R<\lambda\sigma_c$ and $\lambda\sigma_0<\lambda\sigma_c$, but R may be either less than or greater than $\lambda\sigma_0$—less if p_0 lies between .067 and .933, greater if p_0 lies outside those limits unless $t=\lambda^2/n$ be considerable compared with 2. The precise lines of division are $p_0=\frac{1}{2}\pm\frac{1}{2}\sqrt{1-(2+t)^{-2}}$.

The Law of Succession. The law of succession of Laplace states that if we have experienced S successes and F failures out of $S+F=n$ trials, the chance of success on the $(n+1)$st trial is $(S+1)/(n+2)$. Thus the law of succession purports to give the probability from experience not as $p_0=S/n$ but as $p=(S+1)/(n+2)$. This chance is, however, not the true chance of success, because the chance of success p on every trial must be the same. The proof of the law depends on inverse probabilities and in particular on the assumption that all probabilities are *a priori* equi-probable. The proof has been much criticized, for it has been held that the experience $p_0=S/n$ does not permit the assumption that all probabilities are equi-probable, but indicates that those in the neighborhood of p_0 must be much more probable than those remote from p_0. The simplest, if crudest, form of the argument of equi-probability is found in interpreting the formula $(S+1)/(n+2)$ as giving two new trials of which one is assumed to be a success and the other a failure.

If we apply the criterion in terms of the standard deviation as above developed we may state that the center of the range for p is $(p_0+t/2)/(1+t)$. If we now replace t by λ^2/n, and $p_0 n$ by S, the center of the range becomes $(S+\lambda^2/2)/(n+\lambda^2)$, and the probable inference is this: If the true probability lies outside the range

$$\frac{S+\lambda^2/2}{n+\lambda^2}-\lambda\frac{\sqrt{SF/n+\lambda^2/4}}{n+\lambda^2} \text{ and } \frac{S+\lambda^2/2}{n+\lambda^2}+\lambda\frac{\sqrt{SF/n+\lambda^2/4}}{n+\lambda^2},$$

the chance of our having the hard luck to realize the observed value $p_0=S/n$ is less than or equal to P_λ. As the distribution of the chances of an observation is asymmetric, it is perhaps unfair to take the central value of the range as the best estimate of the true probability; but this is what is actually done in practice.

In terms, therefore, of the practical criterion the forecasted value of the true probability is

$$\text{not } \frac{S+1}{n+2}, \text{ nor } \frac{S}{n}, \text{ but } \frac{S+\lambda^2/2}{n+\lambda^2};$$

and the value that should be assigned depends on the value of λ, *i. e.*, on our readiness to gamble on the typicalness of our realized experience. From this viewpoint, only those who believe that their experience is absolutely typical will set $\lambda=0$ and use as a forecast the realized frequency S/n. Those who use the law of succession, set $\lambda^2=2$ and allow a total variation in their experience of 2.8σ, *i. e.*, they wish to assert that they have not had an experience so rare that it or one less probable would arise, on the basis of the probability table as an estimate of P_λ, less than 16 times in 100. Those who make the usual allowance of 2σ for drawing an inference would use $(S+2)/(n+4)$ as a law of succession.

A particularly interesting and instructive case is that in which there has been total failure, $p_0=0$, $\sigma_0=0$. Here clearly the first form of the inference, namely, that the true value of p must lie between $p_0-\lambda\sigma_0=0$ and $p_0+\lambda\sigma_0=0$ is out of the question. The true form states that the experience is not so unusual as P_λ if p is less than $\lambda^2/(n+\lambda^2)n$ or if the expected number of instances is less than $\lambda^2/(n+\lambda^2)$, which for n large is practically λ^2/n. If this were applied to the classic case of determining the chance that the sun should fail to rise, one would take λ very small compared to 1 because general considerations of astronomy make it highly probable that our past experience is very nearly typical. If the application were to the fact that there were no deaths from leprosy in Massachusetts ($n=4{,}000{,}000$) in 1924, λ would also be taken small because leprosy is so rare, perhaps $\lambda=2/3$, meaning that we would take

an even chance. But in the case of paratyphoid fever, we might prefer to use the ordinary criterion with $\lambda = 2$.

Statistical Inference. This brings us to statistical inference which had best be differentiated from probable inference by requiring that something over and above the value of p_0 be known, something that will motivate a choice among values for λ in drawing the inference. It is well known that some phenomena show less and some show more variation than that due to chance as determined by the Bernoulli expansion $(p+q)^n$. The value L of the Lexian ratio is precisely the ratio of the observed dispersion to the value of $(npq)^{1/2}$ or $(pq/n)^{1/2}$ as the case may be. If we have general information which leads us to believe that the variation of a particular phenomenon be supernormal $(L>1)$, we naturally shall allow for some value of L in drawing the inference. Thus if the Lexian ratio is presumed from previous analysis of similar phenomena to be in the neighborhood of 5, we may use $\lambda = 10$ as properly as we should use $\lambda = 2$ if the phenomenon were believed to be normal (Bernoullian).

Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions

Alan AGRESTI and Brent A. COULL

For interval estimation of a proportion, coverage probabilities tend to be too large for "exact" confidence intervals based on inverting the binomial test and too small for the interval based on inverting the Wald large-sample normal test (i.e., sample proportion ± z-score × estimated standard error). Wilson's suggestion of inverting the related score test with null rather than estimated standard error yields coverage probabilities close to nominal confidence levels, even for very small sample sizes. The 95% score interval has similar behavior as the adjusted Wald interval obtained after adding two "successes" and two "failures" to the sample. In elementary courses, with the score and adjusted Wald methods it is unnecessary to provide students with awkward sample size guidelines.

KEY WORDS: Confidence interval; Discrete distribution; Exact inference; Poisson distribution; Small sample; Score test.

1. INTRODUCTION

One of the most basic analyses in statistical inference is forming a confidence interval for a binomial parameter p. Let X denote a binomial variate for sample size n, and let $\hat{p} = X/n$ denote the sample proportion. Most introductory statistics textbooks present the confidence interval based on the asymptotic normality of the sample proportion and estimating the standard error. This $100(1-\alpha)\%$ confidence interval for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \tag{1}$$

where z_c denotes the $1-c$ quantile of the standard normal distribution. This is called the *Wald confidence interval* for p, since it results from inverting the Wald test for p; that is, the interval is the set of p_0 values having P value exceeding α in testing $H_0 : p = p_0$ against $H_a : p \neq p_0$ using the test statistic $z = (\hat{p} - p_0)/\sqrt{\hat{p}(1-\hat{p})/n}$. Historically, this is surely one of the first confidence intervals proposed for any parameter (see, e.g., Laplace 1812, p. 283).

To avoid approximation, most advanced statistics textbooks recommend the Clopper–Pearson (1934) "exact" confidence interval for p, based on inverting equal-tailed binomial tests of $H_0 : p = p_0$. It has endpoints that are the solutions in p_0 to the equations

$$\sum_{k=x}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \alpha/2$$

and

$$\sum_{k=0}^{x} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \alpha/2,$$

except that the lower bound is 0 when $x = 0$ and the upper bound is 1 when $x = n$. This interval estimator is guaranteed to have coverage probability of *at least* $1-\alpha$ for every possible value of p. When $x = 1, 2, \ldots, n-1$, the confidence interval equals

$$\left[1 + \frac{n-x+1}{xF_{2x,2(n-x+1),1-\alpha/2}}\right]^{-1} < p < \left[1 + \frac{n-x}{(x+1)F_{2(x+1),2(n-x),\alpha/2}}\right]^{-1},$$

where $F_{a,b,c}$ denotes the $1-c$ quantile from the F distribution with degrees of freedom a and b. Equivalently, the lower endpoint is the $\alpha/2$ quantile of a beta distribution with parameters x and $n-x+1$, and the upper endpoint is the $1-\alpha/2$ quantile of a beta distribution with parameters $x+1$ and $n-x$. Letters to the editor from J. Klotz and from L. Leemis and K. S. Trivedi in the November 1996 issue of this journal (p. 389) showed how simple it is to calculate this interval using Minitab or S-Plus.

A considerable literature exists about these and other, less common, methods of forming confidence intervals for p. Santner and Duffy (1989, pp. 33-43) and Vollset (1993) reviewed a variety of methods. It has been known for some time that the Wald interval performs poorly unless n is quite large (e.g., Ghosh 1979, Blyth and Still 1983). The Clopper–Pearson exact interval is typically treated as the "gold standard" (e.g., Böhning 1994; Leemis and Trivedi 1996; Jovanovic and Levy 1997; and most mathematical statistics texts). However, this procedure is necessarily conservative, because of the discreteness of the binomial distribution (Neyman 1935), just as the corresponding exact test (without supplementary randomization on the boundary of the critical region) is conservative. For any fixed parameter value, the actual coverage probability can be much larger than the nominal confidence level unless n is quite large, and we believe it is inappropriate to treat this approach as optimal for statistical practice.

A compromise solution is the confidence interval based on inverting the approximately normal test that uses the null, rather than estimated, standard error; that is, its

Alan Agresti is Professor, Department of Statistics, University of Florida, Gainesville, FL 32611-8545 (E-mail: aa@stat.ufl.edu). Brent A. Coull is a post-doc, Department of Biostatistics, Harvard School of Public Health, Boston MA 02115. This work was partially supported by a grant from the National Institutes of Health. The authors thank the referees and Thomas Santner for helpful suggestions.

endpoints are the p_0 solutions to the equations $(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n} = \pm z_{\alpha/2}$. This confidence interval, apparently first discussed by Edwin B. Wilson (1927), has the form

$$\left(\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{[\hat{p}(1-\hat{p}) + z^2_{\alpha/2}/4n]/n}\right)/(1 + z^2_{\alpha/2}/n). \tag{2}$$

This inversion of what is the score test for p is called the *score confidence interval.* (Score tests, and in particular their standard errors, are based on the log likelihood at the null hypothesis value of the parameter, whereas Wald tests are based on the log likelihood at the maximum likelihood estimate; see, e.g., Agresti 1996, pp. 88-95.) This article shows that the score confidence interval tends to perform much better than the exact or Wald intervals in terms of having coverage probabilities close to the nominal confidence level. It can be recommended for use with nearly all sample sizes and parameter values. In addition, we show that a simple adaptation of the Wald interval also performs well even for small samples.

At first glance, the score confidence interval formula seems awkward to interpret, compared to (1). Letting $z = z_{\alpha/2}$, however, the midpoint of this interval is the weighted average

$$\hat{p}\left(\frac{n}{n+z^2}\right) + \frac{1}{2}\left(\frac{z^2}{n+z^2}\right),$$

which falls between $\hat{p}$ and $1/2$, with the weight given to $\hat{p}$ approaching 1 asymptotically. This midpoint shrinks the sample proportion towards .5, the shrinking being less severe as n increases. The coefficient of z in the term that is added to and subtracted from this midpoint to form the score confidence interval has square equal to

$$\frac{1}{n+z^2}\left[\hat{p}(1-\hat{p})\left(\frac{n}{n+z^2}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z^2}{n+z^2}\right)\right].$$

This has the form of a weighted average of the variance of a sample proportion when $p = \hat{p}$ and the variance of a sample proportion when $p = 1/2$, using $n + z^2$ in place of the usual sample size n.

2. COMPARING ACTUAL COVERAGE PROBABILITIES TO NOMINAL CONFIDENCE LEVELS

For a fixed value of a parameter, the actual coverage probability of an interval estimator is the (a priori) probability that the interval contains that value. In many cases, such as with discrete distributions, this varies according to the parameter value. In statistical theory, the confidence coefficient is defined to be the infimum of such coverage probabilities for all possible values of that parameter. Most practitioners, however, probably interpret confidence coefficients in terms of "average performance" rather than "worst possible performance." Thus, a possibly more relevant description of performance is the long-run percentage of times that the procedure is correct when it is used repeatedly for a variety of data sets in various problems with possibly different parameter values.

For any confidence interval procedure for estimating p, the actual coverage probability at a fixed value of p is

$$C_n(p) = \sum_{k=0}^{n} I(k,p)\binom{n}{k} p^k(1-p)^{n-k},$$

where $I(k,p)$ equals 1 if the interval contains p when $X = k$ and equals 0 if it does not contain p. We summarize this, using the alternative description of performance, by averaging over the possible values that p can take. We obtained results $\bar{C}_n = \int_0^1 C_n(p)g(p)dp$ for three beta densities $g(p)$ for this averaging: (1) the uniform distribution (mean = .50, std. dev. = $1/\sqrt{12}$ = .29); (2) bell-shaped with values relatively near the middle (mean = .50, std. dev. = .10); (3) skewed with values relatively near 0 (mean = .10, std. dev. = .05) or, by symmetry, near 1. Due to space considerations, we report results here mainly for the first case, but similar results occurred in the other two cases. Though this evaluation may suggest a Bayesian approach to inference, we restrict attention in this article to comparing the three standard methods decribed previously, in which the user makes no assumption about such a distribution for p.

Table 1 shows the mean of the actual coverage probabilities for the uniform averaging of the parameter values (i.e., $\bar{C}_n$ with $g(p) = 1$, $0 \le p \le 1$) at various sample sizes, for nominal 95% Wald, score, and exact confidence intervals (the three other methods listed in that table are discussed

Table 1. Mean Coverage Probabilities of Nominal 95% Confidence Intervals for the Binomial Parameter p, with Root Mean Square Errors in Parentheses, for Sampling p from a Uniform Distribution

Method	n = 5	n = 15	n = 30	n = 50	n = 100
Exact	.990	.980	.973	.969	.965
	(.041)	(.031)	(.026)	(.022)	(.017)
Score	.955	.953	.952	.952	.951
	(.029)	(.019)	(.014)	(.012)	(.008)
Wald	.641	.819	.875	.901	.922
	(.400)	(.238)	(.170)	(.133)	(.094)
Wald with *t*	.664	.837	.886	.905	.926
	(.391)	(.233)	(.167)	(.131)	(.093)
Mid-*P*	.978	.964	.958	.955	.953
	(.033)	(.021)	(.017)	(.013)	(.010)
Continuity-corrected Score	.987	.979	.973	.969	.965
	(.039)	(.030)	(.025)	(.021)	(.016)

in Section 4). The mean actual coverage probabilities for the Wald interval tend to be much too small. On the other hand, the exact interval is very conservative. For instance, for this method, $\bar{C}_n = .990$ when $n = 5$, .980 when $n = 15$, and .973 when $n = 30$. By contrast, $\bar{C}_n$ for the score method is close to the nominal confidence level, even for $n = 5$ where it is .955. Figure 1, which plots $\bar{C}_n$ as a function of n for the three interval estimators with the uniform and skewed beta weightings, illustrates their performance. Similar results were obtained with the bell-shaped weighting and using .90 nominal confidence coefficient, but are not reported here.

To describe how far actual coverage probabilities typically fall from the nominal confidence level, Table 1 also reports $\sqrt{\int_0^1 (C_n(p) - .95)^2 dp}$, the uniform-weighted root mean squared error of those probabilities about that confidence level. These values indicate that the variability about the nominal level is much smaller for the score confidence interval than for the Wald or exact confidence intervals. The improved performance of the score method relative to the Wald method is no surprise and simply adds to other evidence of this type accumulated over the years (e.g., Ghosh 1979; Vollset 1993). Some readers, though, may be surprised at just how much better the score method does than the exact method. The exact interval remains quite conservative even for moderately large sample sizes when p tends to be near 0 or 1. The Wald interval is also especially inadequate when p is near 0 or 1, partly a consequence of using $\hat{p}$ as its midpoint when the binomial distribution is highly skewed.

Even though the score intervals tend to have considerably higher actual coverage probabilities than the Wald intervals, they are not necessarily wider. In fact, unless the sample proportions fall near 0 or 1, they are shorter. Direct comparison of the formulas for the two interval widths yields that the score interval is narrower than the Wald interval whenever $\hat{p}$ falls within $\sqrt{(n + z^2)/(8n + 4z^2)}$ of 1/2. In particular, since this term decreases in the limit toward $1/\sqrt{8} = .35$ as n increases or $|z|$ decreases, the score interval is narrower than the Wald interval whenever $\hat{p}$ falls in (.15, .85) for any n and any nominal confidence level. See Ghosh (1979) for additional results about the relative lengths of the two types of interval. This comparison has limited relevance, since the actual coverage probabilities of the two methods differ. We mention this, however, to stress that the inadequacy of the Wald approach is not that the intervals are too short.

For fixed n and p, the expected width of an interval estimator is a useful measure of its performance. Figure 2 illustrates the relative sizes of the expected widths for the nominal 95% Wald, score, and exact intervals by plotting them as a function of p, for $n = 15$. For small n, the score intervals tend to be much shorter than exact intervals. The narrowness of the Wald intervals as p approaches 0 or 1 reflects the fact that when $x = 0$ or n, that interval is degenerate at 0 or at 1. By contrast, when $x = 0$, the score interval is $[0, z^2/(n + z^2)] = [0, 3.84/(n + 3.84)]$ and the exact interval is $[0, 1 - (.025)^{1/n}]$, which is approximately $[0, -\log(.025)/n] = [0, 3.69/n]$; the latter shows an extension of the "rule of $3/n$" (Jovanovic and Levy 1997) from the .95 upper confidence bound to .95 confidence limits.

Is anything sacrificed by using the score intervals? Well, since they are not "exact," they are not guaranteed to have coverage probabilities uniformly bounded below by the nominal confidence level, and their actual confidence coefficient (the infimum of such probabilities) is, in fact, well below it. Vollset's (1993) plots of the coverage probabilities as a function of p, for various methods, are illuminating for

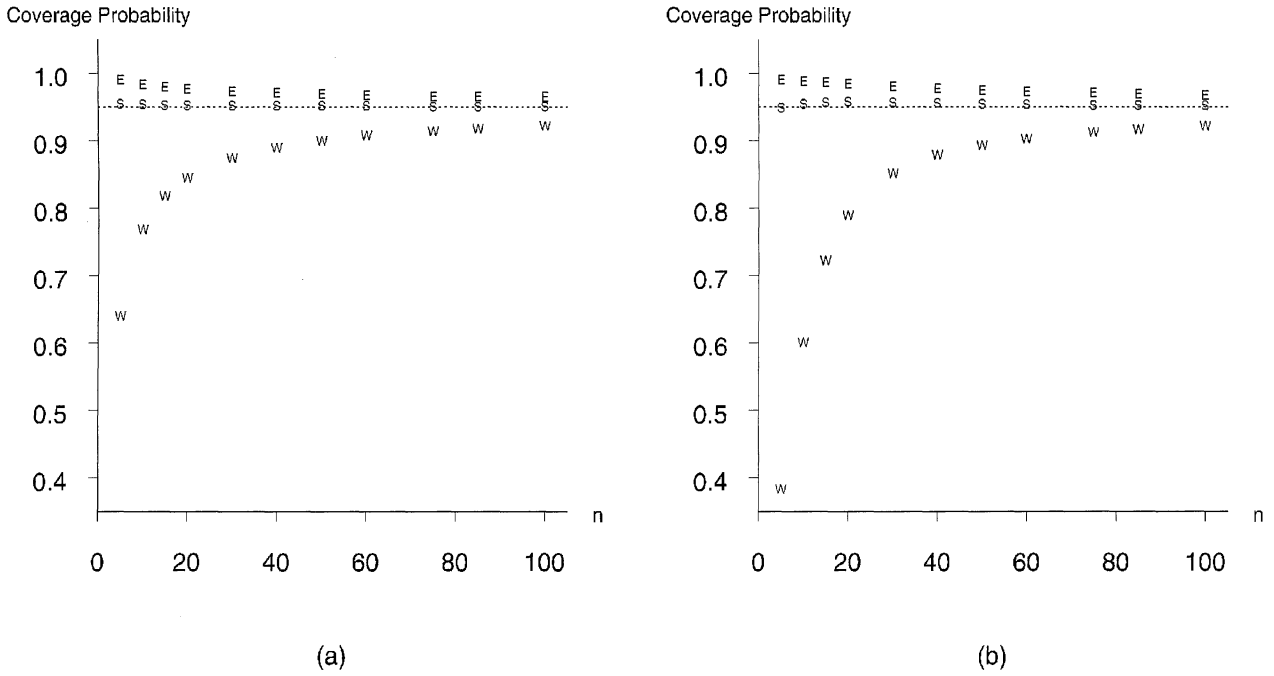


Figure 1. Mean Coverage Probability as a Function of Sample Size for the Nominal 95% Exact (E), Score (S), and Wald (W) Intervals, When p has (a) a Uniform (0,1) Distribution and (b) a Beta Distribution with $\mu = .10$ and $\sigma = .05$.

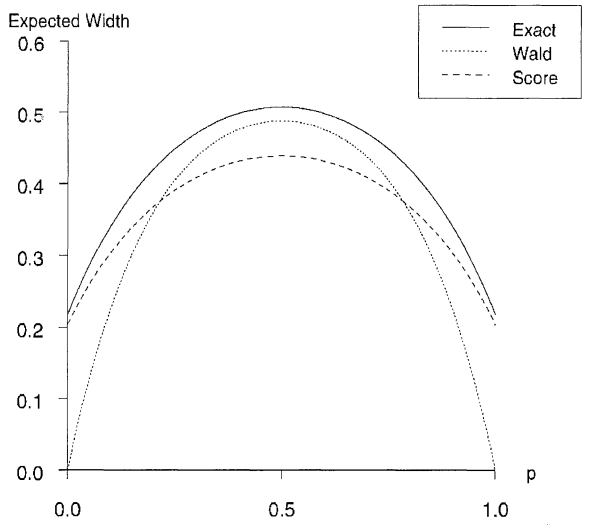


Figure 2. A Comparison of Expected Widths for the Nominal 95% Exact, Wald, and Score Intervals When n = 15.

describing the behavior of the methods. The score method has two very narrow regions of values for p, one near 0 and one near 1, at which the actual coverage probability falls seriously below the nominal confidence level, and this badly affects the actual confidence coefficient. These regions get closer to 0 and to 1 as n increases. For $n = 10$ with nominal 95% confidence intervals, for instance, there is a minimum coverage of .835 at $p = .018$ and $p = .982$, whereas at $n = 100$, there is a minimum coverage of .838 at $p = .002$ and $p = .998$.

We now explain why this happens. There is a region of values $[0, r)$ for p that falls in the score confidence interval only when $X = 0$. The upper bound r of this region is the lower endpoint of the confidence interval when $X = 1$, which for large n is approximately $(1 + z^2/2 - z\sqrt{4 + z^2}/2)/n$. The coverage probability just below r is approximately $P(X = 0) = [1 - (1 + z^2/2 - z\sqrt{4 + z^2}/2)/n]^n \approx \exp\{-(1 + z^2/2 - z\sqrt{4 + z^2}/2)\}$. The analogous remark applies for values of p near 1. This limiting coverage probability is .800 for nominal 90% intervals, .838 for 95% intervals, and .889 for 99% intervals. See Huwang (1995) for related remarks. In particular, the actual confidence level does not converge to the nominal level as n increases.

Though this may seem problematic, the portion of the [0, 1] parameter space over which the actual coverage probability drops seriously below the nominal confidence level is small. Table 2 illustrates. The proportion of the parameter space for which the coverage probability of the nominal 95% score interval falls below .90 is no more than .01 when $n \geq 20$. That table also shows that the proportion of parameter values for which the coverage probability is within .02 of .95 is much higher for the score than the exact interval. In fact, the score coverage probability is closer than the exact coverage probability to .95 over more than 90% of the parameter space, for the sample sizes reported.

3. THE "ADD TWO SUCCESSES AND TWO FAILURES" ADJUSTED WALD INTERVAL

The poor performance of the Wald interval is unfortunate, since it is the simplest approach to present in elementary statistics courses. We strongly recommend that instructors present the score interval instead. Santner (1998) makes the same recommendation. Of course, many instructors will hesitate to present a formula such as (2) in elementary courses. The shrinkage representation of the score approach suggests, however, that for constructing 95% confidence intervals (for which $z^2 = 1.96^2 \approx 4$ and the midpoint of the score interval is $(X + z^2/2)/(n + z^2) \approx (X + 2)/(n + 4)$) an instructor will not go far wrong in giving the following advice: "Add two successes and two failures and then use the Wald formula (1)." That is, this "adjusted Wald" interval uses the usual simple formula presented in such courses, but with $(n + 4)$ trials and point estimate $\tilde{p} = (X + 2)/(n + 4)$.

The midpoint of this interval, $\tilde{p} = (X + 2)/(n + 4)$, is nearly identical to the midpoint of the 95% score interval. It is identical to the Bayes estimate (mean of the posterior distribution) for the beta prior distribution with parameters 2 and 2, which has mean .50 and standard deviation .224 and which shrinks the sample proportion toward .50 somewhat more than does the uniform prior. This simple adjustment to the ordinary Wald interval changes it from highly liberal to slightly conservative, on the average, and a bit more conservative than the score method. Figure 3 illustrates, showing the mean actual coverage probability $\bar{C}_n$ for the nominal 95% Wald and adjusted Wald intervals as a function of n, for the uniform and skewed weightings of p. The adjusted Wald confidence interval behaves surprisingly well, even for very small sample sizes.

Figure 4 shows the actual coverage probabilities as a function of p for the Wald, adjusted Wald, and Clopper–Pearson exact intervals when $n = 5$ and $n = 10$. The im-

Table 2. Proportion of Parameter Space for which (a) Nominal 95% Score Interval has Actual Coverage Probability Below .90; (b) Nominal 95% Score and Exact Intervals Have Actual Coverage Probabilities Between .93 and .97; (c) Actual Coverage Probability is Closer to .95 for Score Interval than Exact Interval

n	Score coverage Prob. below .90	Coverage .93–.97		Coverage closer to .95 for Score than Exact
		Score	Exact	
5	.042	.463	.000	.944
10	.019	.608	.077	.963
20	.010	.792	.297	.925
30	.006	.882	.395	.977
50	.003	.939	.615	.961
100	.002	.968	.830	.961

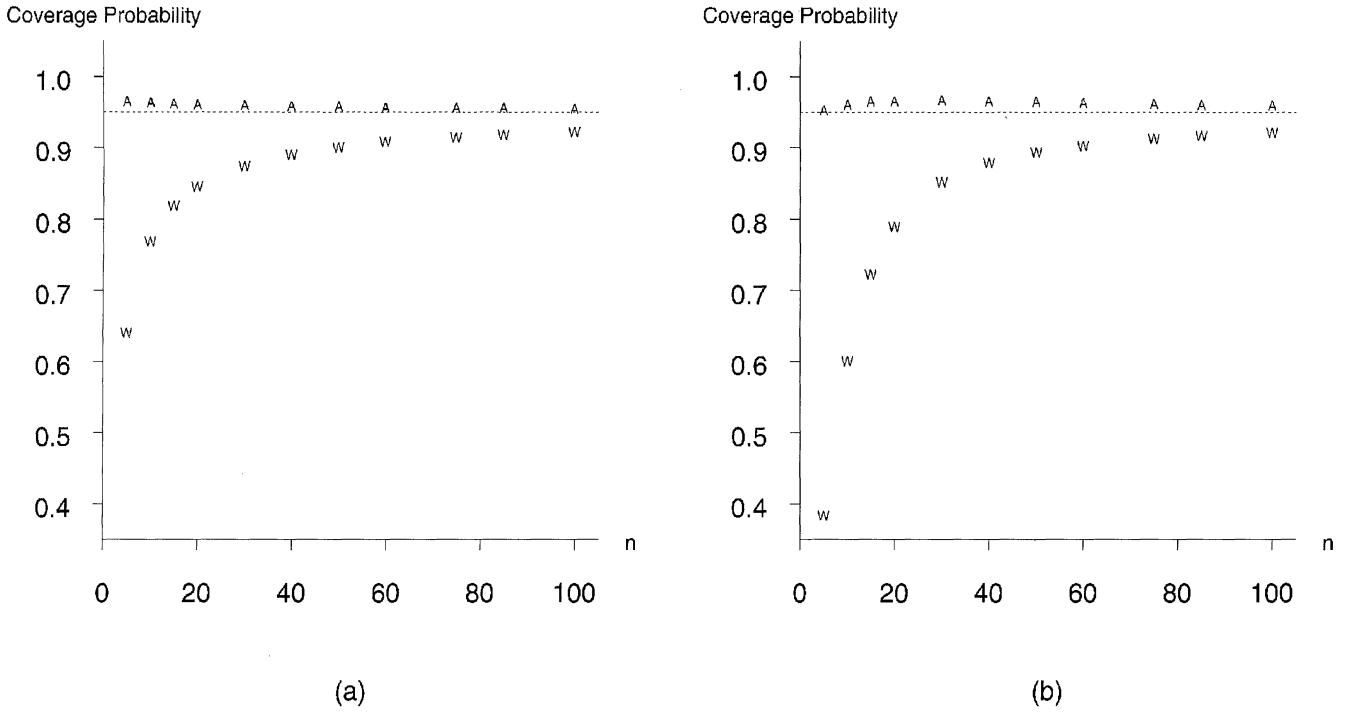


Figure 3. Mean Coverage Probability as a Function of Sample Size for the Nominal 95% Wald (W) and Adjusted Wald (A) Intervals, When p has (a) a Uniform (0,1) Distribution and (b) a Beta Distribution with μ = .10 and σ = .05.

provement of the adjusted Wald interval over the ordinary Wald interval is dramatic. The adjusted Wald interval also has the advantage, relative to the score interval, of not having spikes with seriously low coverage near $p = 0$ and 1. This is because this interval's rather crude bounds contain 0 when $X = 0$ or 1 and contain 1 when $X = n-1$ or n. For

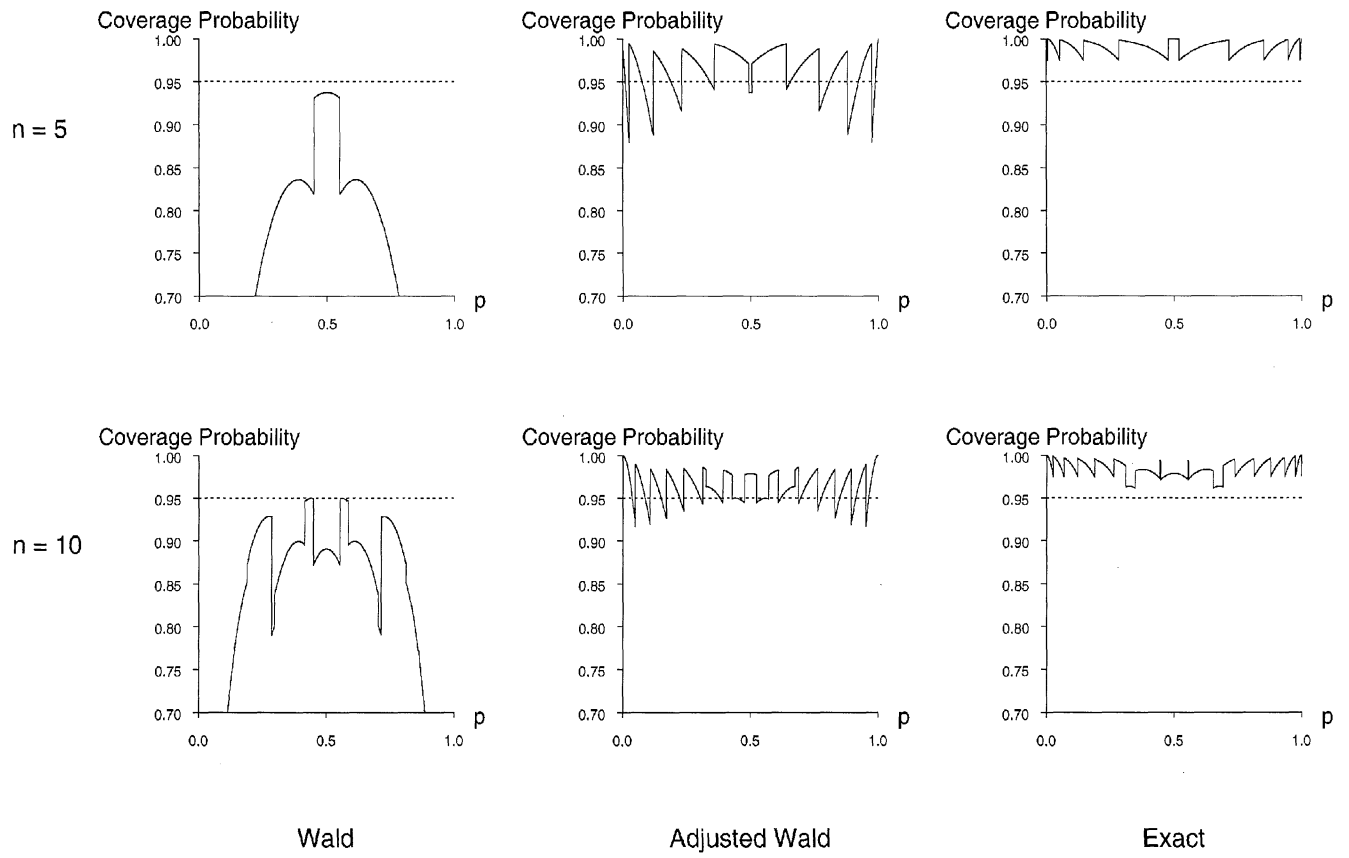


Figure 4. A Comparison of Coverage Probabilities for the Nominal 95% Wald, Adjusted Wald, and Exact Intervals.

instance, the minimum coverage probability for the nominal 95% adjusted Wald interval is .917 for $n = 10$ and never falls below .92 for $n > 10$. The proportion of the parameter space for which the actual coverage probability falls within .02 of .95 is slightly less than reported in Table 2 for the score interval, but the proportion of times its actual coverage probability is closer to .95 than the exact interval is still at least .94 for the sample sizes reported in that table. See Chen (1990) for results about coverage properties of related intervals using Bayes estimates as midpoints.

Introductory statistics textbooks have an awkward time with sample size recommendations for the Wald interval. Most simple recommendations tend to be inadequate (Leemis and Trivedi 1996). Our results suggest that if one tells students to add two successes and two failures before they form the Wald 95% interval, it is not necessary to present such sample size rules, since the "add two successes and two failures" confidence interval behaves adequately for practical application for essentially any n regardless of the value of p.

One can use the adjusted Wald interval without regarding its midpoint $\tilde{p} = (X + 2)/(n + 4)$ as the preferred point estimate of p. However, this rather strong shrinkage toward .5 might often provide a more appealing estimate than $\hat{p}$. The mean square error of $\tilde{p}$ equals $[np(1 - p) + 16(p - .5)^2]/(n + 4)^2$, which is smaller than that of $\hat{p}$ when p is within $\sqrt{3n^2 + 8n + 4}/(6n + 4)$ of .5; this interval of values of p decreases from (.113, .887) to (.211, .789) as n increases. Interestingly, Wilson (1927) mentioned this shrinkage estimator as a reasonable alternative to the sample proportion or the Laplace estimator $(X + 1)/(n + 2)$. Letting S denote X, the number of successes, Wilson stated, "As the distribution of chances of an observation is asymmetric, it is perhaps unfair to take the central value as the best estimate of the true probability; but this is what is actually done in practice. . . . Those who make the usual allowance of 2σ for drawing an inference would use $(S + 2)/(n + 4)$."

In recognition of his pioneering work, predating the famous articles by Neyman and Pearson on confidence intervals, we suggest that statisticians refer to $\tilde{p} = (X + 2)/(n + 4)$ as the Wilson point estimator of p and refer to the score confidence interval for p as the Wilson method. See Stigler (1997) for an interesting summary of Edwin B. Wilson's career. Other highlights included service as the first professor and head of the Department of Vital Statistics at Harvard School of Public Health in 1922, the Wilson–Hilferty normal approximation for the chi-squared distribution in 1931, and the Wilson–Worcester introduction of the median lethal dose (LD 50) in bioassay.

4. OTHER INTERVAL ESTIMATION METHODS FOR p

Although the focus of this article is comparison of the Wald, score, and exact intervals, which are the methods commonly presented in statistics textbooks, we next briefly discuss some alternative methods. Some elementary textbooks (e.g., Siegel 1988), perhaps recognizing the poor performance of the Wald intervals, suggest using ordinary t confidence intervals for a mean for interval estimation of a proportion. These intervals are wider than the Wald intervals, of course, but we found that mean coverage probabilities are still seriously deficient. Table 1 illustrates for the uniform weighting.

Other, more complex, methods exist for constructing exact confidence intervals, such as presented by Blyth and Still (1983) and Duffy and Santner (1987). Our evaluations of these intervals indicated that they perform better than the Clopper–Pearson intervals but not as well as the score intervals, still showing considerable conservatism. To reduce the conservativeness inherent in exact methods for discrete distributions, many authors recommend using tests and confidence intervals based on the mid-P value, namely half the probability of the observed result plus the probability of more extreme results (Lancaster 1961). The mid-P confidence interval is the inversion of the adaptation of the exact test that uses the mid-P value. Results in Vollset (1993) suggest that the mid-P interval tends to perform well but is somewhat more conservative than the score interval, typically having actual coverage probability greater than (and

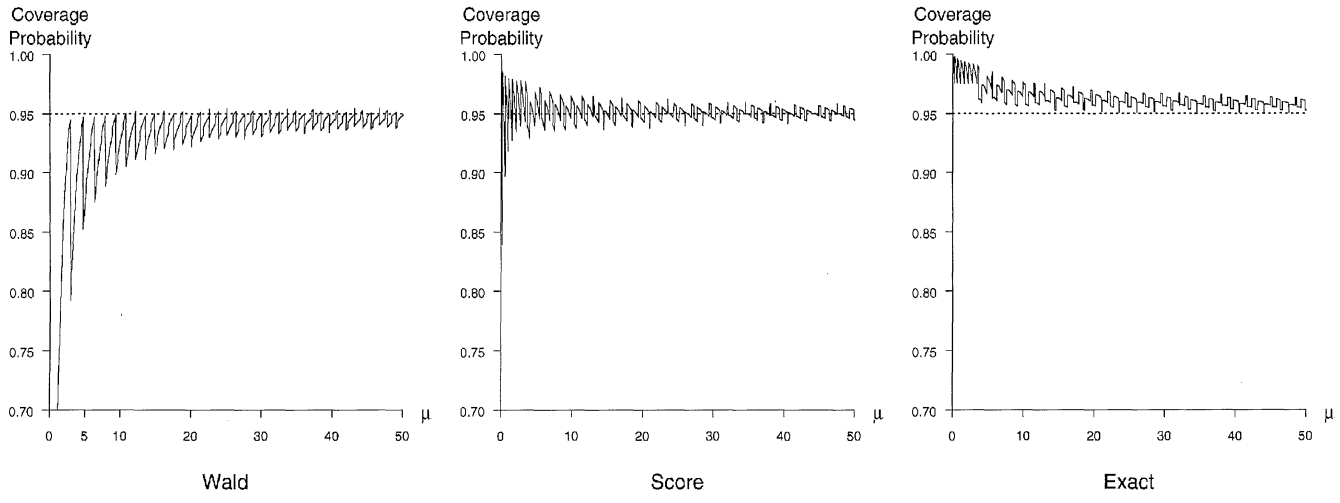


Figure 5. A Comparison of Coverage Probabilities for the Nominal 95% Wald, Score, and Exact Intervals for a Poisson Mean.

never much less than) the nominal confidence level. Our evaluations agreed with this, and are also illustrated in Table 1. We feel this is a reasonable method to use, especially if one is concerned that p may be very close to 0 or 1. It is more complex computationally than the score and adjusted Wald intervals, but like those intervals it has the advantage of being shorter than the exact interval.

Yet another alternative method is a continuity-corrected version of the score interval, based on the normal continuity correction for the binomial. This interval approximates the Clopper–Pearson interval, however, and our evaluations and results in Vollset (1993, Fig. 2) suggest that it is often as conservative as the exact interval itself. Again, Table 1 illustrates, and we do not recommend this approach.

Finally, we mention two other methods that perform well. The confidence interval based on inverting the likelihood-ratio test is similar to the score interval in terms of how it compares with the exact interval, but it is more complex to construct. Not surprisingly, Bayesian confidence intervals with beta priors that are only weakly informative also perform well in a frequentist sense (see, e.g., Carlin and Louis 1996, pp. 117–123).

In deciding whether to use the score interval, some may be bothered by its poor coverage for values of p just below the lower boundary of the interval when $X = 1$ and just above the upper boundary of the interval when $X = n - 1$. One could then use an adapted version that replaces the lower endpoint by $-\log(1-\alpha)/n$ when $X = 1$ and the upper endpoint by $1 + \log(1-\alpha)/n$ when $X = n - 1$. (e.g., at $p = -\log(1-\alpha)/n$, $P(X = 0) = [1 + \log(1-\alpha)/n]^n \approx 1 - \alpha$.) This adaptation improves the minimum coverage considerably. For instance, the nominal 95% interval has minimum coverage probability converging to .895 for large n, which is the large-sample coverage probability at p just below the lower endpoint of the interval when $X = 2$.

5. CONCLUSION AND EXTENSIONS

The Clopper–Pearson interval has coverage probabilities bounded below by the nominal confidence level, but the typical coverage probability is much higher than that level. The score and adjusted Wald intervals can have coverage probabilities lower than the nominal confidence level, yet the typical coverage probability is close to that level. In forming a 95% confidence interval, is it better to use an approach that guarantees that the actual coverage probabilities are *at least* .95 yet typically achieves coverage probabilities of about .98 or .99, or an approach giving narrower intervals for which the actual coverage probability could be less than .95 but is usually quite *close* to .95? For most applications, we would prefer the latter. The score and adjusted Wald confidence intervals for p provide shorter intervals with actual coverage probability usually nearer the nominal confidence level. In particular, even though the score and adjusted Wald intervals leave something to be desired in terms of satisfying the usual technical definition of "95% confidence," the operational performance of those methods is better than the exact interval in terms of how most practitioners interpret that term.

Results similar to those in this article also hold in other discrete problems. For instance, similar comparisons apply for score, Wald, and exact confidence intervals for a Poisson parameter μ, based on an observation X from that distribution. Figure 5 illustrates, plotting the actual coverage probabilities when the nominal confidence level is .95. Here, the score interval for μ results from inverting the approximately normal test statistic $z = (X - \mu_0)/\sqrt{\mu_0}$, the Wald interval results from inverting $z = (X - \mu_0)/\sqrt{X}$, and the endpoints of the exact interval, $(1/2)(\chi^2_{2X,.025}, \chi^2_{2(X+1),.975})$, result from equating tail sums of null Poisson probabilities to .025 (Garwood 1936; for n independent Poisson observations, $X_1, \ldots, X_n$, the same formulas apply if one lets $X = \sum X_i$ and $\mu = E(X) = nE(X_i)$). For another discrete example, see Mehta and Walsh (1992) for a comparison of exact with mid-P confidence intervals for odds ratios or for a common odds ratio in several 2×2 contingency tables.

Exact inference has an important place in statistical inference of discrete data, in particular for sparse contingency table problems for which large-sample chi-squared statistics are often unreliable. However, approximate results are sometimes more useful than exact results, because of the inherent conservativeness of exact methods.

[Received February 1997. Revised November 1997.]

REFERENCES

Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, New York: Wiley.

Blyth, C. R., and Still, H. A. (1983), "Binomial Confidence Intervals," *Journal of the American Statistical Association*, 78, 108–116.

Böhning, D. (1994), "Better Approximate Confidence Intervals for a Binomial Parameter," *Canadian Journal of Statistics*, 22, 207–218.

Carlin, B. P., and Louis, T. A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, London: Chapman and Hall.

Chen, H. (1990), "The Accuracy of Approximate Intervals for a Binomial Parameter," *Journal of the American Statistical Association*, 85, 514–518.

Clopper, C. J., and Pearson, E. S. (1934), "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, 26, 404–413.

Duffy, D. E., and Santner, T. J. (1987), "Confidence Intervals for a Binomial Parameter Based on Multistage Tests," *Biometrics*, 43, 81–93.

Garwood, F. (1936), "Fiducial Limits for the Poisson Distribution," *Biometrika*, 28, 437–442.

Ghosh, B. K. (1979), "A Comparison of Some Approximate Confidence Intervals for the Binomial Parameter," *Journal of the American Statistical Association*, 74, 894–900.

Huwang, L. (1995), "A Note on the Accuracy of an Approximate Interval for the Binomial Parameter," *Statistics & Probability Letters*, 24, 177–180.

Jovanovic, B. D., and Levy, P. S. (1997), "A Look at the Rule of Three," *The American Statistician*, 51, 137–139.

Lancaster, H. O. (1961), "Significance Tests in Discrete Distributions," *Journal of the American Statistical Association*, 56, 223–234.

Laplace, P. S. (1812), *Théorie Analytique des Probabilités*, Paris: Courcier.

Leemis, L. M., and Trivedi, K. S. (1996), "A Comparison of Approximate Interval Estimators for the Bernoulli Parameter," *The American Statistician*, 50, 63–68.

Mehta, C. R., and Walsh, S. J. (1992), "Comparison of Exact, Mid-p, and Mantel-Häenszel Confidence Intervals for the Common Odds Ratio Across Several 2×2 Contingency Tables," *The American Statistician*,

46, 146–150.

Neyman, J. (1935), "On the Problem of Confidence Limits," *Annals of Mathematical Statistics*, 6, 111–116.

Santner, T. J. (1998), "A Note on Teaching Binomial Confidence Intervals," *Teaching Statistics*, 20, 20–23.

Santner, T. J., and Duffy, D. E. (1989), *The Statistical Analysis of Discrete Data*, Berlin: Springer-Verlag.

Siegel, A. F. (1988), *Statistics and Data Analysis*. New York: Wiley.

Stigler, S. M. (1997), "Edwin Bidwell Wilson," in *Leading Personalities in Statistical Sciences*, eds. N. L. Johnson and S. Kotz, New York: Wiley, pp. 344–346.

Vollset, S. E. (1993), "Confidence Intervals for a Binomial Proportion," *Statistics in Medicine*, 12, 809–824.

Wilson, E. B. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, 22, 209–212.

Qualms About Bootstrap Confidence Intervals

NATHANIEL SCHENKER*

The percentile method and bias-corrected percentile method of Efron (1981, 1982) are discussed. When these methods are used to construct nonparametric confidence intervals for the variance of a normal distribution, the coverage probabilities are substantially below the nominal level for small to moderate samples. This is due to the inapplicability of assumptions underlying the methods. These assumptions are difficult or impossible to check in the complicated situations for which the bootstrap is intended. Therefore, bootstrap confidence intervals should be used with caution in complex problems.

KEY WORDS: Bias-corrected percentile method; Nonparametric confidence intervals; Percentile method; Pivotal quantity; Resampling plans.

1. INTRODUCTION

The bootstrap (Efron 1979, 1981, 1982) has been advertised and is widely viewed as a tool that can be used to set nonparametric confidence intervals in complex problems. This article discusses the percentile method and bias-corrected percentile method of Efron (1981, 1982). These methods are shown to perform poorly in the relatively simple problem of setting a confidence interval for the variance of a normal distribution. The purpose of the article is to focus on assumptions underlying the use of bootstrap confidence intervals and to caution practitioners against applying these methods blindly in complex problems.

2. THE BOOTSTRAP

Suppose $X_1, \ldots, X_n$ are iid random variables from a population with unknown cdf F, and suppose the goal is to draw inferences about some parameter θ of the population. Let $\hat{\theta}(X_1, \ldots, X_n)$ be an estimator of θ and let $\hat{F}$ be the sample cdf, that is, the cdf that assigns mass $1/n$ to each X_i. The bootstrap approximates the sampling distribution of $\hat{\theta}$ under F by the sampling distribution of $\hat{\theta}$ under $\hat{F}$. This procedure is usually hard to carry out analytically, and it is often necessary to use Monte Carlo methods as follows (Efron 1981, 1982):

1. Construct $\hat{F}$.
2. Draw a bootstrap sample $X_1^*, \ldots, X_n^*$ iid with cdf $\hat{F}$, and calculate $\hat{\theta}^* = \hat{\theta}(X_1^*, \ldots, X_n^*)$.
3. Independently do step 2 B times (for some large B), obtaining $\hat{\theta}_b^*$, $b = 1, \ldots, B$. The cdf of $\hat{\theta}$ at y is approximated by $\widehat{CDF}(y) = \#\{\hat{\theta}_b^* \leq y\}/B$.

Let F be written as F_θ to signify the dependence of F on θ. The sample cdf $\hat{F}$ is likely to be closer to $F_{\hat{\theta}}$ than to F_θ. For example, if $X_1, \ldots, X_n$ are drawn from a $N(0, 1)$ distribution, then $\hat{F}$ will be better approximated by the $N(\bar{X}, \hat{\sigma}^2)$ cdf, where $\bar{X}$ and $\hat{\sigma}^2$ are the sample mean and variance, respectively, than by the $N(0, 1)$ cdf. Thus the bootstrap procedure just described is likely to approximate the sampling distribution of $\hat{\theta}$ under $F_{\hat{\theta}}$ better than it approximates the sampling distribution of $\hat{\theta}$ under F_θ. This idea will be used in the next section.

3. THE PERCENTILE AND BIAS-CORRECTED PERCENTILE METHODS

Let $z' = \Phi^{-1}(\widehat{CDF}(\hat{\theta}))$, where Φ is the $N(0, 1)$ cdf. Efron's (1981, 1982) bias-corrected percentile method uses $[\widehat{CDF}^{-1}(\Phi(2z' - z_{1-a})), \widehat{CDF}^{-1}(\Phi(2z' - z_a))]$, where $z_p = \Phi^{-1}(p)$, as a nominal $100(1 - 2a)\%$ nonparametric confidence interval for θ. The derivation of this method is based on the assumption that there is a monotone increasing function g such that $g(\hat{\theta}) - g(\theta) \sim N(\eta, \tau^2)$ and $g(\hat{\theta}^*) - g(\hat{\theta}) \sim N(\eta, \tau^2)$ for some constants η and τ^2. (Here, $\sim$ is used to denote the distribution under the repeated bootstrap sampling of Section 2.) Note that the interval for θ does not involve g, so g need not be known; it is only necessary to know that g exists.

The assumption underlying the bias-corrected percentile method is not valid in general, but is approximately valid under the following condition. Suppose there exists a monotone increasing function g such that $g(\hat{\theta}) - g(\theta)$ is a normal pivotal quantity, that is, such that $g(\hat{\theta}) - g(\theta)$ has the same normal distribution for all values of θ. Then since $\hat{F}$ is approximately the same as $F_{\hat{\theta}}$, $g(\hat{\theta}^*) - g(\hat{\theta})$ will have a distribution close to this normal distribution under repeated bootstrap sampling.

When $\widehat{CDF}(\hat{\theta}) = .5$, which should be approximately true if $\hat{\theta}$ is median unbiased for θ (that is, $P[\hat{\theta} \leq \theta] = .5$ for all θ), the bias-corrected percentile method reduces to $[\widehat{CDF}^{-1}(a), \widehat{CDF}^{-1}(1 - a)]$. Efron (1981, 1982) has named this the percentile method interval.

4. EXAMPLE: ESTIMATING THE VARIANCE OF A NORMAL DISTRIBUTION

The techniques of setting nonparametric confidence intervals that were described in the previous section will now be considered for the problem of estimating a normal variance. Suppose $X_1, \ldots, X_n$ are iid from $N(\mu, \sigma^2)$ with μ and σ^2 unknown, and a 90% confidence interval for σ^2 is desired. Let $\hat{\sigma}^2 = \sum_i (X_i - \bar{X})^2/n$ be the estimator of σ^2 used.

In addition to the methods of Section 3, the following procedure will be considered. The cdf of $\hat{\sigma}^2/\sigma^2$ at y will be approximated by $\hat{G}(y) = \#\{\hat{\sigma}_b^{2*}/\hat{\sigma}^2 \leq y\}/B$ (see Section 2 for notation). Then $[\hat{\sigma}^2/\hat{G}^{-1}(1 - a), \hat{\sigma}^2/\hat{G}^{-1}(a)]$ will be used as a nominal $100(1 - 2a)\%$ confidence interval for σ^2. Since $\hat{G}(y) = \widehat{CDF}(\hat{\sigma}^2 y)$, it follows that $\hat{G}^{-1}(\cdot) = \widehat{CDF}^{-1}(\cdot)/\hat{\sigma}^2$, so the interval is just $[\hat{\sigma}^4/\widehat{CDF}^{-1}(1 - a), \hat{\sigma}^4/\widehat{CDF}^{-1}(a)]$. This is not really a nonparametric confidence interval, since knowledge of the parametric family, $N(\mu, \sigma^2)$, was used in choosing the quantity $\hat{\sigma}^2/\sigma^2$ to be bootstrapped; this quantity is pivotal under

* Nathaniel Schenker is with the Statistical Research Division, U.S. Bureau of the Census, Washington, DC 20233. Support for this research was provided by National Science Foundation Grants MCS 81-01836 and SES 83-11428 at the Department of Statistics, University of Chicago. The author thanks Stephen Stigler, David Wallace, and Wing Wong for many fruitful discussions. The comments of the associate editor and the referees are also appreciated.

Table 1. Monte Carlo Coverage Probabilities of Nominal 90% Confidence Intervals for σ^2

n	Percentile Method	Bias-Corrected Percentile Method	Bootstrapping $\hat{\sigma}^2/\sigma^2$
20	.78	.80	.85
35	.82	.85	.86
100	.87	.88	.88

Table 2. Values of $g_n(1) = WH_n(n)$, Where WH_n Is the Wilson–Hilferty Transformation of a χ^2_{n-1} Random Variable

n	$g_n(1)$
20	.27
35	.20
100	.12

the normal family. However, the nonparametric bootstrap step of substituting $\hat{F}$ for F and resampling from $\hat{F}$ (see Section 2) is still used. Thus the performance of this interval will indicate how well the bootstrap performs when applied to a known pivotal quantity.

A simulation study has been conducted for samples of size 20, 35, and 100. Sixteen hundred Monte Carlo trials were used, and the case $\mu = 0$, $\sigma^2 = 1$ was simulated without loss of generality. In computing the bootstrap approximations, $B = 1{,}000$ bootstrap replications were used. Computations were performed using FORTRAN programs on the DECSYSTEM-20 computer of the Graduate School of Business at the University of Chicago. Random variables were generated using the IMSL routines GGUBS and GGNPM.

The proportions $\hat{p}$ of the intervals covering $\sigma^2 = 1$ are given in Table 1. The standard error $(\hat{p}(1 - \hat{p})/1{,}600)^{1/2}$ of each entry is about .01. All of the bootstrap methods have Monte Carlo coverage rates that are reasonably close to the nominal level of 90% when $n = 100$. For smaller n, however, the bootstrap methods do not perform as well.

The intervals based on bootstrapping $\hat{\sigma}^2/\sigma^2$ have the best coverage rates of the three methods; this is expected, since $\hat{\sigma}^2/\sigma^2$ is a pivotal quantity under sampling from $N(\mu, \sigma^2)$. However, this method does not achieve the nominal level of 90%. This is due to the deficiency of the nonparametric step of resampling from the sample cdf $\hat{F}$ in the bootstrap procedure. Since $\hat{\sigma}^2/\sigma^2$ is pivotal under normal sampling, bootstrapping $\hat{\sigma}^2/\sigma^2$ would yield exact 90% intervals if $\hat{F}$ were always a normal cdf. However, $\hat{F}$ is never exactly a normal cdf, and can be much different from one for small n.

Along with the problem associated with resampling from $\hat{F}$, the percentile and bias-corrected percentile methods are based on assumptions that do not hold in this example. Let WH_n denote the Wilson–Hilferty transformation of a χ^2_{n-1} random variable to an approximate $N(0, 1)$ variate (see Kendall and Stuart 1977). Thus, $WH_n(x) = (9(n - 1)/2)^{1/2}((x/(n - 1))^{1/3} - 1 + 2/(9(n - 1)))$. This transformation works very well for degrees of freedom as large as those considered here. Define the function g_n by $g_n(x) = WH_n(nx)$. Algebraic manipulation then shows that $g_n(\hat{\sigma}^2) - g_n(\sigma^2) \sim \sigma^{2/3}N(-g_n(1), 1)$, approximately. Thus, although there exists a transformation to approximate normality, the pivotal quantity discussed in Section 3 does not exist. This is one reason for the poor performance of the bias-corrected percentile method.

Since $g_n(\hat{\sigma}^2) - g_n(\sigma^2)$ has an approximate $\sigma^{2/3}N(-g_n(1), 1)$ distribution, $\hat{\sigma}^2$ is not median unbiased for σ^2. The bias of $g_n(\hat{\sigma}^2)$ for $g_n(\sigma^2)$ in standard deviation units is $-g_n(1)$. Values of $g_n(1)$ are given in Table 2 for $n = 20$, 35, and 100. The poor performance of the percentile method is due in part to the lack of median unbiasedness.

5. CONCLUSIONS

Nonparametric confidence intervals formed using the bootstrap are intended for use in complicated estimation problems. The percentile and bias-corrected percentile methods were examined here in the relatively simple problem of estimating the variance of a normal distribution. The coverage probabilities were well below the nominal level for small to moderate samples. Along with the problems inherent in resampling from the sample cdf, underlying assumptions about pivotal quantities and median unbiasedness were not valid.

If little is known about a problem, it is very difficult or impossible to check the assumptions underlying the use of bootstrap confidence intervals. Therefore, they should be used with caution in complex problems.

[*Received November 1983. Revised October 1984.*]

REFERENCES

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1–26.

——— (1981), "Nonparametric Standard Errors and Confidence Intervals," *Canadian Journal of Statistics*, 9, 139–172.

——— (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, National Science Foundation–Conference Board of the Mathematical Sciences Monograph 38, Philadelphia: Society for Industrial and Applied Mathematics.

Kendall, M. G., and Stuart, A. (1977), *The Advanced Theory of Statistics* (Vol. 1, 4th ed.), New York: Macmillan.

Better Bootstrap Confidence Intervals

BRADLEY EFRON*

We consider the problem of setting approximate confidence intervals for a single parameter θ in a multiparameter family. The standard approximate intervals based on maximum likelihood theory, $\hat{\theta} \pm \hat{\sigma} z^{(\alpha)}$, can be quite misleading. In practice, tricks based on transformations, bias corrections, and so forth, are often used to improve their accuracy. The bootstrap confidence intervals discussed in this article automatically incorporate such tricks without requiring the statistician to think them through for each new application, at the price of a considerable increase in computational effort. The new intervals incorporate an improvement over previously suggested methods, which results in second-order correctness in a wide variety of problems. In addition to parametric families, bootstrap intervals are also developed for nonparametric situations.

KEY WORDS: Resampling methods; Approximate confidence intervals; Transformations; Nonparametric intervals; Second-order theory; Skewness corrections.

1. INTRODUCTION

This article concerns setting approximate confidence intervals for a real-valued parameter θ in a multiparameter family. The nonparametric case, where the number of nuisance parameters is infinite, is also considered. The word "approximate" is important, because in only a few special situations can exact confidence intervals be constructed. Table 1 shows one such situation: the data (y_1, y_2) are bivariate normal with unknown mean vector (η_1, η_2), covariance matrix = **I** the identity; the parameters of interest are $\theta = \eta_2/\eta_1$ and, in addition, $\xi = 1/\theta$. Fieller's construction (1954) gives central 90% interval (5% error in each tail) of [.29, .76] for θ, having observed $\mathbf{y} = (8, 4)$. The corresponding interval for $\xi = 1/\theta$ is the obvious mapping $\xi \in [1/.76, 1/.29]$.

Table 1 also shows the standard approximate intervals

$$\theta \in [\hat{\theta} + \hat{\sigma} z^{(\alpha)}, \hat{\theta} + \hat{\sigma} z^{(1-\alpha)}], \tag{1.1}$$

where $\hat{\theta}$ is the maximum likelihood estimate (MLE) of θ, $\hat{\sigma}$ is an estimate of its standard deviation, often based on derivatives of the log-likelihood function, and $z^{(\alpha)}$ is the $100 \cdot \alpha$ percentile point of a standard normal variate. In Table 1, $\alpha = .05$ and $z^{(\alpha)} = -z^{(1-\alpha)} = -1.645$.

The standard intervals (1.1) are extremely useful in statistical practice because they can be applied in an automatic way to almost any parametric situation. However, they can be far from perfect, as the results for ξ show. Not only is the standard interval for ξ quite different from the exact interval, it is not even the obvious transformation [1/.73, 1/.27] of the standard interval for θ.

Approximate confidence intervals based on bootstrap computations were introduced by Efron (1981, 1982a). Like the standard intervals, these can be applied automatically to almost any situation, though at greater computational expense than (1.1). Unlike (1.1), the bootstrap intervals transform correctly, so the interval for $\xi = 1/\theta$ in the Fieller example is obtained by inverting the endpoints of the interval for θ. They also tend to be more accurate than the standard intervals. In the situation of Table 1 the bootstrap intervals agree with the exact intervals to three decimal places. Efron (1985) showed that this is no accident; there is a wide class of problems for which the bootstrap intervals are an order of magnitude more accurate than the standard intervals.

In those problems where exact confidence limits exist the endpoints are typically of the form

$$\hat{\theta} + \hat{\sigma}(z^{(\alpha)} + A_n^{(\alpha)}/\sqrt{n} + B_n^{(\alpha)}/n + \cdots), \tag{1.2}$$

where n is the sample size (see Efron 1985). The standard intervals (1.1) are *first-order correct* in the sense that the term $\hat{\theta} + \hat{\sigma} z^{(\alpha)}$ asymptotically dominates (1.2). However, the second-order term $\hat{\sigma} A_n^{(\alpha)}/\sqrt{n}$ can have a major effect in small-sample situations. It is this term that causes the asymmetry of the exact intervals about the MLE as illustrated in Table 1. As a point of comparison the Student-t effect is of third-order magnitude, comparable with $\hat{\sigma} B_n^{(\alpha)}/n$ in (1.2). The bootstrap method described in Efron (1985) was shown to be *second-order correct* in a certain class of problems, automatically producing intervals of correct second-order asymptotic form $\hat{\theta} + \hat{\sigma}(z^{(\alpha)} + A_n^{(\alpha)}/\sqrt{n} + \cdots)$.

This article describes an improved bootstrap method that is second-order correct in a wider class of problems. This wider class includes all of the familiar parametric examples where there are no nuisance parameters and where the data have been reduced to a one-dimensional summary statistic, with asymptotic properties of the usual MLE form (see Sec. 5).

Several authors have developed higher-order correct approximate confidence intervals based on Edgeworth expansions (Abramovitch and Singh 1985; Beran 1984a,b; Hall 1983; Withers 1983), sometimes using bootstrap methods to reduce the theoretical computations. There is a close theoretical relationship between this line of work and the current article (see, e.g., Remark G, Sec. 11). However, the details of the various methods are considerably different, and they can give quite different numerical results. An important point, which will probably have to be settled by extensive simulations, is which method, if any, handles best the practical problems of day-to-day applied statistics.

2. OVERVIEW

The standard interval (1.1) is based on taking literally the asymptotic normal approximation

$$(\hat{\theta} - \theta)/\hat{\sigma} \sim N(0, 1), \tag{2.1}$$

* Bradley Efron is Professor of Statistics and Biostatistics, Department of Statistics, Stanford University, Stanford, CA 94305. The author is grateful to Robert Tibshirani, Timothy Hesterberg, and John Tukey for several useful discussions, suggestions, and references.

Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods

Table 1. Central 90% Confidence Intervals for $\theta = \eta_2/\eta_1$ and $\xi = 1/\theta$, Having Observed $(y_1, y_2) = (8, 4)$ From a Bivariate Normal Distribution $\mathbf{y} \sim N_2(\boldsymbol{\eta}, \mathbf{I})$

	For θ	(R/L)	For ξ	(R/L)
Exact interval (also bootstrap)	[.29, .76]	(1.21)	[1.32, 3.50]	(2.20)
Standard approximation (1.1)	[.27, .73]	(1.00)	[1.08, 2.92]	(1.00)
MLE	$\hat{\theta} = .5$		$\hat{\xi} = 2$	

NOTE: The exact intervals are based on Fieller's construction. R/L is the ratio of the right side of the interval, measured from the MLE, to the left side. The exact intervals are markedly asymmetric. The approximate bootstrap intervals of Efron (1982a) agree with the exact intervals to three decimal places in this case.

with the estimated standard error $\hat{\sigma}$ considered to be a fixed constant. In certain cases it is well known that both convergence to normality and constancy of σ can be dramatically improved by considering instead of $\hat{\theta}$ and θ a monotone tranformation $\hat{\phi} = g(\hat{\theta})$ and $\phi = g(\theta)$. The classic example is that of the correlation coefficient from a bivariate normal sample, for which Fisher's inverse hyperbolic tangent transformation works beautifully (see Efron 1982b).

The bias-corrected bootstrap intervals previously introduced by Efron (1981, 1982a), called the BC intervals, assume that normality and constant standard error can be achieved by *some* transformation $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$, say

$$(\hat{\phi} - \phi)/\tau \sim N(-z_0, 1), \tag{2.2}$$

τ being the constant standard error of $\hat{\phi}$. Allowing the bias constant z_0 in (2.2) considerably improves the approximation in many cases, including that of the normal correlation coefficient. Taking (2.2) literally gives the obvious confidence interval $(\hat{\phi} + \tau z_0) \pm \tau z^{(\alpha)}$ for ϕ, which can then be converted back to a confidence interval for θ by the inverse transformation $\theta = g^{-1}(\phi)$. The advantage of the BC method is that all of this is done automatically from bootstrap calculations, without requiring the statistician to know the correct transformation g.

The improved bootstrap method introduced in this article, called BC_a, makes one further generalization on (2.1): it is assumed that for some monotone transformation g, some bias constant z_0, *and some "acceleration constant"* a, the transformation $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$ results in

$$(\hat{\phi} - \phi)/\tau \sim N(-z_0\sigma_\phi, \sigma_\phi^2), \qquad \sigma_\phi = 1 + a\phi. \tag{2.3}$$

Notice that (2.2) is the special case of (2.3), with $a = 0$.

Given (2.3), it is not difficult to find the correct confidence interval for ϕ and then convert it back to an interval for θ by $\theta = g^{-1}(\phi)$. The BC_a method produces this interval for θ automatically, without requiring any knowledge of the transformation to form (2.3). This is the gist of Lemma 1 in Section 3.

The difference between (2.2) and (2.3) is greater than it seems. The hypothesized ideal transformation g leading to (2.2) must be both *normalizing* and *variance stabilizing*, whereas in (2.3) g need be only normalizing. Efron (1982b) shows that normalization and stabilization are partially antagonistic goals in familiar families such as the Poisson and the binomial. Schenker's counterexample to the BC method (1985), which helped motivate this article, is based on a family (discussed in Sec. 3) for which (2.2) fails. The main purpose of this article, to produce automatically intervals that are second-order correct, generally requires assumption (2.3) rather than (2.2).

It is not surprising that a theory based on (2.3) is usually more accurate than a theory based on (2.1). In fact, applied statisticians make frequent use of devices like those in (2.3), transformations, bias corrections, and even acceleration adjustments, to improve the performance of the standard intervals. The advantage of the BC_a method is that it automates these improvements, so the statistician does not have to think them through anew for each new application.

The bootstrap was originally introduced as a nonparametric Monte Carlo device for estimating standard errors. The basic idea, however, can be applied to any statistical problem, including parametric ones, and does not necessarily require Monte Carlo simulations. We will begin our discussion of the BC_a method by considering the simplest type of parametric problem: where the data consists only of a single real-valued statistic $\hat{\theta}$ in a one-parameter family of densities $f_\theta(\hat{\theta})$, say $\hat{\theta} \sim f_\theta$, and where we want a confidence interval for θ based on $\hat{\theta}$.

Sections 3, 4, and 5 describe the BC_a method in this simple setting, show how to calculate it from bootstrap computations, and demonstrate that it gives second-order correct intervals for θ under reasonable conditions.

Of course there is no need for the bootstrap in the simple situation $\hat{\theta} \sim f_\theta$, since then it is usually quite easy to calculate exact confidence intervals for θ. There are three reasons for beginning the discussion with the simple situation: (a) it makes clear the logic of the BC_a method; (b) it makes possible the comparison of BC_a intervals with exact intervals, exact intervals usually not existing in complicated problems; (c) it then turns out to be quite easy to extend the BC_a method to complicated situations, where it is more likely to be needed.

The simple situation $\hat{\theta} \sim f_\theta$ can be made more complicated, and more realistic, in two ways: the data can consist of a vector $\mathbf{y}$ instead of a single summary statistic $\hat{\theta}$, and the parameter can be a vector $\boldsymbol{\eta}$ instead of a single unknown scalar θ. Section 6 considers multiparameter families $f_{\boldsymbol{\eta}}(\mathbf{y})$, where we wish to set an approximate confidence interval for a real-valued function $\theta = t(\boldsymbol{\eta})$.

Our approach is to reduce the problem back to the simple situation. The data vector $\mathbf{y}$ is replaced by an efficient estimator $\hat{\theta}$ of θ, perhaps the MLE, and the multiparameter family $f_{\boldsymbol{\eta}}$ is replaced by a *least favorable* one-parameter family. All of the calculations are handled automatically by the BC_a algorithm. For a class of examples, including the Fieller problem of Table 1, the BC_a method automatically produces second-order correct intervals, but a proof of general second-order correctness does not yet exist for multiparameter situations.

Section 7 returns to the original nonparametric setting of the bootstrap: the data $\mathbf{y}$ is assumed to be a random sample $x_1, x_2, \ldots, x_n$ from a completely unknown probability distribution F. We wish to set an approximate confidence interval for $\theta = t(F)$, some real-valued function of F. The BC_a method extends in a natural way to the

nonparametric setting. In the case where θ is the expectation, theoretical analysis shows the BC_a intervals performing reasonably. Except for the case of the expectation, not much is proved about nonparametric BC_a intervals, though the empirical results look promising. Section 8 develops a heuristic justification for the nonparametric BC_a method in terms of the geometry of multinomial sampling.

In the simple situation $\hat{\theta} \sim f_\theta$ the parametric bootstrap distribution $\hat{\theta}^* \sim f_{\hat{\theta}}$ can often be written down explicitly, or at least approximated by standard parametric devices such as Edgeworth expansions. The number of bootstrap replications of $\hat{\theta}^*$, to use the terminology of previous papers, is then, effectively, infinity. For more complicated situations like the nonparametric confidence interval problem, Monte Carlo sampling is usually needed to calculate the BC_a intervals. How many bootstrap replications are necessary? The answer, on the order of 1,000, is derived in Section 9. This compares with only about 100 bootstrap replications necessary to adequately calculate a bootstrap standard error. Bootstrap confidence intervals require a lot more computation than bootstrap standard errors, if second-order accuracy is desired.

To get the main ideas across, some important technical points are deferred until Sections 10–12.

3. BOOTSTRAP CONFIDENCE INTERVALS FOR SIMPLE PARAMETRIC SITUATIONS

We first consider the simple situation $\hat{\theta} \sim f_\theta$, where we have a one-parameter family of densities $f_\theta(\hat{\theta})$ for the real-valued statistic $\hat{\theta}$. We wish to set a confidence interval for θ having observed only $\hat{\theta}$. The statistic $\hat{\theta}$ estimates θ. Later we will make specific assumptions about the properties of $\hat{\theta}$ as an estimator of θ—essentially that $\hat{\theta}$ behaves like the MLE asymptotically, though $\hat{\theta}$ may be some first-order efficient estimator other than the MLE.

By definition, the parametric bootstrap distribution in this simple situation is

$$\hat{\theta}^* \sim f_{\hat{\theta}}. \tag{3.1}$$

In other words it is the distribution of the statistic of interest when the unknown parameter θ is set equal to the observed point estimate $\hat{\theta}$. We also need to define the cdf of the bootstrap distribution

$$\hat{G}(s) = \int_{-\infty}^{s} f_{\hat{\theta}}(\hat{\theta}^*)d\hat{\theta}^* = \Pr_{\hat{\theta}}\{\hat{\theta}^* < s\}. \tag{3.2}$$

The integral is replaced by a summation in discrete families. The goal of bootstrap theory is to make inferential statements on the basis of the bootstrap distribution. In this article the inferences are approximate confidence intervals for θ.

Example (chi-squared scale family). Suppose that

$$\hat{\theta} \sim \theta(\chi^2_{19}/19), \tag{3.3}$$

the example considered in Schenker (1985). Then

$$f_\theta(\hat{\theta}) = c(\hat{\theta}/\theta)^{8.5}e^{-9.5(\hat{\theta}/\theta)} \quad \text{for } \hat{\theta} > 0 \qquad [c = 9.5^{9.5}/\Gamma(9.5)]. \tag{3.4}$$

Having observed $\hat{\theta}$, the bootstrap distribution $\hat{\theta}^* \sim \hat{\theta}(\chi^2_{19}/19)$ has density $c(\hat{\theta}^*/\hat{\theta})^{8.5}e^{-9.5(\hat{\theta}^*/\hat{\theta})}$ for $\hat{\theta}^* > 0$. The bootstrap cdf is $\hat{G}(s) = I_{9.5}(9.5s/\hat{\theta})$, where $I_{9.5}$ indicates the incomplete gamma function of degree 9.5.

Now suppose that for a family $\hat{\theta} \sim f_\theta$ there exists a monotone-increasing transformation g and constants z_0 and a such that

$$\hat{\phi} = g(\hat{\theta}), \qquad \phi = g(\theta) \tag{3.5}$$

satisfy

$$\hat{\phi} = \phi + \sigma_\phi(Z - z_0), \qquad Z \sim N(0, 1) \tag{3.6}$$

with

$$\sigma_\phi = 1 + a\phi. \tag{3.7}$$

This is of form (2.3), with $\tau = 1$. [Eq. (2.3) can always be reduced to the case $\tau = 1$; see Remark A, Sec. 11.] We will assume that $\phi > -1/a$ if $a > 0$ in (3.7), so $\sigma_\phi > 0$, and likewise $\phi < -1/a$ if $a < 0$. The constant a will typically be in the range $|a| < .2$, as will z_0.

Let Φ denote the standard normal cdf, and let $\hat{G}^{-1}(\alpha)$ denote the $100 \cdot \alpha$ percentile of the bootstrap cdf (3.2).

Lemma 1. Under conditions (3.5)–(3.7), the correct central confidence interval of level 1-2α for θ is

$$\theta \in [\hat{G}^{-1}(\Phi(z[\alpha])), \hat{G}^{-1}(\Phi(z[1 - \alpha]))], \tag{3.8}$$

where

$$z[\alpha] = z_0 + \frac{(z_0 + z^{(\alpha)})}{1 - a(z_0 + z^{(\alpha)})}, \tag{3.9}$$

and likewise for $z[1 - \alpha]$.

The proof of Lemma 1, at the end of this section, makes clear that interval (3.8) is correct in a strong sense: it is equivalent, under assumptions (3.5)–(3.7), to the usual obvious interval for a simple translation problem. Given the bootstrap cdf $\hat{G}(s)$ and values of z_0 and a derived from bootstrap calculations as in the following sections, we can form interval (3.8), (3.9) for θ whether or not assumptions (3.5)–(3.7) apply. This by definition is the BC_a interval for θ.

If z_0 and a equal 0, then $z[\alpha] = z^{(\alpha)}$ and (3.8) becomes $\theta \in [\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1 - \alpha)]$. In this case we just use the obvious percentiles of the bootstrap distribution to form an approximate confidence interval for θ, an approach called the *percentile method* in Efron (1981, 1982a). In general z_0 and a do not equal zero, and formulas (3.8), (3.9) make adjustments to the percentile method that are necessary to achieve second-order correctness.

Continuing example (3.3), the theory of Efron (1982b) shows that for the chi-squared scale family we can find a transformation g very nearly satisfying (3.5)–(3.7). Schenker (1985) proved the same result by a different method. The constants

$$z_0 = .1082, \qquad a = .1077 \tag{3.10}$$

and the transformation g appropriate to family (3.3) are derived in Section 10 and Remark E of Section 11. Simple ways of approximating z_0 and a for general families $\hat{\theta} \sim f_\theta$ are given in Section 4.

Line 2 of Table 2 shows the central 90% BC_a interval, $\alpha = .05$, for family (3.3), with $\hat{G}(s) = I_{9.5}(9.5s/\hat{\theta})$ and z_0 and a as in (3.10). The exact confidence interval is $\theta \in [19\hat{\theta}/\chi^{2(1-\alpha)}_{19}, 19\hat{\theta}/\chi^{2(\alpha)}_{19}]$, where $\chi^{2(\alpha)}_{19}$ is the $100 \cdot \alpha$ percentile point of a χ^2_{19} distribution. Notice how closely the BC_a endpoints match those of the exact interval (see line 1). The standard interval (1.1) is quite inaccurate in this case.

Suppose that we set $a = 0$ in (3.9), so $z[\alpha] = 2z_0 + z^{(\alpha)}$. Interval (3.8) with this definition of $z[\alpha]$ and $z[1 - \alpha]$ is called the *BC interval,* short for bias-corrected bootstrap interval, in Efron (1981, 1982a). In other words, BC = BC_a, with $a = 0$. The constant z_0 is easier to obtain than the constant a, as discussed in the next section, which is why the BC interval might be used. Line 3 of Table 2 shows that for family (3.3) the BC interval is a definite improvement over the standard interval but goes only about half as far as it should toward achieving the asymmetry of the exact interval.

The Fieller situation of Table 1 is an example of a class of multiparameter problems for which $a = 0$, so the BC and BC_a intervals coincide. Efron (1985) showed that the BC intervals are second-order correct for this class, as discussed in Section 6. In general problems the full BC_a method is necessary to get second-order correctness, as shown in Section 5.

Bartlett (1953) and Schenker (1985) discussed problem (3.3). The BC_a method can be thought of as a computer-based way to carry out Bartlett's program of improved approximate confidence intervals without having to do his theoretical calculations.

Proof of Lemma 1. We begin by showing that the BC_a interval for ϕ based on $\hat{\phi}$ is correct in a certain obvious sense: notice that (3.6), (3.7) give

$$\{1 + a\hat{\phi}\} = \{1 + a\phi\}\{1 + a(Z - z_0)\}. \quad (3.11)$$

Taking logarithms puts the problem into standard translation form,

$$\hat{\zeta} = \zeta + W, \quad (3.12)$$

$\hat{\zeta} = \log\{1 + a\hat{\phi}\}$, $\zeta = \log\{1 + a\phi\}$, and $W = \log\{1 + a(Z - z_0)\}$. This example was discussed more carefully in Sections 4 and 8 of Efron (1982b), where the possibility of the bracketed terms in (3.11) being negative was dealt with. Here it will cause no trouble to assume them positive so that it is permissible to take logarithms. In fact the transformation to (3.12) is only for motivational purposes. A quicker but less informative proof of Lemma 1 is possible, working directly on the ϕ scale.

Table 2. Central 90% Confidence Intervals for θ Having Observed $\hat{\theta} \sim \theta\chi^2_{19}/19$

		R/L
1. Exact	$[.631\hat{\theta}, 1.88\hat{\theta}]$	2.38
2. BC_a ($a = .1077$)	$[.630\hat{\theta}, 1.88\hat{\theta}]$	2.37
3. BC ($a = 0$)	$[.580\hat{\theta}, 1.69\hat{\theta}]$	1.64
4. Standard (1.1)	$[.466\hat{\theta}, 1.53\hat{\theta}]$	1.00
5. Nonparametric BC_a	$[.640\hat{\theta}, 1.68\hat{\theta}]$	1.88

NOTE: The BC_a interval, with $a = .1077$, the correct value, is nearly identical to the exact interval. The BC interval, $a = 0$, is only a partial improvement over the standard interval. The nonparametric BC_a interval is discussed in Section 7.

The translation problem (3.12) gives a natural central $1 - 2\alpha$ interval for ζ having observed $\hat{\zeta}$,

$$\zeta \in [\hat{\zeta} - w^{(1-\alpha)}, \hat{\zeta} - w^{(\alpha)}], \quad (3.13)$$

where $w^{(\alpha)}$ is the $100 \cdot \alpha$ percentile point for W, $\Pr\{W < w^{(\alpha)}\} = \alpha$.

We will use the notation $\theta[\alpha]$ for the α-level endpoint of a confidence interval for a parameter θ. For instance (3.13) says that $\zeta[\alpha] = \hat{\zeta} - w^{(1-\alpha)}$, $\zeta[1 - \alpha] = \hat{\zeta} - w^{(\alpha)}$. The interval (3.13) can be transformed back to the ϕ scale by the inverse mappings $\hat{\phi} = (e^{\hat{\zeta}} - 1)/a$, $\phi = (e^{\zeta} - 1)/a$, $(Z - z_0) = (e^W - 1)/a$. A little algebraic manipulation shows that the resulting interval for ϕ has α-level endpoint

$$\phi[\alpha] = \hat{\phi} + \sigma_{\hat{\phi}} \frac{(z_0 + z^{(\alpha)})}{1 - a(z_0 + z^{(\alpha)})}. \quad (3.14)$$

The cdf of $\hat{\phi}$ according to (3.6) is $\Phi((s - \phi)/\sigma_\phi + z_0)$, so the bootstrap cdf of $\hat{\phi}^*$, say $\hat{H}$, is $\hat{H}(s) = \Phi((s - \hat{\phi})/\sigma_{\hat{\phi}} + z_0)$. This has inverse $\hat{H}^{-1}(\alpha) = \hat{\phi} + \sigma_{\hat{\phi}}\{\Phi^{-1}(\alpha) - z_0\}$, which shows that $\hat{H}^{-1}(\Phi(z[\alpha]))$ equals (3.14) [see definition (3.9)]. In other words, the BC_a interval for ϕ, based on $\hat{\phi}$, coincides with the correct interval (3.14), "correct" meaning in agreement with the translation interval (3.13).

The BC_a intervals transform in the obvious way: if $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$, then the BC_a interval endpoints satisfy $\phi[\alpha] = g(\theta[\alpha])$. This follows directly from (3.8), (3.9) and the relationship $\hat{H}(g(s)) = \hat{G}(s)$, equivalently $\hat{H}^{-1}(\alpha) = g(\hat{G}^{-1}(\alpha))$, between the two bootstrap cdf's. Lemma 1 has now been verified: the transformations $\hat{\theta} \to \hat{\phi} \to \hat{\zeta}$ and $\theta \to \phi \to \zeta$ reduce the problem to translation form (3.12); the inverse transformations of the natural interval (3.13) for ζ produce the BC_a interval (3.8), (3.9).

4. THE TWO CONSTANTS

The BC_a intervals require the statistician to calculate the bootstrap distribution $\hat{G}$ and also the two constants z_0 and a. The bootstrap distribution is obtained directly from (3.2). This calculation does not require knowledge of the normalizing transformation g occurring in (3.5). The two constants z_0 and a can also be obtained, or at least approximated, directly from the bootstrap distribution $f_{\hat{\theta}}(\hat{\theta}^*)$. These calculations, which are the subject of this section, assume that a transformation g to form (3.6), (3.7) exists, but do not require g to be known.

In fact the bias-correction constant z_0 is

$$z_0 = \Phi^{-1}(\hat{G}(\hat{\theta})) \quad (4.1)$$

under assumptions (3.5)–(3.7), and so can be computed directly from $\hat{G}$. To verify (4.1) notice that

$$\Pr_\phi\{\hat{\phi} < \phi\} = \Pr\{Z < z_0\} = \Phi(z_0) \quad (4.2)$$

according to (3.6). However, (3.5) gives

$$\Pr_\theta\{\hat{\theta} < \theta\} = \Pr_\phi\{\hat{\phi} < \phi\} = \Phi(z_0) \quad (4.3)$$

for every value of θ. Substituting $\theta = \hat{\theta}$ gives $\hat{G}(\hat{\theta}) = \Pr_{\hat{\theta}}\{\hat{\theta}^* < \hat{\theta}\} = \Phi(z_0)$, which is (4.1).

What about the acceleration constant a? We will show that a good approximation for a is

$$a \doteq \tfrac{1}{6}\,\mathrm{SKEW}_{\theta=\hat{\theta}}(\dot{l}_\theta), \quad (4.4)$$

where $\text{SKEW}_{\theta=\hat{\theta}}(X)$ indicates the skewness of a random variable X, $\mu_3(X)/\mu_2(X)^{3/2}$, evaluated at parameter point θ equal to $\hat{\theta}$, and $\dot{l}_\theta$ is the score function of the family $f_\theta(\hat{\theta})$,

$$\dot{l}_\theta(\hat{\theta}) = \partial/\partial\theta \log f_\theta(\hat{\theta}). \tag{4.5}$$

Formula (4.4) allows us to calculate a from the form of the given density f_θ near $\theta = \hat{\theta}$, without knowing g. Sections 6 and 7 discuss the computation of a in families with nuisance parameters. Section 10 gives a deeper discussion of a and its relationship to other quantities of interest. See also Remark F, Section 11.

Example. For $\hat{\theta} \sim \theta(\chi^2_{19}/19)$, as in Table 2, standard χ^2 calculations give $\text{SKEW}(\dot{l}_\theta)/6 = [8/(19 \cdot 36)]^{1/2} = .1081$, which is quite close to the actual value $a = .1077$ derived in Section 10.

Here is a simple heuristic argument that indicates the role of the constant a in setting approximate confidence intervals. Suppose that $z_0 = 0$ and $a > 0$ in (3.6), (3.7). Having observed $\hat{\phi} = 0$, and noticing $\sigma_\phi = 1$, the naive interval for ϕ [which is almost the same as the standard interval (1.1)] is $\phi \in [z^{(\alpha)}, z^{(1-\alpha)}]$. If, however, the statistician checks the situation at the right endpoint $z^{(1-\alpha)}$, he finds that the hypothesized standard deviation of $\hat{\phi}$ has increased from 1 to $1 + az^{(1-\alpha)}$. This suggests increasing the right endpoint to $z^{(1-\alpha)}(1 + az^{(1-\alpha)})$. Now the hypothesized standard deviation has further increased to $1 + az^{(1-\alpha)}(1 + az^{(1-\alpha)})$, suggesting a still larger right endpoint, and so forth. Continuing on in this way results in formula (3.14), leading to Lemma 1. [Improving the standard interval (1.1) by recomputing $\hat{\sigma}$ at its endpoints is a useful idea. It was brought to my attention by John Tukey, who pointed out its use by Bartlett (1953); see, e.g., Bartlett's eq. (17). Tukey's (1949) unpublished talk anticipated many of the same points.]

We call a the acceleration constant because of its effect of constantly changing the natural units of measurement as we move along the ϕ (or θ) axis. Notice that we can write (3.7) as

$$\sigma_\phi = \sigma_{\phi_0}[1 + a(\phi - \phi_0)/\sigma_{\phi_0}], \tag{4.6}$$

so

$$a = \frac{d(\sigma_\phi/\sigma_{\phi_0})}{d((\phi - \phi_0)/\sigma_{\phi_0})} \tag{4.7}$$

for any fixed value of ϕ_0. This shows that a is the relative change in σ_ϕ per unit standard deviation change in ϕ, no matter what value ϕ has.

The point $\phi_0 = 0$ is favored in definition (3.7), since σ_0 has been set equal to the convenient value 1. There is no harm in thinking of 0 as the true value of ϕ, the value actually governing the distribution of $\hat{\phi}$ in (3.8), because in theory we can always choose the transformation g so that this is the case and, in addition, so that $\sigma_0 = 1$ (see Remark A, Sec. 11). The restriction $1 + a\phi > 0$ in (3.7) causes no practical trouble for $|a| \leq .2$, since it is then at least 5 standard deviations to the boundary of the permissible ϕ region.

The remainder of this section is devoted to verifying (4.4). The discussion is fairly technical and can be deferred until Section 10 at the reader's preference.

If we make smooth one-to-one transformations $\hat{\phi} = g(\hat{\theta})$, $\phi = h(\theta)$, then $\dot{l}_\phi(\hat{\phi}) = \dot{l}_\theta(\hat{\theta})/h'(\theta)$ and $\text{SKEW}(\dot{l}_\phi) = \text{SKEW}(\dot{l}_\theta)$. In other words, the right side of (4.4) is invariant under all mappings of this type. Suppose that for some choice of g and h, we can represent the family of distributions of $\hat{\phi}$ as

$$\hat{\phi} = \phi + \sigma_\phi q(Z), \qquad Z \sim N(0, 1), \tag{4.8}$$

where σ_ϕ and $q(Z)$ are functions of ϕ and z, having at least one and two derivatives, respectively, $q'(Z) > 0$. Situation (4.8), with the added conditions $q(0) = 0$, $q'(0) = 1$, is called a general scaled transformation family (GSTF) in Efron (1982b). [Please note the corrigenda to Efron (1982b).]

Lemma 2. The family (4.8) has score function $\dot{l}_\phi(\hat{\phi})$ satisfying

$$\sigma_\phi \dot{l}_\phi(\hat{\phi}) \sim \left[Z + \frac{q''(Z)}{q'(Z)}\right]\left[\frac{1 + \dot{\sigma}_\phi q(Z)}{q'(Z)}\right] - \dot{\sigma}_\phi, \qquad Z \sim N(0, 1). \tag{4.9}$$

Here $\dot{\sigma}_\phi = d\sigma_\phi/d\phi$ and q' and q'' are the first two derivatives of q.

Before presenting the proof of Lemma 2, we note that it verifies (4.4): in situation (3.6), (3.7), where $\dot{\sigma}_\phi = a$, $q'(Z) = 1$, $q''(Z) = 0$, the distributional relationship (4.9) becomes

$$\sigma_\phi \dot{l}_\phi(\hat{\phi}) \sim (1 - az_0)\left[Z + \frac{a}{1 - az_0}(Z^2 - 1)\right]. \tag{4.10}$$

Let

$$\varepsilon_0 = \frac{a}{1 - az_0}, \tag{4.11}$$

a quantity discussed in Section 10. From the moments of $Z \sim N(0, 1)$, (4.10) gives

$$\frac{\text{SKEW}(\dot{l}_\phi)}{6} = \varepsilon_0 \frac{1 + \frac{4}{3}\varepsilon_0^2}{(1 + 2\varepsilon_0^2)^{3/2}}. \tag{4.12}$$

We will see in Section 10 that for the usual repeated sampling situation both a and z_0 are order of magnitude $O(n^{-1/2})$ in the sample size n. This means that $\varepsilon_0 = a \cdot [1 + O(n^{-1})]$, (4.11), and that $\text{SKEW}(\dot{l}_\theta)/6 = \text{SKEW}(\dot{l}_\phi)/6 = a[1 + O(n^{-1})]$, (4.12), justifying approximation (4.4). The "constant" a actually depends on θ, but substituting $\theta = \hat{\theta}$ in (4.4) causes errors only at the third-order level, like $\hat{\sigma}B_n^{(\alpha)}/n$ in (1.2), and so does not affect the second-order properties of the BC_a intervals.

Proof of Lemma 2. Starting from (4.8), the cdf of $\hat{\phi}$ is $\Phi(q^{-1}((\hat{\phi} - \phi)/\sigma_\phi))$, so $\hat{\phi}$ has density $f_\phi(\hat{\phi}) = \exp(-\frac{1}{2}Z_\phi^2)/(\sqrt{2\pi}\,\sigma_\phi q'(Z_\phi))$, where $Z_\phi \equiv q^{-1}((\hat{\phi} - \phi)/\sigma_\phi)$. This gives log-likelihood function

$$l_\phi(\hat{\phi}) = -\tfrac{1}{2}Z_\phi^2 - \log(q'(Z_\phi)) - \log(\sigma_\phi). \tag{4.13}$$

Lemma 2 follows by differentiating (4.13) with respect to ϕ and noting that $Z_\phi \sim N(0, 1)$ when sampling from (4.8).

5. SECOND-ORDER CORRECTNESS OF THE BC_a INTERVALS

The standard intervals are based on approximation (2.1). The BC_a intervals, which improved considerably on the standard intervals in Tables 1 and 2, are based on the more general approximation (2.3). Is it possible to go beyond (2.3), to find still further improvements over the standard intervals? The answer is no, at least not in terms of second-order asymptotics. The theorem of this section states that for simple one-parameter problems the BC_a intervals coincide through second order with the exact intervals. In terms of (1.2), the BC_a intervals have the correct second-order asymptotic form $\hat\theta + \hat\sigma(z^{(\alpha)} + A_n^{(\alpha)}/\sqrt{n} + \cdots)$.

We continue to consider the simple one-parameter problem $\hat\theta \sim f_\theta$. Suppose that the $100 \cdot \alpha$ percentile of $\hat\theta$ as a function of θ, say $\hat\theta_\theta^{(\alpha)}$, is a continuously increasing function of θ for any fixed α. In this case the usual confidence interval construction gives an exact $1 - 2\alpha$ central interval for θ having observed $\hat\theta$, say $[\theta_{Ex}[\alpha], \theta_{Ex}[1 - \alpha]]$, where $\theta_{Ex}[\alpha]$ is the value of θ satisfying $\hat\theta_\theta^{(1-\alpha)} = \hat\theta$. The exact interval in Table 2 is an example of this construction.

It is not necessary that $\hat\theta$ be the MLE of θ. In (3.6), for instance, $\hat\phi$ is not the MLE of ϕ. The BC_a method is quite insensitive to small changes in the form of the estimator (see Remark B, Sec. 11). It will be assumed, however, that $\hat\theta$ behaves asymptotically like the MLE in terms of the orders of magnitude of its bias, standard deviation, skewness, and kurtosis,

$$\hat\theta - \theta \sim (B_\theta/n,\ C_\theta/\sqrt{n},\ D_\theta/\sqrt{n},\ E_\theta/n). \tag{5.1}$$

Here n is the sample size upon which the summary statistic $\hat\theta$ is based; B_θ, C_θ, D_θ, and E_θ are bounded functions of θ (and of n, which is suppressed in the notation). Then (5.1) says that the bias of $\hat\theta$, B_θ/n, is $O(n^{-1})$, the standard deviation $C_\theta/\sqrt{n}$ is $O(n^{-1/2})$, skewness $O(n^{-1/2})$, and kurtosis $O(n^{-1})$. Higher cumulants, which are typically of order smaller than $O(n^{-1})$, will be assumed negligible in proving the results that follow (see DiCiccio 1984; Hougaard 1982).

In the simple situation $\hat\theta \sim f_\theta$, $\hat\theta$ is a sufficient statistic for θ. Later when we consider more complicated problems we will take $\hat\theta$ to be the MLE of θ. This guarantees that $\hat\theta$ is first-order efficient and asymptotically sufficient (Efron 1975).

The asymptotics of this article are stated relative to the size of the estimated standard error $\hat\sigma$ of $\hat\theta$, as in (1.2). It is often convenient in what follows to have $\hat\sigma$ be $O_p(1)$. This is easy to accomplish by transforming to $\hat\phi \equiv \sqrt{n}\hat\theta$, $\phi \equiv \sqrt{n}\theta$, so (5.1) becomes

$$\hat\phi - \phi \sim (\beta_\phi,\ \sigma_\phi,\ \gamma_\phi,\ \delta_\phi), \tag{5.2}$$

where $\beta_\phi = B_{\phi/n^{1/2}}/n^{1/2}$, $\sigma_\phi = C_{\phi/n^{1/2}}$, $\gamma_\phi = D_{\phi/n^{1/2}}/n^{1/2}$, and $\delta_\phi = E_{\phi/n^{1/2}}/n$. Notice that $\beta_\phi = O(n^{-1/2})$, $\dot\beta_\phi \equiv d\beta_\phi/d\phi = O(n^{-1})$, and so forth. We can just assume to begin with that $\hat\theta$ and θ are the rescaled quantities previously called $\hat\phi$ and ϕ. Then the following orders of magnitude apply:

$$\begin{array}{cccc} O(1) & O(n^{-1/2}) & O(n^{-1}) & O(n^{-3/2}) \\ \sigma_\theta & \dot\sigma_\theta,\ \beta_\theta,\ \gamma_\theta & \ddot\sigma_\theta,\ \dot\beta_\theta,\ \dot\gamma_\theta,\ \delta_\theta & \dddot\beta_\theta,\ \ddot\gamma_\theta,\ \dot\delta_\theta \end{array}. \tag{5.3}$$

Theorem 1. If $\hat\theta$ has bias β_θ, standard error σ_θ, skewness γ_θ, and kurtosis δ_θ satisfying (5.3), then the BC_a intervals are second-order correct.

The theorem states that $\theta_{BC_a}[\alpha]$, the α endpoint of the BC_a interval, is asymptotically close to the exact endpoint,

$$(\theta_{BC_a}[\alpha] - \theta_{Ex}[\alpha])/\hat\sigma = O_p(n^{-1}). \tag{5.4}$$

This is not true for the standard intervals (1.1) or the BC intervals, $a = 0$. The proof of Theorem 1, which appears in Section 12, makes it clear that all three of the elements in (2.3), the transformation g, the bias-correction constant z_0, and the acceleration constant a, make necessary corrections of $O_p(n^{-1/2})$ to the standard intervals.

6. NUISANCE PARAMETERS

The discussion so far has centered on the simple case $\hat\theta \sim f_\theta$, where we have only a real-valued parameter θ and a real-valued summary statistic $\hat\theta$ from which we are trying to construct a confidence interval for θ. We have been able to show favorable properties of the BC_a intervals for the simple case, but of course the simple case is where we least need a general method like the bootstrap.

This section discusses the more difficult situation where there are nuisance parameters besides the parameter of interest θ. Section 7 discusses the nonparametric situation, where the number of nuisance parameters is effectively infinite. Because of the inherently simple nature of the bootstrap it will be easy to extend the BC_a method to cover these cases, though we will not be able to provide as strong a justification for the correctness of the resulting intervals.

Suppose then that the data $\mathbf{y}$ comes from a parametric family $\mathcal{F}$ of density functions $f_{\boldsymbol\eta}$, say $\mathbf{y} \sim f_{\boldsymbol\eta}$, where $\boldsymbol\eta$ is an unknown vector of parameters, and we want a confidence interval for the real-valued parameter $\theta = t(\boldsymbol\eta)$. In Efron (1985), the multivariate normal case $\mathbf{y} \sim N_k(\boldsymbol\eta, \mathbf{I})$ is examined in detail.

From $\mathbf{y}$ we obtain $\hat{\boldsymbol\eta}$, the MLE of $\boldsymbol\eta$, and $\hat\theta = t(\hat{\boldsymbol\eta})$, the MLE of θ. The parametric bootstrap distribution of $\mathbf{y}$ is defined to be

$$\mathbf{y}^* \sim f_{\hat{\boldsymbol\eta}}, \tag{6.1}$$

the distribution of the data when $\boldsymbol\eta$ equals $\hat{\boldsymbol\eta}$. From $\mathbf{y}^*$ we obtain $\hat{\boldsymbol\eta}^*$, the bootstrap MLE of $\boldsymbol\eta$, and then $\hat\theta^* = t(\hat{\boldsymbol\eta}^*)$.

The distribution of $\hat\theta^*$ under model (6.1) is the parametric boostrap distribution of $\hat\theta$, generalizing (3.1). This gives the bootstrap cdf

$$\hat G(s) = \Pr_{\hat{\boldsymbol\eta}}\{\hat\theta^* < s\}, \tag{6.2}$$

as in (3.2). The bias-correction constant z_0 equals $\Phi^{-1}(\hat G(\hat\theta))$, as in (4.1).

To compute the BC_a intervals (3.8), (3.9), we also need to know the appropriate value of the acceleration constant a. We will find a by following Stein's (1956) construction,

which replaces the multiparameter family $\mathcal{F} = \{f_\eta\}$ by a *least favorable* one-parameter family $\hat{\mathcal{F}}$.

Let $\dot{l}_\eta$ be the vector with ith coordinate $\partial/\partial\eta_i \log f_\eta(\mathbf{y})$, so $\dot{l}_{\hat\eta}(\mathbf{y}) = 0$ by definition of the MLE $\hat\eta$, and let $\ddot{l}_{\hat\eta}$ be the $k \times k$ matrix with ijth entry $\partial^2/(\partial\eta_i\partial\eta_j) \log f_\eta(\mathbf{y})|_{\eta=\hat\eta}$. In addition, let $\hat{\nabla}$ be the gradient vector of $\theta = t(\eta)$ evaluated at the MLE, $\hat\nabla_i = (\partial/\partial\eta_i)t(\eta)|_{\eta=\hat\eta}$. The *least favorable direction* at $\eta = \hat\eta$ is defined to be

$$\hat{\mu} \equiv (-\ddot{l}_{\hat\eta})^{-1}\hat{\nabla}. \tag{6.3}$$

Then the least favorable family $\hat{\mathcal{F}}$ is the one-parameter subfamily of $\mathcal{F}$ passing through $\hat\eta$ in the direction $\hat\mu$,

$$\hat{\mathcal{F}} = \{\hat{f}_\lambda(\mathbf{y}^*) \equiv f_{\hat\eta+\lambda\hat\mu}(\mathbf{y}^*)\}. \tag{6.4}$$

Using $\mathbf{y}^*$ to denote a hypothetical data vector from $\hat f_\lambda$ is intended to avoid confusion with the actual data vector $\mathbf{y}$ that gave $\hat\eta$; $\hat\eta$ and $\hat\mu$ are fixed in (6.4), only λ being unknown.

Consider the problem of estimating $\theta(\lambda) \equiv t(\hat\eta + \lambda\hat\mu)$ having observed $\mathbf{y}^* \sim \hat f_\lambda$. The Fisher information bound for an unbiased estimate of θ in this one-parameter family evaluated at $\lambda = 0$ is $\hat\nabla'(-\ddot{l}_{\hat\eta})^{-1}\hat\nabla$, which is the same as the corresponding bound for estimating $\theta = t(\eta)$, at $\eta = \hat\eta$, in the multiparameter family $\mathcal{F}$. This is Stein's reason for calling $\hat{\mathcal{F}}$ least favorable.

We will use $\hat{\mathcal{F}}$ to calculate an approximate value for the acceleration constant a,

$$a \doteq \{\mathrm{SKEW}_{\lambda=0}[\partial \log f_{\hat\eta+\lambda\hat\mu}(\mathbf{y}^*)/\partial\lambda]/6\}. \tag{6.5}$$

This is formula (4.4) applied to $\hat{\mathcal{F}}$, assuming that $\hat\lambda = 0$ (which is the MLE of λ in $\hat{\mathcal{F}}$ when $\mathbf{y}^* = \mathbf{y}$, the actual data vector). See Remark F, Section 11.

Formula (6.5) is especially simple in the exponential family case where the densities $f_\eta(\mathbf{y})$ are of the form

$$f_\eta(\mathbf{y}) = e^{n[\eta'\mathbf{y} - \psi(\eta)]}f_0(\mathbf{y}). \tag{6.6}$$

The factor n in the exponent of (6.6) is not necessary, but it is included to agree with the situation where the data consists of iid observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, each with density $\exp(\eta'\mathbf{x} - \psi(\eta))$, and $\mathbf{y}$ is the sufficient vector $\sum_{i=1}^n \mathbf{x}_i/n$.

Lemma 3. For the exponential family (6.6), formula (6.5) gives

$$a = \frac{1}{6\sqrt{n}} \frac{\hat\psi^{(3)}(0)}{(\hat\psi^{(2)}(0))^{3/2}}, \tag{6.7}$$

where

$$\hat\psi^{(j)}(0) = \frac{\partial^j \psi(\hat\eta + \lambda\hat\mu)}{\partial\lambda^j}\bigg|_{\lambda=0}. \tag{6.8}$$

Proof. We have

$$\frac{\partial \log f_{\hat\eta+\lambda\hat\mu}(\mathbf{y}^*)}{\partial\lambda}\bigg|_{\lambda=0} = n\hat\mu'(\mathbf{y}^* - \hat\psi(\hat\eta)), \tag{6.9}$$

so $\mathrm{SKEW}_{\lambda=0}[(\partial \log f_{\hat\eta+\lambda\hat\mu}(\mathbf{y}^*))/\partial\lambda]$ equals the skewness of $\hat\mu'\mathbf{y}^*$ for $\mathbf{y}^* \sim f_{\hat\eta}$. The fact that $\mathrm{SKEW}(\hat\mu'\mathbf{y}^*)$ equals $[\hat\psi^{(3)}(0)/(\hat\psi^{(2)}(0))^{3/2}]/\sqrt{n}$ is a standard exercise in exponential family theory. Note that Lemma 3 applies to $\mathbf{y} \sim N_k(\eta, \mathbf{I})$, the case considered in Efron (1985), and gives $a = 0$, which is why the unaccelerated BC intervals worked well there.

Table 3 relates to the following example:

$$\mathbf{y} \sim N_4(\eta, \sigma_\eta^2\mathbf{I}), \qquad [\sigma_\eta = 1 + a(\|\eta\| - 8)], \tag{6.10}$$

where we observe $\mathbf{y} = (8, 0, 0, 0)$ and wish to set confidence intervals for the parameter $\theta = t(\eta) = \|\eta\|$. The case $a = 0$ amounts to finding a confidence interval for the noncentrality parameter of a noncentral χ^2 distribution and can be solved exactly. The theory of Efron (1985) applies to the $a = 0$ case, and we see that the BC_0 interval, that is, the BC interval, well-matches the exact interval.

Table 3 shows the result of varying the constant a from .10 to $-.10$. This example has a particularly simple geometry: the sphere $C_{\hat\theta} = \{\eta : \|\eta\| = \hat\theta\}$ is the set of η vectors having $t(\eta)$ equal to the MLE value $\hat\theta = t(\hat\eta)$; the least favorable direction $\hat\mu$ is orthogonal to $C_{\hat\theta}$ at $\hat\eta$; the distribution of $\hat\theta$ is nearly normal (see Efron 1985, Table 2), with standard deviation changing in the least favorable direction at a rate nearly equal to a, as in (4.7). The BC_a intervals alter predictably with a. For instance, comparing the upper endpoint at $a = .10$ with $a = 0$, notice that $(9.70 - 8.00)/(9.44 - 8.00) = 1.18$, closely matching the expansion factor due to acceleration, $1 + .10 \cdot 1.645 = 1.16$.

We could disguise problem (6.10) by making nonlinear transformations

$$\tilde{\mathbf{y}} = g(\mathbf{y}), \qquad \tilde\eta = h(\eta), \tag{6.11}$$

in which case the geometry of the BC_a intervals might not be obvious from the form of the parameter $\theta = t(h^{-1}(\tilde\eta)) = \|h^{-1}(\tilde\eta)\|$ and the transformed densities $\tilde f_{\tilde\eta}(\mathbf{y})$. However, the BC_a method is invariant under such transformations (see Remark C, Sec. 11), so the statistician would automatically get the same intervals as if he knew the normalizing transformations $\mathbf{y} = g^{-1}(\tilde{\mathbf{y}})$, $\eta = h^{-1}(\tilde\eta)$.

Currently we cannot justify the BC_a method as being second-order correct in the multiparameter context of this section, though it seems a likely conjecture that this is so. We know that it is so in the one-parameter case (see Sec. 5) and in the restricted multiparameter case of Efron (1985), where the BC_a and BC methods coincide, and that the BC_a method makes a rather obvious correction to the BC interval in the general multiparameter case.

Table 3. Central 90% Confidence Intervals for $\theta = \|\eta\|$, Having Observed $\|y\| = 8$, From the Parametric Family $\mathbf{y} \sim N_4(\eta, \sigma_\eta^2\mathbf{I})$, With $\sigma_\eta = 1 + a(\|\eta\| - 8)$

	Exact	(R/L)	BC_a	(R/L)	(6.5)
$a = .10$	[6.46, 9.69]	(.96)	[6.47, 9.70]	(.97)	.0984
$a = .05$	[6.32, 9.57]	(.85)	[6.34, 9.56]	(.84)	.0498
$a = 0$	[6.14, 9.47]	(.74)	[6.19, 9.44]	(.75)	0
$a = -.05$	[5.92, 9.38]	(.65)	[6.03, 9.35]	(.66)	$-.0498$
$a = -.10$	[5.62, 9.30]	(.56)	[5.89, 9.27]	(.60)	$-.0984$

NOTE: The standard interval (1.1) is [6.36, 9.64] for all values of a. The last column shows that (6.5) nearly equals the constant a in this case. The exact intervals are based on the noncentral χ^2 distribution.

7. THE NONPARAMETRIC CASE

This section concerns the nonparametric case where the data $\mathbf{y} = (x_1, x_2, \ldots, x_n)$ consist of n iid observations x_i that may have come from any probability distribution F on their common sample space $\mathcal{X}$. There is a real-valued parameter $\theta = t(F)$ for which we desire an approximate confidence interval. We will show how the BC_a method can be used to provide such an interval based on the obvious nonparametric estimate $\hat{\theta} = t(\hat{F})$. Here $\hat{F}$ is the empirical probability distribution of the sample, putting mass $1/n$ on each observed value x_i.

A bootstrap sample $\mathbf{y}^* \sim \hat{F}$ consists in this case of an iid sample of size n from $\hat{F}$, say $\mathbf{y}^* = (x_1^*, x_2^*, \ldots, x_n^*)$. In other words, $\mathbf{y}^*$ is a random sample of size n drawn with replacement from $\{x_1, x_2, \ldots, x_n\}$. The bootstrap sample $\mathbf{y}^*$ gives a bootstrap replication of $\hat{\theta}$, $\hat{\theta}^* = t(\hat{F}^*)$, where $\hat{F}^*$ puts mass $1/n$ on each x_i^*. The bootstrap cdf $\hat{G}(s)$ is the probability that a bootstrap replication is less than s,

$$\hat{G}(s) = \Pr_{\hat{F}}\{\hat{\theta}^* < s\}, \tag{7.1}$$

as in (6.2) and (3.2). The bias-correction constant z_0 equals $\Phi^{-1}(\hat{G}(\hat{\theta}))$, as in (4.1).

For most nonparametric problems the bootstrap cdf $\hat{G}$ has to be determined by Monte Carlo sampling. Section 9 discusses how many Monte Carlo replications of $\hat{\theta}^*$ are necessary. Here we will continue to assume that $\hat{G}$ has been computed exactly—in effect, that we have taken an infinite number of bootstrap replications $\hat{\theta}^*$.

At this point we could use $\hat{G}$ to form the BC interval for θ, but to obtain the BC_a interval (3.8), (3.9) we also need the value of the acceleration constant a. We will derive a simple approximation for a, based on Lemma 3. It depends on

$$U_i = \lim_{\Delta \to 0} \frac{t((1-\Delta)\hat{F} + \Delta\delta_i) - t(\hat{F})}{\Delta}, \qquad i = 1, 2, \ldots, n, \tag{7.2}$$

the *empirical influence function* of $\hat{\theta} = t(\hat{F})$. Here δ_i is a point mass at x_i, so U_i is the derivative of the estimate $\hat{\theta}$ with respect to the mass on point x_i. [Jaeckel's infinitesimal jackknife estimate of standard error for $\hat{\theta}$ is $(\sum_1^n U_i^2)^{1/2}/n$.] Definition (7.2) assumes that $t(F)$ is smoothly defined for choices of F near $\hat{F}$ [see Efron 1982a, (6.3), or Efron 1979, sec. 5]. Note that $\sum_i^n U_i = 0$.

The next section shows that Lemma 3, applied to a family appropriate to the nonparametric situation, gives the following approximation for the constant a,

$$a \doteq \frac{1}{6}\left[\left(\sum_{i=1}^n U_i^3\right) \Big/ \left(\sum_{i=1}^n U_i^2\right)^{3/2}\right]. \tag{7.3}$$

This is a convenient formula since the U_i can be evaluated easily by using finite differences in definition (7.2).

Example 1: The Law School Data. Table 4 shows two indexes of student excellence, LSAT and GPA, for each of 15 American law schools (see Efron 1982a, sec. 2.2). The Pearson correlation coefficient $\hat{\rho}$ between LSAT and GPA equals .776; we want a confidence interval for the true correlation ρ. Table 4 also shows the values of U_i for the statistic $\hat{\rho}$, from which formula (7.3) produces $a \doteq -.0817$. $B = 100{,}000$ bootstrap replications (about 100 times more than actually needed; see Sec. 9) gave $\hat{G}(\hat{\theta}) = .463$, and so $z_0 = -.0927$. Using these values of a and z_0 in (3.8), (3.9) resulted in the central 90% nonparametric BC_a interval [.43, .92] for ρ. The usual bivariate normal interval, based on Fisher's $\tanh^{-1}$ transformation, is [.49, .90]. This is also the *parametric* BC_a interval based on the simple family $\hat{\rho} \sim f_\rho$, where $f_\rho(\hat{\rho})$ is Fisher's density function for the correlation coefficient from bivariate normal data. The standard interval (1.1), $\hat{\rho} \pm 1.645\hat{\sigma}$, using the bootstrap estimate $\hat{\sigma} = .133$, is [.56, .99].

Table 4. The Law School Data and Values of the Empirical Influence Function for the Correlation Coefficient $\hat{\rho}$

i	(LSAT, GPA)	U_i	i	(LSAT, GPA)	U_i
1	(576, 3.39)	−1.507	9	(651, 3.36)	.310
2	(635, 3.30)	.168	10	(605, 3.13)	.004
3	(558, 2.81)	.273	11	(653, 3.12)	−.526
4	(578, 3.03)	.004	12	(575, 2.74)	−.091
5	(666, 3.44)	.525	13	(545, 2.76)	.434
6	(580, 3.07)	−.049	14	(572, 2.88)	.125
7	(555, 3.00)	−.100	15	(594, 2.96)	−.048
8	(661, 3.43)	.477			

Formula (7.3) is invariant under monotone changes of the parameter of interest. This results in the BC_a intervals having correct transformation properties. Suppose, for example, that we change parameters from ρ to $\phi = g(\rho) \equiv \tanh^{-1}(\rho)$, with corresponding nonparametric estimate $\hat{\phi} = g(\hat{\rho})$. The central 90% BC_a interval for ϕ based on $\hat{\phi}$ is then the obvious transformation of the interval for θ based on $\hat{\theta}$, $[g(.43), g(.92)] = [.46, 1.59]$. This compares with Fisher's $\tanh^{-1}$ interval $[g(.49), g(.90)] = [.54, 1.47]$ and the standard interval $\hat{\phi} \pm 1.645\hat{\sigma}_\phi = [.49, 1.59]$. The standard interval is much more reasonable-looking on the $\tanh^{-1}$ scale, as we might expect from Fisher's transformation theory. As commented before, a major advantage of the BC_a method is that the statistician need not know the correct scale on which to work. In effect the method effectively selects the best (most normal) scale and then transforms the interval back to the scale of interest.

Example 2: The Mean. Suppose that F is a distribution on the real line, and $\theta = t(F)$ equals the expectation $E_F X$. The empirical influence function $U_i = (x_i - \bar{x})$, so (7.3) gives

$$\begin{aligned} a &= \tfrac{1}{6}\{\textstyle\sum (x_i - \bar{x})^3 / [\sum (x_i - \bar{x})^2]^{3/2}\} \\ &= (1/6\sqrt{n})(\hat{\mu}_3/\hat{\mu}_2^{3/2}) = \hat{\gamma}/6\sqrt{n}. \end{aligned} \tag{7.4}$$

Here $\hat{\mu}_h = \sum (x_i - \bar{x})^h/n$, the hth sample central moment, and $\hat{\gamma} = \hat{\mu}_3/\hat{\mu}_2^{3/2}$, the sample skewness. It turns out also that $z_0 \doteq \hat{\gamma}/6\sqrt{n}$ in this case, by standard Edgeworth arguments. Both a and z_0 are typically of order $n^{-1/2}$.

Because the sample mean is such a simple statistic, we can use Edgeworth methods to get asymptotic expressions for the α-level endpoint of the BC_a interval:

$$\theta_{BC_a}[\alpha] = \bar{x} + \hat{\sigma}\{z^{(\alpha)} + (\hat{\gamma}/6\sqrt{n})(2z^{(\alpha)^2} + 1) + O_p(n^{-1})\}, \tag{7.5}$$

$\hat{\sigma} \equiv (\hat{\mu}_2/n)^{1/2}$. This compares with

$$\theta_{BC}[\alpha] \doteq \bar{x} + \hat{\sigma}\{z^{(\alpha)} + (\hat{\gamma}/6\sqrt{n})(z^{(\alpha)^2} + 1) + O_p(n^{-1})\}, \tag{7.6}$$

for the BC interval, so the BC_a intervals are shifted approximately $(\hat{\gamma}/6\sqrt{n})z^{(\alpha)^2}$ further right.

Johnson (1978) suggested modifying the usual t statistic $T = (\bar{x} - \theta)/\hat{\sigma}$ to $T_J = T + (\hat{\gamma}/6\sqrt{n})(2T^2 + 1)$ and then considering T_J to have a standard t_{n-1} distribution in order to obtain confidence intervals for $\theta = E_F X$. Efron (1981, sec. 10) showed that this is much like using the bootstrap distribution of $T^* = (\bar{x}^* - \bar{x})/\hat{\sigma}^*$ as a pivotal quantity. Interestingly enough, *the Edgeworth expansion of $\theta_J[\alpha]$, the α endpoint of Johnson's interval, coincides with (7.5).* The BC_a method makes a "t correction" in the case of $\theta = E_F X$, but it is not the familiar Student-t correction, which operates at third order in (1.2), but rather a second-order correction, coming from the correlation between $\bar{x}$ and $\hat{\sigma}$ in nonnormal populations (see Remark D, Sec. 11).

I conjecture that the nonparametric BC_a intervals will be second-order correct for any parameter θ. There is no proof of this, a major difficulty being the definition of second-order correctness in the nonparametric situation. Whether or not it is true, small-sample nonparametric confidence intervals are far from well understood and, as emphasized in Schenker (1985), should be interpreted with some caution.

Example 3: The Variance. Suppose that $\mathcal{X}$ is the real line and $\theta = \text{var}_F X$, the variance. Line 5 of Table 2 shows the result of applying the nonparametric BC_a method to data sets $x_1, x_2, \ldots, x_{20}$, which were actually iid samples from an $N(0, 1)$ distribution. The number .640, for example, is the average of $\theta_{BC_a}[.05]/\hat{\theta}$ over 40 such data sets, $B = 4{,}000$ bootstrap replications per data set. The upper limit $1.68 \cdot \hat{\theta}$ is noticeably small. The reason is simple: the nonparametric bootstrap distribution of $\hat{\theta}^*$ has a short upper tail compared with the parametric bootstrap distribution, which is a scaled χ^2_{19} random variable. The results of Beran (1984a), Bickel and Freedman (1981), and Singh (1981) show that the nonparametric bootstrap distribution is highly accurate asymptotically, but of course that is not a guarantee of good small-sample behavior. Bootstrapping from a smoothed version of $\hat{F}$, as in Efron (1982a, sec. 5.3), alleviates the problem in this particular example.

8. GEOMETRY OF THE NONPARAMETRIC CASE

Formula (7.3), which allows us to apply the BC_a method nonparametrically, is based on a simple heuristic argument: instead of the actual sample-space $\mathcal{X}$ of the data points x_i, consider only distributions F supported on $\hat{\mathcal{X}} = \{x_1, x_2, \ldots, x_n\}$, the observed data set. This is an n-category multinomial family, to which the results of Section 6 can be applied. Because the multinomial is an exponential family, Lemma 3 directly gives (7.3).

We will now examine this argument more carefully, with the help of a simple geometric representation. See Efron (1981, sec. 11) for further discussion of this approach to nonparametric confidence intervals.

A typical distribution supported on $\hat{\mathcal{X}}$ is

$$F(\mathbf{w}) : \text{mass } w_i \text{ on } x_i, \tag{8.1}$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ can be any vector in the simplex $\mathcal{S}_n = \{\mathbf{w} : w_i \geq 0\ \forall i, \sum_1^n w_i = 1\}$. The parameter $\theta = t(F)$ is defined on $\mathcal{S}_n$ by $\theta(\mathbf{w}) = t(F(\mathbf{w}))$. The central point of the simplex,

$$\mathbf{w}^O \equiv \mathbf{1}/n = (1/n, 1/n, \ldots, 1/n), \tag{8.2}$$

corresponds to $F(\mathbf{w}^O) = \hat{F}$, the usual empirical distribution; $\theta(\mathbf{w}^O) = \hat{\theta} = t(\hat{F})$, the nonparametric MLE of θ. The curved surface

$$\mathcal{C}_{\hat{\theta}} = \{\mathbf{w} : \theta(\mathbf{w}) = \theta(\mathbf{w}^O) = \hat{\theta}\} \tag{8.3}$$

comprises those distributions $F(\mathbf{w})$ having $\theta(\mathbf{w}) = \hat{\theta}$. *The vector $\mathbf{U}_i$ is orthogonal to $\mathcal{C}_{\hat{\theta}}$ at $\mathbf{w}^O$,* as shown in Figure 1, which follows from definition (7.2) of the empirical influence function. $\mathbf{U}$ is essentially the gradient of $\theta(\mathbf{w})$ at $\mathbf{w}^O$ (see Efron 1982a, sec. 6.3).

With $\mathbf{w}$ unknown, but $\hat{\mathcal{X}} = \{x_1, \ldots, x_n\}$ considered fixed, one can imagine setting a confidence interval for $\theta(\mathbf{w})$ on the basis of a hypothetical sample $x_1^*, x_2^*, \ldots, x_n^* \overset{\text{iid}}{\sim} F(\mathbf{w})$. A sufficient statistic is the vector of proportions $P_i = \#\{x_j^* = x_i\}/n$, say $\mathbf{P} = (P_1, P_2, \ldots, P_n)$, with distribution

$$\mathbf{P} \sim \text{mult}_n(n, \mathbf{w})/n, \qquad \mathbf{w} \in \mathcal{S}_n. \tag{8.4}$$

The notation here indicates n draws from an n-category multinomial, having probability w_i for category i. We suppose that we have observed $\mathbf{P} = \mathbf{w}^O$ in (8.4), that is, that the hypothetical sample $x_1^*, \ldots, x_n^*$ equals the actual sample $x_1, \ldots, x_n$.

Distributions (8.4) form an n-parameter exponential family (6.6) with $\mathbf{y} = \mathbf{P}$, $\eta_i = \log(nw_i) + c$, and $\psi(\eta) = \log(\sum_1^n \exp(\eta_i)/n)$. Here c can be any constant, since all vectors $\boldsymbol{\eta} + c\mathbf{1}$ correspond to the same probability vector $\mathbf{w}$, namely $w_i = \exp(\eta_i)/\sum_1^n \exp(\eta_j)$.

If one accepts the reduction of the original nonparametric problem to (8.4), with observed value $\mathbf{P} = \mathbf{w}^O$, then it is easy to carry through the least favorable family calculations (6.3)–(6.5): (i) $\hat{\boldsymbol{\eta}} = \mathbf{0}$; (ii) $\hat{\boldsymbol{\mu}} = \mathbf{U}$; (iii) $\hat{f}_\lambda$ is the member of (7.4) corresponding to $\hat{\boldsymbol{\eta}} + \lambda\hat{\boldsymbol{\mu}} = \lambda\mathbf{U}$, namely

$$\mathbf{P}^* \sim \text{mult}(n, \mathbf{w}^\lambda)/n,\ w_i^\lambda = \exp(\lambda U_i) \Big/ \sum_{j=1}^n \exp(\lambda U_j); \tag{8.5}$$

(iv) finally, formula (7.3) follows directly from Lemma 3, by differentiating $\hat{\psi}(\lambda) = \log(\sum_1^n \exp(\lambda' J_j)/n)$ (and remembering that $\sum U_i = 0$).

Only step (ii) is not immediate, but it is a straightforward consequence of definition (6.3) and standard properties of the multinomial. It has already been noted that $\mathbf{U}$ is orthogonal to $\mathcal{C}_{\hat{\theta}}$, so $\mathbf{U}$ is proportional to $\hat{\nabla}$ in (6.3). However, $-\hat{\mathbf{I}}_{\hat{\boldsymbol{\eta}}} = \mathbf{I} - \mathbf{1}\mathbf{1}'/n$, which has pseudo-inverse $\mathbf{I}$. Thus $\hat{\boldsymbol{\mu}}$ is proportional to $\mathbf{U}$. Since (6.7), (6.8) produce the same value of a if $\hat{\boldsymbol{\mu}}$ is multiplied by any constant, this in effect gives $\hat{\boldsymbol{\mu}} = \mathbf{U}$.

An interesting case that provides some support for the

nonparametric BC_a method is that where the sample space is finite to begin with, say $\mathcal{X} = \{1, 2, \ldots, L\}$. A typical distribution on $\mathcal{X}$ is $\mathbf{f} = (f_1, \ldots, f_L)$, where $f_l = \Pr\{x_i = l\}$. The observed sample proportions $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_L)$, $\hat{f}_l \equiv \#\{x_i = l\}/n$, are sufficient, with distribution $\hat{\mathbf{f}} \sim \text{mult}_L(n, \mathbf{f})/n$. This is an L-parameter exponential family, so the theory of Section 6 applies. It turns out that Lemma 3 agrees with formula (7.3) in this case. *Nonparametric BC_a intervals are the same as parametric BC_a intervals when $\mathcal{X}$ is finite.* See remarks G and H of Efron (1979) for the first-order bootstrap asymptotics of finite sample spaces.

Family (8.4) was used in Section 11 of Efron (1981) to motivate a method called *nonparametric tilting*, a nonparametric analog of the standard hypothesis-testing approach to confidence interval construction. The one-parameter tilting family, (11.12) of Efron (1981), is closely related to the least favorable family $\hat{\mathcal{F}}$ in Figure 1. Efron (1981, table 5) considered samples of size $n = 15$ for the one-sided exponential density $f(x) = \exp[-(x + 1)]$ $(x > -1)$. Central 90% tilting intervals for $\theta = E_F X$ were constructed for each of 10 such samples, averaging $[-.34, .50]$. The corresponding nonparametric BC_a intervals averaged $[-.34, .52]$ and were quite similar to the tilting intervals on a sample-by-sample comparison. The nonparametric BC_a method is computationally simpler than nonparametric tilting and seems likely to give similar results in most problems.

We end this section with a useful approximation formula for the bias-correction constant z_0, developed jointly with Timothy Hesterberg. In addition to (7.2) we need the second-order empirical influence function

$$V_{ij} = \lim_{\Delta \to 0} \{[t((1 - 2\Delta)\hat{F} + \Delta\delta_i + \Delta\delta_j) - t((1 - \Delta)\hat{F} + \Delta\delta_i) - t((1 - \Delta)\hat{F} + \Delta\delta_j) + t(\hat{F})]/\Delta^2\}. \quad (8.6)$$

Define $z_{01} \equiv (\frac{1}{6}) \sum_1^n U_i^3/(\sum_1^n U_i^2)^{3/2}$ [approximation (7.3) for a] and

$$z_{02} \equiv \left[\frac{\mathbf{U}'\mathbf{V}\mathbf{U}}{\|\mathbf{U}\|^2} - \text{tr } \mathbf{V}\right]/(2n\|\mathbf{U}\|), \quad (8.7)$$

where $\mathbf{V}$ is the $n \times n$ matrix (V_{ij}).

Lemma 4. The bias-correction constant z_0 approximately equals

$$\Phi^{-1}\{2\Phi(z_{01})\Phi(z_{02})\}. \quad (8.8)$$

For the law school data, Example 1 of Section 7, $z_{01} = -.0817$ and $z_{02} = -.0067$, giving $z_0 = -.0869$ from (8.8), compared with $z_0 = -.0927 \pm .0039$ from $B = 100{,}000$ bootstrap replications.

The term z_{01} relates to skewness in $\hat{\mathcal{F}}$, and z_{02} is a geometric term arising from the curvature of $\mathcal{C}_\theta$ at $\mathbf{w}^o$. It is analogous to formula (A15) of Efron (1985). Lemma 4 will not be proved here but is important in the sample size considerations of Section 9.

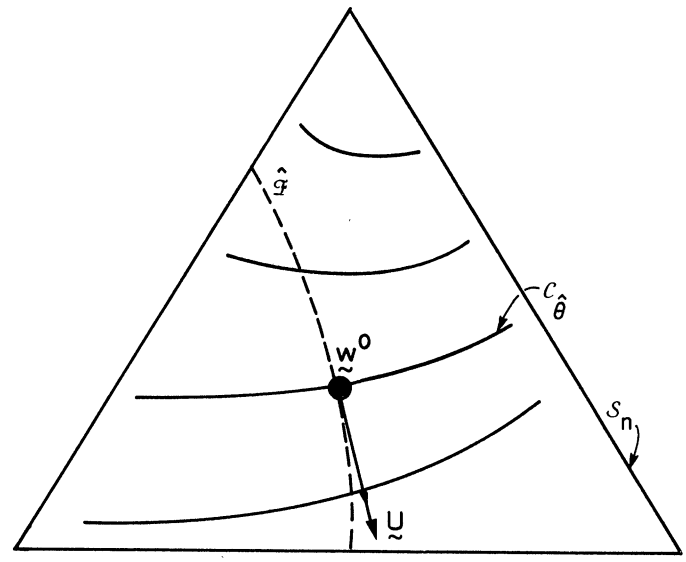


Figure 1. All probability distributions supported on $\{x_1, x_2, \ldots, x_n\}$ are represented as the simplex $\mathcal{S}_n$. The central point $\mathbf{w}^o$ corresponds to the empirical distribution $\hat{F}$. The curves indicate level surfaces of constant value of the parameter θ. In particular $\mathcal{C}_{\hat{\theta}}$ comprises those probability distributions having θ equal to $\theta(\mathbf{w}^o) = \hat{\theta}$, the MLE. The least favorable family $\hat{\mathcal{F}}$ passes through $\mathbf{w}^o$ in the direction $\mathbf{U}$, orthogonal to $\mathcal{C}_{\hat{\theta}}$.

9. BOOTSTRAP SAMPLE SIZES

How many bootstrap replications of $\hat{\theta}^*$ need we take? So far we have pretended that the number of replications $B = \infty$, but if Monte Carlo methods are necessary to obtain the bootstrap cdf $\hat{G}$, then B must be finite, usually the smaller the better. This section gives rough estimates of how small B may be taken in practice. The results are presented without proof, all being standard exercises in error estimation (see, e.g., Kendall and Stuart 1958, chap. 10). They apply to any situation, parametric or nonparametric, where $\hat{G}$ is obtained by Monte Carlo sampling.

First consider the easy problem of estimating the standard error of $\hat{\theta}$ via the bootstrap. The bootstrap estimate based on B replications, $\hat{\sigma}_B = [\sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta}^*)^2/(B - 1)]^{1/2}$, has conditional coefficient of variation (standard deviation divided by expectation)

$$\text{CV}\{\hat{\sigma}_B \mid \mathbf{y}\} \doteq [(\hat{\delta} + 2)/4B]^{1/2}, \quad (9.1)$$

where $\hat{\delta}$ is the kurtosis of the bootstrap distribution $\hat{G}$. The notation indicates that the observed data $\mathbf{y}$ is fixed in this calculation. As $B \to \infty$, then (9.1) $\to 0$ and $\hat{\sigma}_B \to \hat{\sigma}$, the ideal bootstrap estimate of standard error.

Of course $\hat{\sigma}$ itself will usually not estimate the true standard error $\sigma \equiv \text{SD}_\theta\{\hat{\theta}\}$ perfectly. Let $\text{CV}(\hat{\sigma})$ be the coefficient of variation of $\hat{\sigma}$, unconditional now, averaging over the possible realizations of $\mathbf{y}$ [e.g., if $n = 20$, $\hat{\theta} = \bar{x}$, $x_i \overset{\text{iid}}{\sim} N(0, 1)$, then $\text{CV}(\hat{\sigma}) \doteq (1/40)^{1/2} = .16$]. The unconditional CV of $\hat{\sigma}_B$ is then approximated by

$$\text{CV}(\hat{\sigma}_B) \doteq \left[\text{CV}^2(\hat{\sigma}) + \frac{E\hat{\delta} + 2}{4B}\right]^{1/2}. \quad (9.2)$$

Table 5 displays $\text{CV}(\hat{\sigma}_B)$ for various choices of B and $\text{CV}(\hat{\sigma})$, assuming that $E\hat{\delta} = 0$. For values of $\text{CV}(\hat{\sigma}) \geq$

Table 5. Coefficient of Variation of $\hat{\sigma}_B$, the Bootstrap Estimate of Standard Error, as a Function of B, the Number of Bootstrap Replications, and CV($\hat{\sigma}$), the Limiting CV as $B \to \infty$

	$B \to$				
CV($\hat{\sigma}$)	25	50	100	200	∞
.25	.29	.27	.26	.25	.25
.20	.24	.22	.21	.21	.20
.15	.21	.18	.17	.16	.15
.10	.17	.14	.12	.11	.10
.05	.15	.11	.09	.07	.05
0	.14	.10	.07	.05	0

NOTE: These data are based on (9.2), assuming that $E\hat{\delta} = 0$.

.10, typical in practice, *there is little improvement past B = 100*. In fact, B as small as 25 gives reasonable results.

Now we return to bootstrap confidence intervals. In the Monte Carlo situation the bootstrap cdf $\hat{G}$ must be estimated from bootstrap replications $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$, say by

$$\hat{G}_B(s) = \#\{\hat{\theta}_b^* < s\}/B. \tag{9.3}$$

As $B \to \infty$, then $\hat{G}_B \to \hat{G}$, the ideal bootstrap cdf we have been using in the previous sections. Let $\theta_B[\alpha]$ be the level α endpoint of either the BC or BC_a interval obtained from $\hat{G}_B(s)$ by substitution in (3.8), (3.9).

The following formula for the conditional CV of $\theta_B[\alpha] - \hat{\theta}$ assumes that $\hat{G}$ is roughly normal and that z_0 and a are known, for example, from (8.8) and (6.5) or (7.3):

$$\mathrm{CV}\{\theta_B[\alpha] - \hat{\theta} \mid \mathbf{y}\} \doteq \frac{1}{B^{1/2}|z^{(\alpha)}|}\left\{\frac{\alpha(1-\alpha)}{\varphi(z^{(\alpha)})^2}\right\}^{1/2}, \tag{9.4}$$

$\varphi(z) \equiv \exp(-\frac{1}{2}z^2)/\sqrt{2\pi}$. Notice that since we condition on $\mathbf{y}$, the only random quantity on the left side of (9.4) is $\theta_B[\alpha]$. Formula (9.4) measures the variability in $\theta_B[\alpha] - \hat{\theta}$ due to taking only B bootstrap replications, rather than an infinite number.

Here is a brief tabulation of (9.4) $\times B^{1/2}$:

α :	.75	.90	.95	.975
(9.4) $\times B^{1/2}$:	2.02	1.33	1.28	1.36

(9.5)

If $B = 1{,}000$, for instance, then $\mathrm{CV}\{\theta_B[.95] - \hat{\theta} \mid \mathbf{y}\} \doteq 1.28/1000^{1/2} = .040$. Reducing B to 200 increases the conditional CV to .091. This last figure may be too big. The whole purpose of developing a theory better than (1.1) is to capture second-order effects. As the examples have indicated, these become interesting when the asymmetry ratio R/L is larger than say, 1.25, or smaller than .80. In such borderline situations, an extra 9% error in each tail due to inadequate bootstrap sampling may be unacceptable.

If the bias-correction constant z_0 is estimated by Monte Carlo directly from $z_0 = \Phi^{-1}(\hat{G}_B(\hat{\theta}))$, rather than from (8.8), then

$$\mathrm{CV}\{\theta_B[\alpha] - \hat{\theta} \mid \mathbf{y}\} \doteq \frac{1}{B^{1/2}z^{(\alpha)}}\left\{\frac{1}{\varphi(0)^2} - \frac{2(1-\alpha)}{\varphi(0)\varphi(z^{(\alpha)})} + \frac{\alpha(1-\alpha)}{\varphi(z^{(\alpha)})^2}\right\}^{1/2} \tag{9.6}$$

for $\alpha > .50$. This gives larger CV's than (9.4):

α :	.75	.90	.95	.975
(9.6) $\times B^{1/2}$:	3.04	1.97	1.75	1.71

(9.7)

Comparing (9.7) with (9.5) shows that we need B to be about twice as large to get the same CV if z_0 is estimated rather than calculated. Formula (8.8) can be very helpful!

Both (9.4) and (9.6) assume that the bootstrap cdf is estimated by straightforward Monte Carlo sampling, as in (9.3). M. V. Johns (personal communication) has developed importance sampling methods that greatly accelerate the estimation of $\hat{G}$ in some situations.

10. ONE-PARAMETER FAMILIES

We return to the simple situation $\hat{\theta} \sim f_\theta$, where there are no nuisance parameters and where we want a confidence interval for the real-valued parameter θ based on a real-valued summary statistic $\hat{\theta}$. This section gives a more extensive discussion of the acceleration constant a, which has played a basic role in our considerations. Three familiar types of one-parameter families will be investigated: exponential families, translation families, and transformation families.

Efron (1982b) considered the following question: for a given family $\hat{\theta} \sim f_\theta$, do there exist mappings $\hat{\phi} = g(\hat{\theta})$, $\phi = h(\theta)$ such that $\hat{\phi} = \phi + \sigma_\phi q(Z)$, $Z \sim N(0, 1)$, as in (4.8)? This last form, a General Scaled Transformation Family (GSTF), generalizes the concept of the ideal normalization, where $\hat{\phi} = \phi + Z$. [We now add the conditions $q(0) = 0$, $q'(0) = 1$, as in Efron (1982b).]

The question is answered in terms of the diagnostic function $D(z, \theta) \equiv [\varphi(0)/\varphi(z)][\dot{F}_\theta(\hat{\theta}_\theta^{(\alpha)})/\dot{F}_\theta(\mu_\theta)]$. Here $\varphi(z)$ is the standard normal density $(2\pi)^{-1/2}\exp(-z^2/2)$; F_θ is the cdf $F_\theta(s) = \Pr_\theta\{\hat{\theta} \le s\}$; $\dot{F}_\theta(s) = (\partial/\partial\theta)F_\theta(s)$; $\alpha = \Phi(z)$; $\hat{\theta}_\theta^{(\alpha)}$ is the $100 \cdot \alpha$ percentile of $\hat{\theta}$ given θ, $\hat{\theta}_\theta^{(\alpha)} = F_\theta^{-1}(\alpha)$; and μ_θ is the median of $\hat{\theta}$ given θ, $\mu_\theta = \hat{\theta}_\theta^{(.5)} = \hat{F}_\theta^{-1}(.5)$. It is shown that the form of σ_ϕ and $q(z)$ in (4.8) can be inferred from $D(z, \theta)$, the main advantage being that $D(z, \theta)$ is computed without knowledge of the normalizing transformations g, h.

The connection of transformation family theory with the acceleration constant a is the following: define

$$\varepsilon_\theta \equiv (\partial/\partial z)D(z, \theta)|_{z=0}. \tag{10.1}$$

If $q(z)$ in (4.8) is symmetrically distributed about zero, a situation called a symmetric scaled transformation family (SSTF), then

$$\varepsilon_\theta = d\sigma_\phi/d\phi \tag{10.2}$$

(see Efron 1982b, eq. 4.11). A more complicated relationship holds for the GSTF case.

Notice that (10.2) is quite close to our original description of "a" as the rate of change of standard deviation on the normalized scale. As a matter of fact, we can transform (3.6), (3.7) into an SSTF by considering the statistic

$$\tilde{\phi} = \hat{\phi} + \frac{z_0}{1 - az_0}\sigma_{\hat{\phi}} = \hat{\phi} + \frac{z_0}{1 - az_0}(1 + a\hat{\phi}), \tag{10.3}$$

instead of $\hat{\phi}$ itself. Then it is easy to show that

$$\tilde{\phi} = \phi + (1 + \varepsilon_0\phi)Z, \qquad \varepsilon_0 = a/(1 - az_0), \quad (10.4)$$

an SSTF with $\sigma_\phi = 1 + \varepsilon_0\phi$, $\dot{\sigma}_\phi = \varepsilon_0$ for all ϕ. [The quantity ε_0 has the same definition in (10.4) as in (4.11).]

Example. For $\hat{\theta} \sim \theta\chi^2_{19}/19$ as in Table 2, $\varepsilon_\theta = .1090$ for all θ [using Eq. (10.6)]. In addition, $z_0 = \Phi^{-1}\Pr\{\chi^2_{19} < 19\} = .1082$. The relationship $a = \varepsilon_0/(1 + \varepsilon_0 z_0)$ obtained by solving for a in (10.4) gives $a = .1077$, the value used in Table 2. This family is nearly in SSTF (see Remark E, Sec. 11).

We show below that under reasonable asymptotic conditions,

$$\mathrm{SKEW}_\theta(\dot{l}_\theta)/6 \doteq \varepsilon_\theta, \quad (10.5)$$

where $\varepsilon_\theta = (\partial/\partial z)D(z, \theta)|_{z=0}$, as in (10.1). This last definition of ε_θ can be evaluated for any family $\hat{\theta} \sim f_\theta$, assuming only that the necessary derivatives exist. The main point here is that $\mathrm{SKEW}_\theta(\dot{l}_\theta)/6$ always approximates ε_θ (10.1), and in SST families ε_θ has the acceleration interpretation (10.2).

Now to show (10.5). It is possible to reexpress (10.1) as

$$\varepsilon_\theta = -\frac{\varphi(0)}{\dot{\mu}_\theta f_\theta(\mu_\theta)}\,\dot{l}_\theta(\mu_\theta), \quad (10.6)$$

where $\dot{\mu}_\theta = (d/d\theta)\mu_\theta$, the rate of change of the median μ_θ with respect to θ. For notational convenience suppose that $\theta = 0$. Instead of $\hat{\theta}$, consider the statistic $X \equiv \dot{l}_0(\hat{\theta})/i_0$, where i_0 equals the Fisher information $E_0\dot{l}_0(\hat{\theta})^2$. The parameter ε_θ is invariant under one-to-one transformations of $\hat{\theta}$, so we can evaluate the right side of (10.6) in terms of X, $\varepsilon_\theta = -\varphi(0)\dot{l}^X_\theta(\mu^X_\theta)/\dot{\mu}^X_\theta f^X_\theta(\mu^X_\theta)$.

For $\theta = 0$, X has expectation $E_0X = 0$ and standard deviation $\sigma^X_0 = i_0^{-1/2}$; in addition, $\dot{l}^X_0(0) = 0$, since $X = 0$ implies that $\theta = 0$ is a solution of the MLE equation. Assuming the usual asymptotic convergence properties, as in (5.1), (5.3), we have the following approximations: $\dot{\mu}^X_0 \doteq 1$; $\mu^X_0 \doteq -\gamma^X_0 i_0^{-1/2}/6$; $f^X_0(\mu^X_0) \doteq \varphi(0)i_0^{1/2}$; $\dot{l}^X_0(\mu^X_0) \doteq -\sqrt{i_0}\,\gamma^X_0/6$. These are derived from standard Edgeworth and Taylor series arguments, which will not be presented here. Taken together they give $\varepsilon_0 \doteq \mathrm{SKEW}_0(\dot{l}^X_0)/6 = \mathrm{SKEW}_0(\dot{l}_0)/6$, which is (10.5). The quantity $\mathrm{SKEW}_0(\dot{l}_0)/6$ is $O(n^{-1/2})$, and the error of approximation in (10.5) is quite small,

$$\varepsilon_0 = [\mathrm{SKEW}_0(\dot{l}_0)/6][1 + O(n^{-1})]. \quad (10.7)$$

Approximation (10.5) is particularly easy to understand in one-parameter exponential families. Suppose that $x_1, x_2, \ldots, x_n$ are iid observations from such a family, with sufficient statistic $y = \bar{x}$ having density $f_\theta(y) = \exp\{n[\theta y - \psi(\theta)]\}f_0(y)$. In this case formula (10.6) becomes

$$\varepsilon_\theta = \frac{\sigma^Y_\theta\varphi(0)}{\dot{\mu}^Y_\theta f^Y_\theta(\mu^Y_\theta)}\left[\frac{\lambda^Y_\theta - \mu^Y_\theta}{\sigma^Y_\theta}\right], \quad (10.8)$$

where $\lambda^Y_\theta = E_\theta\{y\}$, $\mu^Y_\theta = \mathrm{median}_\theta\{y\}$, $\dot{\mu}^Y_\theta = \partial\mu^Y_\theta/\partial\theta$, and so forth. The term $[(\lambda^Y_\theta - \mu^Y_\theta)/\sigma^Y_\theta] = \gamma^Y_\theta/6[1 + O(n^{-1})]$, and $\sigma^Y_\theta\varphi(0)/\dot{\mu}^Y_\theta f^Y_\theta(\mu^Y_\theta) = 1 + O(n^{-1})$, both of the calculations being quite straightforward. Thus $\varepsilon_\theta = \gamma^Y_\theta/6[1 + O(n^{-1})]$. Since $\dot{l}_\theta(y) = n[y - \lambda_\theta]$, we have $\mathrm{SKEW}_\theta(\dot{l}_\theta(y)) = \mathrm{SKEW}_\theta(y) = \gamma^Y_\theta$, verifying (10.5) for one-parameter exponential families.

Example. If $Y \sim \mathrm{Poisson}(\theta)$, $\theta = 15$, then $\mathrm{SKEW}_\theta(\dot{l}_\theta)/6 = 1/(6 \cdot \theta^{1/2}) = .0430$. For the continued version of the Poisson family used in Efron (1982b; note Corrigenda, p. 1032), $(\partial/\partial z)D(z, \theta)|_{z=0} = .0425$ for $\theta = 15$.

Translation Families. Suppose that we observe a translation family $\hat{\zeta} = \zeta + W$, as in (3.12). Express W as a function $q(Z)$ of $Z \sim N(0, 1)$, for simplicity assuming that $q(0) = 0$ and $q'(0) = 1$, as in Efron (1982b). Then $z_0 = \Phi^{-1}\Pr\{\hat{\zeta} < \zeta\} = 0$. In this case it looks like methods based on the percentiles of the bootstrap distribution must give wrong answers, since if W is long-tailed to the right then the correct interval (3.13) is long-tailed to the left, and vice versa. However, the BC_a method produces at least roughly correct intervals, as we saw in the proof of Lemma 1.

What happens is the following: for any constant A the transformation $g_A(t) \equiv (\exp(At) - 1)/A$ gives $\hat{\phi} = g_A(\hat{\zeta})$, $\phi = g_A(\zeta)$, and $Z_A = g_A(W)$ satisfying

$$\hat{\phi} = \phi + \sigma^A_\phi \cdot Z_A, \qquad \sigma^A_\phi = 1 + A\phi. \quad (10.9)$$

The Taylor series for $W = q(Z)$ begins $W = Z + (\gamma_W/6)Z^2 + \cdots$, where $\gamma_W = \mathrm{SKEW}(W)$. Then $Z_A = Z + (\gamma_W/6)^2Z^2 + (A/2)Z^2 + \cdots$.

The choice $A = a \equiv -\gamma_W/3$ results in $Z_a = Z + cZ^3 + \cdots$, the quadratic term canceling out; Z_a is then approximately normal, so (10.9) is approximately situation (3.6), (3.7), with $z_0 = 0$, $a = -\gamma_W/3$. *But we know that the BC_a intervals are correct if we can transform to situation (3.6), (3.7).* An application of Lemma 2, assuming that $Z_a \sim N(0, 1)$, shows that $a = -\gamma_W/3 \doteq \mathrm{SKEW}(\dot{l}_\zeta(\hat{\zeta}))/6$ for the translation family $\hat{\zeta} = \zeta + W$, reverifying (4.4). [If $Z_a \sim N(0, 1)$ in (10.9), then a must equal ε, the constant value of ε_ζ, (10.1), for the translation family $\hat{\zeta} = \zeta + W$; one can show directly that $\varepsilon \doteq -\gamma_W/3$ for such a family.]

In the example $\hat{\theta} \sim \theta\chi^2_{19}/19$, the two constants z_0 and a are nearly equal. This is no fluke.

Theorem 2. If $\hat{\theta}$ is the MLE of θ in a one-parameter problem having standard asymptotic properties (5.1) or (5.3), then $z_0 \doteq a$,

$$z_0 \equiv \Phi^{-1}\Pr_\theta\{\hat{\theta} < \theta\} = \frac{\mathrm{SKEW}_\theta(\dot{l}_\theta)}{6}[1 + O(n^{-1})]. \quad (10.10)$$

Proof. We follow the notation and results of DiCiccio (1984): thus k_1, k_2, k_3 equal the first three cumulants of $\dot{l}_\theta$ under θ; k_{01}, k_{02}, k_{03} the first three cumulants of $\ddot{l}_\theta$; k_{001}, the first cumulant of $\ddot{l}_\theta$; and $k_{11} = \mathrm{cov}_\theta(\dot{l}_\theta, \ddot{l}_\theta)$. (So $k_2 = i_\theta$, the Fisher information.) All cumulants are assumed to be $O(n)$. Then the relative bias of $\hat{\theta}$ is

$$b \equiv \frac{E_\theta(\hat{\theta} - \theta)}{\mathrm{var}_\theta(\theta)^{1/2}} = \frac{k_{001} - 2k_3}{6k_2^{3/2}} + O(n^{-3/2}), \quad (10.11)$$

and $\hat{\theta}$ has skewness

$$\gamma_\theta = \frac{k_{001} - k_3}{k_2^{3/2}} + O(n^{-3/2}). \tag{10.12}$$

Both b and γ_θ are $O(n^{-1/2})$.

Standard Edgeworth theory now gives

$$\begin{aligned}\Pr_\theta\{\hat{\theta} < \theta\} &= \Phi(-b) - \frac{\gamma}{6}\varphi(b)(b^2 - 1) + O(n^{-3/2}) \\ &= .5 + \varphi(0)\frac{(2k_3 - k_{001}) + (k_{001} - k_3)}{6k_2^{3/2}} \\ &\quad + O(n^{-3/2}) \\ &= .5 + \varphi(0)\frac{k_3}{6k_2^{3/2}} + O(n^{-3/2}).\end{aligned}$$

Since $\mathrm{SKEW}_\theta(\dot{l}_\theta) = k_3/k_2^{3/2}$, this verifies (10.10).

In multiparameter problems it is no longer true that $z_0 \doteq a$. The geometry of the level surface $\mathcal{C}_\theta$ adds another term to z_0, as in (8.8).

11. REMARKS

Remark A. Suppose that instead of (3.6), (3.7) we have $\sigma_\phi = \tau(1 + A\phi)$, so $\sigma_0 = \tau$ ($\tau \neq 1$). The transformations $\hat{\phi}' \equiv \hat{\phi}/\tau$, $\phi' \equiv \phi/\tau$, give $\hat{\phi}' = \phi' + \sigma'_{\phi'}(Z - z_0)$, where $\sigma'_{\phi'} = 1 + a\phi'$ and $a = A\tau$, so we are back in form (3.6), (3.7). Notice that the derivative $d(\sigma_\phi/\sigma_0)/d(\phi/\sigma_0) = a$, as in (4.7). In a similar way we can transform (3.6), (3.7) so that $\sigma_{\phi_0} = 1$ at any point ϕ_0; the resulting value of a satisfies (4.7).

Remark B. Instead of using $\hat{\phi}$ to estimate ϕ in (3.6), (3.7) we might change to the estimator $\hat{\phi}^{(c)} \equiv \hat{\phi} - c\sigma_{\hat{\phi}}$, for some constant c. It turns out that we are still in situation (3.6), (3.7): $\hat{\phi}^{(c)} = \phi + \sigma_\phi^{(c)}(Z - z_0^{(c)})$, where

$$\sigma_\phi^{(c)} = 1 + a^{(c)}(\phi - \phi_0^{(c)}), \qquad \phi_0^{(c)} = c/(1 - ac), \tag{11.1}$$

and $a^{(c)} = a(1 - ac)$, $z_0^{(c)} = z_0 + \phi_0^{(c)}$. The choice $c = -z_0/(1 - az_0)$ gives $z_0^{(c)} = 0$, as in (10.3), (10.4). The choice $c = a$ gives approximately the MLE of ϕ. Interestingly enough, *the BC_a interval for ϕ based on $\hat{\phi}^{(c)}$ is the same for all choices of c.* Minor changes in the choice of estimator seem to have little effect on the BC_a intervals in general, though for computational reasons it is best not to use very biased estimators having large values of z_0.

Remark C. Section 6 uses the MLE $\hat{\theta} = t(\hat{\boldsymbol{\eta}})$. This has one major advantage: *the BC_a interval for θ, based on $\hat{\theta}$, stays the same under all multivariate transformations (6.11).* Stein (1956) noted that the least favorable direction $\hat{\boldsymbol{\mu}}$ transforms in the obvious way under (6.11), $\hat{\tilde{\boldsymbol{\mu}}} = \hat{\mathbf{D}}\hat{\boldsymbol{\mu}}$, where $\hat{\mathbf{D}}$ is the matrix with ijth element $\partial\tilde{\eta}_j/\partial\eta_i|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}}$, from which it is easy to check that formula (6.5) is invariant: the constant a is assigned the same value no matter what transformations (6.11) are applied. The bootstrap distribution $\hat{G}$ is similarly invariant, as shown in Efron (1985), and so is z_0. This implies that the BC_a intervals are invariant under transformations (6.11).

Remark D. The multiparametric theory of Section 5 gives an interesting result when applied to location-scale families; $y = (x, s)$, $\eta = (\theta, \sigma)$, and family of densities $f_\eta(y)$ of the form

$$f_{\theta,\sigma}(x, s) = (1/\sigma^2)f_{01}((x - \theta)/\sigma, s/\sigma), \tag{11.2}$$

$f_{01}(x, s)$ being a known bivariate density function.

Suppose that we wish to set a confidence interval for the location parameter θ on the basis of its MLE $\hat{\theta}$. Parametric bootstrap intervals are based on the distribution of $\hat{\theta}^*$ when sampling from $f_{\hat{\theta},\hat{\sigma}}(x^*, s^*)$. The BC interval essentially amounts to pretending that σ is known (and equal to $\hat{\sigma}$) in (11.2) and that we have only a location problem to deal with, rather than a location-scale problem. In contrast, the BC_a interval takes account of the fact that σ is unknown. In particular the least favorable direction $\hat{\boldsymbol{\mu}}$, plotted in the (θ, σ) plane, is *not* parallel to the θ axis. It has a component in the σ direction, whose magnitude is determined by the correlation between x and s. This means that Stein's least favorable family (6.4) does not treat σ as a constant.

Table 6 relates to the following choice of $f_{01}(x, s)$:

$$x \sim \chi^2_{30}/30 - 1, \qquad s \mid x \sim (1 + x)(\chi^2_{14}/14)^{1/2}, \tag{11.3}$$

the two χ^2 variates being independent. This is a computationally more tractable version of the problem discussed in Efron (1982, tables 4 and 5). Approximate central 90% intervals are given for θ, having observed $(x, s) = (0, 1)$. For any other observed (x, s) the intervals transform in the obvious way, $\theta_{xs}[\alpha] = x + s\theta_{01}[\alpha]$. Line 3 of Table 6 shows the exact interval, based on inverting the distribution of the pivotal quantity $T = (\hat{\theta} - \theta)/\hat{\sigma}$ for situations (11.2), (11.3).

In this case the BC_a method makes a large "second-order t correction," as in Example 2 of Section 6, shifting the BC interval a considerable way rightward and achieving the correct R/L ratio. The length of the BC_a interval is 90% the length of the T interval. This deficiency is a third-order effect, in the spirit of the familiar Student-t correction. It arises from the variability of $\hat{\sigma}$ as an estimate of σ, rather than the second-order effect due to the correlation of $\hat{\sigma}$ with $\hat{\theta}$.

Remark E. Section 3 says that the family $\hat{\theta} \sim \theta\chi^2_{19}/19$ can be mapped into form (3.6), (3.7). What are the appropriate mappings? It simplifies the problem to consider the equivalent family $\hat{\theta} \sim \theta(\chi^2_{19}/c_0)$, where $c_0 = 18.3337 = \mathrm{median}(\chi^2_{19})$. Then $\hat{\zeta} \equiv g_1(\hat{\theta})$, $\zeta \equiv g_1(\theta)$, and $W \equiv g_1(\chi^2_{19}/c_0)$ give a translation family (3.12), with median(W)

Table 6. Central 90% Intervals for θ, Having Observed $(x, s) = (0, 1)$ From the Location-Scale Family (11.2), (11.3) so $\hat{\theta} = 0$ and $\hat{\sigma} = .966$

		RL	Length
1. BC interval	[−.336, .501]	1.49	.837
2. BC_a interval	[−.303, .603]	1.99	.906
3. T interval	[−.336, .670]	1.99	1.006

NOTE: Line 3 is based on the actual distribution of the pivotal quantity $T = (\hat{\theta} - \theta)/\hat{\sigma}$.

$= 0$, for any mapping $g_1(t) = (\log t)/c_1$. Choosing $c_1 = .3292$ results in $W = q(Z)$ having $q(0) = 0$, $q'(0) = 1$, as in the discussion of translation families in Section 10.

Section 10 suggests normalizing a translation family by $g_A(t) = (\exp(At) - 1)/A$, a good choice for A being the constant ε_θ, (10.1), which equals .1090 for all θ in the family $\hat\theta \sim \theta(\chi^2_{19}/c_0)$. The combined transformation $g(t) = g_A(g_1(t))$ is $g(t) = 9.1746[t^{.3311} - 1]$. The transformed family $\hat\phi = g(\hat\theta)$, $\phi = g(\theta)$ is of form (3.6), (3.7),

$$\hat\phi = \phi + (1 + .1090 \cdot \phi)Z,$$
$$Z = 9.1746[(\chi^2_{19}/c_0)^{.3311} - 1]. \tag{11.4}$$

Numerical calculations verify that Z as defined in (11.4) is very close to a standard normal variate. In fact we have automatically recovered, nearly, the Wilson–Hilferty cube root transformation (Johnson and Kotz 1970). Using (11.4), it is not difficult to show that $g(t)$, as defined previously, gives approximately (3.6), (3.7) when applied to the family $\hat\theta \sim \theta(\chi^2_{19}/19)$ considered in Section 3, with constants z_0 and a as stated. Schenker (1985) gave almost the same result.

Remark F. Suppose that $\mathbf{y} = (x_1, x_2, \ldots, x_n)$, where the x_i are an iid sample from a regular one-parameter family $f_\theta(x_i)$, and that $\hat\theta(\mathbf{y})$ is a first-order efficient estimator of θ, like the MLE. The score function $\dot l_\theta$ appearing in (4.4) is that based just on $\hat\theta$, rather than the score function based on the entire data set $\mathbf{y}$. However, it is easy to show from considerations like those in Efron (1975) that the two score functions are asymptotically identical. Their skewnesses differ only by amount $O_p(n^{-1})$. It is often more convenient to calculate a from the score function for $\mathbf{y}$ rather than for $\hat\theta$, as was done, for example, in (6.5).

Remark G. McCullagh (1984) and Cox (1980) gave an interesting approximate confidence interval for θ, which for the simple case $\hat\theta \sim f_\theta$ has endpoint

$$\theta_{\mathrm{APP}}[\alpha] = \hat\theta + 1/\sqrt{\hat k_2} \times \left\{ z^{(\alpha)} + \frac{(3\hat k_2' + 2\hat k_{001}) + \hat k_{001} z^{(\alpha)2}}{6\hat k_2^{3/2}} \right\}. \tag{11.5}$$

Here $\hat\theta$ is the MLE of θ; if $k_2(\theta) = E_\theta \dot l_\theta^2$, the Fisher information, then $\hat k_2 = k_2(\hat\theta)$ and $\hat k_2' = dk_2(\theta)/d\theta|_{\theta=\hat\theta}$; and $\hat k_{001} = (E_\theta \ddot l_\theta)_{\theta=\hat\theta}$. Formula (11.5) is based on higher-order asymptotic approximations to the distribution of the MLE (see also Barndorff-Nielsen 1984).

It can be shown, as indicated in Section 12, that $\theta_{\mathrm{BC}_a}[\alpha]$ also closely matches (11.5), $(\theta_{\mathrm{BC}_a}[\alpha] - \theta_{\mathrm{APP}}[\alpha])/\hat\sigma = O_p(n^{-1})$. We see again that the BC_a method offers a way to avoid theoretical effort, at the expense of increased numerical computations.

12. PROOF OF THEOREM 1

A monotonic mapping $\hat\phi = g(\hat\theta)$, $\phi = g(\theta)$ transforms the exact confidence interval in the obvious way, $\phi_{\mathrm{EX}}[\alpha] = g(\theta_{\mathrm{EX}}[\alpha])$; likewise for the BC_a interval. By using such a mapping we can always make $\hat\phi = 0$ and the distribution of $\hat\phi$ given $\phi = 0$ perfectly normal. Because of (5.3), which says that the distributions of $\hat\theta$ are approaching normality at the usual $O(n^{-1/2})$ rate, the normalizing transformation g is asymptotically linear, $g(\theta) = \theta + c_2\theta^2 + c_3\theta^3 + \cdots$, $c_2 = O(n^{-1/2})$, $c_3 = O(n^{-1})$.

We will assume that the problem is already in the form $\hat\theta = 0$, with the cdf of $\hat\theta$ for $\theta = 0$ normal, say

$$G_0 \sim N(-z_0, 1). \tag{12.1}$$

Here $z_0 = \Phi^{-1}P_0\{\hat\theta < 0\}$ must be included because it is not affected by any monotonic transformations; $z_0 \doteq \gamma_\theta/6$ is $O(n^{-1/2})$ by (5.3). A simple exercise, using the mean value theorem of calculus, shows that if (5.4) is true in the transformed problem (12.1), then it is true in the original problem.

Assuming (5.3), $\hat\theta = 0$, and (12.1), we will show that the exact interval has endpoint

$$\theta_{\mathrm{Ex}}[\alpha] \doteq \frac{z_0 + z^{(\alpha)}}{1 - \dot\sigma_0 z^{(\alpha)} + \dot\beta_0 + (\dot\gamma_0/6)(z^{(\alpha)2} - 1)} + (\ddot\sigma_0/2)(z_0 + z^{(\alpha)})^3, \tag{12.2}$$

compared with

$$\theta_{\mathrm{BC}_a}[\alpha] \doteq \frac{z_0 + z^{(\alpha)}}{1 - \dot\sigma_0(z_0 + z^{(\alpha)})} \tag{12.3}$$

for the BC_a interval. In this section the symbol "$\doteq$" indicates accuracy through $O(n^{-1})$ or $O_p(n^{-1})$, with errors $O(n^{-3/2})$ or $O_p(n^{-3/2})$. Then

$$\frac{\theta_{\mathrm{BC}_a}[\alpha] - \theta_{\mathrm{Ex}}[\alpha]}{\sigma_\theta} \doteq \theta_{\mathrm{BC}_a}[\alpha]\{\dot\sigma_0 z_0 + \dot\beta_0 + (\dot\gamma_0/6)(z^{(\alpha)2} - 1)\} - (\ddot\sigma_0/2)(z_0 + z^{(\alpha)})^3, \tag{12.4}$$

which is $O_p(n^{-1})$, as claimed in Theorem 1.

The proof of (12.2) begins by noting that (12.1) implies that $\beta_0 = -z_0$, $\sigma_0 = 1$, $\gamma_0 = 0$, $\delta_0 = 0$. Then (5.3) gives

$$E_\theta\hat\theta = \theta + \beta_\theta \doteq (1 + \dot\beta_0)\theta - z_0,$$
$$\sigma_\theta \doteq 1 + \dot\sigma_0\theta + \ddot\sigma_0\theta^2/2,$$
$$\gamma_\theta \doteq \dot\gamma_0\theta, \qquad \delta_\theta \doteq 0, \tag{12.5}$$

for $\theta = O(1)$ [i.e., for θ a bounded function of n, in the sequence of situations referred to in (5.3)]. The $100 \cdot \alpha$ percentile of $\hat\theta$ given θ is

$$\hat\theta_\theta^{(\alpha)} \doteq (\theta + \beta_\theta) + \sigma_\theta\{z^{(\alpha)} + (\gamma_\theta/6)(z^{(\alpha)2} - 1)\}$$
$$\doteq [(1 + \dot\beta_\theta)\theta - z_0] + [1 + \dot\sigma_0\theta + (\ddot\sigma_0/2)\theta^2] \times [z^{(\alpha)} + (\dot\gamma_0\theta/6)(z^{(\alpha)2} - 1)], \tag{12.6}$$

using a Cornish–Fisher expansion and (12.5). The θ, however, that has $\hat\theta_\theta^{(\alpha)} = 0$ is by definition $\theta_{\mathrm{Ex}}[1 - \alpha]$. Solving the lower expression in (12.6) for 0 and substituting $1 - \alpha$ for α gives (12.2).

The proof of (12.3) follows from (3.8), (3.9), and (12.1) [which says that $\hat G \sim N(-z_0, 1)$], if we can establish that

$a \doteq \dot{\sigma}_0$. In fact, we show below that

$$\varepsilon_\theta \doteq \dot{\sigma}_0 \quad \text{for} \quad \theta = O(n^{-1/2}), \tag{12.7}$$

which combines with $a = \varepsilon_0/(1 + \varepsilon_0 z_0) \doteq \varepsilon_0$ to give the required result.

Formula (12.7) follows from (12.5), which gives the simpler expressions

$$E_\theta\hat{\theta} \doteq \theta - z_0, \qquad \sigma_\theta \doteq 1 + \dot{\sigma}_0\theta, \qquad \gamma_\theta \doteq 0, \qquad \delta_\theta \doteq 0 \tag{12.8}$$

for $\theta = O(n^{-1/2})$. The cdf of $\hat{\theta}$ given θ is calculated to be

$$G_\theta(\hat{\theta}) \doteq \Phi(z_\theta)\dot{z}_\theta - (\dot{\gamma}_0/6)(z_\theta^2 - 1), \tag{12.9}$$

$z_\theta \equiv (\hat{\theta} - \theta - \beta_\theta)/\sigma_\theta$, $\dot{z}_\theta = (\partial/\partial\theta)z_\theta$. Straightforward expansions give

$$D(z^{(\alpha)}, \theta) \doteq \frac{1 + \dot{\sigma}_0 z^{(\alpha)} + \dot{\beta}_0 + (\dot{\gamma}_0/6)(z^{(\alpha)2} - 1)}{1 + \dot{\beta}_0 - \dot{\gamma}_0/6}, \tag{12.10}$$

from which $\varepsilon_\theta = (\partial/\partial z)D(z, \theta)\,|_{z=0} \doteq \dot{\sigma}_0/(1 + \dot{\beta}_0 - \dot{\gamma}_0/6)$, verifying (12.7), (12.3), and the main result (12.4).

The proof that $\theta_{\mathrm{BC}_a}[\alpha]$ also matches the Cox–McCullagh formula (11.5) is similar to the proof of Theorem 1 and will not be presented here. The main step is an expression for $\theta_{\mathrm{BC}_a}[\alpha]$ involving Lemma 5,

$$\theta_{\mathrm{BC}_a}[\alpha] \doteq z^{(\alpha)} + (\hat{k}_3/6\hat{k}_2^{3/2})\{z^{(\alpha)2} + 1\} + (\hat{k}_3/6\hat{k}_2^{3/2})^2\{2z^{(\alpha)} + z^{(\alpha)3}\}. \tag{12.11}$$

[*Received November 1984. Revised December 1985.*]

REFERENCES

Abramovitch, L., and Singh, K. (1985), "Edgeworth Corrected Pivotal Statistics and the Bootstrap," *The Annals of Statistics*, 13, 116–132.

Barndorff–Nielsen, O. E. (1984), "Confidence Limits From $c|\hat{j}|L$," Report 104, University of Aarhus, Dept. of Theoretical Statistics.

Bartlett, M. S. (1953), "Approximate Confidence Intervals," *Biometrika*, 40, 12–19.

Beran, R. (1984a), "Bootstrap Methods in Statistics," *Jber. d. Dt. Math-Verein*, 86, 14–30.

——— (1984b), "Jackknife Approximations to Bootstrap Estimates," *The Annals of Statistics*, 12, 101–118.

Bickel, P. J., and Freedman, D. A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 9, 1196–1217.

Cox, D. R. (1980), "Local Ancillarity," *Biometrika*, 67, 279–286.

DiCiccio, T. J. (1984), "On Parameter Transformations and Interval Estimation," technical report, McMaster University, Dept. of Mathematical Science.

Efron, B. (1975), "Defining the Curvature of a Statistical Problem (With Applications to Second Order Efficiency)" (with discussion), *The Annals of Statistics*, 3, 1189–1242.

——— (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

——— (1981), "Nonparametric Standard Errors and Confidence Intervals" (with discussion), *Canadian Journal of Statistics*, 9, 139–172.

——— (1982a), "The Jackknife, the Bootstrap, and Other Resampling Plans," CBMS 38, SIAM–NSF.

——— (1982b), "Transformation Theory: How Normal Is a Family of Distributions?," *The Annals of Statistics*, 10, 323–339. (NOTE Corrigenda, *The Annals of Statistics*, 10, 1032.)

——— (1984), "Comparing Non-nested Linear Models," *Journal of the American Statistical Association*, 79, 791–803.

——— (1985), "Bootstrap Confidence Intervals for a Class of Parametric Problems," *Biometrika*, 72, 45–58.

Fieller, E. C. (1954), "Some Problems in Interval Estimation," *Journal of the Royal Statistical Society*, Ser. B, 16, 175–183.

Hall, P. (1983), "Inverting an Edgeworth Expansion," *The Annals of Statistics*, 11, 569–576.

Hougaard, P. (1982), "Parameterizations of Non-linear Models," *Journal of the Royal Statistical Society*, Ser. B, 44, 244–252.

Johnson, N. J. (1978), "Modified *t* Tests and Confidence Intervals for Asymmetrical Populations," *Journal of the American Statistical Association*, 73, 536–544.

Johnson, N. L., and Kotz, S. (1970), *Continuous Univariate Distributions—2*, Boston: Houghton-Mifflin.

Kendall, M., and Stuart, A. (1958), *The Advanced Theory of Statistics*, London: Charles W. Griffin.

McCullagh, P. (1984), "Local Sufficiency," *Biometrika*, 71, 233–244.

Schenker, N. (1985), "Qualms About Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, 80, 360–361.

Singh, K. (1981), "On the Asymptotic Accuracy of Efron's Bootstrap," *The Annals of Statistics*, 9, 1187–1195.

Stein, C. (1956), "Efficient Nonparametric Testing and Estimation," in *Proceedings of the Third Berkeley Symposium*, Berkeley: University of California Press, pp. 187–196.

Tukey, J. (1949), "Standard Confidence Points," Memorandum Report 26, unpublished address presented to the Institute of Mathematical Statistics.

Withers, C. S. (1983), "Expansions for the Distribution and Quantiles of a Regular Functional of the Empirical Distribution With Applications to Nonparametric Confidence Intervals," *The Annals of Statistics*, 11, 577–587.

Comment

NATHANIEL SCHENKER*

It is a pleasure for me to discuss Bradley Efron's scholarly article. Efron proposes the BC_a method, which is an improvement over his previously proposed bootstrap confidence intervals in that it relaxes certain assumptions about pivotal quantities by using the acceleration constant a.

My discussion is divided into three sections. Section 1 relates the BC_a method to other methods of constructing confidence intervals using percentiles of the bootstrap distribution. The assumptions underlying these methods are discussed with special reference to a traditional graphical construction of exact confidence intervals. An alternative to the BC_a method for parametric problems based on directly approximating the graphical construction is described in Section 2. The direct approximation is more general than the BC_a method, and it does not require computer simulation when the family of densities of $\hat{\theta}$ is given as is assumed in most of Efron's theoretical development. For more complicated problems, the direct approximation requires more computing than the BC_a method, and the amount of computation could be prohibitive in multiparameter problems in which the number of parameters is not small. The BC_a method potentially has more value than the direct approximation in such multiparameter problems and in nonparametric problems. I look forward, therefore, to the development of a theoretical justification for the BC_a method in these more complicated situations. Section 3 discusses whether it is necessary to use the bootstrap distribution for all of the BC_a method calculations and suggests a diagnostic tool for the BC_a method.

1. BOOTSTRAP INTERVALS AND THE GRAPHICAL CONSTRUCTION OF EXACT INTERVALS

Suppose that the data $\mathbf{y}$ are a sample from an unknown distribution with cdf F. In parametric problems, where F is known to belong to a family $\{F_\theta\}$, the bootstrap idea is to approximate the sampling distribution of any function $R(\mathbf{y}, F)$ by the bootstrap distribution of $R(\mathbf{y}^*, \hat{F})$, with $\hat{F} = F_{\hat{\theta}}$. The nonparametric bootstrap uses the sample cdf as $\hat{F}$ in calculating the bootstrap distribution. When $R(\mathbf{y}, F) = \hat{\theta} - \theta$, the bootstrap approximates the sampling distribution of $\hat{\theta} - \theta$ by the bootstrap distribution of $\hat{\theta}^* - \hat{\theta}$. Combining this approximation with the usual confidence interval inversion yields

$$[2\hat{\theta} - \hat{G}^{-1}(1 - \alpha), 2\hat{\theta} - \hat{G}^{-1}(\alpha)] \qquad (1)$$

as a nominal $1 - 2\alpha$ confidence interval for θ, where $\hat{G}$ is the bootstrap cdf of $\hat{\theta}^*$. Interval (1) is criticized in Remark D of Efron (1979) in the context of estimating the median and is discussed in general in Schenker (1983), Loh (1984), and Tibshirani (1984). A problem with this interval is that $\hat{F}$ is not the same as F and, depending on how the sampling distribution of $\hat{\theta} - \theta$ changes when F changes, the bootstrap approximation will be better or worse.

A related but more fundamental point is that confidence intervals should be constructed by considering how the distribution of $\hat{\theta}$ changes as F (or θ) changes. The statistician can then determine what values of θ are reasonable given the observed value of $\hat{\theta}$. This approach is described in the parametric context in Cramér (1946, chap. 34). Figure 1 displays graphs of $G_{\hat{\theta}}^{-1}(\alpha)$ and $G_{\hat{\theta}}^{-1}(1 - \alpha)$ versus θ, where G_θ is the cdf of $\hat{\theta}$ given θ; this is a variation on Cramér's (1946) figure 33. Suppose that these quantile graphs are continuous and monotone. If the horizontal line with vertical coordinate $\hat{\theta}$ is drawn through the quantile graphs, an exact $1 - 2\alpha$ confidence interval for θ is obtained as shown in Figure 1. This is the exact interval construction mentioned in Efron's Section 5.

The bootstrap approximation provides only the points of the quantile graphs having horizontal coordinate $\hat{\theta}$ (see Fig. 1). The following question thus arises: How is it possible to draw reasonable inferences about θ based on just the bootstrap distribution? The answer is that extra assumptions are needed to extrapolate from the points of the quantile graphs provided by the bootstrap to the other points of the graphs. For example, the validity of interval (1) depends on $R(\mathbf{y}, F) = \hat{\theta} - \theta$ being a pivotal quantity, that is, having the same sampling distribution for all values of θ. In terms of Figure 1, interval (1) assumes that the quantile graphs are linear with unit slope. Schenker (1983, 1985) and Tibshirani (1984) applied the nonparametric bootstrap to $R(\mathbf{y}, F) = \hat{\theta}/\theta$, where θ is the variance of F (see Efron's Sec. 7, Example 3). The resulting intervals had reasonable coverage properties for simulated $N(0, 1)$ data, since $\hat{\theta}/\theta$ is pivotal under the normal family.

In many applications, the exact form of a pivotal quantity, assuming one exists, is unknown. Efron's (1981, 1982) BC method seeks to alleviate this problem by assuming that for some monotone-increasing transformation g (which need not be known), $R(\mathbf{y}, F) = g(\hat{\theta}) - g(\theta)$ is a normal pivotal quantity; see Efron's (2.2). Thus when the axes of Figure 1 are transformed to the g scale, the quantile graphs are linear with unit slope and the distance between the lines is determined by the normal distribution. The BC interval

$$[\hat{G}^{-1}(\Phi(2z_0 - z^{(1-\alpha)})), \hat{G}^{-1}(\Phi(2z_0 - z^{(\alpha)}))] \qquad (2)$$

* Nathaniel Schenker is with the Undercount Research Staff, Statistical Research Division, U.S. Bureau of the Census, Washington, DC 20233. The author thanks David F. Findley for helpful comments on a previous version of this work, Maureen P. Lynch for preparing Figure 1, and David L. Wallace for several enlightening conversations about inference and the bootstrap.

Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods

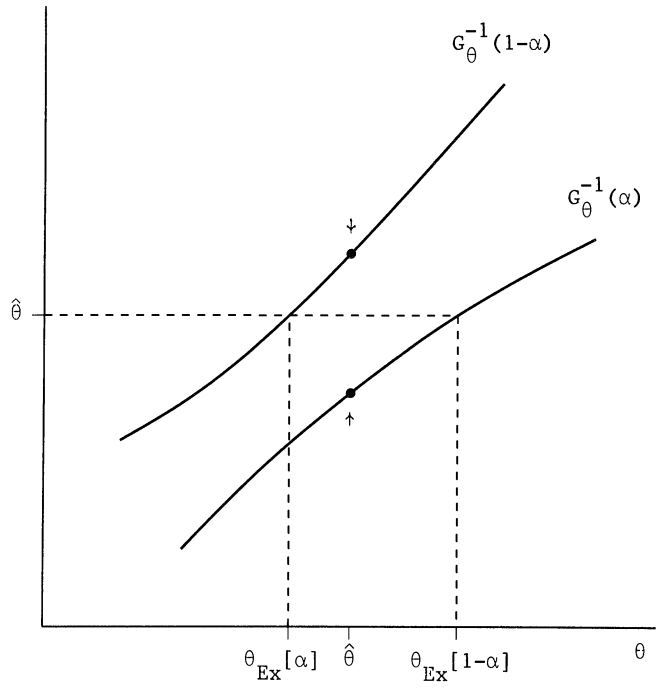


Figure 1. Graphical Construction of the Exact $1 - 2\alpha$ Confidence Interval $[\theta_{Ex}[\alpha], \theta_{Ex}[1 - \alpha]]$ for θ Given $\hat{\theta}$. The arrows show the points of the quantile graphs that are provided by the bootstrap.

(see Efron's Sec. 3) is related to (1) as follows. First, $\hat{\theta}$ is transformed to the normal scale by the mapping $\Phi^{-1}\hat{G}$. Interval (1) is then formed on the normal scale, resulting in $[2z_0 - z^{(1-\alpha)}, 2z_0 - z^{(\alpha)}]$. Finally, the interval is transformed to the original scale by the mapping $\hat{G}^{-1}\Phi$.

Efron's (1981, 1982) percentile method interval

$$[\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1 - \alpha)] \tag{3}$$

(see Efron's Sec. 3) does not explicitly reflect a sampling distribution of $\hat{\theta}$ about θ in any way. It is more similar to a Bayesian posterior interval with the bootstrap distribution of $\hat{\theta}^*$ treated as the posterior distribution of θ. In fact, Rubin (1981) showed that the nonparametric bootstrap distribution of $\hat{\theta}^*$ is quite similar to the posterior distribution for θ obtained from a specific Bayesian analysis involving strong prior assumptions. Rubin (1981) argued that since the validity of inferences based on the nonparametric bootstrap depends on the validity of the prior assumptions, the bootstrap is not a panacea for avoiding sensitivity of inferences to model assumptions.

The percentile method interval (3) is the reflection of interval (1) about $\hat{\theta}$. If $\hat{\theta} - \theta$ is pivotal with distribution symmetric about zero, then $\hat{G}$ represents a distribution symmetric about $\hat{\theta}$ and (1) and (3) coincide. Efron (1981, 1982) pointed out that the percentile method is valid under the more general condition that there exists a monotone increasing transformation g such that $g(\hat{\theta}) - g(\theta)$ is pivotal with distribution symmetric about zero. Thus the method is valid if transforming the axes in Figure 1 to the g scale yields quantile graphs that are linear and equidistant from the line through the origin with unit slope.

Every method discussed so far relies on the existence of a pivotal quantity $R(\mathbf{y}, F)$ for its validity. To bootstrap $R(\mathbf{y}, F)$ directly, the exact form of the pivotal quantity must be known; the sampling distribution of $R(\mathbf{y}, F)$ can remain unknown since it is approximated by the bootstrap. In contrast, the BC and percentile methods impose restrictions on $R(\mathbf{y}, F)$ and its sampling distribution so that the exact form of $R(\mathbf{y}, F)$ need not be known.

Unfortunately, a transformation $g(\hat{\theta}) - g(\theta)$ to a normal pivotal quantity does not always exist. For instance, when θ is the variance of a normal distribution, $\log(\hat{\theta}) - \log(\theta)$ is pivotal but not normal, whereas $\hat{\theta}^{1/3} - \theta^{1/3}$ is approximately normal but not pivotal (see Schenker 1983, 1985). Efron's newly proposed BC_a method allows $R(\mathbf{y}, F)$ to deviate from being a pivotal quantity in a particular way. The assumptions underlying the BC method are relaxed by allowing the scaling of the sampling distributions of $g(\hat{\theta}) - g(\theta)$ to vary linearly with $g(\theta)$ at rate a; see Efron's (2.3). This induces curvature in the quantile graphs of Figure 1 (when the axes are transformed to the g scale), with the curvature greater for larger values of a.

As Efron's article and the discussion in this section show, inferences for θ cannot be drawn solely on the basis of the bootstrap distribution. Assumptions need to be made about the distribution of $\hat{\theta}$ under alternative values of θ. Thus the bootstrap is not an assumption-free method.

2. AN ALTERNATIVE TO THE BC_a METHOD FOR PARAMETRIC PROBLEMS

Most of the theoretical development in Efron's article is based on the family of densities of $\hat{\theta}$, say $\{g_\theta\}$, being given. (I depart from Efron's notation and use g_θ as the density of $\hat{\theta}$ to distinguish it from f_θ, the density of the data.) Given the form of g_θ, either an analytic expression for G_θ is available or G_θ can be computed by numerical integration. This allows the quantile graphs of Figure 1 and the exact interval for θ to be constructed without any bootstrapping or simulation at all. Thus, as pointed out in Efron's Section 2, the bootstrap is not needed when $\{g_\theta\}$ is given. In fact, the logic of using knowledge of $\{g_\theta\}$ in the BC_a method seems circular since the basic purpose of the bootstrap is to approximate G_θ. For these reasons, the potential value of the BC_a method depends on the efficacy of the extensions to more complicated situations discussed in Efron's Sections 6–8 and Remark F of Section 11.

Efron (Sec. 2) describes ways in which the problem of setting a confidence interval for θ can be made more complicated. One way is to assume that only the family of densities $\{f_\theta\}$ of the data is given rather than $\{g_\theta\}$. In such a case, the following computer-intensive method based on the graphical construction in Figure 1 can be used to simulate directly the exact interval for θ without bootstrapping. Let $\theta(b_1)$ $(b_1 = 1, \ldots, B_1)$ be B_1 values of θ spaced throughout a range wide enough to include the exact interval with certainty, say $\theta \in \hat{\theta} \pm 5\hat{\sigma}$. For fixed b_1, draw independent samples $\mathbf{y}^*(b_1, b_2)$ $(b_2 = 1, \ldots, B_2)$ from $f_{\theta(b_1)}$ and compute $\hat{\theta}(\mathbf{y}^*(b_1, b_2))$ for each of the B_2 draws. For every real s, let

$$\hat{G}_{\theta(b_1)}(s) = \frac{\#\{\hat{\theta}(\mathbf{y}^*(b_1, b_2)) \leq s\}}{B_2}.$$

Evaluate $\hat{G}^{-1}_{\hat{\theta}(b_1)}(\alpha)$ and $\hat{G}^{-1}_{\hat{\theta}(b_1)}(1 - \alpha)$ and graph them against $\theta(b_1)$. If this is done for $b_1 = 1, \ldots, B_1$, an approximation to the quantile graphs of Figure 1 is obtained. Thus simulation can be used to approximate the exact interval for θ directly without the extra assumptions underlying the BC_a method.

As $B_1, B_2 \to \infty$, the direct approximation approaches the exact interval. Of course, although B_1 and B_2 can be made very large in principle, they should be kept as small as possible in practice. If the quantile graphs are reasonably smooth, it should be feasible to simulate $G^{-1}_{\hat{\theta}}(\alpha)$ and $G^{-1}_{\hat{\theta}}(1 - \alpha)$ for just a few values of θ (i.e., small B_1) and then fit a curve through the points for α and a curve through the points for $1 - \alpha$. For many situations including those considered by Efron, the quantile graphs should be smooth. Furthermore, B_2 need not be any larger than the number of replications needed for the bootstrap (see Efron's Sec. 9). Thus the direct approximation requires about B_1 times as much computation as the bootstrap, where B_1 is often small.

The other more complicated cases considered by Efron are the multiparameter and nonparametric situations. In principle, the method presented in this section of directly approximating the exact interval for θ can be extended to multiparameter cases. When more than just a few parameters are involved, however, the amount of computing could become prohibitive, even given the current spirit of intensive computation. It does not seem possible to extend the method of direct approximation to nonparametric problems. Thus the BC_a method could prove to have greater value than the method of direct approximation in multiparameter problems and especially in nonparametric problems.

Efron (Sec. 2) states that it is quite easy to extend the BC_a method to complicated situations. In Section 6, he suggests an extension to multiparameter problems that replaces the multiparameter family of densities of the data by its least favorable one-parameter subfamily. For nonparametric problems (Sec. 7 and 8), he applies the multiparameter extension to the multinomial family having support given by the values in $\mathbf{y}$. These are elegant ideas, but it is not clear that the extensions are valid. As Efron points out in Sections 6 and 7, the theory developed in his article showing second-order validity of the BC_a method for the simple case $\hat{\theta} \sim g_\theta$ has not been extended to the multiparameter and nonparametric situations. Since these are the situations in which the BC_a method potentially has more value than the method of directly approximating the exact interval, I look forward to Efron developing the theory for these situations completely.

3. THE ROLE OF THE BOOTSTRAP DISTRIBUTION IN THE BC_a METHOD

A traditional principle of the bootstrap is that inferences are drawn based on quantities computed from the bootstrap distribution of $\hat{\theta}^*$. Efron's current article, for example, describes how to obtain the BC_a interval from such bootstrap calculations. Consideration of Efron's proof of Lemma 1 shows that the use of $\hat{G}^{-1}(\cdot)$ in (3.8) is necessary to avoid having to know the transformation g. It may not be necessary, however, to derive the bias correction z_0 and the acceleration constant a used in (3.9) from the bootstrap distribution, at least in one-parameter situations. Under conditions (3.5)–(3.7), it should be possible to calculate z_0 and a from the sampling distribution of $\hat{\theta}$ under any value of θ. Specifically, (4.3) shows that $z_0 = \Phi^{-1}(G_\theta(\theta))$ for any θ and the argument following the statement of Lemma 2 implies that $\text{SKEW}_{\theta=\hat{\theta}}$ in (4.4) can be replaced by SKEW_θ for any θ.

Perhaps computing z_0 and a from the bootstrap distribution is important in that (3.5)–(3.7) might only be a good approximation to the sampling distributions of $\hat{\theta}$ under values of θ in a neighborhood of $\hat{\theta}$. If this is so, then it may be useful to calculate $\Phi^{-1}(G_\theta(\theta))$ and $\text{SKEW}_\theta(\dot{l}_\theta)$ for a few values of θ in the neighborhood of $\hat{\theta}$ as a diagnostic tool. If either of these quantities were to vary greatly with θ, the use of (3.5)–(3.7) as a local approximation would be called into question.

ADDITIONAL REFERENCES

Cramér, H. (1946), *Mathematical Methods of Statistics,* Princeton: Princeton University Press.

Loh, Wei-Yin (1984), "Estimating an Endpoint of a Distribution With Resampling Methods," *The Annals of Statistics,* 12, 1543–1550.

Rubin, D. B. (1981), "The Bayesian Bootstrap," *The Annals of Statistics*, 9, 130–134.

Schenker, N. (1983), "Qualms About Bootstrap Confidence Intervals," Technical Report 150, University of Chicago, Dept. of Statistics.

Tibshirani, R. (1984), "Bootstrap Confidence Intervals," Technical Report 3, Stanford University, Laboratory for Computational Statistics, Dept. of Statistics.

Advanced Statistics: Bootstrapping Confidence Intervals for Statistics with "Difficult" Distributions

Jason S. Haukoos, MD, MS, Roger J. Lewis, MD, PhD

Abstract

The use of confidence intervals in reporting results of research has increased dramatically and is now required or highly recommended by editors of many scientific journals. Many resources describe methods for computing confidence intervals for statistics with mathematically simple distributions. Computing confidence intervals for descriptive statistics with distributions that are difficult to represent mathematically is more challenging. The bootstrap is a computationally intensive statistical technique that allows the researcher to make inferences from data without making strong distributional assumptions about the data or the statistic being calculated. This allows the researcher to estimate confidence intervals for statistics that do not have simple sampling distributions (e.g., the median). The purposes of this article are to describe the concept of bootstrapping, to demonstrate how to estimate confidence intervals for the median and the Spearman rank correlation coefficient for non-normally-distributed data from a recent clinical study using two commonly used statistical software packages (SAS and Stata), and to discuss specific limitations of the bootstrap. **Key words:** bootstrap; resampling; median; Spearman rank correlation; SAS; Stata; NOSIC Score; confidence intervals. ACADEMIC EMERGENCY MEDICINE 2005; 12:360–365.

The use of confidence intervals in reporting the results of biomedical research has increased dramatically over the past several years. It is well known that confidence intervals provide more information than p-values, and editors of many scientific journals are now requiring or highly recommending their use.[1,2] While a number of articles report methods by which to calculate confidence intervals , they primarily focus on estimating confidence intervals for statistics with mathematically simple distributions, at least when the data themselves have a straightforward sampling distribution (e.g., normal or binomial distribution).[3–6]

In a recent publication, Okada et al. reported confidence intervals around Spearman rank correlation coefficients.[7] The primary objective of their study was to develop and evaluate a neurologic outcome measure, called the Neurologic Outcome Scale for Infants and Children (NOSIC), for pediatric research subjects with neurologic deficits. The NOSIC scale ranges from 3 to 100 and was applied independently by two clinical investigators to a cohort of patients in order to assess its reliability. The first rater (rater 1) applied the NOSIC to 157 patients and the second rater (rater 2) applied it to 84 of the 157 patients. These data are shown in Figures 1–3. It is evident from Figures 1 and 2 that the distributions are highly skewed, making reporting of the medians and Spearman rank correlation coefficient more valid than reporting the means and Pearson correlation coefficient for characterizing each rater's scores and the interrater reliability.

The confidence intervals for the Spearman rank correlation coefficients were estimated using the bootstrap, a statistical method based on resampling that can be used to perform statistical inference.[8] The purpose of this article is to describe the steps in bootstrapping, to demonstrate how to estimate confidence intervals using two commonly used statistical software packages (SAS[9] and Stata[10]) using the data from the Okada study, and to briefly discuss some limitations of the technique.

BOOTSTRAPPING

Bootstrapping was introduced in 1979 as a computationally intensive statistical technique that allows the researcher to make inferences from data without making strong distributional assumptions.[8,11] There are two distributions to consider. The first is the

From the Department of Emergency Medicine, Denver Health Medical Center (JSH), Denver, CO; the Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center (JSH), Denver, CO; the Department of Emergency Medicine, Harbor–UCLA Medical Center (RJL), Torrance, CA; the Los Angeles Biomedical Research Institute at Harbor–UCLA Medical Center (RJL), Torrance, CA; and the David Geffen School of Medicine at UCLA (RJL), Los Angeles, CA.

Received September 19, 2004; revision received October 30, 2004; accepted November 1, 2004.

Series editor: Roger J. Lewis, MD, PhD, Senior Statistical Editor, *Academic Emergency Medicine*, Harbor–UCLA Medical Center, Torrance, CA.

Supported in part by an Individual National Research Service Award from the Agency for Healthcare Research and Quality (F32 HS11509) and a Research Training Grant from the Society for Academic Emergency Medicine to Dr. Haukoos.

Address for correspondence and reprints: Jason S. Haukoos, MD, MS, Department of Emergency Medicine, Denver Health Medical Center, 777 Bannock Street, Mail Code 0108, Denver, CO 80204. Fax: 303-436-7541; e-mail: jason.haukoos@dhha.org.

doi:10.1197/j.aem.2004.11.018

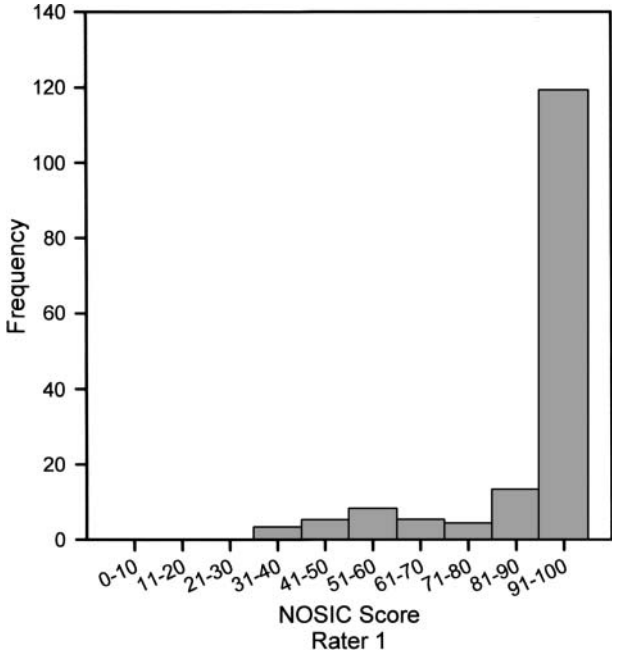


Figure 1. Frequency histogram of scores using the Neurologic Outcome Scale for Infants and Children (NOSIC) for rater 1 (n = 157). The mean value is 90 (standard deviation = 16) and the median value is 97 (interquartile range: 92–100).

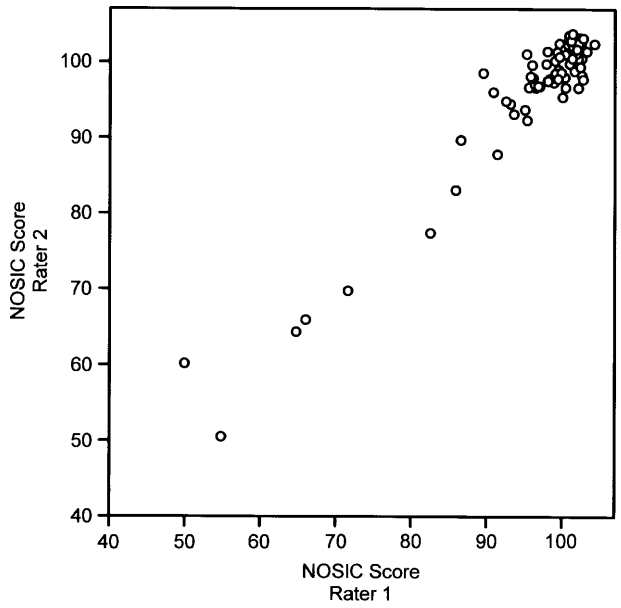


Figure 3. Neurologic Outcome Scale for Infants and Children (NOSIC) scores for rater 1 and rater 2 (n = 84). Data points are ``smeared'' using a normally-distributed random number generator to improve the representation of exactly overlapping data. As a result, some data points exceed 100. The Pearson correlation coefficient is 0.97 and the Spearman correlation coefficient is 0.77.

underlying distribution of the data themselves, which is frequently described as a probability function (e.g., normal, binomial, or Poisson) that shows all the values that the variables can have and the likelihood, or probability, that each will occur. The second is the

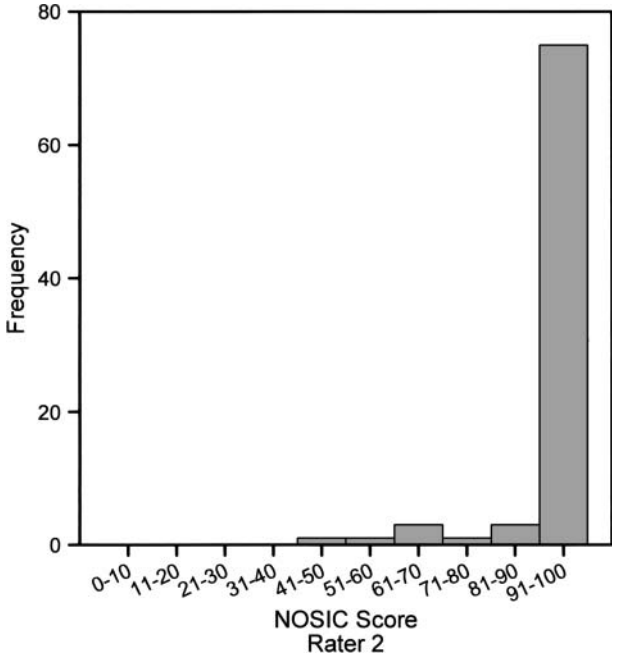


Figure 2. Frequency histogram of scores using the Neurologic Outcome Scale for Infants and Children (NOSIC) for rater 2 (n = 84). The mean value is 95 (standard deviation = 10) and the median value is 98 (interquartile range: 95–100).

distribution of the statistic (e.g., the median) calculated from the data. Both the items of data and the calculated statistic will vary in ways that can be described mathematically under the assumption that new sets of data were obtained or ''sampled'' and, for each set of data, a new statistic was calculated. More precisely, the statistic's sampling distribution is the probability of all possible values of the estimated statistic calculated from a sample of size n drawn from a given population.[12] Bootstrapping uses resampling with replacement (also known as Monte Carlo resampling) to estimate the statistic's sampling distribution. The sampling distribution, if it can be determined, may then be used to estimate standard errors and confidence intervals for that particular statistic.

The steps for estimating confidence intervals using the bootstrap are as follows (Figure 4): First, one uses resampling with replacement to create m resampled data sets (also known as bootstrap samples) that contain the same number of observations (n) as the original data set. To perform resampling with replacement, an observation or data point is randomly selected from the original data set and copied into the resampled data set being created. Although that data point has been ''used,'' it is not deleted from the original data set or, using the usual terminology, is ''replaced.'' Another data point is then randomly selected, and the process is repeated until a resampled data set of size n is created. As a result, the same observation may be included in the resampled data set one, two, or more

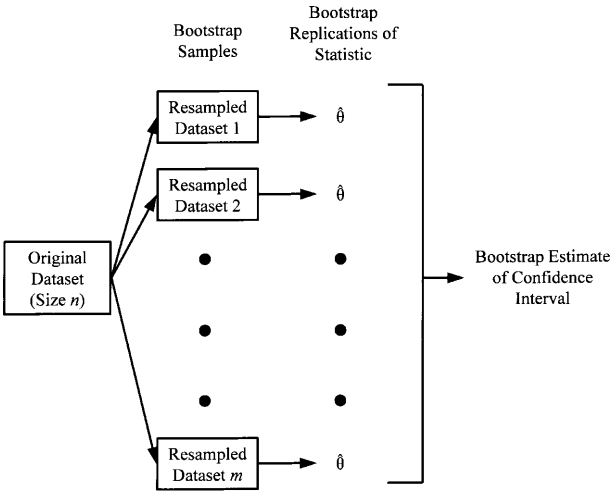


Figure 4. Schematic depiction of the steps in the bootstrap.

times, or not at all. Second, the descriptive statistic of choice is computed for each resampled data set. Third, a confidence interval for the statistic is calculated from the collection of values obtained for the statistic. At this point in the analysis, there are several options for computing confidence intervals, including the normal approximation method, the percentile method, the bias-corrected (BC) method, the bias-corrected and accelerated (BC_a) method, and the approximate bootstrap confidence (ABC) method.[8]

Each bootstrap sample should have the same sample size as the original data set. If the bootstrap sample sizes differ from the sample size of the original data set, the calculated estimation for the confidence interval may be biased.[13] A correction for this bias has been described, although there seems to be no practical advantage gained by performing the analysis in this manner.[14]

The normal approximation method computes an approximate standard error using the sampling distribution resulting from all the bootstrap resamples. The confidence interval is then computed using the z-distribution (original statistic $\pm$ 1.96 $\times$ standard error, for a 95% confidence interval). The percentile method uses the frequency histogram of the m statistics computed from the bootstrap samples. The 2.5 and 97.5 percentiles constitute the limits of the 95% confidence interval. The BC_a method adjusts for bias in the bootstrapped sampling distributions relative to the actual sampling distribution, and is thus considered a substantial improvement over the percentile method.[8] The BC_a confidence interval is an adjustment of the percentiles used in the percentile method based upon the calculation of two coefficients called ''bias correction'' and ''acceleration.'' The bias correction coefficient adjusts for the skewness in the bootstrap sampling distribution. If the bootstrap sampling distribution is perfectly symmetric, then the bias correction will be zero.[8] The acceleration coefficient adjusts for nonconstant variances within the resampled data sets.[8] The ABC method is an approximation of the BC_a method that requires fewer resampled data sets than the BC_a method.[8]

As a general guideline, 1,000 or more resampled data sets should be used when calculating a BC_a confidence interval.[11] As a result of not having to calculate bias correction, a smaller value, in the range of 250, can be used when using the percentile method for estimating a confidence interval.[13] As the number of resampled data sets decreases, more variability is introduced into the confidence interval estimation (i.e., the variability is inversely related to the number of resampled data sets).[8,13]

Example 1: Determining a Confidence Interval around a Median Value. A median value is defined as the observation at the 50th percentile in a set of data ordered from the lowest value to the highest value.[15] This measure of center for a set of values is commonly reported and is considered a more valid definition of center when the frequency distribution of the variable is skewed (i.e., not symmetric around its center). Unlike the mean, there is no simple method for calculating the 95% confidence interval (95% CI) for the median, and it is not valid to use a 95% CI calculated from the standard error to represent the 95% confidence for the median value, unless the distribution of the underlying data is normal. As a result, the bootstrap can be used to estimate the sampling distribution of the median. The central limit theorem states that as the number of resampled data sets increases, the distribution of the resulting statistic, in this case the median, will become approximately normal.[15] This subsequently allows for a relatively unbiased estimation of the confidence interval.

The steps required to bootstrap the 95% CI for a median value are: 1) to resample with replacement from the original data set, creating m bootstrapped data sets; 2) to independently compute the median value for each bootstrapped data set; and 3) to compute the 95% CI from the set of computed median values from the bootstrapped data sets using either the normal approximation method, the percentile method, the BC method, the BC_a method, or the ABC method.

These steps can be accomplished using the SAS software program (SAS Institute, Inc., Cary, NC) as follows. The SAS macro JACKBOOT, which can be obtained from the SAS Web site, must be invoked prior to performing a bootstrap analysis in SAS.[16] A ''macro'' is a program that can be executed by SAS and that may be modified by the user, while a SAS procedure is a ''fixed'' program that performs a specific statistical calculation or other task. The JACKBOOT macro requires another macro (called ANALYZE) to be written that provides it with the procedure

whose result (e.g., the median of the original data set) requires bootstrapping. The univariate procedure (PROC UNIVARIATE) in SAS is used to compute the median value for a group of observations. The following is the ANALYZE macro, modified to bootstrap a 95% CI around a median value for the variable ''normscr1'' (NOSIC score for rater 1):

```
%macro analyze (data=, out=);
proc univariate noprint data=&data;
  output out=&out (drop=_freq_ _type_)
  median=median;
  var normscr1;
  %bystmt;
run;
%mend;
```

In SAS, the ''%macro'' term indicates the beginning of a macro, and is followed by its title (i.e., ''analyze''). The ''%mend'' term indicates the end of a macro, and all text between ''%macro'' and ''%mend'' is called macro text. In this example, PROC UNIVARIATE is invoked with the ''noprint'' option. The ''data=&data'' term references the original data set through the JACKBOOT macro using the ''%boot'' term (see below). The ''output'' statement directs SAS to create a temporary output file for only the median values, as indicated by the term ''median=median,'' for the variable ''normscr1.'' The ''%bystmt'' term references a macro within the JACKBOOT macro that computes a statistic (in this case, the median) for the original data set and for each resampled data set.

The ANALYZE macro is followed immediately by the following bootstrap commands:

```
%boot (data=temp, samples=2500);
%bootci (percentile);
%bootci (bca);
```

In this example, the ANALYZE macro is used by the JACKBOOT macro to apply the statistical procedure (PROC UNIVARIATE) to the original data set (data=temp, referenced in the ''%boot'' statement). The ''%boot'' command invokes the bootstrap procedure, resulting in 2,500 bootstrapped samples, and the ''%bootci'' command invokes the bootstrap confidence interval procedure. The first ''%bootci'' command uses the percentile method to compute a 95% CI for the median and the second ''%bootci'' command uses the BC_a method to compute a 95% CI for the median of the variable ''normscr1.'' The median value was 97 [interquartile range (IQR): 92–100, range 32–100], and the 95% CIs for the median were 96–98 (percentile) and 97 to 98 (BC_a).

Using Stata (Stata Corporation, College Station, TX) to perform the same calculations is substantially simpler. The following Stata commands compute the median value for the variable ''normscr1'' and bootstrap the 95% CIs using the normal, percentile, and BC_a methods using 2,500 resamples[17]:

```
centile normscr1
bs ``centile normscr1'' ``r(c_1)'',
rep(2500)
```

The ''centile'' command calculates the median value for the variable ''normscr1.'' The ''bs'' command calculates a bootstrapped confidence interval for the median value for the variable ''normscr1.'' The primary code appears in the first quotations, '''r(c_1)'' refers to the reference statistic for which the 95% CI will be calculated, and ''rep(2500)'' indicates the number of resampled data sets. After the primary command has been executed, the command ''return list'' can be used to display the codes for each of the resulting statistics for the primary command. In this example, ''c_1'' is the code that refers to the median value calculated by the ''centile'' command.

Example 2: Determining a Confidence Interval around a Spearman Rank Correlation Coefficient. The Spearman rank correlation coefficient is the nonparametric counterpart to the parametric Pearson correlation coefficient.[15] The Pearson correlation coefficient is a valid statistical technique for determining correlation between two normally-distributed continuous variables. On the other hand, the Spearman rank correlation coefficient is a valid statistical technique for determining correlation between two non-normally-distributed continuous variables.

The PROC CORR procedure in SAS is used to compute the Pearson correlation coefficient, and there are two methods for computing the Spearman rank correlation coefficient. The first method simply involves incorporating the option ''spearman'' into the PROC CORR statement. The second method involves ranking the data, using PROC RANK, prior to using PROC CORR.

The following illustrates the ANALYZE macro used by the JACKBOOT macro to perform the bootstrap in SAS:

```
%macro analyze (data=, out=);
proc rank data=&data out=tempdata;
  var normscr1 normscr2;
  %bystmt;
proc corr noprint
  data=tempdata
  out=&out (rename=(_type_=stat
   _name_=with));
  var normscr1 normscr2;
  %bystmt;
run;
%mend;
```

The macro text in this example includes the PROC RANK command for variables ''normscr1'' and ''normscr2.'' This command is followed by the PROC CORR command, which performs correlation of the two ranked variables for each resampled data set. The ''out=tempdata'' term writes a temporary output file

of all ranked resampled data sets. This is read as an input file using the term "data=tempdata" in the PROC CORR command.

Again, the ANALYZE macro is followed immediately by the bootstrap commands:

```
%boot (data=temp, id=stat with,
samples=2500);
%bootci (percentile, id=stat with);
%bootci (bca, id=stat with);
```

In this example, 2,500 bootstrapped samples were created, and the percentile and BC_a methods were used to compute 95% CIs for the Spearman rank correlation coefficient between the variables "normscr1" and "normscr2" (NOSIC score for rater 2). The Spearman rank correlation coefficient was 0.77 for the original data set and the 95% CIs were 0.62–0.88 (percentile) and 0.62–0.87 (BC_a).

Again, it is simpler to perform this calculation using Stata. The following Stata commands compute the Spearman rank correlation coefficient between "normscr1" and "normscr2," and bootstrap the 95% confidence intervals using the normal, percentile, and BC_a methods using 2,500 resamples:

```
spearman normscr1 normscr2
bs ``spearman normscr1 normscr2''
``r(rho)'', rep(2500)
```

The "spearman" command calculates the Spearman rank correlation coefficient for "normscr1" and "normscr2." The primary code appears in the first quotations, "r(rho)" refers to the reference statistic for which the 95% CI will be calculated, and "rep(2500)" indicates the number of resampled data sets.

Limitations of the Bootstrap. Although the idea of the bootstrap has been around for nearly two centuries, theoretical work on the bootstrap is relatively recent and, therefore, the limitations of the bootstrap are not entirely understood.[11] The bootstrap is a tool used, in part, to calculate confidence intervals for point estimates of descriptive statistics. The bootstrap should not be used to compute point estimates themselves, however. The sampling distribution of the bootstrapped statistics is frequently not symmetric. Thus, computing point estimates in this manner may reflect, as opposed to alleviate, biased estimation from the samples.[11] The extent of bias can be estimated but is subject to high variability, making bias correction infeasible.[8]

The most important limitation of the bootstrap is the assumption that the distribution of the data represented by the sample is a reasonable estimate of the population distribution function from which the data are sampled. In other words, the sample must reflect the variety and range of possible values in the population from which it was sampled. If the distribution of data from the sample does not reflect the population distribution function, then the random sampling performed in the bootstrap procedure may add another level of sampling error, resulting in invalid statistical estimations.[18] This emphasizes the importance of obtaining quality data that accurately reflect the characteristics of the population being sampled.

Additionally, the smaller the original sample, the less likely it is to represent the entire population. Thus, the smaller the sample, the more difficult it becomes to compute valid confidence intervals. The bootstrap relies heavily on the tails of the estimated sampling distribution when computing confidence intervals, and using small samples may jeopardize the validity of this computation.[18]

Random sampling performed in the bootstrap procedure also adds another level of potential sampling error. This, as mentioned previously, is reflected in the variation and bias estimates commonly performed during a bootstrap analysis.

CONCLUSIONS

The bootstrap is a relatively simple statistical concept that requires computationally intensive procedures to implement. Modern statistical software packages now allow researchers to employ relatively simple programming to compute confidence intervals for statistics with inconvenient or unknown sampling distributions.

The authors gratefully thank Pamela J. Okada, MD, and Kelly D. Young, MD, MS, for providing the original NOSIC data, and Stephen P. Wall, MD, MPH, for providing programming suggestions in Stata.

References

1. Uniform requirements for manuscripts submitted to biomedical journals. International Committee of Medical Journal Editors. JAMA. 1997; 277:927–34.
2. Cooper RJ, Wears RL, Schriger DL. Reporting research results: recommendations for improving communication. Ann Emerg Med. 2003; 41:561–4.
3. Blyth CR. Approximate binomial confidence limits. J Am Stat Assoc. 1986; 81:843–55.
4. Troendle JF, Frank J. Unbiased confidence intervals for the odds ratio of two independent binomial samples with application to case–control data. Biometrics. 2001; 57:484–9.
5. Young KD, Lewis RJ. What is confidence? Part I: the use and interpretation of confidence intervals. Ann Emerg Med. 1997; 30:307–10.
6. Young KD, Lewis RJ. What is confidence? Part II: detailed definition and determination of confidence intervals. Ann Emerg Med. 1997; 30:311–8.
7. Okada PJ, Young KD, Baren JM, et al. Neurologic outcome score for infants and children. Acad Emerg Med. 2003; 10: 1034–9.
8. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall/CRC, 1998.
9. SAS Version 8.2. SAS Institute, Inc., Cary, NC.
10. Stata Version 8. Stata Corporation, College Station, TX.

11. Mooney CZ, Duval RD. Bootstrapping: A Nonparametric Approach to Statistical Inference. Beverly Hills, CA: Sage Publications, 1993.
12. Levy PS, Lemeshow S. Sampling of Populations: Methods and Applications—3rd edition. New York: John Wiley & Sons, 1999.
13. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci. 1986; 1:54–77.
14. Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap. Ann Stat. 1981; 9:1196–217.
15. Zar JH. Biostatistical Analysis—4th edition. Englewood Cliffs, NJ: Prentice Hall, 1999.
16. Jackknife and bootstrap analyses: SAS jackboot macro. Available at: http://ftp.sas.com/techsup/download/stat/jackboot.html. Accessed Apr 9, 2004.
17. Stata17ase Reference Manual: Volume 1, A-F, Release 8. College Station, TX: Stata Press, 2003.
18. Schenker N. Qualms about bootstrap confidence intervals. J Am Stat Assoc. 1985; 80:360–1.

STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:4279–4292
Published online 31 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2673

The design of simulation studies in medical statistics

Andrea Burton[1,2,*,†], Douglas G. Altman[1], Patrick Royston[1,3] and Roger L. Holder[4]

[1]*Cancer Research UK/NHS Centre for Statistics in Medicine, Oxford, U.K.*
[2]*Cancer Research UK Clinical Trials Unit, University of Birmingham, Birmingham, U.K.*
[3]*MRC Clinical Trials Unit, London, U.K.*
[4]*Department of Primary Care and General Practice, University of Birmingham, Birmingham, U.K.*

SUMMARY

Simulation studies use computer intensive procedures to assess the performance of a variety of statistical methods in relation to a known truth. Such evaluation cannot be achieved with studies of real data alone. Designing high-quality simulations that reflect the complex situations seen in practice, such as in prognostic factors studies, is not a simple process. Unfortunately, very few published simulation studies provide sufficient details to allow readers to understand fully all the processes required to design a simulation study. When planning a simulation study, it is recommended that a detailed protocol be produced, giving full details of how the study will be performed, analysed and reported. This paper details the important considerations necessary when designing any simulation study, including defining specific objectives of the study, determining the procedures for generating the data sets and the number of simulations to perform. A checklist highlighting the important considerations when designing a simulation study is provided. A small review of the literature identifies the current practices within published simulation studies.

KEY WORDS: simulation study; design; protocol; bias; mean square error; coverage

1. INTRODUCTION

Simulation studies use computer intensive procedures to test particular hypotheses and assess the appropriateness and accuracy of a variety of statistical methods in relation to the known truth. These techniques provide empirical estimation of the sampling distribution of the parameters of interest that could not be achieved from a single study and enable the estimation of accuracy measures, such as the bias in the estimates of interest, as the truth is known [1]. Simulation studies are increasingly being used in the medical literature for a wide variety of situations, (e.g. References

*Correspondence to: Andrea Burton, Cancer Research UK/NHS Centre for Statistics in Medicine, Wolfson College Annexe, Linton Road, Oxford OX2 6UD, U.K.
†E-mail: andrea.burton@cancer.org.uk

Contract/grant sponsor: Cancer Research U.K.

Received 15 June 2006
Accepted 6 July 2006

[2–4]). In addition, simulations can be used as instructional tools to help with the understanding of many statistical concepts [5, 6].

Designing high-quality simulations that reflect the complex situations seen in practice, such as in randomized controlled trials or prognostic factor studies, is not a simple process. Simulation studies should be designed with similar rigour to any real data study, since the results are expected to represent the results of simultaneously performing many real studies. Unfortunately, in very few published simulation studies are sufficient details provided to assess the integrity of the study design or allow readers to understand fully all the processes required when designing their own simulation study. Performing any simulation study should involve careful consideration of all design aspects of the study prior to commencement of the study from establishing the aims of the study, the procedures for performing and analysing the simulation study through to the presentation of any results obtained. These are generic issues that should be considered irrespective of the type of simulation study but there may also be further criteria specific to the area of interest, for example survival data.

It is important for researchers to know the criteria for designing a good quality simulation study. The aim of this paper is to provide a comprehensive evaluation of the generic issues to consider when performing any simulation study, together with a simple checklist for researchers to follow to help improve the design, conduct and reporting of future simulation studies. The basic concepts underpinning the important considerations will be discussed, but full technical details are not provided and the readers are directed towards the literature (e.g. References [7, 8]). General considerations are addressed rather than the specific considerations for particular situations where simulations are extremely useful, such as in Bayesian clinical trials designs (e.g. Reference [9]), sample size determination (e.g. References [3, 10]), or in studies of missing data (e.g. Reference [4]). A small formal review of the current practice within published simulation studies is also presented.

2. ISSUES TO CONSIDER WHEN DESIGNING A SIMULATION STUDY

When planning any simulation study, as with randomized controlled trials, a detailed protocol should be produced giving full details of how the study is to be performed, analysed and reported. The protocol should document the specific objectives for the simulation study and the procedures for generating multivariate data sets and, if relevant, with censored survival times. The choices for the different scenarios to be considered, for example different sample sizes, and the methods that will be evaluated should also be included in the protocol together with the number of simulations that will be performed. It is also important to give careful consideration to which data and results will be stored from each run, and which summary measures of performance will be used. If an aim of the study is to judge which is the best of two or more methods, then the criteria should be prespecified in the protocol, where possible. The rationale behind all the decisions made throughout the design stage should be included in the protocol.

Each of the preceding considerations will be discussed in more detail in the following sections. A checklist of the important issues that require consideration when designing a simulation study is provided in Figure 1.

2.1. Clearly defined aims and objectives

Establishing clearly defined aims for the simulation study prior to its commencement is an essential part of any research. This focuses the study and avoids unnecessary repetition and time wasting from having to repeat simulations when new aims are conceptualized.

DOI: 10.1002/sim

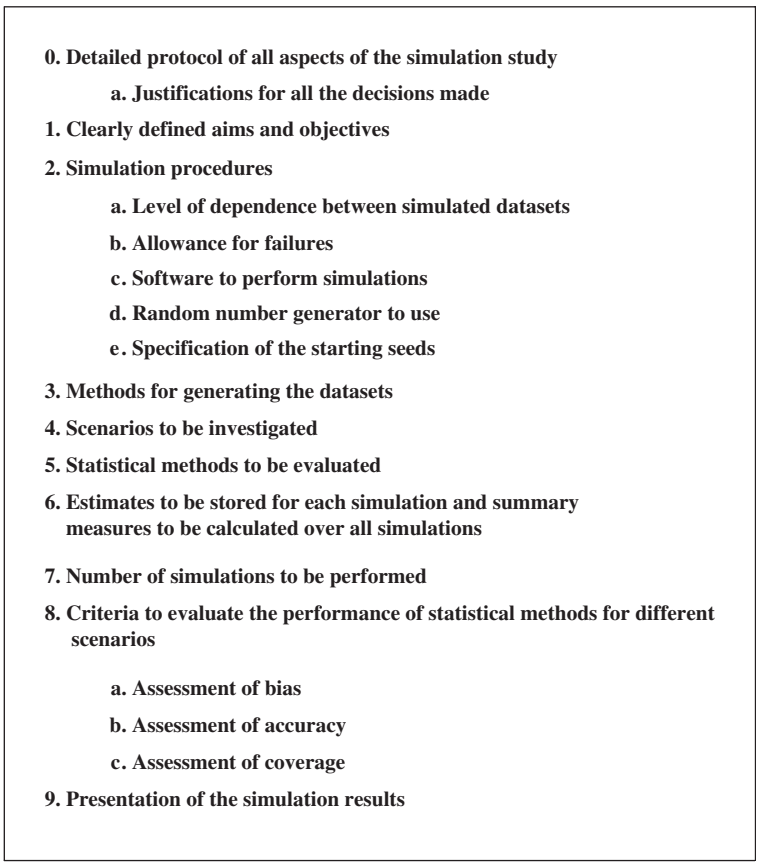


Figure 1. Important considerations when designing any simulation study.

2.2. Simulation procedures

Once the aims and objectives have been formalized, the procedures for performing the simulations can be considered including the level of dependence between simulations, the allowance for failures, the choice of random number generator, starting seeds and the software package to be used. The statistical software package must be able to handle the complexities involved in the proposed simulation study and have a reliable random number generator.

All simulation studies involve generating several independent simulated data sets. These generated data sets should also be completely independent for the different scenarios considered, such as different sample sizes. However, when more than one statistical methodology is being investigated, there is an added complication of determining the level of dependence of the simulated data sets for the different methods, although still retaining independent data sets for each scenario studied. Two feasible simulation strategies are possible. Firstly, fully independent simulated data sets involve generating a completely different set of independent data sets for each method and scenario considered. Secondly, moderately independent simulations use the same set of simulated

DOI: 10.1002/sim

independent data sets to compare a variety of statistical methods for the same scenario, but a different set of data sets is generated for each scenario investigated. These moderately dependent samples are like a matched pair design where the within sample variability is eliminated and therefore are sensitive to detecting any differences between methods. The relationship between the generated samples should form an important consideration when designing the study.

The simulation procedures should have some allowance for failing to estimate the outcome or parameter of interest, e.g. due to rare events or lack of convergence of models, to avoid premature stopping of the study. The simulations can be set up so that a failed sample is discarded and the whole process is repeated. The number of failures that occur should be recorded to gauge how likely this could happen in practice in order to judge whether the applied statistical procedure can reliably be used in the situation being investigated. If many failures occur for a particular scenario causing the early termination of the simulation study, researchers must consider whether in their situation the failures would lead to bias, and hence unacceptable results, or unbiased but imprecise results in order to determine the usefulness of the results from the partial set of completed simulations. Failures for some simulations may result in a *post hoc* change of the protocol to omit scenarios, which cannot be simulated reliably.

2.2.1. Random number generation. A fundamental part of any simulation study is the ability to generate random numbers. The many different types of random number generator have been detailed elsewhere [11, 12]. Any random number generator should be long in sequence before repetition and subsets of the random number sequence should be independent of each other [13]. A variety of statistical tests for randomness exist, including Marsaglia's Diehard battery of tests for randomness [14], which each random number generator must pass before it can be reliably adopted as a means of generating random numbers.

A random number generator must be able to reproduce the identical set of random numbers when the same starting value, known as a seed, is specified [13]. This is also essential when performing simulation studies to enable the generated data sets and hence results to be reproduced, if necessary, for monitoring purposes. The specification of the starting seed also facilitates the choice of simulation strategy. The simulations will be fully independent if completely different starting seeds are used to generate the data sets for each scenario and method combination considered or moderately independent if the same starting seeds are used to compare various methods for the same scenario but different seeds are employed for alternative scenarios. Any simulation strategy involves running several independent simulations for the same scenario, known as parallel simulations, which require independent sequences of random numbers. Random numbers can be generated for parallel simulations by setting different starting values for the individual simulations that are greater than the number of random numbers required for each simulation, which reduces the possibility of correlations between samples [13]. For example, if each simulated data set had a sample size of 500, then each of the 250, say, simulations would require 500 random numbers, therefore the starting seed for each simulation should be separated by at least 500.

2.3. Methods for generating the data sets

The methods for obtaining simulated data sets should be carefully considered and a thorough description provided in both the protocol and any subsequent articles published. Simulating data sets requires an assumed distribution for the data and full specification of the required parameters.

DOI: 10.1002/sim

The simulated data sets should have some resemblance to reality for the results to be generalizable to real situations and have any credibility. A good approach is to use a real data set as the motivating example and hence the data can be simulated to closely represent the structure of this real data set. The actual observed covariate data could be used and only the outcome data generated or just certain aspects, such as the covariate correlation structure, could be borrowed. Alternatively, the specifications could be arbitrary, but the generated data set may be criticized for not resembling realistic situations. The rationale for any choices made regarding the distributions of the data, parameters of any statistical models and the covariate correlation structure used to generate the data set should accompany their specifications. The generated data should be verified to ensure they resemble what is being simulated, for example using summary measures for the covariate distributions, Kaplan–Meier survival curves for survival data or fitting appropriate regression models.

2.3.1. Univariate data. Simple situations may involve generating a vector of random numbers sampled from a known distribution. Demirtas [15] provides procedures for obtaining a variety of univariate distributions from initial values generated from the uniform distribution, if the required distribution is unavailable within the statistical package.

2.3.2. Multivariate data. Generating multivariate data involves the additional specification of correlations between covariates unless the covariates are assumed fully independent, which is unlikely in practice. The specification of the means and associated covariance matrix is more straightforward if based on real data, especially with a large number of covariates, and the generated data will reflect reality. Conversely, the choice of the correlations between covariates can be arbitrary but it is often problematic to determine what are valid relationships. The simplest approach to generate multivariate covariate data with a specified mean and correlation structure is to assume a multivariate normal distribution. Any continuous but non-normally distributed variables in the real data should be transformed to make the assumption of normality more appropriate. Binary variables can be generated as latent normal, i.e. generated as continuous variables and then dichotomized, but the covariate correlation structure used to generate the continuous variable needs to be adjusted to provide the correct correlation with the resulting binary variable [16]. For example, the correction factor for a continuous variable that is to be dichotomized with a 50:50 split is 0.80, suggesting that the correlation between a continuous variable and a binary variable is 20 per cent less than the correlation between two continuous variables [16].

2.3.3. Time to event data. When the outcome is time to an event, such as in prognostic modelling, several additional considerations must be addressed. The simulations require the specification of a model for the multivariate covariate data and a distribution for the survival data, which may be censored. In order to simulate censored survival data, two survival distributions are required, one for the uncensored survival times that would be observed if the follow-up had been sufficiently long to reach the event and another representing the censoring mechanism.

The empirical survival distribution from a similar real data set would provide a reasonable choice for the survival distribution. The uncensored survival distribution could be generated to depend on a set of covariates with a specified relationship with survival, which represents the true prognostic importance of each covariate. Time-dependent covariates could also be simulated and incorporated following the procedures described by Mackenzie and Abrahamowicz [17]. Bender *et al.* [18] discuss the generation of survival times from a variety of survival distributions including

DOI: 10.1002/sim

the exponential for constant hazards, Weibull for monotone increasing or decreasing hazards and Gompertz for modelling human mortality, in particular for use with the Cox proportional hazards model.

Random non-informative right censoring with a specified proportion of censored observations can be generated in a similar manner to the uncensored survival times by assuming a particular distribution for the censoring times, such as an exponential, Weibull or uniform distribution but without including any covariates. Determining the parameters of the censoring distribution given the censoring probability is often achieved by iteration. However, Halabi and Singh [10] provide formulas for achieving this for standard survival and censoring distributions. The censoring mechanism can also be extended to incorporate dependent, informative censoring [19].

The survival times incorporating both events and censored observations are calculated for each case by combining the uncensored survival times and the censoring times. If the uncensored survival time for a case is less than or equal to the censored time, then the event is considered to be observed and the survival time equals the uncensored survival time, otherwise the event is considered censored and the survival time equals the censored time.

2.4. Scenarios to be investigated and methods for evaluation

Simulation studies usually examine the properties of one or more statistical methods in several scenarios defined by values of various factors such as sample size and proportion of censoring. These factors are generally examined in a fully factorial arrangement. The number of scenarios to be investigated and the methods for evaluation must be determined and justifications for these choices provided in the protocol. The scenarios investigated should aim to reflect the most common circumstances and if possible cover the range of plausible parameter values. The number of scenarios and statistical methods to investigate will depend on the study objectives but may be constrained by the amount of time available, the efficiency of the programming language and the speed and availability of several computers to run simulations simultaneously [20].

2.5. Estimates obtained from each simulation

It is essential to plan how the estimates will be stored after each simulation. Storing estimates enables consistency checks to be performed and allows for the identification of any errors or outlying values and the exploration of any trends and patterns within the individual simulations that may not be observed from the summary measure alone. Storing estimates also allows different ways of summarizing the estimates to be calculated retrospectively, if necessary, without the need to repeat all the simulations. A thorough consideration at the design stage of the possible estimates that may be of interest can ensure that all the required estimates are included, analysed and the results stored, and will avoid the risk of needing to repeat simulations. The estimate of interest, $\hat{\beta}_i$, could include the mean value of a variable, the parameter estimate after fitting a regression model, the log hazard ratio for survival models or the log odds ratios for logistic regression models. An associated within simulation standard error (SE) for the estimate of interest, $\mathrm{SE}(\hat{\beta}_i)$, is generally required.

It is also important to establish how to summarize these estimates once all simulations have been performed. Many published simulation studies report the average estimate of interest over the B simulations performed, e.g. $\bar{\hat{\beta}} = \sum_{i=1}^{B} \hat{\beta}_i / B$ as a measure of the true estimate of interest. Simulations are generally designed to mimic the results that could have been obtained from a

DOI: 10.1002/sim

single study and therefore an assessment of the uncertainty in the estimate of interest between simulations, denoted SE($\hat{\beta}$), is usually the empirical SE, calculated as the standard deviation of the estimates of interest from all simulations, $\sqrt{[1/(B-1)]\sum_{i=1}^{B}(\hat{\beta}_i - \bar{\hat{\beta}})^2}$. Alternatively, the average of the estimated within simulation SE for the estimate of interest $\sum_{i=1}^{B} \text{SE}(\hat{\beta}_i)/B$ could be used. The empirical SE should be close to the average of the estimated within simulation SE if the estimates are unbiased [21] and therefore, it may be sensible to consider both estimates of uncertainty. Alternatively, if using the mean and SE of the estimates over all simulations is not considered appropriate then non-parametric summary measures using quantiles of the distribution could be obtained.

2.6. Number of simulations required

The number of simulations to perform can be based on the accuracy of an estimate of interest, e.g. a regression coefficient, as with determining the sample size for any study [22, 23]. The number of simulations (B) can be calculated using the following equation:

$$B = \left(\frac{Z_{1-(\alpha/2)}\sigma}{\delta}\right)^2 \tag{1}$$

where δ is the specified level of accuracy of the estimate of interest you are willing to accept, i.e. the permissible difference from the true value β, $Z_{1-(\alpha/2)}$ is the $1-(\alpha/2)$ quantile of the standard normal distribution and σ^2 is the variance for the parameter of interest [22, 23]. A realistic estimate of the variance may be obtained from real data if the simulations are based on a real data set and are trying to maintain the same amount of variability. If the variance is unknown or cannot be estimated reliably then it may be possible to perform an identical simulation study to obtain realistic estimates for the variance or consider the measure of accuracy as a percentage of the SE. For example, if the variance from fitting a single covariate in a Cox regression model was 0.0166, then the number of simulations required to produce an estimate to within 5 per cent accuracy of the true coefficient of 0.349 with a 5 per cent significance level would be only 209. To estimate the regression coefficient to within 1 per cent of the true value would require 5236 simulations. Alternatively, the number of simulations could be determined based on the power $(1-\theta)$ to detect a specific difference from the true value as significant [22], such that

$$B = \left(\frac{(Z_{1-(\alpha/2)} + Z_{1-\theta})\sigma}{\delta}\right)^2$$

In fact, this formula is equivalent to equation (1) if the power to detect a specified difference is assumed to be 50 per cent.

The number of simulations to perform is thus dependent on the true value of the estimate of interest, the variability of the estimate of interest, and the required accuracy. For example, more simulations are needed if the regression coefficient is small or the estimate has little variability. Increasing the number of simulations will reduce the SE of the simulation process, i.e. $\text{SE}(\hat{\beta})/\sqrt{B}$, but this can be computational expensive and therefore variance reduction techniques could be employed [24]. The rationale for the number of simulations to perform should be included in the protocol.

DOI: 10.1002/sim

2.7. Evaluating the performance of statistical methods for different scenarios

After the simulations have been performed, the required estimates stored after each replication and summary measures calculated, it is necessary to consider the criteria for evaluating the performance of the obtained results from the different scenarios or statistical approaches being studied. The comparison of the simulated results with the true values used to simulate the data provides a measure of the performance and associated precision of the simulation process. Performance measures that are often used include an assessment of bias, accuracy and coverage. Collins *et al.* [4] emphasized the importance of examining more than one performance criterion such as mean square error (MSE), coverage and width of the confidence intervals in addition to bias, as results may vary across criteria. In general, the expectation of the simulated estimates is the main interest and hence the average of the estimates over all simulations is used to calculate accuracy measures, such as the bias. When judging the performance of different methods, there is a trade-off between the amount of bias and the variability. Some argue that having less bias is more crucial than producing a valid estimate of sampling variance [25]. However, methods that result in an unbiased estimate with large variability or conversely a biased estimate with little variability may be considered of little practical use. The most commonly used performance measures are considered in turn. Table I provides a summary of the most applicable performance measures and formulas.

Table I. Performance measures for evaluating different methods.

Evaluation criteria	Formula
BIAS	
Bias	$\delta = \bar{\hat{\beta}} - \beta$
Percentage bias	$\left(\frac{\bar{\hat{\beta}} - \beta}{\beta}\right) * 100$
Standardized bias	$\left(\frac{\bar{\hat{\beta}} - \beta}{\mathrm{SE}(\hat{\beta})}\right) * 100$
ACCURACY	
Mean square error	$(\bar{\hat{\beta}} - \beta)^2 + (\mathrm{SE}(\hat{\beta}))^2$
COVERAGE	Proportion of times the $100(1-\alpha)\%$ confidence interval $\hat{\beta}_i \pm Z_{1-\alpha/2}\mathrm{SE}(\hat{\beta}_i)$ include β, for $i = 1, \ldots, B$.
Average $100(1-\alpha)\%$ confidence interval length	$\frac{\sum_{i=1}^{B} 2Z_{1-\alpha/2}\mathrm{SE}(\hat{\beta}_i)}{B}$

Key: β is the true value for estimate of interest, $\bar{\hat{\beta}} = \sum_{i=1}^{B} \hat{\beta}_i / B$, B is the number of simulations performed, $\hat{\beta}_i$ is the estimate of interest within each of the $i = 1, \ldots, B$ simulations, $\mathrm{SE}(\hat{\beta})$ is the empirical SE of the estimate of interest over all simulations, $\mathrm{SE}(\hat{\beta}_i)$ is the SE of the estimate of interest within each simulation and $Z_{1-(\alpha/2)}$ is the $1-(\alpha/2)$ quantile of the standard normal distribution.

DOI: 10.1002/sim

2.7.1. Assessment of bias. The bias is the deviation in an estimate from the true quantity, which can indicate the performance of the methods being assessed. One assessment of bias is the difference between the average estimate and the true value, i.e. $\delta = \bar{\hat{\beta}} - \beta$ (Table I). The amount of bias that is considered troublesome has varied from $\frac{1}{2}\mathrm{SE}(\hat{\beta})$ [21] to $2\mathrm{SE}(\hat{\beta})$ [26]. Another approach is to calculate the bias as a percentage of the true value (Table I), providing the true value does not equal to zero. The percentage bias could have a detrimental effect on the results if the bias is greater than the amount specified when determining the number of simulations required. Alternatively, the bias as a percentage of the $\mathrm{SE}(\hat{\beta})$ (Table I) can be more informative, as the consequence of the bias depends on the size of the uncertainty in the parameter estimate [4]. A standardized bias of greater than 40 per cent in either direction has been shown to have noticeable adverse impact on the efficiency, coverage and error rates [4].

Testing the significance of the amount of bias in the estimates [27] or obtaining a 95 per cent confidence interval using the average parameter estimate, $\bar{\hat{\beta}}$, seem counterintuitive, since these statistics are based on the number of simulations through the $\mathrm{SE}(\bar{\hat{\beta}}) = \mathrm{SE}(\hat{\beta})/\sqrt{B}$ and hence these statistics can be improved or penalized by changing the number of simulations performed (see Section 2.6). Collins *et al.* [4] warned that with a large number of simulations, the bias may be deemed statistically significant but not be practically significant. Therefore do not rely solely on the p-value but consider the amount of bias as well.

2.7.2. Assessment of accuracy. The MSE provides a useful measure of the overall accuracy (Table I), as it incorporates both measures of bias and variability [4]. The square root of the MSE transforms the MSE back onto the same scale as the parameter [4].

2.7.3. Power, type I and II errors. The empirical power of a test, where relevant, can be determined as the proportion of simulation samples in which the null hypothesis of no effect is rejected at the nominal, usually 5 per cent, significance level, when the null hypothesis is false (e.g. References [3, 28]). Hence the empirical type II error rate is 1-power. The empirical type I error can be calculated as the proportion of p-values from testing the null hypothesis of no difference on each simulated sample that are less than the nominal 5 per cent significance level, when the null hypothesis is true (e.g. Reference [29]).

2.7.4. Assessment of coverage. The coverage of a confidence interval is the proportion of times that the obtained confidence interval contains the true specified parameter value (Table I). The coverage should be approximately equal to the nominal coverage rate, e.g. 95 per cent of samples for 95 per cent confidence intervals, to properly control the type I error rate for testing a null hypothesis of no effect [4]. Over-coverage, where the coverage rates are above 95 per cent, suggests that the results are too conservative as more simulations will not find a significant result when there is a true effect thus leading to a loss of statistical power with too many type II errors. In contrast, under-coverage, where the coverage rates are lower than 95 per cent, is unacceptable as it indicates over-confidence in the estimates since more simulations will incorrectly detect a significant result, which leads to higher than expected type I errors. A possible criterion for acceptability of the coverage is that the coverage should not fall outside of approximately two SEs of the nominal coverage probability (p), $\mathrm{SE}(p) = \sqrt{p(1-p)/B}$ [27]. For example, if 95 per cent confidence intervals are calculated using 1000 independent simulations then $\mathrm{SE}(\hat{p})$ is

DOI: 10.1002/sim

0.006892 and hence between 936 and 964 of the confidence intervals should include the true value.

The average length of the 95 per cent confidence interval for the parameter estimate $\hat{\beta}$ (Table I) is often considered as an evaluation tool in simulation studies (e.g. References [4, 30]). If the parameter estimates are relatively unbiased then narrower confidence intervals imply more precise estimates, suggesting gains in efficiency and power [30].

2.8. Presentation of the simulation results

Simulation studies can generate a substantial amount of results that need to be summarized and displayed in a clear and concise manner for the conclusions to be understood. The appropriate format is highly dependent on the nature of the information presented and hence there is a lack of a consistency in the literature. Structuring a report of any simulation study using separate subheadings for the objectives, methods, results and discussion provides clarity and can aid interpretation.

3. REVIEW OF CURRENT PRACTICE

A small formal review of articles published during 2004 in the Statistics in Medicine journal that included 'simulation' in the title, abstract or as a keyword was carried out to identify the current practices within published simulation studies. Of all 270 articles published in 2004, 58 (21 per cent) were identified as reporting a simulation study; their characteristics are summarized in Table II.

The specifics of the random number generator and the choice of starting seeds were generally omitted from the publications. Only one of the 58 articles explicitly stated the random number generator that was used; drand48 on the Unix/LINUX system [31]. Twenty-two articles gave some indication of the software package that was being used to generate the data or for the analysis, but it was unclear in the remaining 36 articles what statistical package was used to conduct the simulations. The relationship between generated samples was rarely stated within published simulation studies. Only one article stated that the simulations started with different seeds [32], whilst two other articles reported that independent samples were generated but did not explicitly mention anything about the starting seeds.

The number of simulations performed varied from 100 to 100 000 replications, with the most common choices being 1000 (19 articles) and 10 000 (12 articles) replications. It was unclear in four articles how many simulations were performed. Only six of these 58 articles provided any justification for the number of simulations performed. Three articles based their justifications on the expected SE given the number of simulations [33–35]. Two articles provided a justification in terms of the power to detect differences of a specified level from the true value as statistically significant [36, 37]. The last considered the chosen number of simulations to be sufficient, as they were not aiming to estimate any quantities with high accuracy [38].

The distributions and parameter specifications for generating the data were based on a real data set in eight of the simulation studies. In a further 16 articles, the simulated data intended to be typical of real data, although not explicitly based on a particular data set. The remaining 34 articles had no clear justification for the choices of parameters for the specified models used to generate the simulated data sets.

Generally the results from only a small proportion of the scenarios investigated were reported in an article, probably due to the limited space available. The choice of results to publish is fairly

DOI: 10.1002/sim

Table II. Summary of results from review of 58 articles.

Criteria	Frequency
Random number generator	
drand48 on the Unix/LINUX system	1
Not stated	57
Statistical Software used for analysis	
Splus	6
SAS	6
R	3
STATA	1
Mathematica	1
BUGS	1
MLWIN	1
MATLAB	1
Standalone package	2
Not stated	36
Dependence of samples/starting seed	
Samples independent	2
Different seeds used	1
Not stated	55
Number of simulations	
100	6
200	3
400	1
500	8
1000	19
5000	2
10 000	13
50 000	1
100 000	1
Unclear	4
Any justification for number of simulations	
Yes	6
No	52
Justification for data generation	
Based on a real data set	8
Typical of real data	16
Not stated	34

arbitrary and can depend on the important conclusions to be portrayed. However, one article has made available the full set of simulation results, which can be downloaded from a website specified in the article [3].

4. DISCUSSION

The advances in computer technology have allowed simulation studies to be more accessible. However, performing simulations is not simple. In any simulation study, many decision are required

DOI: 10.1002/sim

prior to the commencement of simulations, but there is generally no single correct answer. The choices made at each stage of the simulation process are open to criticism if not supplemented with thorough justifications.

Monte Carlo methods encompass any technique of statistical sampling employed to give approximate solutions to quantitative problems. They include, in addition to simulations, the Monte Carlo Markov chain methods such as Gibbs sampling, which are explicitly used for solving complicated integrals [39, 40]. This paper discusses simulation studies where data sets are formulated to imitate real data. Resampling studies [41, 42], where multiple data sets are sampled from a large real data set, require the same rigorous planning as simulation studies, differing from simulation studies only in terms of the generation of the data sets. Hence, similar considerations as discussed throughout this manuscript are relevant. Simulations are also useful in decision-making and engineering systems, where computer experiments are used to model dynamic processes in order to assess the effects over time and of varying any inputs (e.g. Reference [43]). Specific considerations for designing these studies in terms of formulating the problem, defining and designing the model and the choice of inputs and outputs have been discussed elsewhere (e.g. References [43, 44]).

This paper has discussed the important considerations when designing a simulation study. They include the choice of data to simulate and the procedures for generating the required data. Choices of distributions, parameters of any models, and covariate correlation structures used to generate the data set should be justified. Before commencing simulations, careful consideration should be given to the identification of the estimates of interest, the appropriate analysis, the methods for comparison, the criteria for evaluating these methods, the number of situations to consider, and the reporting of the results. In addition, every simulation study should have a detailed protocol, documenting the specific objectives and providing full details of how the study will be performed, analysed and reported. Modifications of the simulation processes, such as altering the number of simulations or collecting additional parameters or choices of scenarios, as a consequence of emerging data are possible, but can be time-consuming if they require all simulations to be rerun. Therefore, thorough planning at the start of any simulation study can ensure that the simulations are performed efficiently and only the necessary criteria and scenarios assessed. This paper has provided a concise reference (Figure 1) for researchers to follow when designing simulation studies.

A small review of published articles in one journal has suggested that the majority of simulation studies reported in the literature are not providing sufficient details of the simulation process to allow exact replication or clear justifications for the choices made. Future published simulation studies should include details of all the simulation procedures to enable the results to be reproduced. Using separate subheadings for the objectives, methods, results and discussion, irrespective of whether it is the main focus of the article, as in Reference [33], provides clarity and can aid interpretation. In addition, encouraging researcher to consider the suggested criteria (Figure 1) might encourage more sound and reliable simulation studies to be performed and reported with credible results.

ACKNOWLEDGEMENT

Andrea Burton was supported by a Cancer Research U.K. project grant.

REFERENCES

1. De Angelis D, Young GA. Bootstrap method. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: New York, 1998; 426–433.

DOI: 10.1002/sim

2. Kristman V, Manno M, Cote P. Loss to follow-up in cohort studies: how much is too much? *European Journal of Epidemiology* 2004; **19**:751–760.
3. Vaeth M, Skovlund E. A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Statistics in Medicine* 2004; **23**:1781–1792.
4. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 2001; **6**:330–351.
5. Mooney C. Conveying truth with the artificial: using simulated data to teach statistics in social sciences. *SocInfo Journal* 1995; **1**(Part 7):1–5.
6. Hodgson T, Burke M. On simulation and the teaching of statistics. *Teaching Statistics* 2000; **22**:91–96.
7. Morgan BJT. *Elements of Simulation*. Chapman & Hall: London, U.K., 1984.
8. Ripley BD. *Stochastic Simulation*. Wiley: New York, 1987.
9. Whitehead J, Zhou YH, Stevens J, Blakey G, Price J, Leadbetter J. Bayesian decision procedures for dose-escalation based on evidence of undesirable events and therapeutic benefit. *Statistics in Medicine* 2006; **25**:37–53.
10. Halabi S, Singh B. Sample size determination for comparing several survival curves with unequal allocations. *Statistics in Medicine* 2004; **23**:1793–1815.
11. L'Ecuyer P. Random number generation. In *Handbook of Computational Statistics*, Gentle JE, Haerdle W, Mori Y (eds). Springer-Verlag: New York, 2004; 35–70.
12. Marsaglia G. Random number generators. *Journal of Modern Applied Statistical Methods* 2003; **2**:2–13.
13. Masuda N, Zimmermann F. PRNGlib: a parallel random number generator library. *Technical Report*: *TR-96-08*, Swiss Center for Scientific Computing, Switzerland, 1996.
14. Marsaglia G. *The Marsaglia Random Number CDROM with the DIEHARD Battery of Tests of Randomness.* Florida State University: Florida, U.S.A., 1995.
15. Demirtas H. Pseudo-random number generation in R for some univariate distributions. *Journal of Modern Applied Statistical Methods* 2005; **3**:300–311.
16. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. *Psychological Methods* 2002; **7**:19–40.
17. Mackenzie T, Abrahamowicz M. Marginal and hazard ratio specific random data generation: applications to semi-parametric bootstrapping. *Statistics and Computing* 2002; **12**:245–252.
18. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; **24**:1713–1723.
19. Miloslavsky M, Keles S, van der Laan MJ, Butler S. Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society Series B—Statistical Methodology* 2004; **66**:239–257.
20. Sevcikova H. Statistical simulations on parallel computers. *Journal of Computational and Graphical Statistics* 2004; **13**:886–906.
21. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological Methods* 2002; **7**:147–177.
22. Lachin JM. Sample size determination. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: New York, 1998; 4693–4704.
23. Diaz-Emparanza I. Is a small Monte Carlo analysis a good analysis? Checking the size, power and consistency of a simulation-based test. *Statistical Papers* 2002; **43**:567–577.
24. Rubinstein RY. *Simulation and the Monte Carlo Method*. Wiley: New York, 1981.
25. Little RJA, Rubin DB. *Statistical Analysis with Missing Data* (2nd edn). Wiley: New York, 2002.
26. Sinharay S, Stern HS, Russell D. The use of multiple imputation for the analysis of missing data. *Psychological Methods* 2001; **6**:317–329.
27. Tang LQ, Song JW, Belin TR, Unutzer J. A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine* 2005; **24**:2111–2128.
28. Leffondré K, Abrahamowicz M, Siemiatycki J. Evaluation of Cox's model and logistic regression for matched case-control data with time-dependent covariates: a simulation study. *Statistics in Medicine* 2003; **22**:3781–3794.
29. Rempala GA, Looney SW. Asymptotic properties of a two sample randomized test for partially dependent data. *Journal of Statistical Planning and Inference* 2006; **136**:68–89.
30. Chen HY, Little RJA. Proportional hazards regression with missing covariates. *Journal of the American Statistical Association* 1999; **94**:896–908.
31. Kaiser JC, Heidenreich WF. Comparing regression methods for the two-stage clonal expansion model of carcinogenesis. *Statistics in Medicine* 2004; **23**:3333–3350.
32. Kenna LA, Sheiner LB. Estimating treatment effect in the presence of non-compliance measured with error: precision and robustness of data analysis methods. *Statistics in Medicine* 2004; **23**:3561–3580.

DOI: 10.1002/sim

33. Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine* 2004; **23**:1663–1682.
34. Chen YHJ, Demets DL, Lan KKG. Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine* 2004; **23**:1023–1038.
35. Song JW, Belin TR. Imputation for incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in Medicine* 2004; **23**:2827–2843.
36. Austin PC, Brunner LJ. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine* 2004; **23**:1159–1178.
37. Klar N, Darlington G. Methods for modelling change in cluster randomization trials. *Statistics in Medicine* 2004; **23**:2341–2357.
38. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Statistics in Medicine* 2004; **23**:723–748.
39. Gilks WR, Richardson S, Spiegelhalter DJ. Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*, Gilks WR, Richardson S, Spiegelhalter DJ (eds). Chapman & Hall: London, U.K., 1996; 1–19.
40. Robert CP, Casella G. *Monte Carlo Statistical Methods*. Springer-Verlag: New York, 2004.
41. Lunneborg CE. *Data Analysis by Resampling—Concepts and Applications.* Duxborg: Australia, 2000.
42. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993.
43. Santner TJ, Williams BJ, Notz WI. *The Design and Analysis of Computer Experiments*. Springer-Verlag: New York, 2003.
44. Balci O. Guidelines for successful simulation studies. In *Proceedings of the 1990 Winter Simulation Conference*, Balci O, Sadowski RP, Nance RE (eds). IEEE: Piscataway, NJ, 1990; 25–32.

DOI: 10.1002/sim

Chapter 6
Writing a Paper

The title of a book is its special tag, its distinguishing label. The choice of words is limited only by the number of words in the English language, yet how few of them show up in titles.

— ELSIE MYERS STAINTON, *A Bag for Editors* (1977)

Go straight to the point, rather than begin with an historical reference or resounding banality ('There is much research interest at present in the biochemistry of memory').

— BERNARD DIXON, *Sciwrite* (1973)

To write a reference, you must have the work you're referring to in front of you. Do not rely on your memory. Do not rely on your memory. Just in case the idea ever occurred to you, do not rely on your memory.

— MARY-CLAIRE VAN LEUNEN, *A Handbook for Scholars* (1992)

It was said of Jordan's writings that if he had 4 things on the same footing as (a, b, c, d) *they would appear as* a, M_3', ϵ_2, $\Pi_{1,2}''$.

— J. E. LITTLEWOOD, *Littlewood's Miscellany*[9] (1986)

May all writers learn the art (it is not easy) of preparing an abstract containing the essential information *in their compositions.*

— KENNETH K. LANDES, *A Scrutiny of the Abstract. II* (1966)

[9] See [34].

We write scientific papers to communicate our ideas and discoveries. Our papers compete for the readers' attention in journals, conference proceedings and other outlets for scholarly writing. If we can produce well-organized papers that are expressed in clear, concise English, and that convey our enthusiasm and are accessible to people outside our particular speciality, then our papers stand a better than average chance of being read and being referenced. It is generally accepted that the standard of scientific writing is not high[10] [63], [71], [191], [299], so a well-written paper will stand out from the crowd.

In this chapter I examine issues specific to writing a paper, as opposed to the general principles discussed in the previous chapters. Much of the chapter is applicable to writing a thesis (see also Chapter 9), book, or review. After considering the general issues of audience and organization I explore the building blocks of a research paper, from the title to the reference list.

6.1. Audience

Your first task in writing a paper is to determine the audience. You need to identify a typical reader and decide the breadth of your intended readership. Your paper might be written for a mathematics research journal, an undergraduate mathematics magazine, or a book for pre-university students. The formality of the prose and mathematical developments will need to be different in each of these cases. For research journals, at one extreme, you might be writing a very technical paper that builds upon earlier work in a difficult area, and you might be addressing yourself only to experts in that area. In this case it may not be necessary to give much motivation, to put the work in context, or to give a thorough summary and explanation of previous results. At the other extreme, as when you are writing a survey paper, you do need to motivate the topic, relate it to other areas, and explain and unify the work you are surveying. The requirements set forth in the "Guidelines for Authors" for the journal *SIAM Review* (prior to 1998) are even more specific:

> In their introductory sections, all papers must be accessible to the full breadth of SIAM's membership through the motivation, formulation, and exposition of basic ideas. The importance and intellectual excitement of the subject of the paper must

[10] In 1985, the editor of the British Medical Journal said "The 5000 or so articles we see at the BMJ every year are mostly dreadfully written, with numerous faults in English and overall construction" [180, p. 232].

> be plainly evident to the reader. Abstraction and specialization must illuminate rather than obscure. The primary threads of the intellectual fabric of the paper can never be hidden by jargon, notation, or technical detail.

The goals described in the last three sentences are worth striving for, whatever your audience.

The audience will determine the particular slant of your paper. A paper about Toeplitz matrices for engineers would normally phrase properties and results in terms of the physical problems in which these matrices arise, whereas for an audience of pure mathematicians it would be acceptable to consider the matrices in isolation from the application.

The language you use will depend on the audience. For example, whereas for linear algebraists I would write about the least squares problem $\min_x \|Ax - b\|_2$ with its least squares solution x, for statisticians I would translate this to the linear regression problem $\min_b \|Xb - y\|_2$ and the least squares estimate $\widehat{b}$ of the regression coefficients. Failing to use the notation accepted in a given field can cause confusion and can make your work impenetrable to the intended readership.

A good question to ask yourself is why a member of your intended audience should want to read your paper. If the paper is well focused you will find it easy to answer this question. If you cannot find an answer, consider altering the aims of the paper, or doing further work before continuing with the writing.

Whatever your audience, it is worth keeping in mind the words of Ivars Peterson, the editor of *Science News* [223]:

> The format of most journal papers seems to conspire against the broad communication of new mathematical ideas .... The titles, abstracts and introductions of many mathematical papers say: "Outsiders keep out! This is of interest only to those few already in the know."

With a little effort it should be possible to make your work at least partially understandable to non-experts.

6.2. Organization and Structure

At an early point in the writing of your paper you need to think about its high-level organization. It is a good idea to rank your contributions, to identify the most important. This will help you to decide where to put the emphasis and how to present the work, and it will also help you to write the title and abstract.

At the outset you should have some idea of the length of the paper. The larger it is the more important it is that it be well organized. However, it can be harder to write a good short paper than a good long one, for it is difficult to be simultaneously thorough, lucid and concise.

You need to decide in what order to treat topics and how to present results. You should aim to minimize the length by avoiding repetition; to obtain general results that provide others as special cases (even if the latter are more interesting); and to emphasize similarities and differences between separate analyses.

In addition to considering the first-time reader, you must think about the reader who returns to the paper some time after first studying it. This reader will want to skim through the paper to check particular details. Ideally, then, your paper should not only be easy to read through from beginning to end, but should also function as a reference document, with key definitions, equations and results clearly displayed and easy to find.

6.3. Title

According to Kerkut [150], for every person who reads the whole text of a scientific paper, five hundred read only the title. This statistic emphasizes the importance of the title. The title should give a terse description of the content, to help someone carefully scanning a journal contents page or a reference list decide whether to read the abstract or the paper itself. Ideally, it should also be catchy enough to attract the attention of a browser. Achieving a balance between these two aims is the key to writing an effective title. Kelley [147] mentions that he published an abstract with the title "A Decomposition of Compact Continua and Related Results on Fixed Sets under Continuous Mappings". After Paul Halmos suggested to him that it is not a good idea to put the whole paper into the title, he changed it to "Simple Links and Fixed Sets" for the published paper.

A note on the interpolation problem is too vague a title: what is the breakthrough heralded by *A note*, and which *interpolation problem* is under discussion? Similarly, *Approximation by cubic splines* is too vague (except for a thorough survey of the topic), since the approximation problem being addressed is not clear. Here are some examples of real titles, with my comments.

- Computing the eigenvalues and eigenvectors of symmetric arrowhead matrices [D. P. O'Leary and G. W. Stewart. *J. Comp. Phys.*, 90:497–505, 1990]. This is a lively and informative title. It is good to have action words in the title, such as *computing* or *estimating*.

- How and how not to check Gaussian quadrature formulae [Walter Gautschi. *BIT*, 23:209–216, 1983]. "How to" titles immediately arouse the reader's interest.

- Gaussian elimination is not optimal [V. Strassen. *Numer. Math.*, 13: 354–356, 1969]. If you can summarize your paper in a short sentence, that sentence may make an excellent title.

- How near is a stable matrix to an unstable matrix? [Charles F. Van Loan. In *Linear Algebra and Its Role in Systems Theory*, B. N. Datta, editor, volume 47 of *Contemporary Math.*, Amer. Math. Soc., 1985, pages 465–478]. A title that asks a question is direct and enticing.

- *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* [D. A. Belsley, E. Kuh, and R. E. Welsch. Wiley, New York, 1980]. This title has the classic form "general statement followed by colon and more specific information". Dillon [70] defines such a title to have "titular colonicity" and suggests that "To achieve scholarly publication, a research title should be divided by a colon into shorter and longer pre- and postcolonic clauses, respectively, the whole not to fall below a threshhold [sic] of 15–20 words minimum." This paper appears in the August (not the April) issue of *American Psychologist* and appears not to be a spoof.

- ALGOL 68 with fewer tears [Charles H. Lindsey. *Computer Journal*, 15:176–188, 1972]. This is an excellent title—it announces that the paper is about the programming language ALGOL 68 and whets the reader's appetite for a demystifying explanation. This paper has the distinction of being a syntactically valid ALGOL 68 program!

- Nineteen dubious ways to compute the exponential of a matrix [Cleve B. Moler and Charles F. Van Loan. *SIAM Review*, 20(4):801–836, 1978]. A classic paper in numerical analysis, with a memorable title.

- Can you count on your calculator? [W. Kahan and B. N. Parlett. Memorandum No. UCB/ERL M77/21, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, April 1977]. Puns rarely make their way into titles, but this one is effective.

- Performing armchair roundoff analyses of statistical algorithms [Webb Miller. *Comm. Statist. Simulation Comput.*, B7(3):243–255, 1978]. An otherwise drab title transformed by the word *armchair*. It suggests a gentle approach to the error analysis.

- Tricks or treats with the Hilbert matrix [Man-Duen Choi. *Amer. Math. Monthly*, 90:301–312, 1983]. This is another attractive and imaginative title.

- Can one hear the shape of a drum? [Marc Kac. *Amer. Math. Monthly*, 73(4, Part II):1–23, 1966]. The meaning of this attention-grabbing title is explained on the third page of Kac's expository paper:

> You can now see that the "drum" of my title is more like a tambourine (which is really a membrane) and that stripped of picturesque language the problem is whether we can determine Ω if we know all the eigenvalues of the eigenvalue problem
>
> $$\begin{aligned} \tfrac{1}{2}\nabla^2 U + \lambda U = 0 \quad &\text{in} \quad \Omega, \\ U = 0 \quad &\text{on} \quad \Gamma. \end{aligned}$$

  This paper won its author the Chauvenet Prize and the Lester R. Ford Award (see Appendix E). Kac's question is answered negatively in "One cannot hear the shape of a drum" [Carolyn Gordon, David L. Webb, and Scott Wolpert. *Bull. Amer. Math. Soc.*, 27(1):134–138, 1992], and his "hearing" terminology is used in "You can not hear the mass of a homology class" [Dennis DeTurck, Herman Gluck, Carolyn Gordon, and David Webb. *Comment. Math. Helvetici*, 64:589–617, 1989].

- The perfidious polynomial [James H. Wilkinson. In *Studies in Numerical Analysis*, G. H. Golub, editor, volume 24 of *Studies in Mathematics*, Mathematical Association of America, Washington, D.C., 1984, pages 1–28]. A delightful alliteration for a paper that explains why numerical computations with polynomials can be treacherous. This paper won its author the Chauvenet Prize (see Appendix E).

- Fingers or fists? (The choice of decimal or binary representation) [W. Buchholz. *Comm. ACM*, 2(12):3–11, 1959]. An analogy and an alliteration combine here to make an appealing title.

See Appendix E for many more examples of good titles.

A few years ago I submitted for publication a manuscript titled "Least Squares Approximation of a Symmetric Matrix". A referee objected to the title because it does not fully define the problem, so I changed it to "The Symmetric Procrustes Problem", which is even less informative if you are not familiar with Procrustes problems, but is perhaps more intriguing.

Unless the title is short it will have to be broken over two or more lines at the head of the paper (or on the front cover of a technical report). Rules of thumb are that a phrase should not be split between lines, a line should not start with a weak word such as a conjunction, and the lines should not differ too much in length. If the title is to be capitalized then all words except articles, short prepositions and conjunctions should be capitalized. Here are some examples, the first of each pair being preferable. The quotes at the beginning of each chapter provide further examples of the choice of line breaks.

On Real Matrices with
Positive Definite Symmetric Component

On Real Matrices with Positive
Definite Symmetric Component

An Iteration Method for the
Solution of the Eigenvalue Problem of
Linear Differential and Integral Operators

An Iteration Method for the Solution of
the Eigenvalue Problem of
Linear Differential and Integral Operators

Numerically Stable Parallel Algorithms
for Interpolation

Numerically Stable Parallel Algorithms for
Interpolation

In 1851 Sylvester published a paper with the title "Explanation of the Coincidence of a Theorem Given by Mr Sylvester in the December Number of This Journal, with One Stated by Professor Donkin in the June Number of the Same" [268]. Thankfully, titles are generally shorter nowadays.

6.4. Author List

In 1940, over 90% of papers reviewed in *Mathematical Reviews* had one author [120]. Today that figure is about 50%, showing that the proportion of jointly authored works in mathematics has increased greatly.

There are no hard and fast rules about the order in which the authors of a multiply authored paper are listed. Sometimes the person who did the greatest part of the work is listed first. Sometimes the academically senior person is listed first. In some disciplines and institutions the senior person

The Spotlight Factor

It is the custom in the theoretical computer science community to order authors alphabetically. In a tongue-in-cheek article, Tompa [273] defines the *spotlight factor* of the first author of a paper in which the k authors are listed alphabetically to be the probability that if $k-1$ coauthors are chosen independently at random they will all have surnames later in the alphabet than the first author. According to Tompa, the best (smallest) spotlight factor of 0.0255 in theoretical computer science belongs to Santoro, for his paper "Geometric containment and vector dominance" [Nicola Santoro, Jeffrey B. Sidney, Stuart J. Sidney, and Jorge Urrutia. *Theoretical Computer Science*, 53:345–352, 1987]. This value is calculated as

$$(1-.\text{santoro})^3 = \left(1 - \left(\frac{19}{27} + \frac{1}{27^2} + \frac{14}{27^3} + \frac{20}{27^4} + \frac{16}{27^5} + \frac{18}{27^6} + \frac{16}{27^7}\right)\right)^3,$$

where $a = 1, \ldots, z = 26$ and blanks or punctuation are represented by zero. By comparison, of those publications in the bibliography of this book, the best spotlight factors are the 0.1829 of O'Leary [211], the 0.2679 of Strunk [263] and the 0.3275 of Knuth [164].

(typically the director of a laboratory) is listed last. Perhaps most often, the authors are listed alphabetically, which is the practice I favour.

In their book *Computer Architecture* [137], Hennessy and Patterson adopt an unusual solution to the problem of deciding on author ordering: they vary the ordering in the book and in advertisements, even alternating the order when they reference the book! They comment that "This reflects the true collaborative nature of this book ... We could think of no fair way to reflect this genuine cooperation other than to hide in ambiguity—a practice that may help some authors but confuses librarians."

Being the first-named author is advantageous because the first name is easier to find in a reference list and the paper will be associated solely with that name in citations of the form "Smith et al. (1992)". Also, some citation services ignore all names after the first (see §14.3).

Make sure that you use precisely the same name on each of your publications. I declare myself as Nicholas J. Higham, but not N. J. Higham, N. Higham or Nick Higham. If you vary the name, your publications may not all be grouped together in bibliographic lists and indexes and there may be confusion over whether the different forms represent the same person.

Whether to spell out your first name(s) is a matter of personal preference. Chinese and Japanese authors need to decide whether to Westernize their name by putting the surname (family name) after the Christian (given) name, or to maintain the traditional ordering of surname first.

6.5. Date

Always date your work. If you give an unpublished paper to others they will want to know when it was written. You may not be able to distinguish different drafts if they are not dated. The date is usually placed on a line by itself after the author's name, or in a footnote. Spell the month out, rather than using the form "xx-yy-zz", since European and American authors interpret xx and yy in the opposite senses.

6.6. Abstract

The purpose of the abstract is to summarize the contents of the paper. It should do so in enough detail to enable the reader to decide whether to read the whole paper. The reader should not have to refer to the paper to understand the abstract.

Frequently, authors build an abstract from sentences in the first section of the paper. This is not advisable. The abstract is a mini-paper, and should be designed for its special purpose. This usually means writing the abstract from scratch once the paper is written. The shorter it is, the better, subject to the constraint of it being sufficiently informative. Most abstracts occupy one paragraph. Many mathematics journals state a maximum size for the abstract, usually between 200 and 300 words.

Some specific suggestions are as follows.

- Avoid mathematical equations in the abstract if possible, particularly displayed equations. One reason is that equations may cause difficulties for the review services that publish abstracts (though not those, such as *Mathematical Reviews*, that use TEX).

- Do not cite references by number in the abstract, since the list of references will not usually accompany the abstract in the review journals. If a paper must be mentioned, spell it out in full:

  An algorithm given by Boyd [*Linear Algebra and Appl.*, 9:95–101, 1974] is extended to mixed subordinate matrix norms.

- Try to make the abstract easy to understand for those whose first language is not English. Also, keep in mind that the abstract may

be translated into another language, for a foreign review journal, for example.

- Some journals disallow the word *we* in abstracts, preferring the passive voice (usually necessitating *it*). If you are writing for such a journal it pays to adopt the required voice; while a copy editor will convert as necessary, the conversion may lessen the impact of your sentences.
- Obviously, the abstract should not make claims that are not justified in the paper. Yet this does happen—possibly when the abstract is written before the paper is complete and is not properly revised.
- The abstract should give some indication of the conclusions of the paper. An abstract that ends "A numerical comparison of these methods is presented" and does not mention the findings of the comparison is uninformative.
- Your abstract should lay claim to some new results, unless the paper is a survey. Otherwise, if you submit the paper to a research journal, you are making it easy for a referee to recommend rejection.
- Try not to start the abstract with the common but unnecessary phrases "In this paper" or "This paper". Some journals make this request in their instructions to authors.

The suggestions above are particularly relevant for an abstract that is submitted to a conference and appears in a conference programme. Such an abstract will be read and judged in isolation from the paper, so it is vital for it to create a strong impression in isolation.

An intriguing opening paragraph of an abstract is the one by Knuth (1979) in [158]:

ABSTRACT. Mathematics books and journals do not look as beautiful as they used to. It is not that their mathematical content is unsatisfactory, rather that the old and well-developed traditions of typesetting have become too expensive. Fortunately, it now appears that mathematics itself can be used to solve this problem.

If inspiration fails you, you could always use the following generic abstract from [246]:

> ABSTRACT. After a crisp, cogent analysis of the problem, the author brilliantly cuts to the heart of the question with incisive simplifications. These soon reduce the original complex edifice to a [s]mouldering pile of dusty rubble.

6.7. Key Words and Subject Classifications

Some journals list key words supplied by the author, usually after the abstract. The number of key words is usually ten or less. Since the key words may be used in computer searches, you should try to anticipate words for which a reader might search and make them specific enough to give a good indication of the paper's content.

Some journals also require subject classifications. The AMS Mathematics Subject Classifications (1991) divide mathematics into 61 sections with numbers between 0 and 94, which are further divided into many subsections. For example, section 65 covers numerical analysis and has 106 subsections; subsection 65F05 covers direct methods for solving linear systems while subsection 65B10 covers summation of series. The classifications are listed in the *Annual Subject Index* of *Mathematical Reviews* and can be downloaded from the American Mathematical Society's e-MATH service (see §14.1).

Other classification schemes exist, such as the one for computer science from the journal *Computing Reviews*. The Computing Reviews Classification System (1987) is a four-level tree that has three numbered levels and an unnumbered level of descriptors. The top level consists of eleven nodes, denoted by letters A (General Literature) to K (Computing Milieux). An example of a category is

> G.1.3 [Numerical Analysis]: Numerical Linear Algebra—sparse and very large systems.

This specifies the Numerical Linear Algebra node of the Numerical Analysis area under G (Mathematics of Computation), and "sparse and very large systems" is one of several descriptors for G.1.3 listed in the definition of the classification scheme.

6.8. The Introduction

Perhaps the worst way to begin a paper is with a list of notation or definitions, such as *Let G be an abelian group and H be a subgroup of G*, or *Let $\mathcal{F}$ be the complex field $\mathcal{C}$ or the real field $\mathcal{R}$, and let $\mathcal{F}_{m\times n}$ be the linear space of all $m \times n$ matrices over $\mathcal{F}$. If $A \in \mathcal{F}_{m\times n}$ we use A^* to denote the conjugate transpose of A.* It can be argued that the first sentence should be the best in the paper—its job is to entice the uncommitted reader into reading the whole paper. A list of notation will not achieve this aim, but a clear, crisp and imaginative statement may. King [152] gives a slightly different specification for the first sentence: "The first sentence has a dual

function: it must carry some essential information, particularly the problem under consideration, and at the same time gently translate the reader into the body of the article."

Ideally, an introduction is fairly short—say a few hundred words. It should define the problem, explain what the work attempts to do, and outline the plan of attack. Unless there is good reason not to do so, it is advisable to summarize the results achieved. Knowing the problem and the progress made on it, the reader can decide after reading the introduction whether to read the whole paper. You may want to leave the punch-line to the end, but in doing so you risk the reader losing interest before reaching it.

Here is an example of a compelling opening paragraph that strongly motivates the work. It is from Gautschi's paper "How and How Not to Check Gaussian Quadrature Formulae", mentioned on page 81.

> The preparation of this note was prompted by the appearance, in the chemistry literature, of a 16-digit table of a Gaussian quadrature formula for integration with measure $d\lambda(t) = \exp(-t^3/3)\,dt$ on $(0, \infty)$, a table, which we suspected is accurate to only 1–2 decimal digits. How does one go about convincing a chemist, or anybody else for that matter, that his Gaussian quadrature formula is seriously defective?

It can be appropriate to begin with a few definitions if they are needed to state the problem and begin the analysis. As an example, here is the beginning of the paper "Estimating the Largest Eigenvalue of a Positive Definite Matrix" by O'Leary, Stewart and Vandergraft [211].

> Let A be a positive definite matrix of order n with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ corresponding to the orthonormal system of eigenvectors $x_1, x_2, \ldots, x_n$. In some applications, one must obtain an estimate of λ_1 without going to the expense of computing the complete eigensystem of A. A simple technique that is applicable to a variety of problems is the power method.

In the next four sentences the authors define the power method, state that the theory of the method is well understood, and note that convergence of the method can be hindered in two ways, which are then analysed. This introduction is effective because it defines the problem and motivates the analysis without a wasted word, and leads quickly into the heart of the paper. The introduction is not labelled as such: this is a three and a half page paper with no section headings.

A possible way to improve an introduction is to delete the first one or more sentences, which are often unimportant general statements. Try it! As an example, consider the following opening two sentences.

> Polynomials are widely used as approximating functions in many areas of mathematics and they can be expressed in various bases. We consider here how to choose the basis to minimize the error of evaluation in floating point arithmetic.

This opening is not a bad one, but the first sentence is general and unexciting. Under the reasonable assumption that the reader knows the importance of polynomials and need not be reminded, it is better to combine and shorten the two sentences and pose a question:

> In which basis should we express a polynomial to minimize the error of evaluation in floating point arithmetic?

Unless the paper is very short it is advisable to outline its organization towards the end of the introduction. One approach is to write, for each section, a sentence describing its contents. The outline can be introduced by a sentence such as "The outline of this paper is as follows" or "This paper is organized as follows." It is best not simply to list the section titles; instead, give a summary that could only be obtained by reading the sections.

Note that some journals prefer the symbol § to the word section when referring to specific sections: §2.1 is written instead of Section 2.1. The plural of § is §§: "see §§2.1 and 2.2."

6.9. Computational Experiments

Many papers describe computational experiments. These may be done for several reasons: to gain insight into a method, to compare competing methods, to verify theoretical predictions, to tune parameters in algorithms and codes, and to measure the performance of software. To achieve these aims, experiments must be carefully designed and executed. Many decisions have to be made, ranging from what will be the objectives of the experiments to how to measure performance and what test problems to use. One editor of a numerical analysis journal commented that his primary reason for rejecting papers is that the computational experiments are unsound; this underlines the importance of proper design and reporting of experiments.

When you report the results of a computational experiment you should give enough detail to enable the results to be interpreted and the experiment to be reproduced. In particular, where relevant you should state the machine precision (working accuracy) and the type of random numbers used (e.g., normal $(0, 1)$ or uniform $(-1, 1)$). If you wish to display the convergence of a sequence it is usually better to tabulate the errors

rather than the values themselves. Error measures should be normalized. Thus, for an approximate solution $\widehat{x}$ to $Ax = b$, the relative residual $\|b - A\widehat{x}\|/(\|A\|\|\widehat{x}\| + \|b\|)$ is more meaningful than the scale-dependent quantity $\|b - A\widehat{x}\|$. If you are measuring the speed of a numerical algorithm it is important to show that the right answers are being produced (otherwise the algorithm "the answer is 42" is hard to beat).

You may also wish to state the programming language, the version of the compiler used, and the compiler options and optimizations that were selected, as these can all have a significant influence on run-times. In areas where attaining full precision is not the aim (such as the numerical solution of ordinary differential equations), it is appropriate to consider both speed and accuracy (e.g., by plotting cost versus requested accuracy).

One of the difficulties in designing experiments is finding good test problems—ones which reveal extremes of behaviour, cover a wide range of difficulty, are representative of practical problems, and (ideally) have known solutions. In many areas of computational mathematics good test problems have been identified, and several collections of such problems have been published. For example, collections are available in the areas of nonlinear optimization, linear programming, ordinary differential equations and partial differential equations. Several collections of test matrices are available and there is a book devoted entirely to test matrices. For references to test problem collections see [140].

In your conclusions you should make a clear distinction between objective statements and opinions and speculation. It is very tempting to extrapolate from results, but this is dangerous. As you analyse the results you may begin to formulate conclusions that are not fully supported by the data, perhaps because they were not anticipated when the experiments were designed. If so, further experimentation will be needed. When evaluating numerical algorithms I have found that it pays to print out every statistic that could conceivably be of interest; if I decide not to print out a residual or relative error, for example, I often find a need for it later on.

Further guidelines on how to report the results of computational experiments can be found in the journals *ACM Transactions on Mathematical Software* [65] and *Mathematical Programming* [146], and in an article by Bailey [12]. Perhaps the best advice is to read critically the presentation of experimental results in papers in your area of interest—and learn from them.

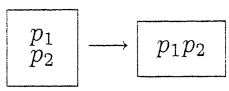
6.10. Tables

Many of the principles that apply to writing apply also to the construction of tables and graphs. However, some particular points should be considered

when designing a table.

To maximize readability the table design should be as simple as possible. Repetition should be avoided; for example, units of measurement or descriptions common to each entry in a column should go in the column header. Compare Tables 6.1 and 6.2. It is best to minimize the number of rules in the table. Two busy examples are shown in Table 6.3 and Table 6.5. The simplified versions, Table 6.4 and Table 6.6, are surely more aesthetically pleasing. Table 6.3 is taken from [159, p. 366],[11] where it is given without any rules and is still perfectly readable.

It is easier to compare like quantities if they are arranged in columns rather than rows. Research reported by Hartley [132], [134] supports this fact, and Tables 6.7 and 6.8 provide illustration, Table 6.8 being the easier to read. The difference between row and column orientation is more pronounced in complex tables. Of course, the orientation may be determined by space considerations, as a horizontal orientation usually takes less space on the page. If a vertical table is too tall, but is narrow, it can be broken into two pieces side by side:



It is also helpful to put columns or rows that need to be compared next to each other.

Only essential information should be included in a table. Omit data whose presence cannot be justified and state only as many digits as are needed (this number is often surprisingly small). In particular, do not state numerical results to more significant figures than are known for the data. As an example, in Table 6.1 there is no need to quote the timings and speedups to six significant figures, so Table 6.2 gives just one decimal place. Note that there is justification for showing so many digits in Tables 6.3 and 6.4: $\pi(x)$ in this table is the number of primes less than or equal to x, and nearly all the digits of $\pi(10^9)$ are needed to show the error in Riemann's formula. Displaying the first one or two digits of the fractional parts of the approximations emphasizes that the approximations are not integers.

If you need to present a large amount of data in tabular form, consider displaying it in an appendix, to avoid cluttering the main text. You could give smaller tables in the text that summarize the data. Large sets of data are often better displayed as graphs, however, particularly if it is the trends rather than the numerical values that are of interest. Tufte [275,

[11] Donald E. Knuth, *The Art of Computer Programming*, vol. 2, ©1981 by Addison-Wesley Publishing Co. Reprinted by permission of Addison-Wesley Publishing Co., Inc., Reading, MA. The form of the original table is not reproduced exactly here.

Table 6.1. Timings for a parallel algorithm.

# processors	Time	Speedup
$p = 1$	28.352197 secs	—
$p = 4$	7.218812 secs	3.9275
$p = 8$	3.634951 secs	7.7999
$p = 16$	1.929347 secs	14.6952

Table 6.2. Timings for a parallel algorithm.

No. of processors	Time (secs)	Speedup
1	28.4	—
4	7.2	4.0
8	3.6	7.8
16	1.9	14.7

Table 6.3. Approximations to $\pi(x)$.

x	$\pi(x)$	$x/\ln x$	$L(x)$	Riemann's formula
10^3	168	144.8	176.6	168.36
10^6	78498	72382.4	78626.5	78527.40
10^9	50847534	48254942.4	50849233.9	50847455.43

Table 6.4. Approximations to $\pi(x)$.

x	$\pi(x)$	$x/\ln x$	$L(x)$	Riemann's formula
10^3	168	144.8	176.6	168.36
10^6	78498	72382.4	78626.5	78527.40
10^9	50847534	48254942.4	50849233.9	50847455.43

Table 6.5. Results for inverting a lower triangular matrix on a Cray 2.

	Mflops			
n	128	256	512	1024
Method 1 ($n_b = 1$)	95	162	231	283
Method 2 ($n_b = 1$)	114	211	289	330
k variant ($n_b = 1$)	114	157	178	191
Method 1B ($n_b = 64$)	125	246	348	405
Method 2C ($n_b = 64$)	129	269	378	428
k variant ($n_b = 64$)	148	263	344	383

Table 6.6. Mflop rates for inverting a lower triangular matrix on a Cray 2.

	n	128	256	512	1024
Unblocked:	Method 1	95	162	231	283
	Method 2	114	211	289	330
	k variant	114	157	178	191
Blocked:	Method 1B	125	246	348	405
($n_b = 64$)	Method 2C	129	269	378	428
	k variant	148	263	344	383

Table 6.7. SI prefixes (10^{-1}–10^{12}). Row orientation.

Multiple	10^{12}	10^{9}	10^{6}	10^{3}	10^{-1}
Prefix	tera	giga	mega	kilo	deci
Symbol	T	G	M	K	d

Table 6.8. SI prefixes (10^{-1}–10^{12}). Column orientation.

Multiple	Prefix	Symbol
10^{12}	tera	T
10^{9}	giga	G
10^{6}	mega	M
10^{3}	kilo	K
10^{-1}	deci	d

p. 56] advises that tables are usually better than graphs at reporting small data sets of twenty numbers or less.

The caption should be informative and should not merely repeat information contained in the table. Notice the simplification obtained by moving the word "Mflops" from the table to the caption in Table 6.6.

Give a clear reference to the table at an appropriate place in the text—you cannot rely on the reader to refer to the table automatically. It is helpful if you explain the salient features of the table in words. The reader will appreciate this guidance, especially if the table contains a lot of data. However, you should not summarize the whole table—if you do, the table might as well be omitted.

Further Reading

The Chicago Manual of Style [58] devotes a whole chapter to tables and offers much useful advice. Another good reference is *A Manual for Writers* [278, Chap. 6]. Bentley [21, Chap. 10] gives a good example of how to redesign a table. References that discuss the preparation of graphs include Bentley [21, Chaps. 10, 11], Hartley [132], [134], MacGregor [187], [188], and Tufte [275], [276], [277].

6.11. Citations

The two main styles of citation in mathematics journals are by number (as used in this book) and by name and year, which is the Harvard system. Examples of the Harvard system are *These results agree with an existing study of Smith* (1990) and *These results agree with an existing study* (*Smith,* 1990). If more than one paper maps to Smith (1990), the papers are distinguished by appending a letter to the year: Smith (1990a), Smith (1990b), and so on. In the number-only system, the number is usually placed in square brackets, though some styles require it to be superscripted.

The main requirement is that a citation does not intrude upon a sentence. For example, *This method was found* [17] *to be unstable* is better written as *This method was found to be unstable* [17]. There are circumstances, however, where a citation has to be placed part-way through a sentence to convey the correct meaning. A good test for whether a citation is well placed is to see whether the sentence reads properly when the citation is deleted. The style of citation inevitably affects how you phrase sentences, so it is worth checking in advance what style is used by the journal in which you wish to publish. Knuth [164] explains that when his paper "Structured programming with go to statements" [*Computing Surveys*, 6:261–301, 1974] was reprinted in a book, he made numerous changes

to make sentences read well with the citation style used in the book.

When you cite by number, it is good style to incorporate the author's name if the citation is more than just a passing one. As well as saving the reader the trouble of turning to the reference list to find out who you are referring to, this practice has an enlivening effect because of the human interest it introduces. Examples:

> Let $A\Pi = QR$ be a QR factorization with column pivoting [10].
> (Passing reference to a textbook for this standard factorization.)
> The rate of convergence is quadratic, as shown by Wilkinson [27]. (Instead of "as shown in [27]".)

The sentence "This question has been addressed by [5]" is logically incorrect and should be modified to "addressed by Jones [5]" or "addressed in [5]".

When you cite several references together it is best to arrange them so that the citation numbers are in increasing order, e.g., "several variations have been developed [2], [7], [13]." Ordering by year of publication serves no purpose when only citation numbers appear in the text. If you want to emphasize the historical progression it is better to add names and years: "variations have been developed by Smith (1974) [13], Hall (1981) [2], and Jones (1985) [7]."

It is important to be aware that the reference list says a lot about a paper. It helps to define the area in which the paper lies and may be used by a reader to judge whether the author is aware of previous work in the area. Some readers look at the reference list immediately after reading the title, and if the references do not look sufficiently familiar, interesting or comprehensive they may decide not to read further. Therefore it is desirable that your reference list contain at least a few of the key papers in the area in which you are writing. Papers should not, however, be cited just for effect. Each citation should serve a purpose within the paper. Note also that if you cite too often (say, for several consecutive sentences) you may give the impression that you lack confidence in what you are saying.

There are several conventions for handling multiple authors using the Harvard system. One such convention is as follows [45], [68, Chap. 12]. For one or two authors, both names are given (e.g., "see Golub and Van Loan (1989)"). If there are three authors, all three are listed in the first citation and subsequent citations replace the second and third names by "et al." (e.g., "see Knuth, Larrabee and Roberts (1989)," then "see Knuth et al. (1989)"). For four or more authors, all citations use the first author with "et al." These conventions can also be used when naming authors in conjunction with the numbered citation system.

If you make significant use of a result from another reference you should give some indication of the difficulty and depth of the result (and give the

author's name). Otherwise, unless readers look up the reference, they will not be able to judge the weight of your contribution.

When you make reference to a specific detail from a book or long paper it helps the reader if your citation includes information that pinpoints the reference, such as a page, section, or theorem number.

For further details on the subtleties of citation consult van Leunen [283].

6.12. Conclusions

If there is a conclusions section (and not every paper needs one) it should not simply repeat earlier sections in the same words. It should offer another viewpoint, discuss limitations of the work, or give suggestions for further research. Often the conclusions are best worked into the introduction or the last section. It is not uncommon to see papers where the conclusions are largely sentences taken from the introduction, such as *We show that X's result can be extended to a larger class ...*; this practice is not recommended.

The conclusions section is a good place to mention further work: to outline open problems and directions for future research and to mention work in progress. Be wary of referring to your "forthcoming paper", for such papers can fail to materialize. A classic example of a justified reference to future work is the following quote from the famous paper[12] by Watson and Crick [290] (*Nature*, April 25, 1953) in which the double helix structure of DNA was proposed:

> It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.
>
> Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

The promise "will be published elsewhere" was fulfilled shortly afterwards, in the May 30, 1953 issue of *Nature*.[13]

6.13. Acknowledgements

Be sure to acknowledge any financial support for your work: grants, fellowships, studentships, sponsorship. A researcher might write "This re-

[12] For some comments by Crick on the writing of this and subsequent papers by Watson and Crick, see [64].

[13] Pyke [231] points out that the words "It has not escaped our notice that" can be removed.

search was supported by the National Science Foundation under grant DCR-1234567"; grant agencies like authors to be this specific. A Ph.D. student supported by the Engineering and Physical Sciences Research Council (UK) might write "This work was supported by an EPSRC Research Studentship." In SIAM journals this type of acknowledgement appears in a footnote on the first page. Other acknowledgements usually appear in a section titled *Acknowledgements* at the end of the paper. (An alternative spelling is *acknowledgments*.)

It is customary to thank anyone who read the manuscript in draft form and offered significant suggestions for improvement (as well as anyone who helped in the research), but not someone who was just doing his or her normal work in helping you (for example, a secretary). The often-used "I would like to thank" can be shortened to "I thank." Note that if you say "I thank X for pointing out an error in the proof of Theorem Y," you are saying that the proof is incorrect; "in an earlier proof" or "in an earlier attempted proof" is what you meant to say.

The concept of anonymous referee sometimes seems to confuse authors when they write acknowledgements. An anonymous referee should not be thanked, as is often the case, for *his* suggestions—it may be a *she*. One author wrote "I thank the anonymous referees, particularly Dr. J. R. Ockendon, for numerous suggestions and for the source of references." Another explained, not realising the two ways in which the sentence can be read, "I would like to thank the unknown referees for their valuable comments."

6.14. Appendix

An appendix contains information that is essential to the paper but does not fit comfortably into the body of the text. The most common use of an appendix is to present detailed analysis that would distract the reader if it were given at the point where the results of the analysis are needed. An appendix can also be used to give computer program listings or detailed numerical results. An appendix should not be used to squeeze inessential information into the paper (though this may be acceptable in a technical report or thesis).

6.15. Reference List

Preparing the reference list can be one of the most tedious aspects of writing a paper, although it is made much easier by appropriate software (see §13.3). The precise format in which references are presented varies among publishers and sometimes among different journals from the same publisher.

Here are four examples.

SIAM journals: J. H. WILKINSON, *Error analysis of floating-point computation*, Numer. Math., 2 (1960), pp. 319–340.

IMA journals: WILKINSON, J. H. 1960 Error analysis of floating-point computation. *Numer. Math.* **2**, 319–340.

Elsevier journals: J. H. Wilkinson, Error analysis of floating-point computation, *Numer. Math.* 2:319–340 (1960).

Springer-Verlag journals: Wilkinson, J. H. (1960): Error analysis of floating-point computation. Numer. Math. **2**, 319–340.

All journals that I am familiar with ask for the use of their own format but will accept other formats and copy edit them as necessary. All publishers have a minimum amount of information that they require for references, as defined in their instructions for authors. It is important to provide all the required information, whatever format you use for the references.

Here are some comments and suggestions on preparing reference lists. For further details I strongly recommend the book by van Leunen [283], but keep in mind that her recommendations may conflict with those of publishers in certain respects.

1. Do not rely on secondary sources to learn the contents of a reference or its bibliographic details—always check the original reference. In studies on the accuracy of citation, the percentage of references containing errors has been found to be as high as 50% [95]. A 1982 paper by Vieira and Messing in the journal *Gene* had been cited correctly 2,212 times up to 1988, but it had also been cited incorrectly 357 times under "Viera"; these errors led to the paper being placed too low in a list of most-cited papers [98]. In another well-documented case in the medical literature, the Czech title "O Úplavici" ("On Dysentery") of an 1887 paper in a Czech medical journal was taken by one writer to be the author's name, and the mistake propagated until it was finally exposed in 1938 [135], [239]. If a secondary source has to be used (perhaps because the reference is unavailable), it is advisable to append to the reference "cited in [ss]", where [ss] is the secondary source.

2. Always provide the full complement of initials of an author, as given in the paper or book you are referencing.

3. Some authors are inconsistent in the name they use in their papers, sometimes omitting a middle initial, for example. In such cases, my

preference is to use the author's full name in the reference list when it is known, to avoid ambiguities such as: Is A. Smith the same author as A. B. Smith?

4. Some sources contain typographical errors or nonstandard usage. Titles should be given unaltered. For example, the title "Van der Monde systems and numerical differentiation" [J. N. Lyness and C. B. Moler, *Numer. Math.*, 8:458–464, 1966] appears to be incorrect because the name is usually written Vandermonde, but it should not be altered (I have occasionally had to reinstate "Van der Monde" in my reference list after a copy editor has changed it). A typographical error in an author's name is rare, but not unknown. It seems reasonable to correct such an error, but to provide some indication of the correction that has been made, such as a note at the end of the reference.

5. Copy bibliographic information of a journal article from the journal pages, not the cover of the journal. The cover sometimes contains typographical errors and you cannot deduce the final page number of the article if the journal puts blank pages between articles or begins articles part-way down a page.

6. Electronic journals do not usually cause any difficulties in referencing, since it is in the journals' interests to make clear how papers should be referenced. For example, the journal *Electronic Transactions on Numerical Analysis* provides papers in PostScript form, and each paper has a clearly defined page range, volume and year; papers are therefore referenced just like those in a traditional journal. It may help readers if a URL for an electronic journal is appended to the reference, but the journal in which you are publishing may delete it to save space.

   It is more difficult to decide how to reference email messages and unpublished documents or programs on the Web. The following suggestions are adapted from those in *Electronic Styles* [178]. I assume that an email address and a URL are both clearly identifiable as such by the `@` and `http`, respectively, so I omit the descriptors "email" and "URL". There are so many different types of item on the Web that no referencing scheme can cover all possibilities.

   (a) A publication available in print and online.

   > Nicholas J. Higham. The Test Matrix Toolbox for MATLAB (version 3.0). Numerical Analysis Report No. 276, Manchester Centre for Computational Mathematics, Manch-

ester, England, Sept. 1995; also available from `ftp://ftp.ma.man.ac.uk/pub/narep/narep276.ps.gz`

(b) A publication available online only.

Melvin E. Page. A Brief Citation Guide for Internet Sources in History and the Humanities (Version 2.1), `http://h-net.msu.edu/~africa/citation.html`, 1996.

(c) A publication on CD-ROM.

A. G. Anderson, Immersed interface methods for the compressible equations. In Proceedings of the Eighth SIAM Conference on Parallel Processing for Scientific Computing (Minneapolis, MN, 1997), CD-ROM, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997.

(d) A piece of software.

Piet van Oostrum. LaTeX package `fancyhdr`. CTAN archive (e.g., `http://www.tex.ac.uk/tex-archive`), `macros/latex/contrib/supported/fancyhdr`.

(e) An item in a discussion list, digest or newsgroup.

David Hough. Random story. NA Digest, 89 (1), 1989. `na.help@na-net.ornl.gov`, `http://www.netlib.org/index.html`

(f) A standard email message. The title is taken from the `Subject:` line.

Desmond J. Higham (`aas96106@ccsun.strath.ac.uk`). Comments on your paper. Email message to Nicholas J. Higham (`higham@ma.man.ac.uk`), August 18, 1997.

(g) A forwarded email message.

Susan Ciambrano (`ciambran@siam.org`). Reader's comments on HWMS. (Original message A. Reader, Handbook of Writing.) Forwarded email message to Nicholas J. Higham (`higham@ma.man.ac.uk`), October 20, 1995.

7. When you reference a manuscript or technical report that is more than a few months old, check to see if it has appeared in a journal. An author will usually be happy to inform you of its status. If it has appeared, check whether the title has changed. The referees may have asked for a better title, or the copy editor may have added a hyphen, combined a hyphenated pair of words, or changed British spelling to American or vice versa.

8. Take care to respect letters and accented characters from other languages. Examples: Å, å, ß, é, ö, ø, Ø.

9. If you maintain a database of references (for example, in BIBTEX format—see §13.3), it is worth recording full details of a reference, even if not all of them are needed for journal reference lists. For a journal article, record the part (issue) number as well as the volume number; this extra information can speed the process of looking up a reference, especially if the journal issues are unbound. For a technical report, the month of publication is useful to know.

10. Be sure you are using the correct journal name and watch out for journals that change their names. For example, the *SIAM Journal on Scientific and Statistical Computing* (1980–1992) became the *SIAM Journal on Scientific Computing* in 1993.

11. In book titles, van Leunen recommends that a colon be added if it is needed to separate a title from a subtitle, and an awkward comma or colon separating a title from a subtitle should be removed. Thus *On Writing Well An Informal Guide to Writing Nonfiction* (as copied from the title page of [304]) needs a colon added after *Well*, and the colon should be removed from *Interpolation Theory:* 5.

12. Van Leunen recommends simplifying the names of major publishers to the bare bones, so that *John Wiley & Sons* becomes *Wiley*, and *Penguin Books* becomes *Penguin*. She also recommends omitting the city for a major publisher; I usually include it because many journals require it. For obscure publishers it is best to give as complete an address as possible.

13. For a book, the International Standard Book Number (ISBN) is worth recording, as it can be used to search library and publishers' catalogues. (Note, though, that hardback and softback editions of a book usually have different ISBNs.) An ISBN consists of ten digits, arranged in four groups whose size can vary. The first group specifies the language group of the publisher (0, 1 = English speaking countries, 2 = French speaking, 3 = German speaking, etc.). The second group (2–7 digits) identifies the publisher (e.g., Oxford University Press is 19) and the third group (1–6 digits) identifies the particular title. The last digit is a checksum. If the ISBN is expressed as $d_1 d_2 \ldots d_{10}$ then

$$d_{10} = \lceil s/11 \rceil * 11 - s, \quad \text{where} \quad s = \sum_{i=1}^{9} (11-i) d_i$$

($\lceil x \rceil$ denotes the smallest integer greater than or equal to x). A value $d_{10} = 10$ is written as "X". This book has the ISBN 0-89871-420-6: 89871 identifies SIAM as the publisher and 420 is the book's individual number. An International Standard Serial Number (ISSN) identifies a serial publication such as a journal, yearbook or institutional report. An ISSN has eight digits.

14. The date to quote for a book is the latest copyright date (excluding copyright renewals)—ignore dates of reprinting. Always state the edition number if it is not the first.

15. Make sure that every reference is actually cited in the paper. Some copy editors check this, as you may see from their pencilled marks on the manuscript when you receive the proofs.

16. Most mathematics journals require the reference list to be ordered alphabetically by author. Many science journals order by citation, so that the nth paper to be cited is nth in the reference list.

17. A list of standard abbreviations for mathematics journals can be found in *Mathematical Reviews* (see §14.3).

6.16. Specifics and Deprecated Practices

Capitalization

References to proper nouns should be capitalized: *See Theorem* 1.5*, the proof of Lemma* 3.4 *and the discussion in Section* 6. References to common nouns (generic objects) should not: *Next we prove the major theorems of the paper.*

Dangling Theorem

The term *dangling theorem* [147] (or *hanging theorem* [121]) refers to a construction such as the following one, where a theorem dangles or hangs from the end of a sentence.

> This result is proved in the following
> **Theorem 3.13**. If f is a twice continuously differentiable function ...

Halmos argues that while the practice can be defended, some readers dislike it, and it is not worth risking annoying them for the sake of avoiding the extra word *theorem*.

The following example does not strictly dangle, but is even more irritating.

> **5.1. Accuracy of the Computed Solution.** It depends on the machine precision and the conditioning of the problem.

Section headings stand alone and should not be taken as part of the text. The obvious solution

> **5.1. Accuracy of the Computed Solution.** The accuracy of the computed solution depends on the machine precision and the conditioning of the problem.

is inelegant in its repetition, but this could be avoided by rewriting the sentence or the title.

Footnotes

Footnotes are used sparingly nowadays in mathematical writing, and some journals do not allow them (see page 78 for an example of a footnote). It is bad practice to use them to squeeze more into a sentence than it can happily take. Their correct use is to add a note or comment that would deflect from the main message of the sentence. Donald W. Marquardt, the author of the 92nd most-cited paper in the Science Citation Index 1945–1988 [An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Indust. Appl. Math.*, 11(2):431–441, 1963], has stated that a critical part of the algorithm he proposed was described in a footnote and has sometimes been overlooked by people who have programmed the algorithm [97].

Numbering Mathematical Objects

Generally, you should number only those equations that are referenced within the text. This avoids the clutter of extraneous equation numbers and focuses the reader's attention on the important equations. Occasionally it is worth numbering key equations that are not referenced but which other authors might want to quote when citing your paper. Except in very short papers it is best to number equations by section rather than globally (equation (2.3) instead of equation (14)), for this makes referenced equations easier to find. The same applies to the numbering of theorems and other mathematical objects. Whether equation numbers appear on the left or the right of the page depends on the journal.

Two possible numbering sequences are illustrated by

> Definition 1, Lemma 1, Theorem 1, Remark 1, Definition 2, Lemma 2, ...
> Definition 1, Lemma 2, Theorem 3, Remark 4, Lemma 5, ...

Opinions differ as to which is the best scheme. The last has the advantage that it makes it easier to locate a particular numbered item, and the equation numbers themselves can even be included in the sequence for complete uniformity. The disadvantage is that the scheme mixes structures of a different character, which makes it difficult to focus on one particular set of structures (say, all the definitions); and on reading Remark 24 (say), the reader may wonder how many previous remarks there have been. A compromise between the two schemes is to number all lemmas, theorems and corollaries in one sequence, and definitions, remarks and so on in another. Some typesetting systems control the numbering of mathematical objects automatically. LaTeX does so, for example, and the numbering sequence for definitions, lemmas, theorems, etc., can be specified by LaTeX commands.

Plagiarism

Plagiarism is the act of publishing borrowed ideas or words as though they are your own. It is a major academic sin. In writing, if you copy a sentence or more you should either place it in quotes and acknowledge the source via a citation, or give an explicit reference such as "As Smith observed ...." In the case of a theorem statement it is acceptable to copy it word for word if you cite the source, but before copying it you should see whether you can improve the wording or make it fit better into your notation and style.

Regarding when to quote and when to paraphrase, van Leunen [283] advises "Quote what is memorable or questionable, strange or witty. Paraphrase the rest." When you wish to paraphrase, it is best to put the source aside, wait a reasonable period, and then rewrite what you want to say in your own words.

If you rework what you yourself have previously published without citing the source, thus passing it off as new, that is self-plagiarism, which is no less a sin than plagiarism.

Plagiarism has led to the downfall of many a career, in academia and elsewhere. Some notable cases are described in Mallon's *Stolen Words* [192] and LaFollette's *Stealing into Print* [169]. The former includes the ironic news, quoted from the *New York Times* of June 6, 1980, that

> Stanford University said today it had learned that its teaching assistant's handbook section on plagiarism had been plagiarized by the University of Oregon. ... Oregon officials apologized and said they would revise their guidebook.

Fraud is another serious malpractice, though apparently and understandably rare in mathematical research. Numerous cases of scientific fraud through history are catalogued in *Betrayers of the Truth* by Broad and Wade [39], while allegations that the psychologist Cyril Burt acted fraudulently are examined carefully in [189].

The Invalid Theorem

Avoid the mistake of calling a theorem into question through sentences such as the following:

> The theorem holds for any continuously differentiable function f. Unfortunately, the theorem is invalid because S is not path connected.

A theorem holds and is valid, by definition. A theorem might be *applicable to any continuously differentiable function* or *its invocation* may be *invalid because S is not path connected.*

"This Paper Proves ..."

In the abstract and introduction it is tempting to use wording such as "this paper proves" or "Section 3 shows" in place of "we prove" or "we show". This usage grates on the ear of some readers, as it is logically incorrect (though "Theorem 2 gives" cannot be criticized). The grating can be avoided by rewriting, but care is required to avoid a succession of sentences beginning "we".

BULLETIN OF THE AMERICAN ASSOCIATION OF PETROLEUM GEOLOGISTS
VOL. 50, NO. 9 (SEPTEMBER, 1966), P. 1992-1999, 2 FIGS.

GEOLOGICAL NOTES

A SCRUTINY OF THE ABSTRACT, II[1]

KENNETH K. LANDES[2]
Ann Arbor, Michigan

ABSTRACT

A partial biography of the writer is given. The inadequate abstract is discussed. What should be covered by an abstract is considered. The importance of the abstract is described. Dictionary definitions of "abstract" are quoted. At the conclusion a revised abstract is presented.

For many years I have been annoyed by the inadequate abstract. This became acute while I was serving a term as editor of the *Bulletin* of The American Association of Petroleum Geologists. In addition to returning manuscripts to authors for rewriting of abstracts, I also took 30 minutes in which to lower my ire by writing, "A Scrutiny of the Abstract."[1] This little squib has had a fantastic distribution. If only one of my scientific outpourings would do as well! Now the editorial board of the Association has requested a revision. This is it.

The inadequate abstract is illustrated at the top of the page. The passive voice is positively screaming at the reader! It is an outline, with each item in the outline expanded into a sentence. The reader is told what the paper is about, but not what it contributes. Such abstracts are merely overgrown titles. They are produced by writers who are either (1) beginners, (2) lazy, or (3) have not written the paper yet.

To many writers the preparation of an abstract is an unwanted chore required at the last minute by an editor or insisted upon even before the paper has been written by a deadline-bedeviled program chairman. However, in terms of market reached, the abstract is *the most important part of the paper*. For every individual who reads or listens to your entire paper, from 10 to 500 will read the abstract.

If you are presenting a paper before a learned society, the abstract alone may appear in a preconvention issue of the society journal as well as in the convention program; it may also be run by trade journals. The abstract which accompanies a published paper will most certainly reappear in abstract journals in various languages, and perhaps in company internal circulars as well. It is much better to please than to antagonize this great audience. Papers written for oral presentation should be *completed prior to the deadline for the abstract*, so that the abstract can be prepared from the written paper and not from raw ideas gestating in the writer's mind.

My dictionary describes an abstract as "a summary of a statement, document, speech, etc. . . ." and that which *concentrates in itself the essential information* of a paper or article. The definition I prefer has been set in italics. May all writers learn the art (it is not easy) of preparing an abstract containing the *essential information* in their compositions. With this goal in mind, I append an abstract that should be an improvement over the one appearing at the beginning of this discussion.

ABSTRACT

The abstract is of utmost importance, for it is read by 10 to 500 times more people than hear or read the entire article. It should not be a mere recital of the subjects covered. Expressions such as "is discussed" and "is described" should *never* be included! The abstract should be a condensation and concentration of the *essential information* in the paper.

[1] Revised from K. K. Landes' "A Scrutiny of the Abstract," first published in the *Bulletin* in 1951 (*Bulletin*, v. 35, no. 7, p. 1660). Manuscript received, June 3, 1966; accepted, June 10, 1966.

Editor's note: this abstract is published together with The Royal Society's "Guide for Preparation and Publication of Abstracts" to give *Bulletin* authors two viewpoints on the writing of abstracts.

[2] Professor of geology and mineralogy, University of Michigan. Past editor of the *Bulletin*.

A scrutiny of the introduction

By JON F. CLAERBOUT
Stanford University
Stanford, California

Abstract

The introduction to a technical paper should be an invitation to readers to invest their time reading it. Typically this invitation has three parts (1) the review, (2) the claim, and (3) the agenda. In the *claim*, the author should say why the paper's *agenda* is a worthwhile extension of its historical *review*. Personal pronouns should be used in the claim and anywhere else the author expresses judgment, opinion, or choice.

Introduction

Throughout the years, I have participated in reading committees of more than a hundred doctoral dissertations. Additionally, reports of the Stanford Exploration Project contain about 60 papers a year, and I am nominally in charge of making them presentable. In all this activity, I have seen many poor abstracts and, in each case, I have spared myself and the author much struggle by referring to the short paper *A scrutiny of the abstract* by Kenneth Landes (*AAPG Bulletin* 1966, 1990), which was formerly distributed by the SEG to all its aspiring authors. I rarely rewrite authors' abstracts any more—it's usually enough to have them read Landes' paper and rewrite it themselves. Landes' own abstract is worth quoting:

> The abstract is of the utmost importance, for it is read by 10 to 100 times more people than hear or read the entire article. It should not be a mere recital of subjects covered. Expressions such as "is discussed" and "is described" should *never* be included. The abstract should be a condensation and concentration of the *essential information* in the paper.

Introductions are not easy to write either. I am pleased to report that in recent years, I have developed a formula for the introduction. With this paper expounding my formula, I am hoping to reduce the need for one-on-one tutoring. You might be able to produce a good introduction without following my formula but if you have trouble producing one *that pleases other people* (and you would like to finish it and get on with your life), then I suggest you follow my formula.

First, I describe the three essential parts of an introduction and then I offer some tips on overall organization. You will see why introductions are so difficult to write once you understand how introductions depend on that most embarrassing of all words, "I."

The body of an introduction

My formula for an introduction is a sequence of three parts. They are (1) the review, (2) the claim, and (3) the agenda.

The review. Pick out about 3-10 papers providing a background to your research and say something about each of them. You could paraphrase a sentence or two from each abstract. The review is not intended to be a *historical* review going back to Newton or Descartes. Try to find a few papers by your office mates, your advisor, your predecessors, or other associates. That way you might find somebody to give you helpful criticism!

Anyone can follow these instructions and write a review that *looks* presentable. Where intelligence and skill are required is in organizing the review so that it leads up to something, namely your *claim*.

The claim. The most important part of the introduction is buried in the middle. It is the *claim*. The claim is where you claim your work is a worthwhile extension of the review you just wrote. If someone says your writing is "unmotivated," they aren't insulting your humanity, it just means they can't find your claim.

In your claim, you should use the personal pronoun "I" (or "we" if you aren't the sole author). The word "I" tells people where common knowledge runs out and your ideas begin. If you are writing a doctoral dissertation or an article for a refereed journal, then you should be making a new contribution to existing knowledge. Your paper is *not acceptable* without an identifiable claim.

Whether your ideas are solid as bedrock or speculative as clouds, you need first-person pronouns. Where your ideas are speculative, the pronouns signal a disclaimer. Where your ideas are solid, the pronouns signal that *you* may be credited for them. When your friends see your personal pronouns, they will know just where they should offer their questions and suggestions.

You may use personal pronouns elsewhere in your paper, too. Generally, you should use a personal pronoun whenever you are *expressing an opinion* or *exercising judgment*. Another time to use "I" is whenever there is a simple matter of *choice*. For example, "To test the theory, I selected some data," or "To examine the theory, I programmed the equations," or "To evaluate the hypothesis, I made some synthetic seismograms."

Good scientific papers contain many types of statements ranging from ancient axioms to common knowledge to speculations

(Scrutiny continued on p. 41)

(Scrutiny continued from p. 39)

and outright guesses. It is the *writer's* fault if a casual reader cannot distinguish these types of statements. Personal pronouns are good words to keep the distinctions clear. Other good words for this purpose are "should, could, would, might, may, can, is, does..." Use them all and pick the best for each purpose.

Some editors of scientific papers mechanically introduce the personal pronoun "I" to avoid the passive voice. I don't agree with them. For example, such editors will change "Substitution of equation (1) into equation (2) gives..." into "Substituting equation (1) into equation (2), I find..." The first wording states a simple fact but the second wording hints that someone else might get a different result.

The agenda. An agenda is found at the end of many introductions. It summarizes what you will show the reader as your paper progresses. Your agenda will be dull if it is merely a recital of the topics you will cover. Instead, it should tell how your paper works to fulfill your claim. In this way, your agenda should clarify your claim.

The agenda is not as important as the review and the claim. Keep it short.

Occasionally, you will be fortunate enough to be writing about something in which some of your conclusions can be made in simple statements. If so, state them early, right after your agenda. You aren't trying to write a mystery! Many more people will *begin* reading your paper than will *finish* reading it. Motivate them to finish! Unfortunately, many technical papers do not lend themselves to early conclusions.

After the introduction

Of course, you want people to read beyond your introduction, too. So think carefully about the order of your material and how you say things. (Notice this tiny paragraph is a small abstract of what follows.)

Order of material. You could write your paper so that each part builds on earlier parts. Like the axiomatic approach to geometry, you could refuse to refer to things not yet proven. But, rather than write your paper that way, it is wiser to maximize your readership. Since many more people will begin your paper than will plow through all the way to the end, try to state results before you prove them. Put off complicated derivations and digressions until the end. Complicated mathematical derivations, especially if marginal to your main thesis, should be relegated to appendices.

What is central and what is peripheral? In your paper, you might want to include digressions, possible applications, etc. That's nice. But be sure to include language that labels them as peripheral. If you don't, you may find people (and not just critics and enemies) missing your main point.

Conclusion

This short article is not a typical technical paper, but you might like to look back at the introduction to see if I follow my own advice. LE

ENVIRONMETRICS
Environmetrics 2009; **20**: 172–180
Published online 24 April 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.919

Simple confidence intervals for lognormal means and their differences with environmental applications

G. Y. Zou[1,2,*,†], Cindy Yan Huo[3] and Julia Taleban[1]

[1]*Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada*
[2]*Robarts Clinical Trials, Robarts Research Institute, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada*
[3]*Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada*

SUMMARY

The lognormal distribution has frequently been applied to approximate environmental data, with inference focusing on arithmetic means. Confidence interval estimation involving lognormal means in small to moderate sample sizes has received much attention over the years without a simple procedure in sight. We therefore propose a closed-form procedure for constructing confidence intervals for a lognormal mean and a difference between two lognormal means. The advantage of our procedure is that it only requires confidence limits for a normal mean and variance. The results of a numerical study show that our method performs as well as the generalized confidence interval (GCI) approach, which relies completely on computer simulation. Two real datasets are used to illustrate the methodology.

KEY WORDS: generalized confidence interval; log-normal; coverage; bootstrap

1. INTRODUCTION

It has become a tradition to fit the lognormal distribution to empirical data in environmental sciences (e.g., El-Shaarawi and Viveros, 1997; El-Shaarawi and Lin, 2007), due largely to the multiplicative central limit theorem (Limpert *et al*., 2001) in the sense that multiplication of a large number of random variables will result in a composite variable which can be approximated by the lognormal distribution.

A simple approach to analyzing lognormal data would be to log-transfer the data prior to employing standard statistical methods. The resultant inference would then be in terms of the median, which is less than the mean, and thus may provide substantial underestimates if the mean is the parameter of interest.

Inference in terms of lognormal means has received widespread attention in the literature, with two volumes devoted to the topic (Aitchison and Brown, 1957; Crow and Shimizu, 1988). Statistical methods for inference involving lognormal means have also appeared frequently in this journal, ranging from

*Correspondence to: G. Y. Zou, Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada N6A 5C1.
†E-mail: gzou@robarts.ca

Received 28 August 2007
Accepted 2 March 2008

computationally intensive methods such as the Gibbs sampler and bootstraps (Wild *et al*., 1996) to a *t*-distribution-based method (El-Shaarawi and Lin, 2007). It seems evident that the results for single lognormal means are not entirely satisfactory. Furthermore, there has not been much discussion on methods of comparing two lognormal means.

The purpose of this paper is to present a closed-form confidence interval procedure for a single lognormal mean and a difference between two lognormal means. We show that this closed-form procedure, requiring only confidence limits for a normal mean and variance, performs at least as well as the generalized confidence interval (GCI) approach which relies entirely on computer simulation.

The rest of the paper is structured as follows. Section 2 presents the new procedure, after summarizing the GCI (Krishnamoorthy and Mathew, 2003) and the modified Cox method (Armstrong, 1992; El-Shaarawi and Lin, 2007). In Section 3, we perform simulation studies to compare the performance of our method with previous ones. Two real datasets in an environmental context are used to illustrate the methods in Section 4. The paper closes with a discussion.

2. METHODS

Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed (iid) as lognormal with parameters μ and σ^2. This is to say that the log-transformed variables $X_1 = \ln Y_1, X_2 = \ln Y_2, \ldots, X_n = \ln Y_n$ are iid normal, denoted here as $N(\mu, \sigma^2)$. It is well known that the lognormal mean is $M = E(Y) = \exp(\mu + \sigma^2/2)$, estimated by

$$\widehat{M} = \exp\left(\bar{x} + s^2/2\right)$$

where $\bar{x}$ and s^2 are the sample mean and variance obtained using the log-transformed observations. Note that $\bar{x}$ and s^2 are independent of each other.

2.1. *Confidence interval for a single lognormal mean*

2.1.1. Existing methods. Land (1971) proposed an exact confidence interval by inverting the uniformly most powerful unbiased test. The procedure is computationally tedious and requires extensive tables. Thus, Land (1972) searched for simple approximate approaches and ended up with the one suggested by DR Cox in a personal communication showing promising performance. This method uses the property that $\bar{x}$ and s^2 are independent, with respective variances given by s^2/n and $s^4/[2(n-1)]$. Thus, as n becomes large, the $100(1-\alpha)\%$ confidence limits for $\mu + \sigma^2/2$ are given by

$$\left[\bar{x} + s^2/2\right] \pm z_{1-\alpha/2}\sqrt{s^2/n + s^4/[2(n-1)]}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. These limits can then be exponentiated to obtain a confidence interval for $\exp(\mu + \sigma^2/2)$.

As pointed by Land (1972), this method is not entirely satisfactory, particularly in the case of small n or large σ^2. To improve the performance in small samples, Armstrong (1992) and El-Shaarawi and Lin (2007) suggested replacing $z_{1-\alpha/2}$ with critical values from the *t*-distribution. This approach ignores the fact that the sampling distribution for s^2, which is distributed as chi-squared, is right-skewed.

DOI: 10.1002/env

Recently, a computer simulation-based method termed GCI appeared to perform very well. Krishnamoorthy and Mathew (2003) provide an algorithm as follows:

1. Obtain $\bar{x}$ and s^2 from log-transformed data.
2. Compute

$$T = \exp\left(\bar{x} - \frac{Z}{U/\sqrt{n-1}} \cdot \frac{s}{\sqrt{n}} + \frac{s^2}{2U^2/(n-1)}\right)$$

   where Z and U^2 are random numbers generated independently from the standard normal and chi-square distribution with $n-1$ degrees of freedom, respectively.
3. Repeat step 2 a large number (m) of times.
4. Sort the values of T. The $m(\alpha/2)^{\text{th}}$ and $m(1-\alpha/2)^{\text{th}}$ values are the $100(1-\alpha)\%$ confidence limits for $\exp(\mu + \sigma^2/2)$.

2.1.2. The proposed method. Before presenting our method for a single lognormal mean, we propose a general approach to setting confidence limits for a sum of two parameters, $\theta_1 + \theta_2$.

The conventional $100(1-\alpha)\%$ two-sided limits are

$$\widehat{\theta}_1 + \widehat{\theta}_2 - z_{1-\alpha/2}\sqrt{\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2)}$$

and

$$\widehat{\theta}_1 + \widehat{\theta}_2 + z_{1-\alpha/2}\sqrt{\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2)}$$

assuming $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are independent of each other. Besides the application of the central limit theorem, these limits are immediate results of assuming $\widehat{\theta}_i$ ($i = 1, 2$) and $\text{var}(\widehat{\theta}_i)$ are statistically independent of each other. Except for a normal mean $\bar{x}$, this is unlikely to hold in general.

Our idea is to exploit the dependence between $\widehat{\theta}_i$ and $\text{var}(\widehat{\theta}_i)$ in confidence interval construction. Specifically, we strive to estimate the variance of $\widehat{\theta}_1 + \widehat{\theta}_2$ in the vicinity of the limits (L, U) for $\theta_1 + \theta_2$.

By the duality between hypothesis testing and confidence interval construction, we recognize L as the minimum and U as the maximum value of $\theta_1 + \theta_2$ such that

$$\frac{\left[\widehat{\theta}_1 + \widehat{\theta}_2 - (\theta_1 + \theta_2)\right]^2}{\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2)} \approx z^2_{1-\alpha/2} \tag{1}$$

Thus, we should estimate the variances for $\widehat{\theta}_1$ and $\widehat{\theta}_2$ in the vicinity of $\min(\theta_1 + \theta_2)$ for L and that of $\max(\theta_1 + \theta_2)$ for U.

Now suppose the confidence limits for θ_i are readily obtained as (l_i, u_i), for $i = 1, 2$. Among the plausible values provided by the two pairs of confidence limits (l_1, u_1) and (l_2, u_2), the plausible minimum is $l_1 + l_2$ and the plausible maximum is $u_1 + u_2$. This implies that to obtain L, we need to estimate $\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2)$ under the condition $\theta_1 = l_1$ and $\theta_2 = l_2$. Similarly, to obtain U, we need to estimate $\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2)$ under the condition $\theta_1 = u_1$ and $\theta_2 = u_2$.

DOI: 10.1002/env

Again by the duality between hypothesis testing and confidence interval construction, l_i is $\min(\theta_i)$ satisfying

$$\frac{\left(\widehat{\theta}_i - l_i\right)^2}{\mathrm{var}(\widehat{\theta}_i)} \approx z^2_{1-\alpha/2}$$

which results in the estimated variance $\widehat{\mathrm{var}}(\widehat{\theta}_i)$ under the condition $\theta_i = l_i$ of

$$\widehat{\mathrm{var}}_l(\widehat{\theta}_i) \approx \frac{\left(\widehat{\theta}_i - l_i\right)^2}{z^2_{1-\alpha/2}}$$

Similarly, the estimated variance $\widehat{\mathrm{var}}(\widehat{\theta}_i)$ under the condition $\theta_i = u_i$ is

$$\widehat{\mathrm{var}}_u(\widehat{\theta}_i) \approx \frac{\left(u_i - \widehat{\theta}_i\right)^2}{z^2_{1-\alpha/2}}$$

Substituting these variance estimates back into Equation (1) yields the confidence limits (L, U) for $\theta_1 + \theta_2$ as

$$\begin{cases} L = \widehat{\theta}_1 + \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (\widehat{\theta}_2 - l_2)^2} \\ U = \widehat{\theta}_1 + \widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + (u_2 - \widehat{\theta}_2)^2} \end{cases} \tag{2}$$

These limits can now be applied in the current context, where $\theta_1 = \mu$ and $\theta_2 = \sigma^2/2$, with respective confidence intervals given by $(l_1, u_1) = (\bar{x} - z_{1-\alpha/2}\sqrt{s^2/n}, \bar{x} + z_{1-\alpha/2}\sqrt{s^2/n})$ and $(l_2, u_2) = \left[\frac{(n-1)s^2}{2\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{2\chi^2_{\alpha/2,n-1}}\right]$. Exponentiation of these limits yields a confidence interval for the lognormal mean. Specifically, the limits (*LL*, *UL*) are given by

$$LL = \widehat{M} \exp\left[-\left(\frac{z^2_{1-\alpha/2}s^2}{n} + \left(\frac{s^2}{2} - \frac{(n-1)s^2}{2\chi^2_{1-\alpha/2,n-1}}\right)^2\right)^{1/2}\right] \tag{3}$$

and

$$UL = \widehat{M} \exp\left[\left(\frac{z^2_{1-\alpha/2}s^2}{n} + \left(\frac{(n-1)s^2}{2\chi^2_{\alpha/2,n-1}} - \frac{s^2}{2}\right)^2\right)^{1/2}\right] \tag{4}$$

DOI: 10.1002/env

The Cox method can be obtained the same way by replacing the confidence interval for $\sigma^2/2$ with

$$(l_2, u_2) = \left(s^2 \left[\frac{1}{2} - z_{1-\alpha/2} \sqrt{\frac{1}{2(n-1)}} \right], s^2 \left[\frac{1}{2} + z_{1-\alpha/2} \sqrt{\frac{1}{2(n-1)}} \right] \right)$$

This is equivalent to treating the confidence interval of σ^2 as symmetric, indicating that for $n < 8$ a 95% confidence interval contains negative variance values. Replacing $z_{1-\alpha/2}$ with the t-value will not reduce the problem since the $t_{1-\alpha/2,n-1}$ is larger than that of a Normal distribution.

2.2. *Confidence intervals for a difference between two lognormal means*

Denoting a difference between two lognormal means as

$$\Delta = \exp(\mu_1 + \sigma_1^2/2) - \exp(\mu_2 + \sigma_2^2/2)$$

the correspondent estimator is

$$\widehat{\Delta} = \exp(\bar{x}_1 + s_1^2/2) - \exp(\bar{x}_2 + s_2^2/2)$$

with $(\bar{x}_1, s_1^2)$ and $(\bar{x}_2, s_2^2)$ computed from the log-transformed observations from two independent samples.

2.2.1. Generalized confidence interval approach. Krishnamoorthy and Mathew (2003) proposed the following algorithm for obtaining a $100(1-\alpha)\%$ confidence interval for Δ:

1. Compute $(\bar{x}_1, s_1^2)$ and $(\bar{x}_2, s_2^2)$.
2. Compute

$$T_\Delta = \exp\left(\bar{x}_1 - \frac{Z_1}{U_1/\sqrt{n_1 - 1}} \cdot \frac{s_1}{\sqrt{n_1}} + \frac{s_1^2}{2U_1^2/(n_1)} \right) - \exp\left(\bar{x}_2 - \frac{Z_2}{U_2/\sqrt{n_2 - 1}} \cdot \frac{s_2}{\sqrt{n_2}} + \frac{s_2^2}{2U_2^2/(n_2)} \right)$$

   where Z_i and U_i^2 are random numbers generated independently from the standard normal and chi-squared distribution with $n_i - 1$ degrees of freedom from two independent samples ($i = 1, 2$);
3. Repeat step 2 a large number of, say m, times.
4. Sort the T_Δ values from step 3. The confidence limits are given by the $m(\alpha/2)^{\text{th}}$ and $m(1-\alpha/2)^{th}$ T_Δ values.

2.2.2. The proposed method. Our alternative is first to obtain confidence limits for $M_1 = \exp(\mu_1 + \sigma_1^2/2)$ and $M_2 = \exp(\mu_2 + \sigma_2^2/2)$ using Equations (3) and (4), then to treat M_1 as θ_1 and $-M_2$ as θ_2 in the application of Equation (2). Note here that the limits for M_2, obtained using Equations (3) and (4), must be multiplied by -1 and then switched positions before plugging into Equation (2).

DOI: 10.1002/env

Straightforward algebra yields the $100(1-\alpha)\%$ confidence interval (L_Δ, U_Δ) for the difference between two lognormal means as

$$L_\Delta = \widehat{M}_1 - \widehat{M}_2 - \sqrt{(M_1 - LL_1)^2 + (UL_2 - M_2)^2}$$

and

$$U_\Delta = \widehat{M}_1 - \widehat{M}_2 + \sqrt{(UL_1 - M_1)^2 + (M_2 - LL_2)^2}$$

where

$$LL_i = \widehat{M_i} \exp\left[-\left(\frac{z_{1-\alpha/2}^2 s_i^2}{n_i} + \left(\frac{s_i^2}{2} - \frac{(n_i-1)s_i^2}{2\chi_{1-\alpha/2,n_i-1}^2}\right)^2\right)^{1/2}\right]$$

and

$$UL_i = \widehat{M_i} \exp\left[\left(\frac{z_{1-\alpha/2}^2 s_i^2}{n_i} + \left(\frac{(n_i-1)s_i^2}{2\chi_{\alpha/2,n_i-1}^2} - \frac{s_i^2}{2}\right)^2\right)^{1/2}\right]$$

for $i = 1, 2$.

3. SIMULATION

The confidence interval procedures described above are all asymptotic, meaning that their performance such as average percentage and tail errors may depend on sample size and parameter values. Before making any recommendations, we must evaluate their performance in finite sample sizes. For this purpose, we use Monto Carlo simulations to compare the procedures for the 95% confidence interval in terms of the percentage of times the interval contains the parameter value (coverage%). For a given parameter value, we assess the performance of a procedure using the percentage of times the confidence interval lies completely below or above the parameter value, termed left and right tail errors, respectively. We used 10 000 replicates for each parameter combination, with 10 000 resamples for the GCI approach. Using two standard errors of the nominal coverage rate as the criterion, we regarded coverage as within $(.95 \pm 2\sqrt{0.95 \times 0.05/10\,000})$, or (94.6–95.4) as adequate.

The second criterion is the balance between left and right tail errors (Jennings, 1987; Efron, 2003). We used confidence width as the third criterion to distinguish procedures satisfying the first and second criteria equally. Without loss of generality (Land, 1972, p. 147), we set $\mu = -\sigma^2/2$ in the simulation study.

For a single lognormal mean, we considered $n = 10$, 15, 25, and 50; $\sigma^2 = 0.1$, 0.5, 1.0, 1.5, and 2.0. The performance of the modified Cox method, our proposed method, and the generalized confidence interval are shown in Table 1. These results indicate that all three methods have acceptable coverage percentages. As expected, the modified Cox method has unbalanced tail errors, while the other two methods deliver reasonably balanced tail errors, with the proposed method showing consistently narrower average width.

DOI: 10.1002/env

Table 1. Comparative performance of three procedures for constructing a 95% two-sided confidence interval for a lognormal mean with $\mu = -\sigma^2/2$ based on 10 000 runs

		$n = 10$		$n = 15$		$n = 25$		$n = 50$	
σ^2	Method	Cover (ML , MR)	W	Cover (ML , MR)	W	Cover (ML , MR)	W	Cover (ML , MR)	W
0.1	MCox	95.23 (3.31, 1.46)	0.46	95.33 (3.03, 1.64)	0.36	94.90 (3.15, 1.95)	0.27	95.08 (2.88, 2.04)	0.18
	Proposed	93.27 (3.87, 2.86)	0.44	93.85 (3.53, 2.62)	0.34	94.13 (3.32, 2.55)	0.26	94.55 (2.96, 2.49)	0.18
	GCI	95.10 (2.20, 2.70)	0.50	95.00 (2.27, 2.73)	0.37	94.88 (2.38, 2.74)	0.27	94.92 (2.33, 2.75)	0.19
0.5	MCox	94.88 (4.48, 0.64)	1.29	94.93 (4.29, 0.78)	0.94	94.84 (3.87, 1.29)	0.68	94.44 (3.97, 1.59)	0.46
	Proposed	94.50 (3.24, 2.26)	1.67	94.54 (3.30, 2.16)	1.05	94.79 (2.97, 2.24)	0.71	94.54 (3.20, 2.26)	0.47
	GCI	94.84 (1.94, 3.22)	1.90	94.76 (2.33, 2.91)	1.14	95.03 (1.95, 3.02)	0.75	94.57 (2.62, 2.81)	0.48
1.0	MCox	93.99 (5.82, 0.19)	2.53	94.44 (5.17, 0.39)	1.67	94.69 (4.70, 0.61)	1.14	94.89 (3.89, 1.22)	0.73
	Proposed	94.68 (3.39, 1.93)	5.36	94.89 (3.12, 1.99)	2.28	95.02 (3.18, 1.80)	1.30	94.87 (2.80, 2.33)	0.77
	GCI	94.49 (2.40, 3.11)	6.12	94.42 (2.23, 3.35)	2.49	94.86 (2.42, 2.72)	1.37	94.77 (2.29, 2.94)	0.79
1.5	MCox	93.76 (6.19, 0.05)	4.84	94.24 (5.58, 0.18)	2.60	94.07 (5.40, 0.53)	1.63	95.00 (4.05, 0.95)	1.01
	Proposed	95.37 (2.94, 1.69)	24.34	95.28 (2.89, 1.83)	4.48	94.74 (3.08, 2.18)	2.06	94.99 (2.64, 2.37)	1.11
	GCI	94.89 (2.06, 3.05)	27.52	95.18 (2.04, 2.78)	4.91	94.53 (2.34, 3.13)	2.17	95.04 (2.11, 2.85)	1.14
2.0	MCox	93.32 (6.65, 0.03)	10.63	93.71 (6.16, 0.13)	4.14	94.58 (5.02, 0.40)	2.27	94.72 (4.50, 0.78)	1.31
	Proposed	95.15 (2.98, 1.87)	497.08	94.83 (3.09, 2.08)	9.84	95.24 (2.73, 2.03)	3.19	94.82 (2.70, 2.48)	1.49
	GCI	94.71 (2.15, 3.14)	899.26	94.38 (2.41, 3.21)	10.80	95.07 (2.14, 2.79)	3.38	94.64 (2.41, 2.95)	1.53

MCox, the modified Cox method; GCI, generalized confidence interval; ML, the confidence interval lies completely below the parameter; MR, the confidence interval lies completely above the parameter; W, average interval width.

For a difference between two lognormal means, we considered $n_1 = 10, 15, 20, 25$, and 50; $n_2 = 10, 20, 25$, and 50; $\sigma_1^2 = 0.1, 0.5, 1.0, 1.5, 2.0$; $\sigma_2^2 = 0.5, 1.5$, and 2.0. The performance of the generalized confidence interval method and the proposed method with modified Cox method for single means for these 300 parameter combinations are presented using summary statistics (Table 2). These results clearly show that the Modified Cox method provides severely unbalanced tails with coverage percentage ranging from 93.17 to 98.11%. Our proposed method is very competitive with the computer simulation-based GCI, both having coverage rates outside the range of 94.6 to 95.4% when $n \leq 15$.

Table 2. Comparative performance of three procedures for constructing a 95% two-sided confidence interval for a difference between two lognormal means with $\mu_i = -\sigma_i^2/2$, $i = 1, 2$ (summary of 300 parameter combinations with 10 000 runs for each combination)

Method		Mean	Min	10th pctl	25th pctl	50th pctl	75th pctl	90th pctl	Max
MCox	Cover	95.57	93.17	94.55	95.09	95.63	96.13	96.52	98.11
	ML	1.86	0.04	0.25	0.66	1.50	2.85	3.98	6.19
	MR	2.57	0.07	0.49	1.09	2.42	3.86	4.90	6.76
	Width	2.36	0.49	1.05	1.46	2.11	2.96	3.90	8.25
Proposed	Cover	95.32	94.26	94.92	95.13	95.32	95.52	95.75	96.32
	ML	2.28	1.74	1.97	2.12	2.27	2.42	2.59	3.17
	MR	2.40	1.69	2.06	2.19	2.36	2.59	2.84	3.42
	Width	4.11	0.49	1.13	1.73	2.92	5.32	9.54	29.67
GCI	Cover	95.25	94.29	94.86	95.03	95.23	95.48	95.71	96.18
	ML	2.40	1.76	2.04	2.18	2.35	2.59	2.83	3.40
	MR	2.34	1.82	2.06	2.16	2.32	2.51	2.68	3.25
	Width	4.47	0.51	1.18	1.83	3.07	5.91	10.81	33.10

MCox, the modified Cox method; GCI, generalized confidence interval; ML, the confidence interval lies completely below the parameter; MR, the confidence interval lies completely above the parameter.

DOI: 10.1002/env

4. ILLUSTRATIVE EXAMPLES

As an example of a simple lognormal mean, we consider air lead levels (μ g/m^3) of $n = 15$ sites at the Alma American Labs, Fairplay, Colorado on 23 February 1989 (Krishnamoorthy *et al*., 2006): 200, 120, 15, 7, 8, 6, 48, 61, 380, 80, 29, 1000, 350, 1400, 110. The lognormal distribution was found to fit the data well. Log-transformation of the data yields $\bar{x} = 4.333$ and $s = 1.739$. Therefore, we have the 95% confidence limits for $\theta_1 = \mu$ given by $[4.333 - 1.96 \times 1.739/\sqrt{15}, 4.333 + 1.96 \times 1.739/\sqrt{15}]$, i.e., (3.452584, 5.213141) and that for $\theta_2 = \sigma^2/2$ given by:

$$\frac{1}{2}\left[\frac{(15-1)\times 1.739^2}{\chi^2_{0.975,14}}, \frac{(15-1)\times 1.739^2}{\chi^2_{0.025,14}}\right]$$

that is, (0.8108892, 3.762765). Substituting these limits into Equations (3) and (4) yields the 95% two-sided confidence interval for $\exp(\mu + \sigma^2/2)$ as (112, 3873), comparable with the GCI of (122, 4280) based on 100 000 simulations.

As an example for a difference between two lognormal means. We consider a dataset from the Data and story Library (http://lib.stat.cmu.edu/DASL). In April–May 1993, an oil refinery near San Francisco submitted $n = 31$ daily CO emission measurements from its stacks to the Bay Area Air Quality Management District for establishing a baseline. It was of interest to see whether the refinery had over-measured CO emission, as compared to nine measurements taken by the Management District person between September 1990 to March 1993. The data are given as:

Refinery ($n_1 = 31$): 45, 30, 38, 42, 63, 43, 102, 86, 99, 63, 58, 34, 37, 55, 58, 153, 75, 58, 36, 59, 43, 102, 52, 30, 21, 40, 141, 85, 161, 86, 71.

District management ($n_2 = 9$): 12.5, 20, 4, 20, 25, 170, 15, 20, 15.

Recognizing the temporal dependence among the measurements, we nevertheless treat them as independent for illustration purposes. The lognormal distribution fits both dataset well (Krishnamoorthy and Mathew, 2003), with $\bar{x}_1 = 4.074252$, $s_1^2 = 0.252081$, $\bar{x}_2 = 2.963333$, and $s_2^2 = 0.949618$. Using our approach, the estimated mean and 95% confidence interval of the refinery data are given by 66.70583 (55.57714, 81.69155) and that of the district Management data are given by 31.12906 (15.66019, 128.6178). Application of our procedure yields the difference and 95% confidence interval of 35.58 (−62.55, 57.11). Again, comparable with those from the GCI (−79.15, 57.47) based on 100 000 simulations.

5. DISCUSSION

We have presented a simple approach to confidence interval estimation concerning lognormal means. The resultant procedures for a single lognormal mean and a difference between two lognormal means are in closed-form, requiring only methods found in introductory textbooks. The performance of our procedure has been shown to do at least as well as the GCI approach, which relies on computer simulation. Moreover, although exact in theory, even with the same dataset the latter approach may result in different answers from different analysts or the same analyst performing analyses at different times.

We note that the method we described here can be readily applied to lognormal regression models (Bradu and Mundlak, 1970; El-Shaarawi and Viveros, 1997; El-Shaarawi and Lin, 2007). Exten-

DOI: 10.1002/env

sions and applications of this method in other contexts can be found elsewhere (Zou, 2007; Zou and Donner, 2008).

We did not consider bootstrap methods for lognormal data, as it has been revealed that such methods fail even for a normal variance (Schenker, 1985). It is then inevitable for bootstrap to fail for the lognormal mean because it is a function of the normal mean and variance. We refer to Zhou and Dinh (2005) for simulation results showing that bootstrap methods fail terribly in the case of lognormal data. Interestingly, many papers have appeared by merely implementing a bootstrap method, as if it is the gold standard. This practice is a result of overlooking the fact that bootstrap is also asymptotically reliable and requires evaluation on a case-by-case basis (DiCiccio and Efron, 1996).

ACKNOWLEDGEMENTS

Dr Zou is a recipient of an Early Researcher Award from Ontario Ministry of Research and Innovation, Canada. His research is also supported partially by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Aitchison J, Brown JAC. 1957. *The Lognormal Distribution*. Cambridge University Press: Cambrige.

Armstrong BG. 1992. Confidence intervals for arithmetic means of lognormally distributed exposures. *American Industrial Hygiene Association Journal* **53**: 481–485.

Bradu D, Mundlak T. 1970. Estimation in lognormal linear models. *Journal of the American Statistical Association* **65**: 198–211.

Crow EL, Shimizu K. 1988. *Lognormal Distributions: Theory and Applications*. Dekker: New York.

DiCiccio TJ, Efron B. 1996. Bootstrap confidence intervals. *Statistical Science* **11**: 189–228.

Efron B. 2003. Second thoughts on the bootstrap. *Statistical Science* **18**: 135–140.

El-Shaarawi AH, Lin J. 2007. Interval estimation for log-normal mean with applications to water quality. *Environmetrics* **18**: 1–10.

El-Shaarawi AH, Viveros R. 1997. Inference about the mean in log-regression with environmental applications. *Environmetrics* **8**: 569–582.

Jennings DE. 1987. How do we judge confidence-interval adequacy? *The American Statistician* **41**: 335–337.

Krishnamoorthy K, Mathew T, Ramachandran G. 2006. Generalized P-values and confidence intervals: a novel approach for analyzing lognormally distributed exposure data. *Journal of Occupational and Environmental Hygiene* **3**: 642–650.

Krishnamoorthy K, Mathew TP. 2003. Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. *Journal of Statistical Planning and Inference* **115**: 103–121.

Land CE. 1971. Confidence intervals for linear functions of the normal mean and variance. *Annals of Mathematical Statistics* **42**: 1187–1205.

Land CE. 1972. An evaluation of approximate confidence interval estimation methods for lognormal means. *Technometrics* **14**: 145–158.

Limpert E, Stahel WA, Abbt M. 2001. Log-normal distributions across the sciences: keys and clues. *BioScience* **51**: 341–352.

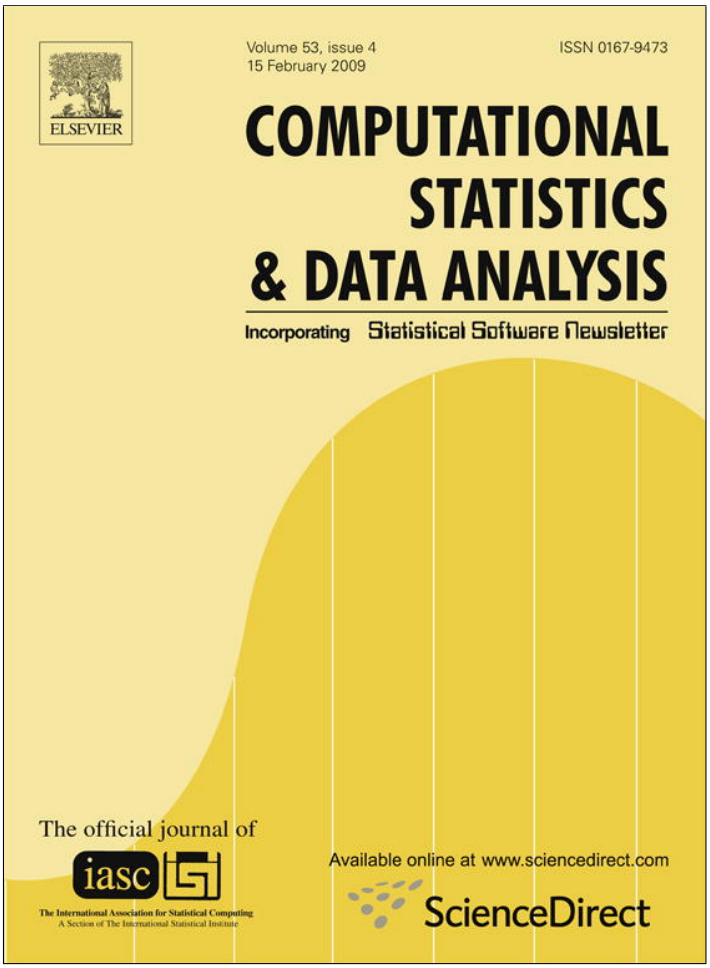
Schenker N. 1985. Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association* **80**: 360–361.

Wild P, Hordan R, Leplay A, Vincent R. 1996. Confidence intervals for probabilities of exceeding threshold limits with censored log-normal data. *Environmetrics* **7**: 247–259.

Zhou XH, Dinh P. 2005. Nonparametric confidence intervals for the one- and two-sample problems. *Biostatistics* **6**: 187–200.

Zou GY. 2007. Toward using confidence intervals to compare correlations. *Psychological Methods* **12**: 399–413.

Zou GY, Donner A. 2008. Construction of confidence limits about effect measures: a general approach. *Statistics in Medicine* **27**: http://dx.doi.org/10.1002/sim.3095



Computational Statistics and Data Analysis 53 (2009) 1080–1085

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A note on confidence interval estimation for a linear function of binomial proportions

Guang Yong Zou [a,b,*], Wenyi Huang [a], Xiaohe Zhang [c]

[a] Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada N6A 5C1

[b] Robarts Clinical Trials, Robarts Research Institute, University of Western Ontario, London, Ontario, Canada N6A 5K8

[c] Department of Medicine, McMaster University, Population Health Research Institute, Hamilton, Ontario, Canada L8L 2X2

ARTICLE INFO

Article history:
Received 23 May 2008
Received in revised form 24 September 2008
Accepted 25 September 2008
Available online 11 October 2008

ABSTRACT

The Wilson score confidence interval for a binomial proportion has been widely applied in practice, due largely to its good performance in finite samples and its simplicity in calculation. We propose its use in setting confidence limits for a linear function of binomial proportions using the method of variance estimates recovery. Exact evaluation results show that this approach provides intervals that are narrower than the ones based on the adjusted Wald interval while aligning the mean coverage with the nominal level.

1. Introduction

There exists a large literature on confidence interval estimation involving binomial proportions. For a single proportion, there are several choices. The first is given by adding to and subtracting from the maximum likelihood estimator the standard normal quantile multiplied by its estimated standard error. This procedure is commonly referred to as the Wald method. The second is the interval based on inverting the approximate normal test that uses the standard errors estimated at the lower and upper limits. This procedure is commonly referred to as Wilson score method (Wilson, 1927). The score confidence interval has now become very popular, especially after the expositions by Agresti and Coull (1998) and Newcombe (1998b). With an attempt to ease classroom teaching, Agresti and Coull (1998) suggested an adjusted Wald method by adding two successes and two failures and then using the Wald formula. Despite the terminology, the adjusted Wald method is actually an approximation of the score method.

The superior performance of the Wilson method has been carried over to cases of a difference between two proportions (Newcombe, 1998a) and a difference between two differences (Newcombe, 2001). It is interesting to note that this seemingly *ad hoc* procedure has become more popular than the rigorous score interval for a difference between two proportions (Mee, 1984; Miettinen and Nurminen, 1985; Gart and Nam, 1990), caused largely by the computation involved in obtaining the latter.

Due to its important practical value, confidence interval construction for a linear function of binomial proportions has received some attention recently (Price and Bonett, 2004; Tebbs and Roths, 2008). The purpose of this note is to extend the argument of Zou and Donner (2008) to a linear function of parameters, and in particular to binomial proportions. Since our main idea is to recover variance estimates from readily available confidence limits for single parameters, we refer to the approach as the MOVER, the method of variance estimates recovery. As shown below, the MOVER will not only shed some light to Newcombe (1998a) and Newcombe (2001) but also provide an alternative to Price and Bonett (2004) who proposed

* Corresponding address: Robarts Clinical Trials, Robarts Research Institute, P. O. Box 5015, 100 Perth Drive, London, Ontario, Canada N6A 5K8. Tel.: +1 519 663 3400x34092; fax: +1 519 663 3807.
E-mail address: gzou@robarts.ca (G.Y. Zou).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.09.033

a procedure for a linear function of proportions based on the adjusted Wald interval for single proportions (Agresti and Coull, 1998). We will show that the confidence interval for a linear function of binomial proportions based on the Wilson method is narrower than that of Price and Bonett (2004). We will not consider the approach by Tebbs and Roths (2008) because of its inherent drawbacks such as involved computation, restriction in parameter ranges, and undercoverage.

2. The MOVER and its application to linear functions of binomial proportions

Suppose we wish to construct an approximate $100(1-\alpha)\%$ two-sided confidence interval for $\theta_1+\theta_2$, where the estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are assumed to be independent. By the central limit theorem, the lower limit (L) is given by

$$L = \widehat{\theta}_1 + \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{\theta}_1) + \mathrm{var}(\widehat{\theta}_2)}. \tag{1}$$

Inspired by the score method for interval estimation (Bartlett, 1953; Gart and Nam, 1990), we can estimate the variance needed for L at $\theta_1+\theta_2 = L$. This has at least one disadvantage that it is in general an iterative procedure, which can be an obstacle to wide application in practice as what happened to the score interval for a difference between two proportions. Therefore, we proceed with estimating the variance in the neighborhood of L.

Now, suppose that the $100(1-\alpha)\%$ two-sided confidence intervals (l_i, u_i) for single parameters θ_i, $i = 1, 2$ are available. Note that there is no need to specify the approaches taken to obtain (l_i, u_i). Among all the plausible parameter values of θ_1 provided by (l_1, u_1) and that of θ_2 provided by (l_2, u_2), $l_1 + l_2$ is usually closer to L than $\widehat{\theta}_1 + \widehat{\theta}_2$. As a result, for L, we can estimate $\mathrm{var}(\widehat{\theta}_1)$ at $\theta_1 = l_1$ and $\mathrm{var}(\widehat{\theta}_2)$ at $\theta_2 = l_2$.

Furthermore, we can recover the required variance estimates from $\hat{\theta}_i(l_i, u_i)$, $i = 1, 2$, as follows. By the central limit theorem and letting $z_{\alpha/2}$ be the upper $\alpha/2$ quantile of the standard Normal distribution, we have

$$l_i = \widehat{\theta}_i - z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_i)},$$

which gives a variance estimate for $\widehat{\theta}_i$ at $\theta_i = l_i$ as

$$\widehat{\mathrm{var}}_l(\widehat{\theta}_i) = (\widehat{\theta}_i - l_i)^2/z^2_{\alpha/2}$$

and

$$u_i = \widehat{\theta}_i + z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_i)},$$

which gives a variance estimate at $\theta_i = u_i$ as

$$\widehat{\mathrm{var}}_u(\widehat{\theta}_i) = (u_1 - \widehat{\theta}_i)^2/z^2_{\alpha/2}.$$

Note that the recovered variance estimates $\widehat{\mathrm{var}}_l(\widehat{\theta}_i)$ and $\widehat{\mathrm{var}}_u(\widehat{\theta}_i)$ are different, except when the interval (l_i, u_i) is symmetric about $\widehat{\theta}_i$. Symmetric intervals are known to perform poorly in finite samples for most problems in practice. In fact, it was stated (Efron and Tibshirani, 1993, p. 180) that symmetry is the most serious error in confidence interval construction. The Wald interval for a binomial proportion is a perfect example. In contrast, the Wilson interval is asymmetric as a consequence of estimating variances at the lower and upper limits separately.

Plugging the recovered variance estimates into Eq. (1) results in

$$\begin{aligned} L &= \widehat{\theta}_1 + \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{\theta}_1) + \mathrm{var}(\widehat{\theta}_2)} \\ &= \widehat{\theta}_1 + \widehat{\theta}_2 - z_{\alpha/2}\sqrt{(\widehat{\theta}_1 - l_1)^2/z^2_{\alpha/2} + (\widehat{\theta}_2 - l_2)^2/z^2_{\alpha/2}} \\ &= \widehat{\theta}_1 + \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (\widehat{\theta}_2 - l_2)^2}. \end{aligned}$$

Analogous steps with the notion that $u_1 + u_2$ is in the vicinity of U yield the upper limit U as

$$U = \widehat{\theta}_1 + \widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + (u_2 - \widehat{\theta}_2)^2}.$$

Rewriting $\theta_1-\theta_2$ as $\theta_1+(-\theta_2)$ and noting that the confidence limits for $-\theta_2$ are given by $(-u_2, -l_2)$, we obtain confidence limits for $\theta_1 - \theta_2$ as

$$L = \widehat{\theta}_1 - \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (u_2 - \widehat{\theta}_2)^2}$$

and

$$U = \widehat{\theta}_1 - \widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + (\widehat{\theta}_2 - l_2)^2}.$$

This confidence interval, apparently first presented by Howe (1974), has been applied by Newcombe (1998a) and by Newcombe (2001) to binomial proportions. There has been no analytic justification for its general applicability until recently (Zou and Donner, 2008).

Regarding $\theta_1+\theta_2$ and $\theta_1-\theta_2$ as $c_1\theta_1+c_2\theta_2$, where c_1 and c_2 are constants, we can rewrite the intervals as

$$L=c_1\widehat{\theta}_1+c_2\widehat{\theta}_2-\sqrt{[c_1\widehat{\theta}_1-\min(c_1l_1,c_1u_1)]^2+[c_2\widehat{\theta}_2-\min(c_2l_2,c_2u_2)]^2}$$

and

$$U=c_1\widehat{\theta}_1+c_2\widehat{\theta}_2+\sqrt{[c_1\widehat{\theta}_1-\max(c_1l_1,c_1u_1)]^2+[c_2\widehat{\theta}_2-\max(c_2l_2,c_2u_2)]^2}.$$

For a $100(1-\alpha)\%$ confidence interval for $\sum_{i=1}^{g}c_i\theta_i$, where $g>2$, an application of mathematical induction results in

$$\begin{cases} L=\sum_{i=1}^{g}c_i\widehat{\theta}_i-\sqrt{\sum_{i=1}^{g}\left[c_i\widehat{\theta}_i-\min(c_il_i,c_iu_i)\right]^2} \\ U=\sum_{i=1}^{g}c_i\widehat{\theta}_i+\sqrt{\sum_{i=1}^{g}\left[c_i\widehat{\theta}_i-\max(c_il_i,c_iu_i)\right]^2}. \end{cases} \tag{2}$$

Because L and U are derived using the recovered variance estimates, we can refer to the method as the MOVER, standing for method of variance estimates recovery. A further extension of the MOVER to incorporate dependence between $\widehat{\theta}_i$ and $\widehat{\theta}_j$ $(i\neq j)$ has been applied to measures of additive interaction in epidemiology (Zou, 2008).

We can now apply the confidence interval in (2) to linear functions of binomial proportions. Since there are at least three intervals for a single proportion, i.e., Wald, adjusted Wald (Agresti and Coull, 1998) and Wilson, we end up with three procedures for linear functions of binomial proportions.

Specifically, let Y_i $(i=1,2,\ldots,g)$ be independent binomial variates with parameters (n_i,p_i), and let $\widehat{p}_i=Y_i/n_i$ be the sample estimates for p_i. A linear function of binomial proportions may be defined as $\sum_{i=1}^{g}c_ip_i$, where the c_i are known constants. Using the equations in (2), the $100(1-\alpha)\%$ Wald confidence interval can be obtained by setting $\widehat{\theta}_i=\widehat{p}_i=Y_i/n_i$, $l_i=\widehat{p}_i-z_{\alpha/2}\sqrt{\widehat{p}_i(1-\widehat{p}_i)/n_i}$, and $u_i=\widehat{p}_i+z_{\alpha/2}\sqrt{\widehat{p}_i(1-\widehat{p}_i)/n_i}$.

The Wilson interval for $\sum_{i=1}^{g}c_ip_i$ may be obtained by setting $\widehat{\theta}_i=\widehat{p}_i=Y_i/n_i$,

$$l_i,u_i=\left(\widehat{p}_i+z^2_{\alpha/2}/(2n_i)\mp z_{\alpha/2}\sqrt{[\widehat{p}_i(1-\widehat{p}_i)+z^2_{\alpha/2}/(4n_i)]/n_i}\right)/(1+z^2_{\alpha/2}/n_i).$$

The adjusted Wald interval for $\sum_{i=1}^{g}c_ip_i$ (Price and Bonett, 2004) may be obtained by setting $\widehat{\theta}_i=\tilde{p}_i=(Y_i+2/k)/(n_i+4/k)$ (where k is the number of nonzero elements in c_i), $l_i=\tilde{p}_i-z_{\alpha/2}\sqrt{\tilde{p}_i(1-\tilde{p}_i)/n_i}$, and $u_i=\tilde{p}_i+z_{\alpha/2}\sqrt{\tilde{p}_i(1-\tilde{p}_i)/n_i}$. Note that the adjusted Wald method for a single proportion is an approximation of the Wilson score method for 95% interval, see Agresti and Coull (1998) for its motivation and derivation. We also must point out that this method has the potential to provide confidence limits that are out of parameter space.

It is fair to say that the superior performance of Newcombe (1998a) originates from that of the Wilson method for a single proportion (Agresti and Coull, 1998; Newcombe, 1998b). On the same token, we can postulate that applying Wilson interval for cases of more than two binomial proportions will be very competitive to that of Price and Bonett (2004).

To evaluate this claim, we conducted a numerical study to compare the performance of these two procedures in finite samples for 90%, 95%, and 99% two-sided confidence intervals, in terms of mean coverage, minimum coverage, and mean interval width as defined here.

For a $100(1-\alpha)\%$ interval (L,U) for $\sum_{i=1}^{g}c_ip_i$, the coverage is defined by

$$\text{Coverage}=100\sum_{y_1=0}^{n_1}\cdots\sum_{y_g=0}^{n_g}\prod_{i=1}^{g}\binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i}I\left(L<\sum c_ip_i<U\right),$$

where $I(.)$ is an indicator function which takes values of 1 or 0 as the event in the brackets is true or not.

The expected interval width is defined as

$$\text{Width}=\sum_{y_1=0}^{n_1}\cdots\sum_{y_g=0}^{n_g}\prod_{i=1}^{g}\binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i}(U-L).$$

We conducted the evaluation by first randomly sampling 1000 sets of p_i's from the uniform (0,1) distribution, and then applied the above two definitions to each set. We did not arbitrarily truncate the adjusted Wald confidence limits when they fell out of the parameter space. With respect to each method, we obtained the mean coverage, minimum coverage, and the mean interval width using these 1000 sets of values for coverage and width.

Table 1
Performance of the method of variance estimates recovery in constructing two-sided confidence intervals (CI) for a linear function of binomial parameters, $\sum_{i=1}^{3} c_i p_i$, using confidence limits for single proportions obtained by the adjusted Wald and Wilson methods. Entries in each row are based on 1000 sets of p_i's randomly sampled from uniform (0,1), and each set evaluated by exact calculation.

Group sizes	90% CI		95% CI	
$n_1/n_2/n_3$	Adjusted Wald	Wilson	Adjusted Wald	Wilson
$c = (1/3, 1/3, 1/3)$				
5/5/5	92.04 (80.18, 0.32)*	90.58 (82.60, 0.31)	96.12 (91.24, 0.38)	94.99 (86.34, 0.36)
5/5/10	91.65 (83.59, 0.29)	90.45 (85.36, 0.29)	95.95 (88.12, 0.35)	95.07 (89.55, 0.33)
5/10/15	91.53 (84.60, 0.25)	90.69 (87.42, 0.25)	95.86 (92.93, 0.30)	95.26 (90.65, 0.29)
5/10/20	91.63 (86.89, 0.25)	90.80 (87.37, 0.24)	95.88 (92.51, 0.29)	95.31 (91.02, 0.28)
5/15/20	91.57 (84.34, 0.23)	90.80 (84.39, 0.23)	95.83 (90.95, 0.28)	95.30 (90.51, 0.27)
5/20/20	91.53 (83.99, 0.23)	90.72 (87.15, 0.22)	95.73 (84.97, 0.27)	95.27 (91.50, 0.26)
$c = (1, -1/2, -1/2)$				
5/5/5	92.29 (80.22, 0.67)	90.82 (85.45, 0.65)	96.19 (86.68, 0.79)	95.14 (89.23, 0.75)
5/5/10	91.97 (84.26, 0.64)	90.84 (84.63, 0.62)	95.91 (89.26, 0.77)	95.27 (89.48, 0.72)
5/10/15	92.00 (82.28, 0.60)	91.06 (86.44, 0.58)	95.87 (86.43, 0.72)	95.31 (91.04, 0.67)
5/10/20	92.00 (82.74, 0.60)	91.06 (85.77, 0.57)	95.83 (87.82, 0.71)	95.33 (91.62, 0.66)
5/15/20	92.00 (82.28, 0.59)	91.06 (86.34, 0.56)	95.75 (88.50, 0.70)	95.27 (91.63, 0.65)
5/20/20	92.13 (81.69, 0.58)	91.13 (85.68, 0.55)	95.80 (86.75, 0.69)	95.27 (91.15, 0.64)
$c = (-1, 1/2, 2)$				
5/5/5	92.08 (79.06, 1.25)	90.94 (86.88, 1.20)	95.91 (86.89, 1.49)	95.19 (89.05, 1.39)
5/5/10	91.52 (85.49, 1.00)	90.67 (86.58, 0.98)	95.81 (91.28, 1.19)	95.24 (89.05, 1.15)
5/10/15	91.33 (85.15, 0.88)	90.64 (87.14, 0.87)	95.66 (91.82, 1.05)	95.29 (88.53, 1.02)
5/10/20	91.35 (85.32, 0.82)	90.72 (87.51, 0.81)	95.71 (92.89, 0.98)	95.32 (90.85, 0.95)
5/15/20	91.29 (82.82, 0.81)	90.65 (87.98, 0.80)	95.67 (92.39, 0.97)	95.23 (91.42, 0.94)
5/20/20	91.29 (85.58, 0.81)	90.65 (86.88, 0.80)	95.66 (90.11, 0.96)	95.30 (90.18, 0.93)
$c = (1, 1, -1)$				
5/5/5	92.04 (80.36, 0.95)	90.56 (81.21, 0.93)	96.19 (92.21, 1.13)	95.15 (86.22, 1.08)
5/5/10	91.74 (85.12, 0.88)	90.64 (85.70, 0.86)	95.96 (89.78, 1.04)	95.18 (89.86, 1.00)
5/10/15	91.49 (84.88, 0.76)	90.71 (87.33, 0.74)	95.78 (87.97, 0.90)	95.26 (90.45, 0.87)
5/10/20	91.49 (85.56, 0.74)	90.76 (86.80, 0.72)	95.80 (92.41, 0.88)	95.29 (90.54, 0.84)
5/15/20	91.42 (85.05, 0.70)	90.69 (86.68, 0.69)	95.70 (90.10, 0.84)	95.20 (91.24, 0.80)
5/20/20	91.59 (85.28, 0.68)	90.82 (87.43, 0.66)	95.82 (88.32, 0.81)	95.26 (90.77, 0.78)

∗ Mean coverage % (minimum coverage %, mean confidence interval width) based on 1000 sets of proportion parameters randomly sampled from uniform (0,1) distribution.

For linear functions of 3 binomial proportions, results in Table 1 show consistently that the intervals for linear functions based on the Wilson score method have mean coverage closer to the nominal levels, with narrow average width. For group sizes considered, the minimum coverage for the adjusted Wald can be as low as 79.06% for 90% nominal level, and 84.97% for 95% nominal level. For confidence interval based on the Wilson method, the minimum coverage can be as low as 81.21% for 90% nominal level, and 86.22% for 95% nominal level. Results from constructing confidence intervals for linear functions of 4 binomial proportions in Table 2 show again that the procedure based on Wilson score method performed better in terms of mean coverage and interval width, as well as minimum coverage. For example, the minimum coverage for the adjusted Wald can be as low as 77.16% for 90% nominal level, compared to that of 83.23% for Wilson score method. Similar trends were observed with nominal level of 99% (results not shown). One possible explanation for our results is that the adjusted Wald method was proposed to approximate the Wilson score method at 95% level, on the rationale that the middle point of Wilson interval is a weighted average of $\widehat{p}$ and 0.5, and that $1.96^2 \approx 4$ (Agresti and Coull, 1998, p. 122).

3. Examples

In the light of the above numerical results, we now compare confidence intervals using two examples from Price and Bonett (2004).

Example 1. This data set arose from a study in which rats are fed with different types of diets. The diets are controlled by two factors, namely fiber and fat. Each rat is observed to determine if it has developed a tumor during the study period. The outcome of the experiment is summarized in Table 3 (each group had 30 rats). It is of interest to construct confidence intervals for the main effects of fiber and fat, as well as their interaction. Here we can obtain the 95% confidence intervals using the MOVER for the linear functions of proportions. The results are shown in Table 3, which shows that the intervals obtained using the Wilson method for single proportions are narrower than those using the adjusted Wald method for single proportion. This is consistent with the results in our evaluation study. In fact, the Wilson method based intervals are all contained in that based on the adjusted Wald method for single proportions in this moderate size study.

Example 2. This example arose from the Framingham heart study. As an alternative to conventional generalized linear model with logistic link function, Price and Bonett (2004) approached the problem with a linear function of binomial

Table 2
Performance of the method of variance estimates recovery in constructing two-sided confidence intervals (CI) for a linear function of binomial parameters, $\sum_{i=1}^{4} c_i p_i$, using confidence limits for single proportions obtained by the adjusted Wald and Wilson methods. Entries in each row are based on 1000 sets of p_i's randomly sampled from uniform (0,1), and each set evaluated by exact calculation.

Group sizes	90% CI		95% CI	
$n_1/n_2/n_3/n_4$	Adjusted Wald	Wilson	Adjusted Wald	Wilson
$c = (1/4, 1/4, 1/4, 1/4)$				
5/5/5/5	91.27 (81.65, 0.28)*	90.24 (83.23, 0.27)	95.63 (92.16, 0.33)	94.86 (88.56, 0.31)
5/5/10/10	91.03 (87.49, 0.24)	90.33 (85.65, 0.24)	95.48 (93.03, 0.29)	95.07 (90.18, 0.28)
5/5/15/15	91.01 (87.61, 0.23)	90.53 (86.71, 0.22)	95.45 (92.91, 0.27)	95.13 (91.33, 0.26)
5/5/15/20	91.11 (88.33, 0.22)	90.67 (87.18, 0.22)	95.50 (93.10, 0.26)	95.25 (91.75, 0.26)
5/10/15/20	90.84 (86.26, 0.20)	90.57 (87.66, 0.20)	95.33 (90.52, 0.24)	95.27 (91.64, 0.23)
$c = (-1, 1, -1, 1)$				
5/5/5/5	91.33 (85.72, 1.10)	90.29 (82.48, 1.08)	95.70 (92.91, 1.31)	95.09 (88.82, 1.25)
5/5/10/10	91.03 (83.19, 0.96)	90.37 (85.96, 0.95)	95.49 (92.68, 1.15)	95.09 (90.56, 1.11)
5/5/15/15	91.11 (87.96, 0.91)	90.67 (86.77, 0.89)	95.50 (92.51, 1.08)	95.24 (91.47, 1.04)
5/5/15/20	91.08 (87.41, 0.89)	90.71 (87.05, 0.88)	95.50 (92.45, 1.06)	95.26 (91.74, 1.02)
5/10/15/20	90.86 (87.58, 0.81)	90.51 (87.68, 0.80)	95.37 (91.86, 0.97)	95.17 (91.40, 0.94)
$c = (1/3, 1/3, 1/3, 1)$				
5/5/5/5	91.33 (77.16, 0.63)	90.87 (86.18, 0.61)	95.29 (84.48, 0.75)	95.15 (89.89, 0.70)
5/5/10/10	90.82 (86.51, 0.49)	90.52 (87.43, 0.49)	95.23 (90.73, 0.59)	95.16 (91.24, 0.57)
5/5/15/15	90.64 (87.56, 0.44)	90.29 (87.33, 0.43)	95.22 (93.25, 0.52)	95.10 (91.60, 0.51)
5/5/15/20	90.74 (88.44, 0.40)	90.25 (86.91, 0.40)	95.34 (93.10, 0.48)	95.07 (90.91, 0.47)
5/10/15/20	90.48 (88.76, 0.39)	90.29 (87.89, 0.38)	95.12 (93.65, 0.46)	95.15 (92.17, 0.45)
$c = (-3, -1, 1, 3)$				
5/5/5/5	91.34 (83.03, 2.44)	90.89 (83.44, 2.38)	95.49 (90.28, 2.90)	95.20 (89.81, 2.76)
5/5/10/10	90.94 (83.95, 2.14)	90.80 (87.55, 2.10)	95.23 (88.85, 2.55)	95.32 (91.50, 2.44)
5/5/15/15	91.01 (82.16, 2.01)	90.80 (87.95, 1.96)	95.24 (86.35, 2.40)	95.22 (90.66, 2.29)
5/5/15/20	90.92 (83.77, 1.96)	90.77 (88.24, 1.90)	95.13 (86.91, 2.33)	95.19 (91.50, 2.22)
5/10/15/20	91.08 (83.32, 1.91)	91.02 (88.56, 1.86)	95.15 (85.88, 2.27)	95.36 (91.06, 2.16)

* Mean coverage % (minimum coverage %, mean confidence interval width) based on 1000 sets of proportion parameters randomly sampled from uniform (0,1) distribution.

Table 3
Confidence intervals for effects of factors in the diet–tumor study.

Fiber	Fat	$\widehat{p}_i$	c_i		
			Fiber × Fat	Fiber	Fat
Yes	High	20/30	1	1/2	1/2
	Low	14/30	−1	1/2	−1/2
No	High	27/30	1	−1/2	1/2
	Low	19/30	−1	−1/2	−1/2
Interval for $\sum c_i p_i$:					
		Adj Wald	−0.3806, 0.2516	−0.3516, 0.0355	0.0677, 0.3839
		Wilson	−0.3790, 0.2386	−0.3459, 0.0375	0.0691, 0.3773

Table 4
Framingham heart study.

Systolic BP	Number of subjects	Number with heart disease
115	156	3
121	252	17
131	284	12
141	271	16
151	139	12
161	85	8
176	99	16
190	43	8

proportions. Specifically, if the population proportion of heart disease is considered a linear function of systolic blood pressure, the slope is $\sum c_i p_i$, which is a linear function of the proportions p_i of heart disease of systolic blood pressure groups, where $c_i = (x_i - \sum x_i/g)/\sum(x_i - \sum x_i/g)^2$ and x_i is the value of the quantitative factor in group i. Using the data in Table 4, we obtained the 95% confidence interval for the population slope using the adjusted Wald method as 0.0010 to 0.0032, comparable to that of using the Wilson method as 0.0012 to 0.0034 in such a large study.

4. Concluding remarks

The confidence interval for a general linear function of binomial proportions introduced here is a simple application of a more general idea presented by Zou and Donner (2008). The basic idea is to recover variance estimates needed for linear functions of proportions from the confidence limits for single proportions. Since the Wilson interval procedure has been strongly recommended for single proportions (Agresti and Coull, 1998; Newcombe, 1998b; Santner, 1998), it is thus natural to extend it to linear functions of binomial proportions. By use of the MOVER, we have provided a very competitive procedure to that of Price and Bonett (2004), whose procedure can be seen as an application of the MOVER based on the adjusted Wald method for single proportions. The MOVER has also provided an analytic justification for Newcombe (1998a, 2001).

It should also be noted that the derivation of the MOVER relies only on the validity of confidence limits for single parameters such that variance estimates can be recovered by normal distributions. The direct implication is that one can apply the MOVER to linear functions of other discrete distribution parameters, e.g., Poisson rates (Stamey and Hamilton, 2006; Tebbs and Roths, 2008), and linear functions of normal mean and variance, e.g., lognormal means (Zou and Donner, 2008).

Acknowledgments

The authors gratefully acknowledge the comments from two anonymous reviewers which led to the insight that the adjusted Wald method is actually an approximation of Wilson score confidence interval for a single binomial proportion. Guang Yong Zou is a recipient of the Early Researcher Award, Ontario Ministry of Research and Innovation, Canada. His work was also partially supported by an Individual Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Agresti, A., Coull, B., 1998. Approximate is better than "exact" for interval estimation of binomial proportions. American Statistician 52, 119–126.
Bartlett, M.S., 1953. Approximate confidence intervals. 2. More than one unknown parameter. Biometrika 40, 306–317.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC, New York.
Gart, J.J., Nam, J.M., 1990. Approximate interval estimation of the difference in binomial parameters - correction for skewness and extension to multiple tables. Biometrics 46, 637–643.
Howe, W.G., 1974. Approximate confidence limits on the mean of $X + Y$ where X and Y are two tabled independent random variables. Journal of the American Statistical Association 69, 789–794.
Mee, R.W., 1984. Confidence bounds for the difference between two probabilities. Biometrics 40, 1175–1176.
Miettinen, O., Nurminen, M., 1985. Comparative analysis of two rates. Statistics in Medicine 4, 213–226.
Newcombe, R.G., 1998a. Interval estimation for the difference between independent proportions: Comparison of eleven methods. Statistics in Medicine 17, 873–890.
Newcombe, R.G., 1998b. Two-sided confidence intervals for the single proportions: Comparison of seven methods. Statistics in Medicine 17, 857–872.
Newcombe, R.G., 2001. Estimating the difference between differences: Measurement of additive scale interaction for proportions. Statistics in Medicine 20, 2885–2893.
Price, R.M., Bonett, D.G., 2004. An improved confidence interval for a linear function of binomial proportions. Computational Statistics & Data Analysis 45, 449–456.
Santner, T.J., 1998. Teaching large-sample binomial confidence intervals. Teaching Statistics 20, 20–23.
Stamey, J., Hamilton, C., 2006. A note on confidence intervals for a linear function of Poisson rates. Communications in Statistics–Simulation and Computation 35, 849–856.
Tebbs, J.M., Roths, S.A., 2008. New large-sample confidence intervals for a linear combination of binomial proportions. Journal of Statistical Planning and Inference 138, 1884–1893.
Wilson, E.B., 1927. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22, 209–212.
Zou, G.Y., 2008. On the estimation of additive interaction using the four-by-two table and beyond. American Journal of Epidemiology 168, 212–224.
Zou, G.Y., Donner, A., 2008. Construction of confidence limits about effect measures: A general approach. Statistics in Medicine 27, 1693–1702.

Giving an Effective Presentation

David Giltinan (ENAR), *Member of the Local Organizing Committee of IBC2000*

Introduction

Several articles in the statistical literature contain tips on giving an effective statistical presentation. An excellent recent example, by Becker and Keller-McNulty, appears in *The American Statistician* (1996, pages 112-115). So why did I agree to attempt another essay on this topic, for speakers intending to present at IBC2000? In part, because my experience at recent statistical meetings strongly suggests that most of us could benefit from a reminder of the common pitfalls that can mar a presentation. While I have been lucky to attend some excellent talks at recent meetings, these have not been in the majority. None of the following advice is novel, but I hope that a short review of common presentation mistakes may be helpful. If there is one essential message, it can be summarized in this exhortation to speakers — "always be considerate of your audience".

To avoid a monotonous litany of "do's and don'ts", I have tried to inject some humor into the following remarks. This does not mean that I think the generally low prevailing standard of statistical presentations is not a serious matter. On the contrary, I believe effective presentation is one of the most important challenges facing any statistician. Until clear communication becomes a top priority, we cannot hope to achieve the degree of influence, or make the type of effective contribution, that society needs from our profession.

Today's airport bookshop is typically stocked with a plethora of titles along the lines of "*Jesse Ventura's eight secrets for charismatic communication*" or "*Darth Vader's seven steps to effective leadership*". It appears that the modern business professional expects advice to be packaged in snappy bite-sized nuggets, suitable for digestion on a plane. Accordingly, this essay follows the organizational structure: "*Ten tips for a truly dreadful presentation*". Those who aspire to the status of truly dreadful presenter (abbreviated as TDP from here on) should try to implement as many of these tips as possible. Speakers interested in improving the quality of their presentations, on the other hand, would be better served by rigorous avoidance of the types of misbehavior described in this essay.

I have grouped these into categories, only one of which is specific to statistical pre-

sentations. For concreteness, illustrations below assume use of overhead transparencies; however, most points apply equally to other types of visual aids.

Ten tips for a truly dreadful presentation

Sensory deprivation

1. Small is beautiful

A key component of this technique is information overload. Here, the defining characteristic is to cram as many words/numbers/symbols onto each overhead as possible. Audience members will be delighted by the wealth of detail and the resulting chance to practice their speed-reading skills. Handwritten overheads should aspire to a cramped, wobbly, style that evokes the drama of an airplane flying through extreme turbulence. This effect is harder to achieve using presentation software, but much can be accomplished by creative use of novel font styles and tiny font sizes.

2. Confusion through color

For handwritten overheads, the optimal choice of pen color is clearly yellow, as it can generally be relied on to yield text which is not just unreadable, but also virtually invisible. Other color options may achieve a similar effect, though some experimentation may be needed to find the best combination (light and pastel shades hold the most promise). If forced to use dark-colored pens when preparing overheads, the experienced TDP will know to choose non-waterproof pens, sometimes known as "smudgies". The resulting combination of densely written material, ambient humidity and/or perspiration during presentation virtually guarantees enhanced illegibility through smudging. Initially, this option might seem to be available only for handwritten overheads. However, trial and error should reveal the potential of one's software package to generate color combinations for which distinguishing text from background is an impossibility. No self-respecting TDP will leave this potential unrealized.

3. The human shield

Occasionally, one may be provided with clear, legible presentation aids, prepared by someone else. Without some neutralizing tactic, this carries a genuine risk of conveying information clearly to the audience. A simple countermeasure in this situation is the "human shield" approach, wherein the presenter blocks all visibility by standing directly in front of the projector while speaking. Static implementation of this tactic can be challenging, as it may be hard to ignore the progressively louder bleats of protest from the audience. A preferred alternative is thus the so-called "random Wimbledon" variation, in which the speaker darts randomly from one blocking position to another. This has the added benefit of keeping audience members alert, while giving a good calisthenic workout to their neck muscles.

Audience alienation

4. Cultural insensitivity

Opening with a sexist joke can usually be relied on to alienate most of the audience. An alternative tactic is the consistent use of gendered language to perpetuate some demeaning stereotype of women's roles and abilities; for further discussion, see the 1997 article by Hammer (*The American Statistician*, pages 13-18). "Humor" that reinforces some other negative cultural stereotype or ethnic prejudice may be effective in offending remaining audience members.

5. Avoid eye contact

Making eye contact with individual audience members is discouraged for several reasons. It could be taken as indicating a genuine desire to communicate. Worse, it could provide a real-time check on audience reaction to the presentation, which, if acted upon, could slow progress through the remaining overheads. Finally, it is particularly critical to avoid eye contact with the session chair, who may be actively trying to put a premature end to your presentation.

Continued on p. 14

Giving an Effective Presentation

Continued from p. 8

6.The illiterate audience

A tactic which never fails to amuse is to present certain overheads in a manner which conveys the obvious belief that members of the audience can't read.Typically, this calls for the speaker to read aloud each and every word of text shown overhead, at an excruciatingly slow pace, adopting the tone of a particularly conscientious kindergarten teacher. Clearly, the audience irritation potential of this tactic is quite wasted if it is deployed simultaneously with the human shield technique described above. Judicious alternation of the two methods, on the other hand, may have the potential for a superadditive irritant effect.

Presentation style

7. Keep them on their toes

- Give no context, background, or motivation for the problem you discuss.You don't want to deprive your audience of the fun of trying to puzzle it out.
- Similarly, provide no clues about the relative importance of different parts of your talk. Alert listeners should be able to distinguish the important from the trivial without your help.
- Waste no time on 'signposting' devices such as a presentation outline, or subdivision of your talk into sections. They'll know you have finished when you sit down.
- Adopt a variable pacing strategy, alternating between vastly accelerated and excruciatingly slow delivery. Devote least time to the overheads with the highest density of content.

8. Rehearsal is for amateurs

Conscientious amateurs, worried about such petty trivia as time restrictions, abstract notions of "fairness" towards the session chair and other speakers, and consideration for the audience, may feel impelled to practice their presentations several times beforehand. Some fanatics have even been known to seek input from colleagues on issues such as emphasis, organization, clarity, length, potential "early stopping points", etc.This type of weakness is for lesser mortals – remember, you are a professional. Nothing as mundane as rehearsal should be allowed to interfere with the delightful spontaneity which is the hallmark of your oratory. As for time constraints, these are imposed with other, less experienced, speakers in mind. The session chair should understand that they do not apply to you. If not, simply cease to acknowledge the chair's existence.

Statistical specialties

9.The power of notation

Although most audiences are familiar with the conventional deployment of Greek symbols in statistics, the experienced TDP still has no difficulty in harnessing the full power of notation to bewilder and confuse. This is possible, even when sticking to notational con-

ventions which are technically legitimate. Strategies sure to amuse and challenge one's audience include

- Refusal to be bound by conventions of "standard" usage. General recognition of a particular choice of symbols as conventional does not make alternative choices invalid. There's nothing illegal about denoting the mean by s and the standard deviation by μ.
- Giving equal time to the lesser-known Greek letters. Consider using your talk as a vehicle for pursuing the rehabilitation of ζ, ξ, ϖ, ψ, and ν, as partial redress for years of neglect.
- Including each and every detail of the technical conditions needed for your main convergence result, especially those ugly higher-moment assumptions. After all, the audience deserves nothing less than the complete story.
- Subtly changing the symbol for a key parameter half-way through your presentation.

10. Tables and graphs

When deciding how to summarize information for one's presentation, it may be helpful to remember these general points about communicating information intelligibly

- The more densely packed with information, the harder a table is to assimilate, particularly if displayed for a maximum of 30 seconds.
- Mislabeling, or failure to label, rows and columns of a tabular display can greatly enhance audience confusion.
- Most simulations defy clear, concise summarization. Their potential for audience confusion thus greatly exceeds that of real data examples.
- Graphs generally provide more audience-friendly summaries than tables.
- The potential superiority of graphical displays to communicate information is easily sabotaged by techniques such as (i) mislabeling axes (ii) omitting axis labels altogether (iii) using a microscopic font for axis labels (iv) misleading choice of scale (v) confusing choice of symbols, connecting lines, shading patterns etc. This list is in no way exhaustive.
- Including lots of irrelevant detail makes both tables and graphs harder to understand.
- Use of the "show 'n whisk" presentation style to tease audience members (for instance, by limiting display time for information-laden overheads to 30 seconds) can greatly reduce their chances of assimilating the information displayed, whether summarized in tabular or graphical form.

Presenting effectively

Giving an effective presentation can indeed be difficult. However, if you can resist the temptation to misbehave in the various ways described in this essay, you will be well on your way. Looking forward to some truly excellent talks at IBC2000!

Reference sheet for `natbib` usage

(Describing version 8.1 from 2007/10/30)

For a more detailed description of the `natbib` *package, LaTeX the source file* `natbib.dtx`*.*

Overview

The `natbib` package is a reimplementation of the LaTeX `\cite` command, to work with both author–year and numerical citations. It is compatible with the standard bibliographic style files, such as `plain.bst`, as well as with those for `harvard`, `apalike`, `chicago`, `astron`, `authordate`, and of course `natbib`.

Loading

Load with `\usepackage[`*options*`]{natbib}`. See list of *options* at the end.

Replacement bibliography styles

I provide three new `.bst` files to replace the standard LaTeX numerical ones:

`plainnat.bst` `abbrvnat.bst` `unsrtnat.bst`

Basic commands

The `natbib` package has two basic citation commands, `\citet` and `\citep` for *textual* and *parenthetical* citations, respectively. There also exist the starred versions `\citet*` and `\citep*` that print the full author list, and not just the abbreviated one. All of these may take one or two optional arguments to add some text before and after the citation.

`\citet{jon90}`	⇒	Jones et al. (1990)
`\citet[chap.~2]{jon90}`	⇒	Jones et al. (1990, chap. 2)
`\citep{jon90}`	⇒	(Jones et al., 1990)
`\citep[chap.~2]{jon90}`	⇒	(Jones et al., 1990, chap. 2)
`\citep[see][]{jon90}`	⇒	(see Jones et al., 1990)
`\citep[see][chap.~2]{jon90}`	⇒	(see Jones et al., 1990, chap. 2)
`\citet*{jon90}`	⇒	Jones, Baker, and Williams (1990)
`\citep*{jon90}`	⇒	(Jones, Baker, and Williams, 1990)

Multiple citations

Multiple citations may be made by including more than one citation key in the `\cite` command argument.

`\citet{jon90,jam91}`	⇒	Jones et al. (1990); James et al. (1991)
`\citep{jon90,jam91}`	⇒	(Jones et al., 1990; James et al. 1991)
`\citep{jon90,jon91}`	⇒	(Jones et al., 1990, 1991)
`\citep{jon90a,jon90b}`	⇒	(Jones et al., 1990a,b)

Numerical mode

These examples are for author–year citation mode. In numerical mode, the results are different.

`\citet{jon90}`	⇒	Jones et al. [21]
`\citet[chap.~2]{jon90}`	⇒	Jones et al. [21, chap. 2]
`\citep{jon90}`	⇒	[21]
`\citep[chap.~2]{jon90}`	⇒	[21, chap. 2]
`\citep[see][]{jon90}`	⇒	[see 21]
`\citep[see][chap.~2]{jon90}`	⇒	[see 21, chap. 2]
`\citep{jon90a,jon90b}`	⇒	[21, 32]

Suppressed parentheses

As an alternative form of citation, `\citealt` is the same as `\citet` but *without parentheses*. Similarly, `\citealp` is `\citep` without parentheses.

The `\citenum` command prints the citation number, without parentheses, even in author–year mode, and without raising it in superscript mode. This is intended to be able to refer to citation numbers without superscripting them.

`\citealt{jon90}`	⇒	Jones et al. 1990
`\citealt*{jon90}`	⇒	Jones, Baker, and Williams 1990
`\citealp{jon90}`	⇒	Jones et al., 1990
`\citealp*{jon90}`	⇒	Jones, Baker, and Williams, 1990
`\citealp{jon90,jam91}`	⇒	Jones et al., 1990; James et al., 1991
`\citealp[pg.~32]{jon90}`	⇒	Jones et al., 1990, pg. 32
`\citenum{jon90}`	⇒	11
`\citetext{priv.\ comm.}`	⇒	(priv. comm.)

The `\citetext` command allows arbitrary text to be placed in the current citation parentheses. This may be used in combination with `\citealp`.

Partial citations

In author–year schemes, it is sometimes desirable to be able to refer to the authors without the year, or vice versa. This is provided with the extra commands

`\citeauthor{jon90}`	⇒	Jones et al.
`\citeauthor*{jon90}`	⇒	Jones, Baker, and Williams
`\citeyear{jon90}`	⇒	1990
`\citeyearpar{jon90}`	⇒	(1990)

Forcing upper cased names

If the first author's name contains a *von* part, such as "della Robbia", then `\citet{dRob98}` produces "della Robbia (1998)", even at the beginning of a sentence. One can force the first letter to be in upper case with the command `\Citet` instead. Other upper case commands also exist.

when	`\citet{dRob98}`	⇒	della Robbia (1998)
then	`\Citet{dRob98}`	⇒	Della Robbia (1998)
	`\Citep{dRob98}`	⇒	(Della Robbia, 1998)
	`\Citealt{dRob98}`	⇒	Della Robbia 1998
	`\Citealp{dRob98}`	⇒	Della Robbia, 1998
	`\Citeauthor{dRob98}`	⇒	Della Robbia

These commands also exist in starred versions for full author names.

Citation aliasing

Sometimes one wants to refer to a reference with a special designation, rather than by the authors, i.e. as Paper I, Paper II. Such aliases can be defined and used, textual and/or parenthetical with:

`\defcitealias{jon90}{Paper~I}`		
`\citetalias{jon90}`	⇒	Paper I
`\citepalias{jon90}`	⇒	(Paper I)

These citation commands function much like `\citet` and `\citep`: they may take multiple keys in the argument, may contain notes, and are marked as hyperlinks.

Selecting citation style and punctuation

Use the command `\setcitestyle` with a list of comma-separated keywords (without spaces) as argument.

Citation mode: `authoryear` or `numbers` or `super`
Braces: `round` or `square` or `open={`*char*`},close={`*char*`}`
Between citations: `semicolon` or `comma` or `citesep={`*char*`}`
Between author and year: `aysep={`*char*`}`
Between years with common author: `yysep={`*char*`}`
Text before post-note: `notesep={`*text*`}`

Defaults are `authoryear`, `round`, `comma`, `aysep={;}`, `yysep={,}`, `notesep={, }`

Example 1, `\setcitestyle{square,aysep={},yysep={;}}` changes the author–year output of

```
\citep{jon90,jon91,jam92}
```

into [Jones et al. 1990; 1991, James et al. 1992].

Example 2, `\setcitestyle{notesep={; },round,aysep={},yysep={;}}` changes the output of

```
\citep[and references therein]{jon90}
```

into (Jones et al. 1990; and references therein).

Other formatting options

Redefine `\bibsection` to the desired sectioning command for introducing the list of references. This is normally `\section*` or `\chapter*`.

Define `\bibpreamble` to be any text that is to be printed after the heading but before the actual list of references.

Define `\bibfont` to be a font declaration, e.g. `\small` to apply to the list of references.

Define `\citenumfont` to be a font declaration or command like `\itshape` or `\textit`.

Redefine `\bibnumfmt` as a command with an argument to format the numbers in the list of references. The default definition is `[#1]`.

The indentation after the first line of each reference is given by `\bibhang`; change this with the `\setlength` command.

The vertical spacing between references is set by `\bibsep`; change this with the `\setlength` command.

Automatic indexing of citations

If one wishes to have the citations entered in the `.idx` indexing file, it is only necessary to issue `\citeindextrue` at any point in the document. All following `\cite` commands, of all variations, then insert the corresponding entry to that file. With `\citeindexfalse`, these entries will no longer be made.

Use with `chapterbib` package

The `natbib` package is compatible with the `chapterbib` package which makes it possible to have several bibliographies in one document.

The package makes use of the `\include` command, and each `\included` file has its own bibliography.

The order in which the `chapterbib` and `natbib` packages are loaded is unimportant.

The `chapterbib` package provides an option `sectionbib` that puts the bibliography in a `\section*` instead of `\chapter*`, something that makes sense if there is a bibliography in each chapter. This option will not work when `natbib` is also loaded; instead, add the option to `natbib`.

Every `\included` file must contain its own `\bibliography` command where the bibliography is to appear. The database files listed as arguments to this command can be different in each file, of course. However, what is not so obvious, is that each file must also contain a `\bibliographystyle` command, with possibly differing arguments.

As of version 8.0, the citation style, including mode (author–year or numerical) may also differ between chapters. The `\setcitestyle` command can be issued at any point in the document, in particular in different chapters.

Sorting and compressing citations

Do not use the `cite` package with `natbib`; rather use one of the options `sort`, `compress`, or `sort&compress`.

These also work with author–year citations, making multiple citations appear in their order in the reference list.

Long author list on first citation

Use option `longnamesfirst` to have first citation automatically give the full list of authors.

Suppress this for certain citations with `\shortcites{`*key-list*`}`, given before the first citation.

Local configuration

Any local recoding or definitions can be put in `natbib.cfg` which is read in after the main package file.

Options that can be added to `\usepackage`

`round` (default) for round parentheses;

`square` for square brackets;

`curly` for curly braces;

`angle` for angle brackets;

`semicolon` (default) to separate multiple citations with semi-colons;

`colon` the same as `semicolon`, an earlier mistake in terminology;

`comma` to use commas as separators;

`authoryear` (default) for author–year citations;

`numbers` for numerical citations;

`super` for superscripted numerical citations, as in *Nature*;

`sort` orders multiple citations into the sequence in which they appear in the list of references;

`sort&compress` as `sort` but in addition multiple numerical citations are compressed if possible (as 3–6, 15);

`compress` to compress without sorting, so compression only occurs when the given citations would produce an ascending sequence of numbers;

`longnamesfirst` makes the first citation of any reference the equivalent of the starred variant (full author list) and subsequent citations normal (abbreviated list);

`sectionbib` redefines `\thebibliography` to issue `\section*` instead of `\chapter*`; valid only for classes with a `\chapter` command; to be used with the `chapterbib` package;

`nonamebreak` keeps all the authors' names in a citation on one line; causes overfull hboxes but helps with some `hyperref` problems.

Words and expressions: Less is more

July 15, 2009

[1]

~~a considerable amount of~~ → much
~~a great deal of~~ → much
~~absolutely essential~~ → essential
~~accounted for by the fact~~ → because
~~adjacent to~~ → near, next to
~~along the lines of~~ → like
~~as a consequence of~~ → because
~~as a matter of fact~~ → in fact
~~as a result of~~ → because
~~as to~~ → about
~~at present~~ → now
~~based on the fact that~~ → because
~~because of the fact that~~ → because
~~by means of~~ → by, with
~~causal factor~~ → cause
~~cognizant~~ → aware of
~~completely full~~ → full
~~contingent upon~~ → depend on
~~despite the fact that~~ → although
~~due to the fact that~~ → because
~~during the course of~~ → during, while
~~elucidate~~ → explain
~~employ~~ → use
~~end result~~ → result
~~endeavor~~ → try
~~fabricate~~ → make
~~facilitate~~ → help
~~first of all~~ → first
~~firstly~~ → first
~~for the purpose of~~ → for
~~for the reason that~~ → because
~~from the point of view of~~ → for
~~give an account of~~ → describe
~~give rise to~~ → cause
~~has been engaged in a study~~ → has studied
~~has the capability of~~ → can
~~has the potential to~~ → can, may
~~have the appearance of~~ → look like, resemble
~~in case~~ → if
~~in close proximity to~~ → close, near
~~in light of the fact that~~ → because
~~in only a small number of cases~~ → rarely
~~in order to~~ →to
~~in relation to~~ → toward, to
~~in respect to~~ → about
~~in terms of~~ → about
~~in the absence of~~ → without
~~in the event that~~ → if
~~in this day and age~~ → today
~~in view of the fact that~~ → because
~~inasmuch as~~ → for, as
~~initiate~~ → begin, start
~~it has been reported by Smith~~ → Smith reported
~~it is apparently that~~ → apparently, clearly
~~it is believed that~~ → I think
~~it is my understanding that~~ → I understand that
~~it is often the case~~ → often
~~it is worth pointing out in this context that~~ → note that
~~it may be that~~ → I think
~~it may, however, be noted that~~ → but
~~join together~~ → join
~~lacked the ability to~~ → could not
~~met with~~ → met
~~needless to say~~
~~new initiative~~ → initiative

[1] Compiled from *How to write and publish a scientific paper (6th ed)* by Day and Gastel, 2006

~~no latter than~~ → by
~~of great theoretical and practical~~ → useful
~~on behalf of~~ → for
~~on the basis of~~ → by
~~on the grounds that~~ → because
~~owing to the fact that~~ → because
~~perform~~ → do
~~pooled together~~ → pooled
~~referred to as~~ → called
~~so as to~~ → to
~~take into consideration~~ → consider
~~the reason is because~~ → because
~~the vast majority of~~ → most, almost all
~~there is reason to believe~~ → I think
~~through the use of~~ → by, with
~~utilize~~ → use
~~we wish to thank~~ → thank
~~whether or not~~ → whether
~~with a view to~~ → to
~~with regard to~~ → concerning, about
~~with respect to~~ → about
~~with the exception of~~ → except
~~with the result that~~ → so that