Reprinted from Investigative Radiology, January 1989
Vol. 24, No. 1
© J. B. Lippincott Co. Printed in U.S.A.

Special Series: Statistics in Radiology

HAROLD L. KUNDEL, MD, *Editor*

The Place of Statistical Methods in Radiology (and in the Bigger Picture)

JAMES A. HANLEY, PhD

In the January issue of Investigative Radiology, we begin a series of invited articles about statistics in radiology. The articles are not intended as substitutes for textbook chapters or consultation with statisticians. Neither are they statistics cookbooks. Rather, we hope that they will become gateways to a better understanding of the use of statistics in laboratory and clinical research. Statistics and its companion medical discipline epidemiology are part of the basic science infrastructure of radiology. Yet, although most radiologists are familiar with fundamental statistical principles, few have gone beyond the subject matter of entry-level courses. As a consequence, our profession lacks statistical maturity—the ability to select the right approach at the right time—and this is sometimes reflected in the reports we publish.[1]

The author of our first article, James A. Hanley of the Department of Epidemiology and Biostatistics at McGill University, has made substantial contributions to our understanding of the statistics of observer performance evaluation in imaging.[2] *However, for this series he was asked to deal with more general issues and has provided some provocative insights into the use of statistical methods. The minimalists among us will appreciate his approach.*

In the future, there will be articles about ROC analysis and cost-effectiveness analysis and, if there is a positive response from the readership, about other topics, as well.

Harold L. Kundel, MD

References

1. Cooper LS, Chalmers TC, McCally M, et al. The poor quality of early evaluations of magnetic resonance imaging. JAMA 1988; 259:3277–3280.
2. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.

Hanley JA. The place of statistical methods in radiology (and in the bigger picture). Invest Radiol 1989;24:10–16.

Key words: statistics, appropriate uses, basic principles; p-values; confidence intervals.

STATISTICAL METHODS are being used increasingly in medical research. Microcomputers and friendly software have made statistical computation much more accessible to the nonstatistician, although some long-time observers consider this increased availability of automated computation a mixed blessing.[1] At the insistence of editors and referees, statistical methods also are being reported in much more detail. The International Committee of Medical Journal Editors, which earlier set down uniform requirements for manuscript presentation, recently added guidelines for presenting and writing about statistical aspects of research.[2] These guidelines have been expanded and explained in a useful companion article.[3]

Despite these generally favorable trends in statistical

From the Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada.

Supported by grants from the Natural Sciences and Engineering Research Council of Canada and Fonds de la Recherche en santé du Quebec.

Michael Kramer, Susan Scott, and Stanley Shapiro provided helpful comments.

Reprint requests: James A. Hanley, PhD, Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, PQ, H3A 1A2, Canada.

Received July 25, 1988, and accepted for publication August 4, 1988.

usage and reporting, there are still many misconceptions about statistical methods; as one author put it, many researchers use them as the drunk uses a lamp-post—more for support than illumination. The purpose of this article, then, is to broadly describe the role of, and the philosophy behind, statistical methods; to indicate what they can and cannot do; to discuss recent trends toward more sensible statistical usage; and to provide references to useful reading material. Just as others have done,[4] I will stay with fundamental inferential statistical principles; I will try to step back from the details and show how statistical methods fit into the "bigger picture." The principles apply not just to radiologic research but to research in general.

What Statistical Methods Are Commonly Used?

The results of an inquiry into what statistical methods are heavily used in general medical journals are reassuring.[5] A reader who is conversant with just descriptive statistics (percentages, means, and standard deviations) has statistical access to more than half of the articles; knowledge of t-tests, chi-square tests, and simple linear regression (topics that are covered in the usual one-semester course in statistical methods) would increase access to more than 80%. These latter methods tend to be used a little more in specialty journals. For example, in the 50 research articles in a recent volume of *Investigative Radiology*, only 30% of the statistical presentations involved nothing more complicated than descriptive statistics; half of the articles used analysis of variance and/or linear regression; the remainder involved comparisons of proportions and assessments of observer performance with diagnostic tests. (Incidentally, several articles compared time curves; statistical techniques for doing so generally are not well covered in a useful way in statistical texts).

I doubt if we need to add substantially to this basic statistical repertoire needed for producers and consumers of medical research reports; rather, it makes more sense to try to understand better what is behind these techniques.

The Purpose of Statistical Methods

When Statistical Methods Can Help

Statistical methods can be used in two ways. First, they can be used as a descriptive tool to quantify variability and to summarize qualitative and quantitative data, be the data from a "universe," or from a sample of it. However, because one usually can study only a sample of the units (eg, cells, organs, or intact patients) in the universe of interest, clinical research must project results obtained in the sample of units that were studied to the universe of "all other or all future units," ie, to the similar units who were not studied. Therefore, the second use of statistical methods is as an inferential tool, to judge the contribution that sampling variability makes to the uncertainty of numerical estimates derived from samples. Statistical laws allow us to quantify the likely (and unlikely) amount of sampling variation that can be present in statistical estimates and to judge them accordingly. The main ways of assessing them are through statistical tests and calculation of confidence intervals. Both use the concept of a sampling distribution; standard errors (or their equivalents) play a central role in this quantification.[6] Even when data are more complex, the same principles apply.[7]

When Statistical Methods Cannot Help

As just explained, statistical methods help to predict and quantify the amount of sampling variation, and therefore the degree of uncertainty, one should and should not expect in estimates derived from samples of the units of interest. (Unless we make repeat measurements on the units, we will not be able to distinguish the true interunit variation from any intraunit variation attributable to biologic variation, random error of measurement, etc. In other words, the interunit and intraunit sources become one overall source of variation.) Therefore, statistical methods allow us, for example, to calculate that large numerical differences between the mean levels of X in two groups are, to put it loosely, greater than those that usually would arise from sampling fluctuation alone. However, in such a situation, they do not provide an explanation for the observed difference; they simply rule out chance (sampling variation) as the sole contributor to the difference. In reality, any observed difference is an unknown mixture of (1) real differences, ie, differences produced by the factor under study (the factor presumably present in different amounts in the two groups); (2) sampling variation; and (3) other extraneous causes that bias the comparison. Proper statistical design and attention to measurement issues help reduce component (2) and statistical methods help quantify that sampling variability that remains. It is the investigator's task to design the investigation and analyze the data so that component (3) also is minimized or, ideally, ruled out. The degree to which this can be achieved depends on the context; as one moves "upward" from the cell to the intact human being, the challenges and constraints of controlling systematic extraneous variation (and even in obtaining reproducible and meaningful measurements) become more difficult. Unfortunately, the magnitude of the threats from these extraneous sources to the validity of a comparison cannot be quantified with the same precision that the cumulative effects of random variation (the grist of statistical laws) can. Therefore, even if sampling variation is minimized, observed differences can be judged only by experts in the subject who understand the other factors that influence the outcomes and the assessments of them. If the observed difference is rewritten as the equation this point becomes more obvious.

$$\underset{(1)}{\begin{array}{l}\text{true}\\ \text{effect of}\\ \text{study factor}\end{array}} = \begin{array}{l}\text{observed}\\ \text{difference}\\ \text{in samples}\end{array} - \left[\underset{(2)}{\begin{array}{l}\text{sampling}\\ \text{variability}\end{array}} + \underset{(3)}{\begin{array}{l}\text{effect of}\\ \text{extraneous}\\ \text{factors}\end{array}}\right]$$

Can Statistical Tests Be Applied to Nonrandom Samples?

The correct answer is, "Yes, provided we take them for what they are." Statistical tests calculate a simple probability in answer to a simple (and usually hypothetical) question: how likely is it that we could observe such a big difference if the only factor operating were (random) sampling variability. Many users and readers, even though they know that the comparison may be a biased one, are impressed by the seeming exactness of statistical procedures and tend to forget the conditional or hypothetical or "what if" nature of this question.[8] Because we usually cannot quantify component (3) in exact numerical terms, we are forced to perform the numerical calculations assuming it is zero. Thus, it is legitimate to carry out the calculations on the hypothesis that the samples were matched and randomly chosen "just to assess how big (or how small) the sampling variation component might have been." However, obtaining the hoped for P value is not an excuse to forget that it was a hypothetical calculation and to forget to consider the other components of the observed difference.

Statistics and the Individual Patient: Clinical Epidemiology

Many clinicians see statistical methods as dealing with characteristics of aggregates or groups and not with the individual. They see these methods helping administrators or others who need to know only the bottom line (be it the total number of dollars spent, the number of items used in a certain year, or the average patient throughput in the radiology department), for which statistical techniques can be valuable: precise estimates of these quantities can be obtained by sampling methods at a fraction of the cost of doing a "census." Moreover, even if the distribution of the individual observations or measurements is decidedly nonnormal, the "Central Limit Theorem" allows the uncertainty of an estimate to be made via a normal distribution. In assessing the individual patient, however, these powerful mathematical laws are of little comfort to the clinician, and many of the methods of statistics, because they do best when dealing with aggregates, seem not to apply. Instead, in diagnosis for example, one must rely on probabilities and on the rules for updating and revising them as more information becomes available. However, groups of patients have to be used to obtain (uncertain) estimates of these probabilities (a cartoon in the Wall Street Journal showed a meteorologist studying his data and saying, "I figure there's a 40% chance of showers, and a 10% chance we know what we're talking about").

In the last few years, those who use probabilities for assessing individual patients and those who use them for assessing group characteristics have moved a little closer to each other. One text,[9] has shown how to apply probability and statistics for the care of the individual patient, emphasizing their use in problems of diagnosis, treatment, and follow-up. Indeed, all of the familiar statistical concepts (chi-square and t-tests, confidence intervals, the binomial and normal distributions) are motivated and introduced in the context of the individual patient. Likewise, the clinical trial, long thought of (and for this reason, often scoffed at) as applying only to group of patients, has found a new application in the management of the individual patient.[10,11] There also have been some excellent texts covering the gamut of clinical epidemiology.[12,13] At least two journals, the *Journal of Clinical Epidemiology* (previously the *Journal of Chronic Diseases*) and *Medical Decision Making*, are devoted to this emerging "basic clinical science of medicine." In a valuable article, one author has made explicit for medical researchers the strong analogy between diagnostic tests and statistical tests.[14]

Cutting Down on Statistical Tests: Using Confidence Intervals

Over the past decades there has been an increasing preoccupation in research reports with statistical tests and P values. This began when journals began to insist that claims of purported differences be backed by significance levels. In part, this is understandable: one would like to dampen the enthusiasm of those who were trying to claim that one treatment will definitely produce a higher response rate than another on the basis of an observed success rate of 67% (two responses in a sample of three patients) with one versus 33% with the other (one response in three patients). The use of statistical testing has now reached epidemic proportions,[1] and abstracts and articles have, as one commentator put it, become adorned with more stars than the Michelin Guide (in their special communication, Browner et al[14] note that in the four original contributions in the same issue [presumably an unbiased sample], "the authors report the results of statistical tests of 76 hypotheses"). Just as with multichannel laboratory tests, when one performs so many tests, one is hard pressed to distinguish the true positives from the false ones. Also, certain journals accept only studies that show positive statistical tests. How does one interpret published findings when a survey of one volume of a journal in the behavioral sciences showed that 105 of the 106 papers had results that were significant at the 5% level?[15]

Fortunately, this trend toward unthinking statistical testing is now being slowly reversed. The comments of Ross, although directed at social scientists, apply just as well to medicine.[16] Thanks to the efforts of an associate editor, the American Journal of Public Health refuses articles that give P values without estimates of the magnitude of the effects found.[17] (The correspondence that followed this editorial decision is enlightening.) Others also have pointed to the folly of "science by P value."[18,19] Several journals have adopted a policy to reduce the emphasis on P values and to increase the use of confidence intervals when appropriate. Most notable is the British Medical Journal,[20–22] which over the past two years also has run a special series on how to calculate confidence intervals for estimates of various parameters.[23]

What Is Wrong with the P Value Approach

The biggest objection to a statistical test is that it answers with a "yes" or a "no" an overly simplistic question: Is there some difference? The emphasis on the significant difference, and indeed the choice of the word "significant," distracts from the real issues, which are how big is the difference and how much of it is likely to be attributable to the factor under study? The P value loses much of its significance when one realizes that (if the sample is large enough) a study can show a statistically significant difference that is clinically trivial or that, although it could not be accounted for by chance alone, could easily be accountable for by biases in selection or evaluation. Even worse, relying solely on the P value makes "not statistically significant" differences (often associated with small samples) even more troublesome to interpret. Whatever the situation, one always needs to examine the location of the point estimate and the size of the confidence interval, which provides a measure of the uncertainty or noise in the estimation process. A "nonsignificant" difference with a wide confidence interval means that "trivial differences" and "certain differences that all would agree are of some clinical import" cannot be distinguished on the basis of this study. On the other hand, all other biases having been ruled out, a "nonsignificant" difference with a narrow confidence interval can be taken as a definitive negative study in the sense that the real difference, even allowing that it could be masked by some sampling error, is trivial.

Definitive Negative Studies

An extreme example of a statistically nonsignificant difference that is also literally nonsignificant is a study of starch blockers performed on five persons.[24] In a double-blind crossover study with controlled calorie intake and a careful calorie-balance technique, the average number of calories blocked was not significantly different when they were consumed with starch blocker tablets (78 kcal) than when they were not (80 kcal). This failure to demonstrate an effect was not attributable to the small sample of 5. To establish this formally, one usually would have to resort to calculations of statistical power, a topic that few investigators and even fewer readers fully understand. A much simpler way is to look at the confidence interval around the measured "blockage" of −2 kcal. It doesn't take a fancy spreadsheet or statistical package, and it does not matter whether Student's t tables with 4 degrees of freedom are entirely appropriate; by any stretch of the truth, and by any best-case scenario, ie, believing the top end of the confidence interval more than the middle, the calories blocked by this technique in this type of situation are unlikely to average more than 10, a far cry from the 400 kcal claimed by the manufacturer. In most clinical settings, such tight control over random and other sources of variation is not possible, and larger numbers (of persons or occasions or both) would be needed to measure the difference with such precision.

Why Do We Calculate So Many P *Values?*

Cynics would answer that it is because that is one of the tasks statisticians are trained to do early in their careers and it is a practice that they perpetuate in the textbooks and statistical packages they produce. (Unlike confidence intervals, p-values usually are output from programs by default.) Statisticians often are eager to calculate probabilities that are much smaller than they should be.[25,26] This preoccupation with statistical testing was sadly brought home to me at a recent statistics conference. After I had presented a simple nomogram for calculation of a confidence interval for a statistical index used in observer agreement studies, I was asked by a new PhD in statistics, "But do you have a test?" Investigators often are to blame, too, and tend to use P values to distract readers from other scientific weaknesses in their studies.

To be fair, a second more legitimate reason is that there are some situations, such as in many nonparametric tests, that by their nature lend themselves more to probability calculations than to estimating meaningful summary parameters.

In Defense of (Some) Small Studies

Two examples will help illustrate how simply using a confidence interval can settle arguments that could not have been settled by a P value approach. The first is given in a valuable paper dealing with small studies.[27] The authors present three small clinical trials, "none of which shows conventionally significant results and none of which, given the present climate of opinion, would be likely to be accepted for publication in their own right." However, only large differences in outcome were relevant; thus, with the appropriate statistical analysis, including calculation of confidence intervals, it was possible to make important practical decisions from them.

A second example was brought to me by a radiologist who had evaluated a nuclear medicine examination as a possible screening tool to exclude patients from a more definitive but long, painful and low-yield neurologic assessment involving spinal anethesia. He assessed 25 patients by both procedures. The screening procedure identified three potential cases, whereas the definitive procedure yielded four cases with the neurologic abnormality of interest. However, none of the four were among the three identified by the screening examination—it "missed" all four. At this point, most readers would agree that this is an insensitive screening test and would not consider it further. Not so the referees and the editor who read his report. Unconvinced, they requested that he either "increase the size of his series or consult a statistician."

Why are they not convinced, when the rest of us (without any formal statistical calculations) are? I believe that it is because of an over-reliance on the *P* value paradigm, an approach that seldom translates the observed result into operational terms. A conventional *P* value approach fails, as does any attempt to compute statistical power. In fact, from one perspective, the above problem was too easy: it's just one sample, there is no null hypothesis.

The key is to consider what Moses[4] calls the "infinite data case." Imagine that the referees had a large series, so large that one patient more or one less in the numerator would be inconsequential. In such a situation, although opinions might vary somewhat, they would have little trouble (even before seeing the data) in deciding what was a minimally acceptable sensitivity; thus the task would simply be to compare the actual performance with the minimally acceptable one. This minimally acceptable performance does not change because the series is smaller, only the uncertainty about the measured performance does. In this example, computing a 95% confidence interval from the observed binomial proportion 0/4, the "best case" or upper limit on the sensitivity is a mere 60%, surely well below what even our defensive referees would consider acceptable. The insight is provided by asking, how sensitive do you estimate the screening test to be? Such a straightforward question deserves an equally straightforward answer: "I'm not sure exactly, but it's almost certainly not more than 60%." There is nothing wrong about not being able to say precisely how sensitive it is; in fact, in this particular situation, it might be unethical to strive for any more precision.

Confidence Intervals: Thinking of Studies as Measurements

The above examples emphasize the concept of measurement rather than statistical testing, Samples should be regarded as providing measurements (albeit uncertain ones) on a parameter. Like any other measurements, they are not entirely repeatable; however, as with any scientific instrument, the repeatability (or conversely their uncertainty) can be estimated. The opinion polls are a good example of this; indeed, the uncertainty statements (in the form of confidence intervals or "margins of error") that now accompany most polls are written in such clear layman's language that the scientific community would do well to copy them. Using confidence intervals would emphasize that the point of publishing the results in your sample of patients is only as a guide or projection to what other readers can expect with their patients. We hear readers quote, with seemingly great precision, that "76.2% of the patients responded" when the response was 16/21. What is important to the reader is not that the observed proportion was exactly 76.2% but rather that this observed proportion projects to somewhere between 53% and 92% (if using a 95% confidence interval) of "patients elsewhere." Remembering just the 76.2% is like putting a very exact mark on a graph when in reality a thick felt tip marker, or possibly even a paintbrush, was all that was needed. The real meaning of the 76.2% is that the percentage probably is between 60% and 80% or 'at worst' somewhere between 50% and 90%, ie, that the chances for an individual patient are almost certainly better than 50:50. (Of course, even after that, the reader still has to decide whether the patients the investigator studied are sufficiently like his that he can even make the projection.) This idea of projecting to a universe is made in the nomograms such as those illustrated in the confidence charts of Ingelfinger et al,[9] or in statistical tables, or in the article dealing with inferences from a zero numerator.[28]

Beyond the Single Research Study: The Even Bigger Picture

Nowhere does the role of statistical methods in the bigger picture become more apparent than when the results of several studies dealing with the same question are abstracted and listed in a review article. While most of the statistical techniques in a single article ideal with the internal validity of the comparisons, the results of different studies can be quite disparate (much more than would have been projected from the standard errors). Recently, some investigators have begun to mathematically analyze, and sometimes combine, the results of several studies in an exercise that had been termed "meta- analysis." Whether or not one agrees with this mathematical analysis of what some would call a literature review, the process does emphasize that any one study, no matter how seemingly precise, is still just one study. As one professor tells his students, "The world did not begin with your study, and it will not end with it." Scientific knowledge is accumulated slowly through the sifting of patterns and outliers. To this end, a full description of the setting and the methods of each study can be very valuable. Mainland's plea[1] for less derived and more raw data (although it referred to the data of individual patients) also is relevant at a study level: "I thought of the loss to other readers who wished to form their own opinions from the recorded observations, perhaps to answer questions not raised by the authors, and to seek exceptions and individual peculiarities, so fundamental in medicine."

Concluding Remarks

The intent of this article is to encourage more rational use of statistical methods. One does not need a large statistical repertoire but rather the knowledge and the confidence to use the basic techniques wisely and sparingly. Above all, the techniques should be used to communicate, and not, as is so often done, to obfuscate or to seem erudite. I hope readers find that the references help with this communication.[29]

Annotated References

1. Mainland D. Statistical ritual in clinical journals: is there a cure?-I Br Med J 1984;288:841–843.
   Some sobering reflections on the current state of statistical presentations in medical journals and a plea for "more thinking and less arithmetic."
2. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. Ann Intern Med 1988;108:258–265.
   Fifteen statements.
3. Bailar JC, Mosteller F. Guidelines for statistical reporting in articles for medical journals: amplifications and explanations. Ann Intern Med 1988;108:266–273.
   An expansion of the above; also contains material to help investigators in early stages make critical decisions about research approaches and protocols.
4. Moses LE. Statistical conbepts fundamental to investigations. N Engl J Med 1985;312:890–897.
   Very valuable overview of statistical principles. No formulae.
5. Emerson JD, Colditz GA. Use of statistical analysis in the New England Journal of Medicine. N Engl J Med 1983;309:709–713.
   Surprising results of survey of statistical usage in a general medical journal. This article, along with that of Moses, and eighteen other articles on statistics in medical research, have been compiled into a valuable single publication, "Medical Uses of Statistics," 1986, 424 pages, edited by Bailar JC and Mosteller F. Available from NEJM Books, Waltham, Massachusetts. It is regularly advertised in the New England Journal of Medicine.
6. Brown GW. Standard deviation, standard error: which 'standard' should we use? Am J Dis Child 1982;136:937–941.
   One of the better attempts to distinguish and demystify these two quantities.
7. Hanley JA. Appropriate uses of multivariate analysis. Annu Rev Public Health 1983;4:155–180.
   An overview, mostly without formulae, of the logic of multiple regression and other multivariate techniques (for which multivariate is defined as an analysis "involving three or more variables").
8. Mainland D. Statistical ritual in clinical journals: is there a cure?-II. Br Med J 1984;288:920–922.
   Deals with the uses and abuses of many of the commonly used techniques.
9. Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH. Biostatistics in Clinical Medicine, ed. 2. New York: Macmillan, 1988;316.
   A novel approach: Uses management of single patient to explain statistical concepts and does it well.
10. Guyatt G, Sackett D, Taylor DW, Chong J, Roberts R, Pugsley S. Determining optimal therapy: randomized trials in individual patients. N Engl J Med;314:889–892.
11. McLeod RS, Taylor DW, Cohen Z, Cullen JB. Single patient randomized clinical trials. Lancet 1986;1:726–728.
12. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little Brown and Co., 1985.
13. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology: the essentials, ed 2. Baltimore: Williams and Wilkins, 1988.
   Chapters on abnormality, diagnosis, frequency, and prognosis.
14. Browner WS, Newman TB. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. JAMA 1987;257:2459–2463.
   Highly recommended; makes the most of the connection; gives a greater understanding of what the P *value is and is not.*
15. Sterling TD. Publication decisions and their possible effects on inferences drawn for tests of significance—or vice versa. Journal of the American Statistical Association 1959;54:30–34.
   These data are quoted in the paper "Publication Bias," by Begg and Berlin, soon to appear in the Journal of the Royal Statistical Society.
16. Ross J. Misuse of statistics in social sciences. Nature 1985;318:514.
   Some harsh truths about P *values and science. "The hypothesis of no effect is the correct one for improbable claims of psychokinesis or water divining, for example. But it is not correct when we have a body of knowledge to draw on and a theory designed to predict not just that something will happen but what and how much."*
17. Rothman KJ, Yankauer A. Confidence intervals vs. significance tests; quantitative interpretation (editors' note). Am J Public Health 1986;76:587–588.
   One of the first journals to take a stand against P *values.*
18. Rosenthal R, Rubin DB. Statistical analysis: summarizing evidence versus establishing facts. Psychol Bull 1985;97(3):527–529.
   Two views of how science progresses.
19. Rothman KJ. Significance questing. Ann Inter Medicine 1986;105(3): 445–447.
   Easy to understand; difficult to disagree with.
20. Langman MJS. Towards estimation and confidence intervals. Br Med J 1986;292:716.
   Launching the British drive toward "measuring rather than testing."
21. Gardner MJ, Altman DG. Confidence intervals rather than *P* values: estimation rather than hypothesis testing. Br Med J 1986;292:746–750.
   Expanding on why the change in editorial policy.
22. Gardner MJ, Altman DG. Estimating with confidence. Br Med J 1988;296:1210–1211.
   Introducing further articles in the Statistics in Medicine series.
23. Br Med J 292:746–750; volume 296:1238–1242, 1313–1316, 1369–1371. See also the earlier BJM series "Statistics and ethics in medical research," which ran from 1 November 1, 1980 to January 3, 1981, and the article "Statistical guidelines for contributors to medical journals in the issue of May 7, 1983."
24. Bo-Linn GW, Santa Ana CA, Morawski SG, Fordtran JS. Starch blockers: their effect on calorie absorption from a high-starch meal? N Engl J Med 1982;307:1413–1416.
   One of the more convincing small studies I have encountered. Of course, in line with comments in the text, one also should assess the appropriateness of the experimental setup to mimic real-life use.
25. Hanley JA. Lotteries and probabilities: three case reports. Teaching Statistics 1984;6:88–92.
   Statisticians sometimes jump to co-incidences.
26. Samuels SM, McCabe GP. More lottery repeaters are on the way. New York Times 1986; Feb 27:A22.
   Some p-values are just so small that they must have been incorrectly calculated.
27. Powell-Tuck J, MacRae KD, Healy MJR, Lennard-Kones JE, Parkins RA. A defense of the small clinical trial: evaluation of three gastroenterological studies. Br Med J 1986;292:599–602.
   Sensible approach to small studies that gave an answer that was precise enough for the purposes at hand.
28. Hanley JA, Lippmann-Hand A. If nothing goes wrong is everything all right? Interpreting zero numerators. JAMA 1983;249(13):1743–1745.
   Uses an extreme case to show how "the confidence interval translates the result of a single sample not into a single number, but rather into a range that is quite likely to contain the rate that is characteristic of the population." Makes the link between confidence intervals and statistical tests.
29. Additional reading:
   Swinscow TDV. Statistics at square one. London: British Medical

Journal Publishing, 1983;86.

Cookbook approach; based on series of articles in BJM.

Castle W. Statistics in small doses. Edinburg: Churchill Livingstone, 1976;220.

As the name suggests.

Colton T. Statistics in medicine. Boston: Little, Brown and Co. 1974;372.

A good introductory textbook for those who want more than just a cookbook.

Armitage P, Berry G. Statistical methods in medical research , 2 ed. Oxford: Blackwell, 1987;559.

Revised edition of a standard text on medical statistics. Comprehensive; many beginners find it difficult but appreciate it more after they have covered Colton.

Norman GR, Streiner DL. PDQ statistics. Burlington, Ontario: B.C. Decker Inc. 1986;172.

From introductory statistics to multivariate methods in 160 pages; irreverent; amusing glossary; Sometimes oversimplified or even misleading (as when describing nonparametric statistics) but a useful overview.

Glantz SA. A primer of biostatistics, 2 ed. Oxford: Blackwell, 1987;379.

A slightly different approach; some good diagrams; introduces inferential procedures, and particularly statistical tests, very early on. Good chapter on statistical power.

Kong A, Barnet O, Mosteller F, Youtz C. How medical professionals evaluate expressions of probability. N Engl J Med 1986;315:740–744.

Interesting survey of the words and phrases physicians use to describe (un)certainty.

Feinstein AR. The *t* test and the basic ethos of parametric statistical inference. Clinical biostatistics article LV. Clin Pharmacol Ther 1981;29:548–560.

"A colleague who wanted to know more about the basic essentials of statistical principles challenged me to compose a simple, clear explanation of how the t-test is mathematically derived, how it works, what it accomplishes, and how it is used or abused." The explanation runs to many pages, but is worth it. Many of the articles in this series were later reprinted as a text called Clinical Biostatistics, by A. R. Feinsten, St. Louis: C. V. Mosby Co., 1977.