Journal of Scientific Exploration, Vol. 17, No. 4, pp. 593–600, 2003 0892-3310/03

COLUMN

Inference and Scientific Exploration

MIKEL AICKIN

Center for Health Research
Kaiser Permanente Northwest Region
Portland, OR

Statistical inference is a bane and a boon to empirical science. On the one hand, it supplies a link connecting the partial, limited data of individual study with the much larger world we want to know about. On the other hand, it is a slippery, unsatisfactory, tedious area of activity, being susceptible to contentious debate and the generation of misleading conclusions.

My intent in this column is to open the doors of discourse on both the blessings and curses of statistical inference, with particular attention to how they have an impact on the type of science that appears in *JSE*. After observing the uses of statistical reasoning in a wide variety of disciplines over the past thirty years, I have come to the conclusion that most scientists acquire their statistical lore more from the pages of the journals they read than from any textbook or college lecture. This would seem to suggest that every disciplinary journal should have some sort of regularly appearing statistics section to ensure that the methods used in that discipline do not become insular and idiosyncratic. Very few journals actually do.

The absence of such sections has taken its toll. A number of years ago an accountant friend of mine explained to me that there was something called GAAP (generally accepted accounting principles). I had always thought that accounting was simply a matter of toting up income and expenditures, with little left to discretion. To the contrary, the variety of different conventions one can set up in describing a complex financial entity means that it can appear either healthy or sick, depending on how one wants to look at it. The GAAP more or less induces all accountants to use the same assumptions, so that readers of financial reports can have some confidence that they understand them. Many of the scientists I have interacted with behave as if there were a GASP (generally accepted statistical principles) that plays the same role for inference as GAAP does for accountancy. The overwhelming problem is that each of these scientists believes that the conventions of statistics in their particular discipline are what defines GASP—and of course these different canons for inference tell mutually inconsistent stories. There is no GASP.

There are good reasons, however, why scientists would want a GASP, even if it were only generally accepted in their own discipline. When they design and carry out experiments, they want to report the results and interpret their meaning. They do not want to get into debates about whether the results are "real" or due to "chance", which is exactly the kind of discussion that statistical inference threatens to lead us to. If one just had a conventional set of inferential principles that could be applied uniformly, then certainly the fruitless statistical battles could be avoided.

An example, although perhaps an excessive one, is epidemiology. This is the study of the relationships between diseases and their potential causes. One classical statistical device is the "2-by-2 table", a square arrangement of four cells containing numbers of people who do or do not have the disease (the rows), and who were or were not exposed to the putative cause (the columns). So long as the disease and cause are both clearly either present or absent, then the 2-by-2 table is appropriate. Moreover, it is an extremely handy method in epidemiology, since the conventional statistical association measure (the "odds ratio") applies both to prospective and retrospective studies, the main two designs used in analytic epidemiology. The perversity came in when epidemiologists started taking measures of the cause (such as blood pressure in mmHg, or serum cholesterol concentrations in mg/dl) and dividing them into the high and low groups. The only evident reason for taking a perfectly good continuous measure and degrading it to a high/low categorization is to make it possible to cram the data into a 2-by-2 table. This strategy exonerates the epidemiologist from ever having to think about statistical issues—and indeed one can argue that this is the real reason for the convention. The articles pointing out the inferential problems caused by mindless categorization appear in statistics journals, unread by epidemiologists.

A number of more amusing examples of this phenomenon appear in various social science journals. One occasionally sees a "methodologist" publishing a collection of "guidelines" for carrying out statistical procedures. The implication is often that potential authors had better follow the guidelines if they want to be published in that journal. They would be more accurately called "dogmatic rules", of course, but this phrase would fail to disguise the iron fist within the velvet glove. It should go without saying that these guidelines contain a good deal of common sense, which the would-be authors are well-advised to follow. To the professional statistician, however, they also contain annoyingly large amounts of misconception and bad methodology. A set of guidelines is never a substitute for discourse and explication.

And that returns me to what I would like to accomplish in this column. First, I would like the discourse to be reader-driven. For me (or others) to write about what vexes them the most would no doubt be therapeutic for us, but perhaps unlikely to benefit *JSE* readers or authors. Therefore, I welcome questions, comments, raising of issues—anything that leads to a fruitful discussion of inferential problems. I

will do the best I can to respond dispassionately and even-handedly, and to turn the column over to others when the topic goes beyond my expertise.

Secondly, I would like the content to remain relevant to *JSE* research. There are, to be sure, many topics that are completely general, but also important to novel or unconventional research. There are also specialized issues that are far more salient to this journal than to any other scientific journal now being published. It is, therefore, hard to find good discussions of them anywhere: here is where they belong.

Thirdly, I think it is a good idea to avoid dogmatism. I have my own set of beliefs about inference, some weakly and some strongly held, but I don't see this as the place for me to try to impose them on other people. In the process of honing my own opinions I have had to understand the arguments of those who think differently, and so I believe that I can (for the most part) fairly portray procedures that I do not prefer. Statistical inference is a language that stretches across the entire vast landscape of empirical science, and so it is inevitable that, like any other language, it will develop dialects. So long as they can be rationally argued, they deserve a place here.

The Implications of Method

An interesting theme that runs strongly through "hard" science is that statistics don't mean very much. Particularly bench-scientists tend to believe that if you need a statistical test to prove something, then it probably wasn't real in the first place. In doing this, they succumb to their own disciplinary GASP: you don't actually need statistical inference at all. If one looks at the kinds of science these people do, one finds that the natural variability is small (or sometimes artifactually removed), so that the results appear clear-cut. Their attitude fits what they do. The problem is, of course, that their GASP doesn't fit what others do.

In most areas of science nowadays the natural variability in the measurements virtually requires that the results of a study be interpreted, at least to some degree or another. It is variability, more than any other factor, that drives the need for statistical inference. Whether a result is "statistically significant" is largely determined by how much unwanted variability the measurements contain. Consequently, for most scientists there is simply no way to avoid the necessity of adopting a method for converting noisy results into assertions or conclusions. Because GASP doesn't exist, their choices are not dictated, and can be subjective or even arbitrary.

Why does any of this matter? In the biomedical sciences, where I work, the results of research have an influence on what physicians believe, and therefore on what they practice. The results often have an impact on the beliefs of administrators of hospitals and health plans, which in turn determine practice guidelines or health care policy. All of these beliefs-translated-into-actions have consequences for people who need care. If the results are wrong, or misleading, then the untoward consequences multiply rapidly and widely through the

mechanism of policy. We usually think of malpractice as a concept that applies to a particular, incompetent physician, but in actual fact an incompetent researcher has the power to do infinitely more damage to people's health than any individual physician does.

What does this have to do with *JSE*? The answer is that in many areas of science the issue is a question within the context of that science, but not whether the science itself has any basis. One can argue for or against a particular drug on a number of grounds, but no one tests the hypothesis that drug therapy in general is useless. In many frontier areas of science, the reverse is true: the question is whether the underlying scientific concept has any validity. Psi phenomena are notoriously difficult to replicate reliably, for example. There are studies which suggest that the beliefs or intentions of the research investigator might influence the results of their studies—not through overt manipulation, but through some unknown mechanism. Evidence exists for some means through which human beings can influence physical random number generators. These are far from conventional science, and no doubt many disbelieve they are real. If the methodology suggesting that they are true is responsible for misleading results, then the effort pursuing them has been wasted, and even worse, the errors of the past will be relived as future researchers continue the pursuit. On the other hand, researches by skeptics that cast doubt on such phenomena, when misleading, discourage research paths that should be taken. While issues like these are not life-threatening, they should be of concern to *JSE* readers nevertheless.

Does choice of method really have much impact? Let me give one example that makes several points. In comparing two conditions, A and B, a ''cross-over design'' assigns half of the participants to receive them in the order AB, the other half in the order BA. The great advantage of this design is that A-B comparison is made within individuals, which is far more efficient than making comparisons between individuals. The great drawback is the possibility of ''carry-over''. In the AB group the comparison is between B-preceded-by-A versus A-preceded-by-nothing, while in the BA group it is between A-preceded-by-B versus B-preceded-by-nothing. Because these may not be parallel experiments, it might be dangerous to mix them. For a very long time conventional statistical wisdom said that cross-over designs should be analyzed as follows. First, test for an interaction between condition and order of administration. If you confirm no interaction, then conclude the two versions of the experiment are parallel, and simply ignore treatment order. If you find interaction, then compare the A-first group with the B-first group, ignoring the second stage. On the surface, this seems quite sensible. If there is no interaction, you get the benefit of the efficient cross-over design, and if there is, you drop back to the less desirable, but still valid two-group comparison.

It was not until rather recently that the problem with this design was published. The difficulty arises because the interaction test generally has inadequate statistical power. This means that it too often fails to find the

interaction, even when one really exists. Thus, the researcher will too often be led to pool the data (ignore order of treatment) when the experiments are not actually parallel. This can easily result in a biased result (one that systematically gives the wrong message). In other words, the conventional statistical analysis promulgated for more than two generations promotes bias. When this was published, all of the literature containing cross-over designs became suspect overnight. Who wants to go back through all the articles to see whether they had policy implications that might be harming people? Who wants to cast doubt on a whole area of frontier research because the early studies were cross-overs?

The second point is, how could this happen? The original advice came from highly respected professional statisticians. It was published in top journals, and appeared endlessly in textbooks and applications. This leads to a very important fact about the production of statistical knowledge in our society. Professional advancement in statistics comes from publishing new methods, or improvement of existing methods—the more mathematical, the better. It does not come from debunking well-accepted methods published by famous people, particularly if you have to stoop to brute-force methods, such as simulations. Very few young statistical researchers will set themselves on a course of examining the shibboleths of their profession, when it is both easier and more lucrative to find as small an area as one can, no matter how specialized and remote from actual application, and devote oneself to becoming an expert there. So the simple fact was that no one saw enough professional reward from revisiting the cross-over design, until someone finally stumbled on the problem through chance.

The third point is inertia. Even though it has now been over a decade since the article uncovering the cross-over problem was published, this result is still almost completely unknown among biomedical scientists. One still sees cross-over designs in grant proposals, with no hint that dealing with carry-over might be a problem. A recent volume on clinical research in complementary and alternative medicine recommended the cross-over design as a method of choice in CAM studies, without mentioning either the conventional method of analysis, nor the carry-over problem.

But the situation is actually even worse. The fundamental point of the cross-over example is that using a statistical test of little power, to decide which model to use for the primary analysis, promotes bias. The reason is that in most cases using a model that is too simple results in bias, and the powerless test leads to using a too-simple model too often. This basic inferential strategy is nonetheless very widely used throughout all sciences. For example, I and some colleagues recently submitted an article to a leading journal, in which we were told that we could not present results by age, gender, or ethnicity subgroups, unless we first performed interaction tests to demonstrate that the within-group results differed. Since our sample size barely gave us enough power to detect within-group main effects, we clearly had too little power for the interactions. The editor was simply requiring us, out of his/her concept of GASP, to promote bias in our analyses.

(Moreover, he/she was also requiring us to violate NIH guidelines, which stipulate that results must be reported by subgroups—especially gender and race—to make it possible to cumulate them in meta-analyses.) The sad fact is that the news on improvements of statistical methods gets out very, very slowly. Most textbooks are at least one full generation out of date on the day they are published.

Errors and Opportunities

The main kinds of errors resulting from methodology can be roughly classified as overt, subtle, and trivial. Overt errors are those that can be discerned from the published article itself. A common example is to treat observations that are probably inter-correlated as if they were independent. This is not done by asserting the improbable lack of correlation, but instead by using a statistical procedure whose validity depends on the observations being independent, without ever alluding to the fact. Another overt error occurs when the data in the table or figure do not match the p-values that announce statistical significance, strongly suggesting that something is wrong. It is very common in some journals nowadays to encourage displays or tables that actually hide important features of the data, which I would regard as an overt error.

Trivial errors probably occur very frequently. They often consist of not following some reasonable rule, or using a statistical procedure under conditions that depart from its usual assumptions, the validity of which has never been researched. These occasions are often the result of sloppiness or ignorance, but they do not result in a misleading message. Many times the results were sufficiently overwhelming that it would be difficult to misuse a statistical test badly enough to obscure them.

The subtle errors are the most troubling. Since they are not overt, they cannot be detected in the publication, and since they are not trivial they can mislead. In my opinion the most prevalent of the subtle errors has to do with selecting the specific results to present, while (artfully or obliviously) not presenting other results. In studies of any complexity, there are many variables, and many possible scenarios for linking them together in analyses. There are also plenty of computer packages for performing a blizzard of different procedures. In my experience almost all scientists investigate many variables, and try varying procedures, in order to tease-out the fundamental story that they are sure their data are trying to tell them. They have no trouble recognizing the correct message when it emerges. Their publications leave no doubt of the straightforward, direct path that led to their conclusions. The fact that their data also had many other stories to tell, that many of these stories had statistical claims equal to the one that was published, and that some of those stories contradict or diminish the published story, all these facts lie about in wreckage after the passage of the methodological tornado.

In my opinion, frontier research is particularly vulnerable to selection effects. Since they are universally under-funded, frontier projects will often arise from

observations made in a more casual mode than is usual in expensive studies. Observations may be gleaned over a period of time, and when they seem to have reached the point that they have a story to tell, they are analyzed and sent off for publication. The fact that there are other notebooks, with other gleanings, which have not borne fruit, is important for interpreting the analyses that are published, but of course there is no section in the modern research article for the author to give a description of the context in which the project was conceived. In our mass publication society, the sound-bite research article leaves no room for the kind of detail and completeness that were the hallmark of scientific articles of the nineteenth century. One might argue that this is not a statistical issue, but it is an inferential issue.

It is common for statistical critics to focus on errors, but very few make mention of an equally important topic, lost opportunities. Very many researchers believe that they have enough grasp (or maybe GASP) of statistics to do their own analysis. Even when this is true, there is a hidden problem, that they will design their research studies to fit within their limited repertoire of analyses. Perhaps this is nowhere more evident than in psychology, in which one sees experiments crammed over and over into the MANOVA (multivariate analysis of variance) framework, irrespective of whether there are other designs that would have worked better. Epidemiologists do the same thing with regard to 2-by-2 tables. The irony is, that on one side of the room we have statisticians constantly trying (occasionally succeeding) to devise new, elegant, efficient designs, with their corresponding analyses, to help scientists push their research programs forward, and on the other side of the room we have the scientists themselves, using outdated and simplistic designs, denying themselves the pleasures of reaping greater scientific rewards, but not the pleasure of denying the statisticians employment. That statisticians can be perverse and unhelpful at times certainly does nothing to improve the situation, but regardless of the ultimate causes, loss of opportunity to do excellent science is something we ought to struggle against.

To Summarize

The purpose of this column is to discuss issues of statistical inference, defined broadly. This includes topics such as the validity of specific statistical procedures, answering questions about the meaning or interpretation of jargon terms and mathematical concepts, reporting on the spectrum of expert advice from other literature on thorny problems, advertising new methodological developments, and (in the absence of reader response) long, philosophical ruminations.

As an example of a technical issue, in comparing pre-post changes on some outcome measure, should one (1) use the standard t-test, (2) use the "nonparametric" alternative—the signed rank test, (3) use an even more nonparametric alternative—the sign test, (4) use analysis of covariance, or (5)

use an exact permutation test? Most disciplinary GASPs press one toward one or another of these approaches, but I maintain that you cannot tell what to do without considering the context, and that reasoned discourse on how to do this is worthwhile. On a more strategic issue, when presenting multiple hypothesis tests, should one be compelled to adjust the p-values, and if so, how? On a design issue, sometimes it is impossible to randomize participants to treatment conditions—can anything be done about this? Because I am a hands-on biostatistician, I would also welcome data sets, with queries about analysis that are specific rather than general.

As I hope is now obvious, I do not intend this column to be a tutorial, nor to recapitulate material that can be found well-covered elsewhere. It's also not my intention to produce a GASP for *JSE*. What I do want to offer is an awareness of the principles of inference, which can and should be argued as part of the presentation of frontier science results. The focus that I would like to maintain throughout is on the aims, strategy, and tactics of inference in scientific exploration, dedicated to the service of the readership of *JSE*.