COMMENTARIES

Commentaries are informative essays dealing with viewpoints of statistical practice, statistical education, and other topics considered to be of general interest to the broad readership of *The American Statistician*. Commentaries are similar in spirit to Letters to the Editor, but they involve longer discussions of background, issues, and perspectives. All commentaries will be refereed for their merit and compatibility with these criteria.

The Religion of Statistics as Practiced in Medical Journals

DAVID S. SALSBURG*

The way in which hypothesis tests are often used as the sole tool of statistics in medical research is satirized. This is followed by a suggestion for reform.

After 17 years of interacting with physicians, I have come to realize that many of them are adherents of a religion they call *Statistics*. It bears some resemblance to the mathematical theories and practices of statistics as described in journals like this one, using many of the same words, but it reflects activity in only a small portion of the statistical world—the use of hypothesis tests. To the physician who practices this religion, Statistics refers to the seeking out and interpretation of p values. Like any good religion, it involves vague mysteries capable of contradictory and irrational interpretation. It has a priesthood and a class of mendicant friars. And it provides Salvation: Proper invocation of the religious dogmas of Statistics will result in publication in prestigious journals. This form of Salvation yields fruit in this world (increases in salary, prestige, invitations to speak at meetings) and beyond this life (continual references in the citation indexes).

There are some who manage to publish without Statistics. They calculate averages and show plots of these averages versus doses or bar charts of averages for various subgroups. As long as they avoid the magical word "significance" and make no attempt to state that one group is different from another with respect to anything other than averages, there are many editors who will allow publication without invocations to the gods of Statistics. There is a middle ground, too, where it is considered good taste to publish these averages along with a symbol that looks like this: (± 43). It is not necessary to identify the meaning of this appendage, though some writers refer to it as the "S.D." and others as the "S.E.," two apparently interchangeable labels. Most reviewers and readers appear to ignore them, anyway. The few reviewers who do examine them appear to prefer small numbers after the $\pm$, so it is useful when using a standard packaged computer program (more on this later) to pick the smallest of such numbers if several are offered.

*David S. Salsburg is Research Advisor, Department of Clinical Research, Pfizer Central Research, Pfizer, Inc., Groton, CT 06340.

RELIGIOUS PRACTICES

The previous paragraph is all I intend to devote to nonpractitioners of the religion of Statistics, for I find the religion itself a fascinating study. The practitioner engages in a ritual known as "hunting for p values." He manipulates data derived from an experiment according to a set of apparently irrational rules. To do this, he can use standard packages of statistical routines (such as BMDP and SAS) on mainframe computers, smaller packages on desktop computers (often written by local computer programmers who are equally ignorant of the theories of statistics), or predeveloped keys on hand calculators; or he can even resort to tally sheets with formats prepared by somebody else from the arcane mathematical formulas found in elementary statistics textbooks. Once the calculations are completed, the computer output (or the screen of the minicomputer, the crystal liquid diodes of the calculator, or the final line of the tally sheets) will present the practitioner with a p value.

At this point, the practitioner must be prepared to suffer the wrath of the angry gods of Statistics. If the p value is bigger than .05, he will not be allowed to publish. It may even mean running another experiment. If he is clever, the practitioner may find ways to modify the original data (leaving out numbers that are obviously wrong is the most common practice) and invoke the gods again. The gods are a bit stupid. Even if you run various modifications of the data through the same computer program again and again, the gods never catch on and keep presenting you with new p values. Sometimes, however, no manipulation of the data short of outright fraudulent misrepresentation will produce a p value less than .05. The sensible practitioner will remember that we live in an unfair and irrational world and accept his defeat. Salvation comes to an elect few, but the religion is not unrelenting. Perhaps it will present the practitioner with a p value less than .05 for the next experiment.

RELIGIOUS FIGURES

There are a body of priests for the religion. I am one of them, as are many other readers of this journal. One seeks out the priests at one's own peril. First, there are very few of them, so it is often hard to get an appointment. Second, the priest will usually do his best to confuse the issue. He will ask irrelevant and unexpected questions, such as, "Why did you run the experiment to begin with?" He will usually not understand the urgent need for Salvation and will try to

involve the practitioner in an arcane theological discussion. It is better to leave the priests to discuss among themselves how many angels can dance within the limit of a sequence of nested sigma fields.

More useful than priests are the mendicant friars. Any important researcher will usually have a large number of subordinates (called graduate students, interns, or fellows). At least one of those subordinates can often be induced to study elementary statistics textbooks or learn how to run packages like SAS and BMDP. Graduate students have usually taken a vow of poverty and penitence, and some of them are so fond of the hair shirt that they will eagerly take to such tasks. The important researcher can now depend on his own personal "statistician" to rummage through his experimental data and search for p values.

Actually, the search for p values can be an exciting ritual. In clinical trials, it is often possible to run the trial long enough so that most measures will change over time. Then in spite of the fact that the study may be a controlled parallel study of two groups, one can run something called "paired-t" tests on differences from baseline to final. This will usually produce a rich collection of small p values, often making the article acceptable for publication even if the p values that result from comparing the two treatments never reach the magical .05. If the practitioner runs such paired-t tests, however, generally accepted procedures require that he keep the published information to a minimum. He is allowed to show the averages and standard deviations (or standard errors) of the baseline values and of the final values, but he is not allowed to show the standard deviation of the differences. It is also not considered good taste to print the exact value of the t statistic (because no one will look at it, and because a passing priest might be able to compute the standard deviation of the paired differences from it). Nor is it appropriate to publish the exact p value. Instead, one should use the deep mysterious symbols of the religion, NS, *, **, and (mirabile dictu) ***.

ANALYSIS OF VARIANCE

When it comes to comparing more than two groups in an experiment, the search for p values is a much chancier thing. Suppose that we have three or more groups, for instance. The ritual for this type of festival is called "Analysis of Variance." However, finding out if two groups differ "significantly" with this ritual is difficult. Most computer programs print out "Duncan's Multiple Range Test." This is best taken at face value. There are other options called "Contrasts" but the computer program often asks for something else when you type in "contrasts." If you give it a set of numbers that are as many as there are groups, you will get a p value. If the journal editor asks what this means, it is best to go back to Duncan's multiple range test.

There is something else called analysis of variance that asks for blocks and treatments. It is dangerous to use this one, however, because you can get a small p value that is associated with "interactions." This is one case in which a small p value brings anything but Salvation. Editors scream that your experiment is no good. Having a significant interaction is a little like eating chicken with your fingers in public or wearing track shoes to a wedding. Somehow it is all your fault, and you are not quite sure what you have done wrong.

There is an important test that does not fit into any of these categories but is widely used in medical journals. This is called the "Chi-Squared Test." It is an omnibus test, and it is not necessary to accompany its use with any numbers. One merely writes, "The effect was significant ($p < .05$, chi-squared test)." Most practitioners who use this test go out of their way to be vague about what effect was tested under what circumstances.

HERESIES

There are certain heresies that occur in this religion. One of them is the Neymanian heresy. It is firmly established by almost all practitioners that one should use "two-tailed tests" because they are conservative (whatever that means). Those who practice the Neymanian heresy will sometimes use "one-tailed tests" by thinking up good reasons, after the fact, why treatment A should be better than treatment B. To the more orthodox practitioners this is a very suspicious procedure, since it converts unacceptable p values, such as .08, to p values capable of providing Salvation.

Finally, we should consider the subclass of practitioners who are "more holy than the Pope," so to speak. To these practitioners, the whole purpose of the religion of Statistics is to maintain the sanctity of the alpha level (which is another name for .05). No activity that appears to involve looking at the data for sensible combinations or for interesting effects is allowed. It is forbidden, in fact, to do anything more than compute the p value using a method determined in advance of the experiment and fully documented at that time. An example of this subculture occurs in the "intent to treat" paradigm for medical trials (Sackett and Gent 1979). Here one randomly assigns patients to one of two or more groups and follows them for a period of time. The act of randomization invokes the gods of Statistics. Exactly what treatment is given to each patient is irrelevant. If a patient is assigned treatment A and drops out of the study because treatment A had serious side effects, that patient is followed to the end of the period of time and all measurements from that patient are assigned to treatment A. If the pharmacy made a mistake and gave bottles of treatment B to a patient assigned to treatment A, then the numbers accumulated on that patient are assigned to treatment A. No accident of fact is allowed to interfere with the computation of a p value, lest it cast a shadow of doubt on the sanctity of the alpha level.

SERIOUS CONSIDERATIONS

In spite of the flippant tone of the previous paragraphs, I am convinced that there is a serious problem here, not only for statistics but for medicine. The emphasis on hypothesis tests and on the absoluteness of specific alpha levels (such as .05 and .01) have distorted the interpretation of clinical trials.

Great effort has been expended on the planning and execution of large-scale studies in the U.S. and elsewhere (e.g., CDPA Research Group 1976, MRFIT Research Group

1982, and PARIS Research Group 1980), and the result has been the accumulation of banks of well-structured data from carefully controlled and followed experiments. However, the published articles put the greatest, and often the only, emphasis on a clutter of significance tests, slicing this way and that way through the data, hunting for the rare mice of acceptable p values. In some cases (CDPA Research Group 1976), the influence of good biostatistical advice appears when the descriptive significance levels are displayed and the article is careful to note that the exact interpretation of these p values must be adjusted because of the multiplicity of testing. In none of these studies, however, will the authors cast off the awkward cloak of hypothesis testing and treat the data as an exercise in estimation of parameters and the identification of reasonable subsets of patients. One result has been a perceived failure of these studies to influence medical practice (Banta et al. 1983).

CONSUMERS OF INFORMATION AND THEIR NEEDS

If one thinks about it, there are two major consumers of the information derived from these clinical studies, the practicing physician and the formulator of public health policy. To the practicing physician, there is little value in knowing that treatment A is significantly better than placebo. He needs answers to questions like the following:

1. If a patient is going to respond to treatment A, how long will it take for the response to manifest itself and what can be monitored to know whether such a response has occurred?
2. What patient characteristics are there that will identify patients most likely to respond and patients most likely to suffer adverse reactions? (This assumes, of course, that there are some patients who will respond, so it might be appropriate to apply a preliminary hypothesis test before chasing down will-o'-the-wisps via regression analyses.)
3. If an adverse reaction is going to occur, what are its early manifestations, and what is the general pattern of the hazard function (increasing, decreasing, or constant over time)?

The framer of public policy needs to know answers to questions like these:

1. If treatment A replaces treatment B in general, what are the overall differences in cost to the public? in lost work hours because of illness? in utilization of scarce facilities?
2. How can the long-term effects of treatment A be best monitored?

[There are, of course, secondary consumers of the information accumulated in clinical studies. These include other researchers who wish to use the results for planning additional studies, teachers who seek examples for pedagogical purposes, patients, and other members of the lay public. However, the expectations of these groups (can it produce an exciting new article in a popular magazine that will send suffering patients to demand the new medication from their doctors?) have been created by the way in which studies are currently analyzed and reported. The complexities of medicine are such that simple yes–no answers created by hypothesis tests often do more damage than good. For instance, it is of little use to society to have a simple yes or no to the question of whether a new treatment is "dangerous" or whether a new treatment "works."]

A finding of significant differences (or a failure to find such) is irrelevant to the questions noted above. We need, instead, to use these studies to identify subsets of patients with specific response patterns and to estimate degrees of effect and the time course of effect.

THE MRFIT STUDY

An instructive example of the misuse of hypothesis testing is in the MRFIT study (MRFIT Research Group 1982). In this study, 12,866 men identified as being at high risk for coronary heart disease were randomly assigned to one of two regimes. For one regimen, they were subjected to intensive counseling to modify their life styles and reduce the risk of coronary heart disease. For the other regimen, they were merely identified as being at high risk and allowed their usual medical follow-up practices. The write-up of the study was influenced by the "intent to treat" paradigm, and great care was taken to identify specific hypothesis tests that would be run in advance. For none of these was there a significant difference between groups. In addition, they ran hypothesis tests on subsets of patients identified after looking at the results—tests for which they stated, "It must be emphasized that this kind of analysis does not preserve the randomized controlled design of the MRFIT [multiple risk factor intervention trial] and must be interpreted with regard for the possibility of confounding by many factors" (p. 1473).

Among these hypothesis tests they found one small mouse of significance. Approximately 1,200 men in each group were hypertensive at baseline and showed electrocardiogram (ECG) abnormalities when at rest. Of these, 29% died of coronary disease in the intensive intervention group and only 18% died of coronary disease in the usual care group. Without a complete reexamination of the data, it was presumed that the patients undergoing special intervention would be more likely to have been treated for hypertension and that given the usual practice in the U.S., the most widely used drugs were probably diuretics. Thus on the basis of this one small significance and speculation about what might have happened, it has become a standard piece of the current medical mythology that the MRFIT study showed that the use of diuretics can be dangerous for patients with abnormal resting ECG traces.

A far more sensible analysis of the data would have resulted from looking at its purpose. Patients were assigned to special intervention or usual care after having been identified as being at high risk. The study obviously asks whether it is worthwhile to expend additional resources on intensive counseling for such patients. One way to answer this is to compute the exact "costs" associated with each patient (assigning cost figures for work loss because of illness and for death, along with the calculable medical costs of treatment and counseling). Then we need only compare the distributions of the two sets of costs and compute confidence bounds on the mean difference. Thus the study could tell framers of public policy that widespread use of intensive

counseling would result in an overall savings (or loss) of $XXX,XXX if applied to the entire population.

UNNECESSARY CONSERVATISM

In addition to answering irrelevant questions, the widespread use of hypothesis tests invokes a degree of conservatism that is foreign to most medical research. We tend to deal with events surrounded by a considerable degree of random noise in clinical studies. It is often necessary to decide about something with less than "95% confidence." If we use hypothesis tests, we tend to defer decision or go with the "null hypothesis" unless there is fairly strong evidence against it. However, any question one might pose of the data that has a numerical answer can be thought of as a functional on the distribution of the random variables observed. Thus we can think of the answer as a parameter of that distribution and can compute confidence bounds on that parameter. Even if the mathematics are relatively intractable for that computation, we can always use the Bootstrap (Efron 1979). Suppose that we consistently compute three levels of confidence with coverage of 50%, 80%, and 99%. We can think of the 50% confidence interval in terms of the "probable error" of old-fashioned statistics. We are more sure that the true value of the parameter is within that interval than outside it. The 80% interval contains information of which we are reasonably sure. The 99% interval contains information of which we are quite sure. The legal profession has long used such a three-tiered concept of evidence. In criminal trials, there is "probable reason," "clear and convincing evidence," and conclusions that can be made "beyond a reasonable doubt." A suspect can be bound over for trial on the basis of the first. Noncapital offences can be decided on the basis of the second. The last is reserved for decisions with very serious consequences.

Surely a medical treatment can be considered worth using on the basis of a 50% interval, with wide use if it does not involve serious adverse consequences on the basis of a 75–80% interval, with 99% intervals left for issues in which the use of the treatment might entail very serious consequences.

Most readers of this journal will recognize the limited value of hypothesis testing in the science of statistics. I am not sure that they all realize the extent to which it has become the primary tool in the religion of Statistics. Since the practitioners of that faith seem unable to cure their own folly, it is time we priests of the faith brought them around to realizing that there are more appropriate ways to get to useful answers.

[*Received July 1984. Revised December 1984.*]

REFERENCES

Banta, H. D., Behnez, C. D., Gelband, H., et al. (1983), *The Impact of Randomized Clinical Trials on Health Policy and Medical Practice*, Background Paper, Office of Technology Assessment.

CDPA Research Group (1976), "Aspirin in Coronary Heart Disease," *Journal of Chronic Disease*, 29, 625–642.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1–26.

MRFIT Research Group (1982), "Multiple Risk Factor Intervention Trial," *Journal of the American Medical Association*, 248, 1465–1477.

PARIS Research Group (1980), "Persantine and Aspirin in Coronary Heart Disease," *Circulation*, 62, 449–460.

Sackett, D. L., and Gent, M. (1979), "Controversy in Counting and Attributing Events in Clinical Trials," *New England Journal of Medicine*, 30, 1410–1412.