0148-396X/81/0904-0414$02.00/0
NEUROSURGERY

Vol. 9, No. 4, 1981
Printed in U.S.A.

Six Statistical Suggestions for Surgeons

Stephen J. Haines, M.D.

Department of Neurological Surgery, University of Pittsburgh, School of Medicine, Pittsburgh, Pennsylvania

Statistical analysis has become very important in medical research. The large number and variety of statistical techniques required to appropriately handle different types of medical research make it impossible for most physicians to acquire enough statistical expertise to analyze critically the details of many reports. However, a basic understanding of certain fundamental principles of statistical analysis is vital if statistical errors and misapplications are to be identified and prevented. The basic principles underlying six common statistical errors are discussed and a guide to pertinent literature is provided so that the practicing physician without special statistical knowledge can be in a better position to understand and interpret statistical analysis in the medical literature. (*Neurosurgery* 9:414–418, 1981)

Key words: Statistics

The reader of medical literature is increasingly faced with unfamiliar statistical analyses. For example, Volume 47 of the *Journal of Neurosurgery* confronted him with Fisher's exact test, the Mann-Whitney U test, the Wilcoxon rank sum test, the method of Dixon, and the Pearson coefficient of contingency with Cramer's modification, in addition to the usual chi square and *t*-test. Often the conclusions drawn from a study depend heavily on the interpretation of such statistical analyses. The clinician who may apply these results to his patients must have some fundamental understanding of the principles underlying these tests so that he will not be led astray by incorrectly applied or interpreted statistics. That such errors of application or interpretation frequently creep into the medical literature is well documented (12, 16, 17, 25, 30, 32).

The usual approach to the statistical education of physicians is to teach a few basic principles and methods of calculating some elementary statistics. Several excellent reviews along these lines have been published (4, 6, 13, 28, 33, 34). An outstanding effort to combine methodological instruction with a discussion of underlying concepts has been published by Peto et al., but it is restricted to a single type of study (22). There are certain conceptual errors that appear repeatedly in published medical reports and involve concepts not usually included or sufficiently emphasized in introductory statistical education. Some of these concepts are considered advanced because the technical calculations that they require are complicated. However, they can be discussed from a theoretical viewpoint without requiring advanced technical statistical skills. An increasing recognition that the understanding of these basic principles is vital is evident in the literature (1, 11, 16, 21). Six common statistical conceptual errors will be discussed with the intention of helping the reader to be more critical in his interpretation of the medical literature.

A certain fundamental familiarity with statistics is assumed in the following pages. A statistic is a number that summarizes a group of numbers. It may be used merely to describe that group of numbers or to draw some inference about a larger group of numbers. When such inferences are drawn, the risk that the inference is in error must be calculated. This risk is expressed as a probability of error or *P* value. The statistical significance of an inference is assessed from the magnitude of the *P* value. These concepts are adequately explained elsewhere (1, 13, 28, 33, 34). The reader unfamiliar with them should consult these sources before continuing.

THE ORPHAN *P*

A common error that makes the interpretation of a statistical test impossible and that is easily correctable is to report a *P* value without the statistic used to derive it. The *P* value is entirely dependent on the statistic to which it refers. If the statistic is appropriate to the problem at hand, the *P* value is meaningful. If an inappropriate statistic is used, however, even a *P* value indicating a very low risk of error has no meaning: the error has been introduced in the selection of the statistic. Indeed, an unscrupulous researcher could invent a statistic that could confer meaningless "significance" to his data. Reporting the orphan *P* without its parent, the test statistic, is roughly analogous to saying "Mr. Jones' blood chemistry is 142 mg/dl," but not specifying whether this is the sodium, in which case it is normal, or the blood urea nitrogen, in which case a serious problem is present, or the creatinine, in which case a laboratory error is quite likely. No meaning should be ascribed to statistical analyses that make this error of omission.

Our first simple rule is: Never *P* alone.

TOO MANY *P*'s IN THE POD

When one discovers a good thing, there is a natural tendency to use it repeatedly. The calculation of statistics and *P* values is good to the extent that it allows quantification of the risk of error inherent in the conclusions reached. Therefore, it is not unusual to find scientific articles liberally sprinkled with *P* values that are intended to bolster the reader's confidence in the author's conclusions.

The interpretation of a single *P* value is relatively straightforward. It represents the chance that the difference found in the data is the result of random variation when there is no true difference in the population from which the data came. For example, a *P* value of 0.05 indicates 1 chance in 20 that the trend shown in the data is the result of random variation. Frequently, such a risk is accepted for a single test, but what are the implications for this interpretation when a number of tests are done? If 20 tests are performed, the risk that at least 1 of them is positive as a result of random variation must, intuitively and mathematically, be substantial (it is in fact about 64%). If 100 are done, it is almost certain that some tests are spuriously positive. The problem is that we do not know which tests are the spurious ones, so our confidence in the results of all of the tests must be reduced somewhat. This problem appears in medical reports in two ways.

Multiple analyses of accumulating data

There is a strong temptation to analyze data as they accumulate rather than to wait until the end of a study. If this is not allowed to influence the study in any way, one has little practical objection to the practice. However, too often the duration of the study will be determined by the results of the

early analyses. The study may be stopped as soon as statistical significance is reached. In these cases, multiple analyses affect the interpretation of significance and may render the data meaningless.

As data accumulate, they tend to fluctuate around the true population value. At any one time, there is a chance that statistical significance will be reached because of these random fluctuations. The more often the data are analyzed, the greater the probability that such a point will be found and, therefore, the less confidence there can be in the conclusion reached. This problem has been succinctly discussed by McPherson (20). Sequential statistical analysis techniques exist to circumvent these problems, but are seldom used (2).

Multiple analyses of one data set (fishing for significance)

When large amounts of data have been collected, it is difficult to refrain from running as many different analyses as possible on the data so as to make the best use of them. This can lead to the calculation of an enormous number of statistics and P values, a certain number of which will attain statistical significance by chance. The problem again is not knowing which numbers are real and which are spurious, so confidence in all of them is reduced.

This method can be quite useful in searching for potentially important associations among variables but, because of the effect of multiple analyses on the interpretation of significance, the results must be used cautiously. There are two approaches to the problem (29). By using stricter criteria for statistical significance (i.e., $P = 0.01$ rather than $P = 0.05$), the multiplicity effect is compensated. However, the adjustment required rapidly becomes unreasonable. For 20 "significant" P values to have a joint probability of 0.05 requires a P value for each separate test of 0.003 (16). The other alternative is to accept a certain laxity in the criterion of statistical significance and use the results of such a "fishing expedition" as merely a suggestion deserving confirmation. Such confirmation comes from a controlled study designed to answer only one or two specific questions, thus keeping the number of significant tests to an interpretable minimum. In special instances, a single analysis of variance allows these assessments of multiple factors without distorting the interpretation of significance.

Our second simple rule is: As the number of significant differences rises, the significance of the different numbers falls.

DO THE N's JUSTIFY THE MEANS?

For the title of this section, we are indebted to "Student," who used it in a different context (26). It introduces the complex but very important question of sample size and its influence on statistical significance.

The magnitude of a statistic is directly dependent on the size of the sample (n) on which it is calculated. To demonstrate this, one formula for chi square for a two by two table is:

$$\chi^2 = n \cdot \frac{([ad \cdot bc] - n/2)^2}{(a + b)(a + c)(b + d)(c + d)} \qquad \text{(Ref. 4)}$$

and one form of Student's t is:

$$t = \sqrt{n}\,(x - \mu)/s \qquad \text{(Ref. 5)}$$

In both cases, as N increases so does the value of the statistic. Because the P value of the statistic depends on the magnitude of that statistic, increasing the sample size while keeping proportions in the data constant will lead to smaller P values and greater inferred significance.

This means that one can create significance from trivia. For example, Keen was able to find in the literature 101 cases of nephrorrhaphy for "floating kidney" (15). These are categorized in Table 1. Testing for the association between the success of the operation and the type of procedure, the chi square is 2.62, $P > 0.05$, with no significant difference shown. However, if Keen had been able to wait until 1010 cases were available and if the same proportions fell into each category, suture of the renal parenchyma could have been said to be significantly superior to suture of the capsule alone (chi square = 26.6; $P < 0.001$).

All statistical analyses are subject to this kind of false significance. Fortunately, sample size inflation is limited by the natural consequences of increasing expense and effort. The best guard against such faulty interpretation of significance, however, is the reader's knowledge and good sense.

Remember Rule 3, a saying attributed to Gertrude Stein: "A difference, to be a difference, must make a difference."

THE POWER OF PERSUASION

Just as one can create significance from nothing by increasing the sample size, one can hide significance by using too small a sample. This fact is known to clinical investigators, who are apt to see a trend in their data that supports their initial hypothesis but fails to reach statistical significance. The reader is then treated to statements such as "the data, while not statistically significant, are in agreement with the work of . . . ."

A more subtle and potentially more serious misinterpretation that may result from hiding significance in a small sample is that of assuming that the failure to attain statistical significance implies that there is actually no difference in the data.

The interpretation of statistical significance depends on the acceptance of the known risk of error. The usual type of significance testing considers only the risk of calling a difference significant when in fact there is no difference and chance factors have created the difference in the sample. This corresponds to a false-positive diagnostic test and is referred to as the Type I or α error. When one uses data to demonstrate that there is no difference between groups, one is taking a different risk: that a true difference has been masked by chance factors in choosing the sample. This corresponds to a false-negative diagnostic test. Statisticians refer to this as Type II or β error.

In usual statistical analyses, only the risk of α error is specified. This is what we refer to in saying "chi square is significant at the 0.05 level." The risk of Type I error is 0.05 or 1 in 20. If we wish to interpret negative results, the risk of Type II error must be specified. Otherwise, we can say only "no significant difference had been demonstrated," not "there is no significant difference between the groups." That this error in interpretation frequently is found in medical reports has been well documented (9, 18, 27). For example, a study of prophylactic antibiotics for ventriculoperitoneal shunts has the results shown in Table 2 (31). It was concluded that there was no

TABLE 1
Success of Nephrorrhaphy According to Procedure

Procedure	Cured or Improved	Failed	Total
Suture of capsule	30	12	42
Suture of parenchyma	50	9	59
Total	80	21	101

TABLE 2
Shunt Infection Related to Treatment with or without Antibiotics

	Infection	No Infection	Total
Antibiotic	1	14	15
No antibiotic	0	15	15
Total	1	29	30

difference between the treated and untreated groups and therefore that prophylactic antibiotics were not necessary. Obviously, such a conclusion was unwarranted. With only one infection in the entire study, a difference between the groups could not have been detected even if actually present.

The ability of a statistical test to discover a difference between groups is referred to as its power. Numerically, the power is $1 - \beta$ (the risk of Type II error). As we have seen, the power will be dependent on the sample size. It will also depend on how large a difference we wish to detect. Bigger differences are easier to find. One should expect reports of statistical analyses that contain negative results to specify, if possible, the power of the tests used so that the negative results may be properly interpreted. This concept will become increasingly important as more studies are designed to assess the cost and benefits of certain procedures. Here the demonstration of a significant lack of difference between groups will be very important and the power of such tests must be specified.

If one can specify α, β, and the size of difference to be detected (δ) during the design of an experiment, the approximate sample size required can be calculated. This is most helpful as it allows rational planning before committing resources to projects that may have no hope of producing interpretable results. By choosing sample size in this way, the problems of creating or hiding significance by sample size manipulation can be minimized. Calculations can be complex, but tables to aid in determining approximate sample size are available (1, 3, 8, 10, 19, 23, 24).

This problem is summarized as Rule 4: Absence of proof does not prove absence.

THE HOLEY ASSUMPTION

The use of mathematical techniques to derive general conclusions from specific data requires certain assumptions. These assumptions are commonly made and are too often unstated, and errors in statistical application and interpretation may result when assumptions are violated. Three common assumptions will be discussed.

Random sampling

The objective of statistical analysis is frequently to make inferences about a large population from the study of a subset or a sample of that population. All statistical methods in general use make the assumption that the sample is randomly drawn from the population by one of a variety of statistically acceptable procedures (6). When this is not true, the possibility of systematic bias in the selection of the sample makes the application of statistical inference techniques inappropriate. The estimates of error calculated from such data have no meaning. The *Literary Digest* presidential poll that predicted Landon's victory over Roosevelt in the 1936 presidential election is a good example (6). Based on the sample of 2 million, the statistical precision of the estimates was excellent. However, the nonrandom selection of the sample led to an intellectual and economic bias that rendered the statistical analyses meaningless, as shown by the results of the actual election.

As a matter of practical necessity, this assumption is almost universally ignored in clinical investigations. Without a centralized registry and coordinating center for clinical investigations, random samples of patient populations of sufficient size for appropriate analysis cannot be generated. It is important to remember that this assumption is being violated and to scrutinize carefully every study for the possibility of bias in sample selection. It is essential that published reports of clinical studies include enough descriptive information about the patient population that differences between the study population and the populations to which the results may be generalized can be taken into account.

Independent observations

The laws of probability from which statistical procedures are derived rely heavily on the assumption of independence of observations. This simply means that the chance of one observation being included in the sample remains the same regardless of what the other observations are. In general, this assumption is satisfied by making each observation on different, unrelated individuals. Occasionally, however, this requirement is forgotten. In one example, tests of skin reactivity to chemicals were performed on a group of patients by applying both chemicals to the forearm of each patient (14). The chi square was calculated to test for an association between the reactivities and gave a statistically significant result. However, as the reactions were tested on the same individuals, the observations were not independent—patients who reacted to one chemical might be more likely to react to the other. When the appropriate test for these paired observations (McNemar's test) was applied to the data, no significant association was found.

Another situation in which the assumption of independence may be violated is when multiple observations of a variable are made on each patient in the study. For example, assume that some base line chemical value is measured, a treatment is applied, and the chemical measurement is performed daily for 3 days after treatment. Not only must one be careful to use an appropriate analysis for the paired data when comparing the pre- and post-treatment values, but the multiple post-treatment values cannot be treated as independent. Values obtained from the same person are more likely to be similar than values obtained from different people. This similarity will tend to produce spurious associations unless taken into account. Furthermore, the multiple observation values, if treated as independent observations, inflate the apparent sample size without adding new information. If this were a legitimate technique, it would be possible to produce samples of a thousand values from a single patient and create impressive statistical significance from very small numbers of subjects.

Normal distribution

A source of unending confusion is the use of the word "normal." It may describe single values (a "normal" sodium), a range of values (the x-ray findings are within "normal" limits), or an entire population (the height of physicians has a "normal" distribution) (7). Used in the last sense, the "normal" distribution (the familiar "bell-shaped curve") has a specific mathematical meaning. It implies that most values fall near the mean and a few values are either much greater or much less than the mean. It also implies that the distribution of values is symmetrical: that there are as many values above the mean as below it. If the values being studied approximate this distribution, a number of powerful statistical techniques may be used for analysis. Too frequently the assumption of normality is made without confirming its validity.

Something as simple as the interpretation of the standard deviation depends heavily upon the assumption of normality. In a "normally" distributed population, approximately two-thirds of the values fall within 1 standard deviation and 95% fall within 2 standard deviations of the mean. This is true only if the distribution is approximately normal. For skewed or otherwise "abnormal" distributions, this interpretation of the standard deviation has no meaning. Weech cites an example in which the standard deviation exceeded the mean in all but 1 of 14 groups of measurements of biochemical constituents of

tissue (30). For example, the mean and standard deviation of the percentage of glucose in the tissue were 1.25 and 2.8, respectively. If these data were normally distributed, this would imply that about one-third of the glucose percentages in the population were less than 0, which is clearly nonsense. It is clear that the population data were not normally distributed. Nonetheless, the author went on to apply the *t*-test (which assumes normality) to his data and arrived at erroneous conclusions.

When statistical tests more complicated than the standard deviation are used in the absence of the assumed normal distribution, the usual interpretation may not apply. Student's *t*-test, the F-test and the analysis of variance schemes based on it, correlation coefficients, and confidence limits for regression coefficients are among the commonly used statistics that depend on the assumption of normality.

Mild deviations from normality in the population have little effect on the results and, in the opinion of many statisticians, allow the use of these powerful statistical techniques in many circumstances. Statistics that are relatively insensitive to deviations from normality are called "robust." One must be sure, however, that major departures from normality do not exist in the data. Scales or arbitrary scores constructed by the investigator may deviate markedly from a normal distribution. The Apgar score, for example, in the population of all newborns is heavily skewed toward scores of 7 to 10. The *t*-test on such data may lead to erroneous conclusions (26). Before using statistical methods that assume a normal distribution of the data, it is important to look at the actual distribution of the data, perhaps as a graph, to be sure that it approximates normality.

Statistical techniques that do not depend on the assumption of normality have been developed. These are often called "nonparametric" statistics. Chi square is the best known of this group. Standard statistical texts describe several others. The use of these tests initially was limited by the tedious nature of the calculations required for finding *P* values, but the ready availability of computers has eliminated this obstacle and increasing use of these techniques is being made.

This section is summarized in Rule 5: Ignorance of assumption leads to assumption of ignorance.

FISHY SCALES

The observations that are collected for statistical analysis may be measured on three different types of scales: nominal, ordinal, and interval. A nominal scale is merely a set of categories with names. It is not possible to rank the categories objectively with respect to each other. Variables with categories such as red, white, blue, and other are nominal. An ordinal scale is one that ranks the responses but does not allow measurement of the distance between the ranks. The Apgar score, the Botterell classification of clinical status after subarachnoid hemorrhage, the Glasgow coma scale, and scales such as (excellent, good, fair, poor) are ordinal scales. Interval scales are ordinal scales on which the distance between each ranking is numerically equal. This allows mathematical manipulation of the rankings. The most familiar interval scales are physical measurements such as height, weight, serum sodium, and so on. Interval scales may be discrete or continuous. A continuous scale allows any value to be taken on by the variable. There are an infinite number of theoretically possible values between any two actual values on the scale. A discrete scale, on the other hand, has fixed increments between which no other value may appear. Weight is measured on a continuous scale, whereas number of fingers is measured on a discrete scale.

The type of scale on which data are measured influences the type of statistical analysis that may be performed. At the simplest level, the calculation of a mean for nominal data is "meaningless": The appropriate measure of central tendency is the mode, the most frequent category. For ordinal data, the median, which makes use of the ranking of the data but does not require arithmetic manipulations, is most appropriate. Only when the data are measured on an interval scale can the arithmetic mean be used as a measure of central tendency. Similar considerations apply to measures of variants and association and to perimetric estimation and hypothesis testing.

Many of the more powerful statistical techniques such as *t*-tests, analysis of variance, regression, and the usual forms of correlation assume interval (and usually continuous) data. Other tests are designed for ordinal data (or interval data that have been grouped into discrete rankings). Such tests include chi square, many of the nonparametric tests, and measures of association such as Goodman and Kruskal's gamma. Because of their nature, nominal data do not lend themselves to much in the way of sophisticated statistical analysis.

One must be careful that analyses are not applied to data measured on inappropriate scales. Numerical results may be generated, but they may be totally meaningless; too much weight is given to the numbers. Especially dangerous in this regard are "pseudointerval" scales such as the Apgar score. Assigning numbers to rankings may lead to arithmetic manipulations that give meaningless results, such as an Apgar of 6.7 or the idea that a score of 8 is twice as good as a score of 4 (26). Similar difficulties arise with the classification of subarachnoid hemorrhage. A mean grade of 2.3 implies a degree of precision that is clinically unobtainable. Further manipulations depending upon the mean will only further confuse the interpretation.

Misapplications of this sort are generally easy to uncover, and the reader who finds such mismatching of scale and statistic must be especially cautious in interpreting the results.

Rule 6: Don't break the scale.

SUMMARY

Several basic principles underlying statistical analysis have been examined and may be summarized as follows:

1. Always state the statistical test used to obtain the reported *P* value.
2. Beware of multiple analyses. The reported significance level is probably much too small.
3. Significance depends on sample size.
4. Failure to disprove the null hypothesis does not prove the null hypothesis.
5. There are many assumptions underlying statistical techniques that may have been ignored.
6. Different kinds of data require different types of analysis.

By keeping these rules in mind and frequently consulting statistical colleagues, the physician can more accurately interpret the significance of the multitude of reports to which he is exposed.

Reprint requests: Stephen J. Haines, M.D., Department of Neurosurg, University of Minnesota Health Sciences Center, 420 Delaware Street, Southeast, Minneapolis, Minnesota 55455.

REFERENCES

1. Altman DG: Statistics and ethics in medical research: Misuse of statistics is unethical. Br Med J 281:1182–1184, 1980; Study design. Br Med J 281:1267–1269, 1980; III. How large a sample? Br Med J 281:1336–1338, 1980; Collecting and screening data. Br Med J 281:1399–1401, 1980; V. Analysing data. Br Med J 281:1473–1475,

1980; VI. Presentation of results. Br Med J 281:1542–1544, 1980; VII. Interpreting results. Br Med J 281:1612–1614, 1980; VIII. Improving the quality of statistics in medical journals. Br Med J 282:44–47, 1981.
2. Armitage P: *Sequential Medical Trials.* Oxford, Blackwell Scientific Publications, 1972, ed 2.
3. Boag JW, Haybittle JL, Fowler JF, Emery EW: The number of patients required in a clinical trial. Br J Radiol 44:122–125, 1971.
4. Dixon WJ, Massey FJ: *Introduction to Statistical Analysis.* New York, McGraw-Hill, 1969, ed 3, p 242.
5. Dixon WJ, Massey FJ: op cit, p 98.
6. Feinstein AR: Clinical biostatistics: VII. The rancid sample, the tilted target, and the medical poll-bearer. Clin Pharmacol Ther 12: 134–150, 1971.
7. Feinstein AR: Clinical biostatistics: XXVII. The derangements of the 'range of normal.' Clin Pharmacol Ther 15:528–540, 1974.
8. Feinstein AR: Clinical biostatistics: XXXIV. The other side of 'statistical significance': alpha, beta, delta and the calculation of sample size. Clin Pharmacol Ther 18:491–505, 1975.
9. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR: The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. N Engl J Med 299:690–694, 1978.
10. George SL, Desu MM: Planning the size and duration of a clinical trial studying the time to some critical event. J Chronic Dis 27:15–24, 1974.
11. Glantz SA: Biostatistics: How to detect, correct and prevent errors in the medical literature. Circulation 61:1–7, 1980.
12. Gore SM, Jones IG, Rytter EC: Misuse of statistical methods: Critical assessment of articles in *BMJ* from January to March 1976. Br Med J 1:85–87, 1977.
13. Hill GB: The statistical analysis of clinical trials. Br J Anaesth 39: 294–310, 1967.
14. Hoffman JIE: The incorrect use of chi-square analysis for paired data. Clin Exp Immunol 24:227–229, 1976.
15. Keen WW: Nephrorrhaphy. Trans Am Surg Assoc 8:181–204, 1890.
16. Lee KL, McNeer JF, Starmer CF, Harris PJ, Rosati RA: Clinical judgement and statistics: Lessons from a simulated randomized trial in coronary artery disease. Circulation 61:508–515, 1980.
17. Mainland D: The use and misuse of statistics in medical publications. Clin Pharmacol Ther 1:411–422, 1960.
18. Mainland D: The significance of 'nonsignificance.' Clin Pharmacol Ther 4:580–586, 1963.
19. Mainland D: Statistical ward rounds—14. Clin Pharmacol Ther 10:272–281, 1969.
20. McPherson K: Statistics: The problem of examining accumulating data more than once. N Engl J Med 290:501–502, 1974.
21. Mosteller F: Problems of omission in communications. Clin Pharmacol Ther 25:761–764, 1979.
22. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG: Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I. Introduction and design. Br J Cancer 34:585–612, 1976; II. Analysis and examples. Br J Cancer 35:1–39, 1977.
23. Schlesselman JJ: Planning a longitudinal study: I. Sample size determination. J Chronic Dis 26:553–560, 1973.
24. Schneider MA: The proper size of a clinical trial: 'Grandma's strudel method.' J. New Drugs 3:3–11, 1953.
25. Schor S, Karten I: Statistical evaluation of medical journal manuscripts. JAMA 195:1123–1128, 1966.
26. "Student": When do the Ns justify the means? Clinical biostatistics gone astray. Pediatrics 51:758–759, 1973.
27. Sundaresan N, Voorhies R, Kwok K-L, Thaler HT: Hypothesis testing in neurosurgical trials. J Neurosurg 54:468–472, 1981.
28. Swinscow TDV: *Statistics at Square One.* London, British Medical Association, 1980, ed 6.
29. Tukey JW: Some thoughts on clinical trials, especially problems of multiplicity. Science 198:679–684, 1977.
30. Weech AA: Statistics: Use and misuse. Aust Paediatr J 10:328–333, 1974.
31. Weiss SR, Raskind R: Further experience with the ventriculoperitoneal shunt: Prophylactic antibiotics. Int Surg 53:300–303, 1970.
32. White SJ: Statistical errors in papers in the *British Journal of Psychiatry.* Br J Psychiatry 135:336–342, 1979.
33. Worcester J: The statistical method. N Engl J Med 274:27–36, 1966.
34. Zivin JA, Bartko JJ: Statistics for disinterested scientists. Life Sci 18:15–26, 1976.