Statistics, The Literature, Hospital Data And Patient Profiles: A Survival Guide

J Civetta

Citation

J Civetta. *Statistics, The Literature, Hospital Data And Patient Profiles: A Survival Guide*. The Internet Journal of Anesthesiology. 1998 Volume 3 Number 4.

Abstract

INTRODUCTION

Continuing self-education and appraisal of the medical literature are simultaneously a responsibility and a joy for the physician who intends to keep abreast of new developments and incorporate new research findings into clinical practice. It is a responsibility, given the tremendous accrual of new knowledge. It is a joy because this stimulus keeps one's mind alive and renews the original feelings that prompted the study of medicine.([1]) Feinstein notes, ``The statistician brings a long tradition of intellectual neglect of the significance of management of clinical data. He finds as his collaborators, clinicians who have a long tradition of intellectual fear of statistics.'‘([2]) He also noted that the numerical method was introduced into clinical medicine in 1836 by Pierre Louis, who helped end the popularity of blood-letting by counting and comparing the results of patients treated in various ways. Louis was both attacked and vilified according to Feinstein who refers to Louis' experience as, “A caveat for any clinician who questions an accepted therapy of his era and who urges his colleagues to make better use of their senses and statistics in evaluating therapy.'‘ He also quotes Louis, ``Let those, who engage hereafter in the study of therapeutics, pursue an opposite course to that of their predecessors. Let them labor to demonstrate rigorously the influence and the degree of influence of any therapeutic agent, on the duration, progress, and termination of a particular disease.'‘ Thus, one faces strong positive and negative stimuli to pursue the goals of this chapter.

The medical literature contains reports of well-conceived, well-conducted, and well-analyzed studies in addition to reports of ill-conceived, poorly conducted, and inadequately analyzed studies. How can one tell the good from the bad? Although this will remain a continuing and perplexing problem for the clinician, we hope to provide some assistance in separating the wheat from the chaff. Although the reader may make some inferences, guided by prior knowledge of the quality of a journal, for instance,we should approach the evaluation of an existing medical report and the creation of our own research in a more formal and structured manner. Although it is tempting to read the title, abstract, and conclusions and then rush to apply the new methodology or treatment, we should spend the time to evaluate the entire article. This has the dual advantages of training our often capricious minds and also avoiding errors due to our precipitous acceptance of the new finding. Remember, ``jumping to conclusions seldom leads to happy landings.'‘ ([3]) The high-quality journal that uses vigorous peer review and meticulous editorial appraisal will likely publish good, solid, and sound reports of medical studies. But even with such journals, the occasional poor paper can slip through the peer review editorial process and achieve the ``importance'‘ of publication. With the proliferation of journals, it is also true that articles that fail to pass the stringent peer review of prestigious journals may also achieve publication in another journal that is struggling to fill its allotted pages. Hence, it is important that the clinician, in assessing medical reports in areas of particular interest, has the ability to assess critically and formulate judgments concerning the strengths and weaknesses of published medical reports.

CRITICAL READING OF THE MEDICAL LITERATURE

A formalized process aids particularly in the evaluation of a medical report to ensure that the conclusions are well supported and to further the development of one's own critical acumen. These are general questions that apply to virtually any research report involving the collection and analysis of data.

OBJECTIVE AND HYPOTHESIS

Obviously, the most pertinent starting point is an understanding of the investigator’s objective. The investigator has the obligation to state clearly and specifically the purpose of the study conducted, but this may be difficult to discern. In such cases, we may question whether the author had, indeed, a clear objective. ``Fishing expeditions,’ ‘ that is, extensive data collection projects with the intention of exploring and identifying important relationships, achieve success when the captain knows where the fish are. In other words, the so-called gold mine of data does not guarantee that statistical search will lead to pay dirt and reveal important new relationships. The author, or we as researchers, must formulate specific objectives and a clear-cut hypothesis for testing. Lack of an understanding of objectives handicaps both the reader and the author in any assessment or interpretation of the results.

A more specific and somewhat more subtle question in assessing objectives is classification of a study as descriptive and exploratory versus analytic. Using epidemiologic terminology, descriptive studies are those that ``describe’ ‘ diseases, characterize disease patterns, and explore relationships, particularly in regard to person, place, and time. Such studies mainly serve the purpose of “hypothesis generation.’ ‘ The specific hypothesis can then be tested by means of an analytic study, one whose primary objective involves the test of a specific hypothesis.

To illustrate this distinction, a descriptive study reported the use of high-level positive end-expiratory pressure (PEEP) in acute respiratory insufficiency in patients who developed severe, progressive, acute respiratory insufficiency despite aggressive application of conventional respiratory therapy. (4) Later, the term ``optimal PEEP,’ ‘ introduced in the first study, was updated in another descriptive study of 421 patients reported in 1978. (5) The second study entailed treatment of a large group with respiratory failure using titration of PEEP in conjunction with intermittent mandatory ventilation (IMV) but using cardiovascular interventions to support cardiac function until a preselected end point of 15% shunt could be achieved. The first study represented a description of the development of a treatment regimen; in the second study, refinements in this treatment regimen were applied to a broader population. Later, a hypothesis was constructed to test whether, in moderate arterial hypoxemia, there was any improvement in patient outcome or resource utilization using ``optimal PEEP’ ‘ compared with similar modalities of therapy, with an end point defined as achievement of nearly complete arterial oxygen saturation at nontoxic inspired oxygen fractions.

The hypothesis that PEEP titration to achieve an intrapulmonary shunt of less than 20% would have a better outcome or would achieve faster resolution of the disease process could not be substantiated in the analytic study. (6) The two descriptive studies 4,5 served to identify a specific hypothesis that the third or analytic study tested.

STUDY DESIGN

The reader should consider carefully the definitions of the groups studied and the population to which the investigators intend to refer their findings. For instance, in the three studies quoted, one might assume that the failure to prove the hypothesis in the third study invalidated the findings of the two earlier descriptive studies. The third, an analytic study, however, involved only patients with early and moderate arterial hypoxemia. The original group of patients who were studied specifically excluded these patients and concentrated on developing therapy for those who had persistent hypoxemia despite aggressive application of conventional respiratory therapy. Thus, a technique that reversed hypoxemia in patients who were refractory to the then ``conventional therapy’ ‘ of acute respiratory insufficiency was found to be not useful in another population that had only moderate hypoxemia and did not have true adult respiratory distress syndrome (ARDS). If the authors do not state clearly the populations with which they are dealing, the readers can easily lose this important distinction. This has even greater importance in review articles that may omit the important qualifiers or modifiers found in the original reports. The fact that a particular form of therapy useful in advanced disease has no particular advantage in patients with mild disease indicates that we should restrict therapy to those patients who can benefit rather than arrive at some alternative conclusion that titration of PEEP to preselected end points has no advantage.

The reader should examine carefully the Materials and Methods section for a description of the study design. Epidemiologically, there are two major classifications of study design: experimental and observational. Loosely defined, an experimental study is one in which the investigator has control over or can manipulate the major factor under study. The epitome of the experimental study is the randomized controlled trial in which the investigator demonstrates ``control’ ‘ over the factor under study by randomizing patients to various regimens. Many prophylactic and therapeutic studies tend to be experimental in design. One cannot assume that just because a study was experimental and the investigator may have randomized patients that the study was well done and its conclusions are valid. Experimental studies are prone to various sources of bias and to poor execution. The label ``randomized'‘ is not equivalent to assurance of high quality, nor does it alone add validity to the study. Thus, randomized studies also need careful assessment of their design, methods, analyses, and conclusions. One other factor, “blinding” is often viewed as an attribute of the highest quality studies. If there are subjective elements used to judge the effectiveness of treatment,there is a compelling rationale to blind the investigators. If there are subjective assessments of the patients’ response,there is a compelling rationale to blind the subjects. If all of the outcome variables are objective, blinding, strictly speaking, is unnecessary. Thus in the assessment of a new medication to relieve pain, double blinding (both subjects and investigators) is necessary.

When the investigator cannot manipulate the major factor under study, he or she must rely on what has been observed; this study is an observational study. We should not view observational studies as being inferior to experimental studies. Clearly, a tight, well-designed, well-executed experimental study carries the greatest strength of evidence, but observational studies can also provide substantial, sound medical evidence. In fact, a well-planned and well-executed observational study can be much more informative than a weakly designed and poorly executed randomized study. There are various approaches to the design of observational studies, such as cross-sectional, case control, prospective cohort, and retrospective cohort. The interested reader should consult basic epidemiology or statistics textbooks for further descriptions of these various design strategies as well as the relative strengths and weaknesses of each design format. ([7,8])

With respect to observational studies, one should determine whether the data collection was prospective or retrospective. The principal advantage of prospective data collection is that the researcher, having clearly identified the objectives, can ensure collection of this relevant information in a manner that he or she can determine. Retrospective analysis of medical records depends on what happens to appear in the record, often with no indication of the manner in which the information was obtained. For example, sex, age, and hospital outcome (survival or death) are key data elements that may not appear for every patient in a retrospective chart review. Clearly, without a specified protocol, one cannot anticipate that a daily blood gas, serum creatinine, or any other intermittent measurement dependent on a specific order will appear in the chart. Everyone should attempt a retrospective study (at least once) to learn the pitfalls and the impossibility of obtaining a complete data base. This would enable each of the then-frustrated researchers to read other retrospective studies both with a great deal of deserved skepticism and with empathy for the difficulties with such research.

Selection of the study group is another important step. One should look for possible sources of selection that would make the sample atypical or nonrepresentative. It is interesting that even such seemingly ``random'‘ allocation of cases such as alternate days may introduce an unappreciated bias. For instance, the Trauma Service at the University of Miami/Jackson Memorial Medical Center had two separate teams that alternated coverage every 24 hours. Patients admitted on alternate days, therefore, are cared for by different teams of physicians. A study that entailed alternate-day assignment to treatment groups would entail, as well, the factor of differences in physician practice style, a factor that one could not disentangle in analysis of study results.

We must also consider the nature of the control group or standard of comparison. We frequently encounter the ``historical control'‘ group that, almost always, has a ``poorer'‘ result than the contemporary group. The problem, of course, is that the basic assumption that the modality of treatment under investigation is the only cause for the difference in results is clearly erroneous. It has been tempting to ascribe the remarkable reduction in wartime mortality from World War II to Korea to Vietnam to the marked diminution in delay between injury and treatment. However, the entire surgical training experience changed during that time, an almost completely new pharmacopoeia was available in Vietnam, and, most assuredly, many other variables are yet unaccounted for between the two eras. In fact, the principal reason for randomization in a study is to attempt to distribute the unknown and potentially important variables equally among groups to avoid selection bias. We may also see this effect if patients accrue slowly and the study thus runs over many years. Other aspects of therapy may change and have a greater impact on outcome than the original variable selected for study.

VALIDITY AND GENERALIZABILITY

Two aspects of clinical research that sometimes perplex beginning researchers and inexperienced readers are validity and generalizability. Validity deals with the ability of a study to give a scientifically sound answer to the question posed. Insofar as possible, this answer should be free from bias, uninfluenced by the effects of other related or confounding variables and with good statistical precision. Only then is there a basis for a valid study result.

Generalizability deals with extrapolation of study findings to a larger population or to other groups. Assessment of generalizability depends on the degree to which the study subjects are representative of some larger target population, how well the selection of study subjects simulates the process of drawing a random sample from a population.

The ideal is for studies to be both valid and generalizable. In practice, this is rarely the case. In the design of clinical research, investigators face many situations in which they must choose between validity and generalizability. When faced with a choice, undoubtedly they should opt for validity. Without a valid study, an investigator has little or nothing of scientific merit. The investigator may have actually drawn a random sample from a larger population and have virtually ideal generalizability. But, if in the process, validity was threatened or compromised, the findings are worthless. With findings of questionable or doubtful validity, there is nothing of value to generalize. Generalizability plays a subordinate role and, in fact, should not even surface until firm establishment of validity. Often it is left to the reader to assume the onus of assessment of generalizability and of whether findings can be extrapolated to other populations.

METHODOLOGY AND OBSERVATIONS

In the reporting of research results, clarity in the definitions of the terms and measurements made has great importance. The more clearly the authors (or we as potential researchers) define the terms, including diagnostic criteria, measurements made, and the criteria of outcome, the more likely it is that we, the readers, can interpret the findings correctly and gain a proper perspective. For instance, in the field of invasive catheter-related infection, terms such as ``colonization,'‘ ``contamination,'‘ and ``infection of the catheter'‘ abound. Authors often use these terms differently, leading to great difficulty in interpretation and synthesis of results from different studies. Furthermore, a ``positive culture'‘ may represent different bacteriologic methodologies: some authors use a semiquantitative culture of an intracutaneous catheter segment, ($_9$) whereas others use blood cultures aspirated through the catheter. ($_{10}$) Clearly, results from one methodology may not be comparable to another, and interpretations based on differing methodologies may lead to different conclusions.

We must also try to evaluate the methods of classification or of measurement. The essential question is to assess whether inconsistencies in observation or evaluation could have sufficient impact to influence materially the results of the study. We also must try to evaluate the reliability and reproducibility of the observations. This is more difficult to assess. Frequently some clues inform the reader of the author's concern with and awareness of reproducibility and reliability. When a subjective element enters into an assessment, an author often refers to and sometimes provides data on the results of evaluations by independent observers and their degree of agreement. Interrater reliability would refer to the ability of two or more independent raters to make the same observations. Intrarater reliability would refer to an observation made by the same rater over two or more different times. With respect to abstracting information from charts, interrater and intrarater reliability is usually in the range of only 80% to 90%. An author who devotes some attention to issues concerning measurement or laboratory error would seemingly be cognizant of the importance of reproducibility and reliability. It is well to be suspicious of results from a study that seems entirely devoid of concern with these elements, especially if some subjective element is clearly involved in either diagnosis, observation, or assessment of outcome.

PRESENTATION OF FINDINGS OR RESULTS

Authors must walk the fine line of clear and concise data presentation in the Results section without editorializing or drawing conclusions from the data they presented. Remember, the facts should be able to speak for themselves. The author must still detour into enough necessary detail for the reader to judge the importance of the data. Important findings require proper documentation. If a small number of patients are presented, a table listing the important demographic characteristic is useful so that the reader has a clear understanding of the population studied.

It is surprising how often numerical inconsistencies are contained within papers published in even the most reputable medical journals. This may be due, in part, to the many drafts and revisions compounded by textual proofreading, computational and tabular proofreading, and other processes. Because of the frequency of these errors, the reader may wish to use some quick checks: columns and rows should add up to their indicated totals; percentages of mutually exclusive categories should add up to 100%; numbers in tables and figures should agree with those in the text; and totals in various tables describing the same population should agree. With the ubiquitous presence of hand-held calculators and personal computers, we can even run some of our own statistical tests, especially when the reported results appear incompatible with our quick mental assessment or even personal bias!

Clarity and precision are important criteria to judge the overall scientific validity of an article. Assessments, comparisons, and judgments belong in the Discussion section. However, when these are enthusiastically included in the Results section, they strongly suggest bias in the author's approach. Strictly speaking, an investigator should undertake an analytic study when he or she can wholeheartedly support affirmation or rejection of the hypothesis under test. Thus, inclusion of subjective opinions (``markedly improved outcome'‘) in the Results section may be a subtle indication that the investigator performed the study to confirm his or her preexisting personal view.

DATA ANALYSIS

In reality, the first question we, as readers, should ask is ``Are the data worthy of statistical analysis?'‘ We must then examine the methods of statistical analysis to determine whether they were appropriate to the source and nature of the data and whether the analysis was correctly performed and interpreted. These questions are difficult to answer. However, we recognize that this is an entire field to itself for which this chapter should serve as a stimulus to pursue more vigorous study.

One of the first issues that should cross the reader's mind is to ask whether the observed and reported finding could be due simply to chance, the luck of the draw, or sampling variation. There is an arsenal of statistical methodology available ranging from simple (e.g., t-test, chi-square test) to sophisticated (multiple logistic regression, Cox proportional hazards model) to examine the role of chance in the analysis of study results. Each medical reader may not have sufficient expertise to assess whether the investigators have chosen their methodology appropriately and have correctly performed the statistical analyses. We may hope that the journal's peer review process will have included some form of assessment of the statistical aspects of the paper. Until we, the readers, learn enough, we must solicit expert biostatistical assistance. However, there are three points to remember. First, it is the author's responsibility to provide the reader with information on the specific statistical analysis used in the assessment of the role of chance. Second, whatever the level of significance reported, no matter how small the p value is, we can never rule out chance with certainty. An exceedingly small p value (1 instance in 1000) denotes that chance is a most unlikely explanation of the result, but there remains the possibility, although unlikely, that this is indeed that one instance in 1000. The third point is that a statistically significant result is not necessarily important or even indicative of a real effect, only that an effect of chance has been ruled out with some reasonable certainty. Often we must apply context or perspective to the author's work. We have discussed the importance of reliability and reproducibility.

As clinicians, we know that measurements of pulmonary artery occlusion pressure (PAOP) differ among observers. For instance, estimation of PAOP from a visual inspection of the oscilloscope tracing may be 3 - 4 mmHg different from the results calculated electronically and displayed in digital form on the monitor. In reviewing the effects of a drug, however, some investigators may interpret a change of the same magnitude (3---4 mm Hg) as an ``effect'‘ of the therapy. Thus, in addition to deciding whether a particular result is ``statistically significant,'‘ that is, if it represents a real event (or is due to chance), we must decide whether it has any real clinical, biologic meaning.

Furthermore, in our interpretation of study results we must, with reasonable certainty, rule out the possibility of bias and confounding. A result may be highly significant statistically but the study design and conduct could lead to a substantially biased result, or there may be some other related variable that also explains statistically significant results.

Confounding refers to effects of one or more related variables. In its strict epidemiologic definition, a confounding variable is one that is associated with both the ``exposure'‘ or independent variable and with the ``outcome'‘ or dependent variable under study. For example, in an observational study that compared the mortality experience of two modalities of treatment for head injuries, an obvious ``confounding'‘ variable would be the severity of the injury. Clearly, the severity of the injury relates to the dependent variable under study--mortality. The injury severity, however, may also have an association with the independent variable--the choice of the particular modality of treatment. Thus, any finding of a difference in mortality between modalities of treatment, no matter how statistically significant the difference is, might be explained by the confounding effects of the severity of injury.

The important point is to judge whether the authors have considered all the pertinent known confounding variables in their analyses and have taken proper steps to account for their effects. The reader, without substantive knowledge of the particular field of study, may be unable to delineate what pertinent potential confounding variables should have been considered. We (authors and readers) must cautiously proceed with forming conclusions.

Bias refers to a systematic departure from the truth. Bias may exist in many forms, and many statistical and epidemiologic adjectives can precede the word ``bias'‘ to denote some specific hazard or snag that can lead to a departure from the truth. Sackett provides a useful compendium of the various biases that lurk to ensnare the unwary investigator, as well as the unwary reader, in the conduct of biomedical research. ($_{11}$) For our purposes, we shall use the three adjectives: selection, observation, and analysis.

Selection bias refers to how subjects got into the study. Is the manner of selection of persons for study such that the study will result in substantial distortion of the truth? As a simple example, consider a study to compare outcome of surgery in patients who agree and volunteer to undergo the operation with those who refuse. Those who choose surgery may be better operative risks (at least from their own perception) probably with less co-morbid disease than that found in the nonoperated group. Of course, other factors may have influenced the other group to refuse surgery. Still, the difference in the outcome of surgery might be more likely to result from the selective nature of the groups rather than from any real effect of the surgical procedure.

Observation bias refers to the methodology for handling and evaluating subjects during the course of the study. If a therapeutic intervention group receives more attention, more supportive therapy, and more intense scrutiny than a control group, an observed difference in outcome might more likely be explained by observation bias rather than by any real effects of the intervention. Retrospective studies are particularly prone to observation bias.

Analysis bias refers to fallacies that exist in the choice of statistical methods to analyze data. An example is the ``average-age-at-death'‘ fallacy. Calculation of average-age-at-death among decedents does not measure longevity; it reflects mainly the age composition of the total members of the groups, mostly those who are alive. For example, consider a newspaper report of a study that compared the average age at death of U.S. professional football players with professional baseball players. ($_{12}$) The report stated that football players died, on average, 7 years earlier than baseball players. It would be erroneous to conclude that this differential reflects the more hazardous and traumatic aspects of professional football compared with professional baseball. The fact of the matter is that professional football is a much newer sport (dating from the mid-1920s) than professional baseball, dating from the 1860s. Consequently, the total group of professional baseball players is considerably older than the total group of professional football players. As an extreme example of this average-age-at-death bias, consider the result anticipated in a comparison of the average-age-at-death in a children’s hospital with that in a retirement community hospital.

When, in the assessment of a study, we can rule out with reasonable certainty that the finding is not due to chance, bias, or confounding, we are well on the road to determining a real and meaningful effect.

Finally, it is important to emphasize that the interpretation of ``statistical significance'‘ does not in and of itself connote medical or biologic importance. Correlation and regression illustrate this distinction and the misinterpretation that often occurs. Correlations are used to describe the degree of association between independent variables, such as blood pressure and age. With regression, this implies that we can choose one variable as dependent and one (or more) as independent. With regression, our concern is the relationship between an independent and dependent variable, namely, what happens to the dependent variable as we alter the independent variable. We use the linear regression equation, Y = a + bx, to predict values of the dependent variable Y from the independent variable, X, where b is the slope of the regression line and a is the intercept. Both correlation and regression may be presented familiarly as a scatter diagram and usually involve some analysis of the ``statistical significance'‘ of this relationship. The relation of a correlation coefficient in the case of a perfect linear relationship would yield a value of 1. Correlation coefficients of 0.5 or less might achieve ``statistical significance'‘ at a p value of less than 0.001, particularly when the sample size is large, (i.e., there are a large number of data points). The square of the correlation coefficient (and multiplication by 100) indicates the percentage of the variance of the dependent variable explained by the variance in the independent variable. Given a regression coefficient of 0.5, 25% of the dependent variable’s variance would be explained by changes in the independent variable, although 75% remains unexplained. This may be a highly statistically significant relationship, with, say, p <lt> 0.001; that is, there is a likelihood of less than 1 in 1000 that this has occurred by chance. Many authors and readers do not understand clearly the separation of the statistical reliability of the relationship and the importance of the statistical relationship. They often place a high weight of importance on relationships of minor causal significance (low r value) only because it is highly unlikely to be due to chance (low p value). To illustrate this point, some years ago the hemodynamic and respiratory data for numerous disparate patients were combined and subjected to regression analysis. (13)

Venous oxygen content and intrapulmonary shunt had a statistically significant relation (p 0.01) with an r value of 0.46. Such a low value indicates that the individual data points are like a scatter diagram Therefore, other factors must have an even greater weight to explain the other 79% percentage of variance (r = 0.46, r2 = 0.21, or 21% of variance). Obviously, the degree of pulmonary disease in this instance would have the greatest effect in determining the intrapulmonary shunt. Thus, there is a weak though causal link between the two variables (low r value) and a strong likelihood that there is a relationship, that is, that the result was not due to chance (low p value). The problem occurs when we or other authors are searching for a relationship and infer an important causal physiologic relationship based on ``statistical significance.’‘ Thus, we must remember that the p value, the significance of the statistical result, reflects merely the likelihood of chance at the level specified rather than the precision of the result; a lower p value, then, does not make the result more biologically important or clinically ``significant.’‘

DISCUSSION AND CONCLUSIONS

In the Discussion section, the author can attempt to provide an interpretation of findings. Here the author can attach clinical relevance to the reported statistically significant findings. The findings may be compared with those of other studies and interpretations. Possible explanations for results can be postulated and differences from other reports in the literature explained. One would hope that the author bases the conclusions on the findings. This is not always the case. When we discuss the results, we should consider whether they have any meaning in the real world of bedside practice. A ``significant’‘ but relatively small difference in cardiac performance discovered only in carefully controlled circumstances has little resemblance to the constantly changing status of the critically ill patient in whom such a finding may not have any real import. We must ask ourselves whether the demonstrated result is important in influencing or directing bedside practice. We must retain our skepticism and use it to balance enthusiasm.

We will discuss later the concept of power relating to the necessary size of the sample. Authors who conclude that results would have been statistically significant if only a larger sample had been available display their lack of foresight and preparation; clearly, the time to discover the proper sample size is at the outset, the study planning phase. Rather, it would be refreshing to encounter conclusions that forthrightly admitted that the hypothesis was incorrect, that the study showed that therapy did not lead to improvement, or that the investigator headed off on the wrong track. ``Negative reports’‘ of this sort will prevent other investigators from pursuing ideas that turn out to be flawed and can also serve to direct investigators, including themselves, along more fruitful pathways.

The reporting of negative studies has recently been addressed from an editorial standpoint. (14) Angell states, ``... it is widely believed that reports of negative studies are less likely to be published than those of positive studies and some data have been put forth to support this belief ... it is assumed that editors and reviewers are biased against negative studies, considering them less inherently interesting than positive studies. However, a bias against publishing negative studies would distort the scientific literature.’‘ Although she believes that the New England Journal of Medicine publishes fewer negative reports than positive ones, it is not a matter of policy. She asks, ``Does it deal with an important question? Is the information new and interesting? Was the study well done? ... We feel a particular obligation to publish a negative study when it contradicts an earlier study we have published and is of a similar or superior quality. When a good study addresses an important question, the answer is interesting and the work deserves publication whether the result is positive or negative.’‘14

Finally, we should consider whether the conclusions are relevant to the questions posed by the investigators. Far too many papers seem to begin with ``unwarranted assumptions'‘ in the introduction, end with ``foregone conclusions'‘ in the discussion, and contain in between a mass of barely relevant data. If we care to spend the time necessary to review published papers and, in particular, to do the preparation necessary before we embark on our own clinical investigations, such discouraging assessments will occur much less frequently than they now do.

ABSTRACT OR SUMMARY

Although we know that we should spend time in analyzing the medical literature, it is quite clear that, given the pressures of everyday life and the journals that appear with seemingly increasing frequency on our desks each month, we are often tempted to read only the title and the abstract. One final caveat: there may be important disparities among the results, discussion, and abstract. One memorable paper compared two forms of fluid resuscitation. Three patients in one group had been given from two to three times the amount specified in the protocol. With exclusion of these patients properly in the data analysis, as noted in the results section, there were no differences between the two groups. With inclusion of patients with protocol violations, there was a ``statistically significant'‘ difference. The abstract cited the ``statistically significant'‘ analysis without any reference to the patients who should have been excluded. The authors' conclusion of a statistically significant difference in treatment modalities was, in fact, denied by their own results. If you are in a hurry, do not just read the abstract and move on, come back and read the article properly when you have enough time.

FUNDAMENTALS OF BIOSTATISTICS IN MEDICAL RESEARCH

This section is intended to present some of the fundamentals of descriptive statistics and an introduction to inferential statistics primarily to alert the readers to certain important caveats in the interpretation of results of statistical analyses. The author readily acknowledges the limitations of the fundamentals presented here. However, I hope that I will stimulate the reader both to seek expert assistance initially and to pursue further study.

Two distinctive components to statistical methods are

- Descriptive statistics--methods that deal with the organization, summarization, and presentation of data;
- Inferential statistics--methods that deal with the drawing of inferences about populations from results obtained in study samples.

The main contemporary interest in statistics for clinical research is inferential statistics. This leads to the tests of significance, p values, and confidence limits that pervade the medical literature.

DESCRIPTIVE STATISTICS

1) TYPES OF DATA

We encounter two major classes of data: quantitative and categorical. Quantitative data deal with measured quantities such as age, blood pressure, or arterial or oxygen tension. Quantitative data may be subdivided into discrete when only certain values are possible such as the number of patients currently in the ICU or continuous when any value is possible. An example is serum sodium values which could be reported as whole numbers such as 140 meQ/l or in decimals; 140.6 meQ/l, depending on the accuracy of the measuring instrument. Categorical data deal with attributes such as living or dead, hypertension or not, blood type. Similarly, categorical or qualitative variables may be subdivided into ordinal and nominal categories. The foregoing examples such as living or dead or blood type are nominal or named variables. Yes and no categories such as the presence of hypertension or not may be called dichotomous. Ordinal variables consist of numbers, such as a ranking scale for pain. Because these are subjective assessments rather than objective measurements, they are not considered continuous quantitative variables.

2) FOURFOLD TABLE FOR QUANTAL DATA

Many medical investigations involve the comparison of two groups: a ``study group'‘ and a ``comparison group.'‘ When the data are categorical, the fourfold or 2 2 table is a convenient device for summarizing the results. The fourfold table has two columns, corresponding to the study and comparison groups, and two rows, corresponding to the dichotomous data collected (e.g., success or failure, live or die, respond or not respond). The four cells in the body of the table provide the study results.

The obvious descriptive measure with quantal data is the percentage with the attribute. From a fourfold table, one wishes to compare the percentages with the attribute in the study and comparison groups. For simplicity, we will choose examples from one study to illustrate various points.6 The study design consisted of randomizing patients into groups with two different end points for the titration of PEEP. In group I, the authors used 5 cm H2O PEEP to maintain PaO2 greater than 65 mm Hg at an inspired oxygen fraction (FIO2) of 0.45. If PaO2 fell below this level, PEEP was titrated until PaO2 rose to the treatment goal, 65 mm Hg. In group II, the authors increased PEEP until physiologic shunt was reduced to less than 0.2. The authors compared the two groups for outcome and resources used: 6 (33%) of 18 patients in group I died; 5 (25%) of 20 patients in group II died. This can be displayed as a fourfold table.

As we shall see later, depending on the situation, one may wish to compare the percentages by examining their difference, their ratio, or, in epidemiology, their odds ratio. Authors should always report the numbers used to calculate percentages. In fact many journals and abstract rules require that the basic data be reported.

3) DESCRIPTIVE MEASURES FOR QUANTITATIVE DATA

The two descriptive characteristics of prime importance are location and spread.

For location, the most common measures used are the mean (arithmetic average), median (middlemost value), and mode (most frequently occurring). For purposes of statistical inference, the mean and median are most frequently used. The mode is rarely used in clinical research.

For spread, the most obvious is the range (highest minus lowest), but it has limited use for inference. The measure of spread ubiquitous in research is the standard deviation (SD), defined as

SD =__(x x)2/n 1

A convention many investigators follow with studies involving quantitative data is to summarize results as mean SD.

WARNING

There is another convention that involves expression of results as mean standard error (SE) where SE is some other quantity (we shall discuss this later). The author’s onus in a medical paper is to indicate whether the number following the is SD or SE.

4) NORMAL OR GAUSSIAN DISTRIBUTION

The normal or gaussian distribution is a theoretical mathematical curve (i.e., one can write an equation for it) that has particular importance in statistical work. It has the familiar bell-shaped appearance. The distribution can be described entirely by its mean and SD.

Empirically, the distribution of many quantitative measurements tends to approximate a normal distribution in shape (e.g., weight, examination scores, IQs).

Particular properties of the normal distribution are that its

- Mean <pm> 1 SD encompasses about 65% of the measurements;
- Mean <pm> 2 SD encompasses about 95% of the measurements;
- Mean <pm> 3 SD encompasses almost all measurements.

Unfortunately, many populations in critical care are not normal in the sense that the distribution of many quantitative measurements does not tend to approximate a Gaussian distribution. For instance, in any clinical series, a few patients tend to remain in the hospital far longer and tend to skew the distribution to the right on a scale of increasing duration of hospital stay. Since hospital charges are proportional to the time spent in the hospital, one might anticipate that the distribution of hospital charges would not be normal. In the PEEP study, the hospital charges in group I patients were $32,000 $18,000 (mean SD).6 Clearly, hospital charges do not approximate a normal distribution. If we had naively calculated mean 2 to encompass 95% of the patients, we would see that the lowest value is an impossible negative charge, $4000. Whether or not data under study follow an approximately normal distribution can influence the choice of method for analysis. In the above example, to compare charges in the two study groups--rather than use the t-test, which has an implicit assumption that the data follow an approximate normal distribution, the Mann---Whitney U test was used, which does not entail any assumption regarding the shape of the distribution of the data. With the wide availability of statistical computer packages, one can readily calculate means and standard deviations, regardless of whether the data are normally distributed (even with ordinal data such as ranking scales). If all of the data are of similar sign, (that is all positive or all negative) and the standard deviation is greater than half of the mean, the data clearly will not be normally distributed. The median and range can be used as descriptive statistics and, in addition to the Mann-Whitney U test for testing differences between two unpaired treatments, other nonparametric tests of significance are available. These include the Wilcoxon signed rank test which tests for differences between two paired treatments, the Kruskal-Wallis analysis of variance (see below) for testing differences between more than two treatments, and the Spearman rank correlation which can be used to test the strength of association between two variables.

The more important use of the normal distribution in statistics is in the process of statistical inference, which we are now ready to discuss.

STATISTICAL INFERENCE

1) BACKGROUND AND OBJECTIVES

The ultimate result of statistical tests of significance is presentation in a report of a ``p value'‘ and a claim that findings were ``statistically significant'‘ or ``not statistically significant.'‘ These claims result from a series of numerical calculations. In today’s world of electronic aids we can deemphasize the calculation ritual and emphasize the underlying rationale for the statistical test of significance and the proper interpretation of its result.

We will focus on the comparison of two groups of observations: a study group and a comparison group. The data under consideration could be categorical or quantitative. In fact, the nature of the data involved in the comparison dictates the statistical methods for analysis. Categorical data lead to chi-square tests (sometimes cited with Yates’ correction), whereas quantitative data lead to t-tests (or, as occur alternatively, nonparametric Wilcoxon rank tests or the Mann---Whitney test).

2) CONCEPT OF SAMPLING VARIATIONS AND DEFINITION OF SE

Basic to statistical inference is the underlying concept of sampling variation, namely, that any quantity calculated in a sample (proportion, mean, median, SD, and so on) will differ with different samples (of the same size) from an underlying population. The variation in this quantity from sample to sample is sampling variation.

For example, consider quantal data and calculation in a sample of the proportion of persons who have the attribute under study. Think of many repeat samples of the same size from this population with, in each sample, determination of the proportion with the attribute. The variation among the proportions in the various repeat samples is the sampling variation of a proportion.

As a second example, consider quantitative data and the calculation of a mean for the sample. Think of many repeat samples of the same size from this population and, with each sample, determination of the mean. The variation among means in the repeat samples is the sampling variation of a mean.

If we now consider use of the SD to describe variation, we come to the definition of SE, the SD for sampling variation of some quantity (proportion, mean, and so on) calculated in each of the repeat samples. For our first example, the SD among proportions for repeat samples of the same size is the SE of a proportion. For our second example, the SD among means for repeat samples of the same size is the SE of a mean.

Finally, without going through the laborious process of empirically obtaining repeat samples, statistical theory indicates that a for proportions, the SE of a proportion (SEp) is

SEp = _P(1 P)/n

where n = sample size and P = proportion in the population with the attribute; and b for means, the SE of a mean (SE) is

SEx = s/n

where n = sample size and s = SD in the population.

Notes. When the SD (s) in the entire population is unknown, which is almost always the case, one uses the SD in the sample, s, as an estimate of s. Thus,

est SEx = s/_n

The choice of mean SD or mean SE depends on the purported use of the data. If the intention is to indicate variation of individual values about a mean, the choice is meanl SD. If the intention is to indicate sampling variation in many different samples or stability of the mean and to conduct a statistical inference regarding the mean, then the choice is mean SE.

The calculation of SE results in a smaller number than SD because it is SD divided by the square root of the sample size. Sometimes, however, we will find mean SE used to indicate variation of individual values about a mean in a single sample even though the choice should be mean SD. This could be a mistake in transforming or selecting the data, but another inference is that the authors believe this somehow ``tightens'‘ the data by appearing to mask the actual variation of individual values in the sample.

An amazing but mathematically provable feature is that if a distribution of the means or proportions from repeat samples of the same size is plotted, the distribution tends to follow a normal or Gaussian curve. The mathematical derivation of this is called the Central Limit Theorem. The fact that sampling distributions of a mean and a proportion are normal is the reason why the normal distribution is the basis for many of the methods of statistical inference we encounter in research.

3) SPECIFICATIONS FOR A TEST OF SIGNIFICANCE

Three specifications are necessary to perform a test of significance. One must specify a null hypothesis, set a significance level, and determine whether one wishes to conduct a one- or two-sided test.

NULL HYPOTHESIS

The null hypothesis generally states that there is no difference between two groups or no effect of treatment. In the comparison of two groups with categorical data, the null hypothesis concerns the proportions with the attributes in the two larger ``study'‘ and ``comparison'‘ populations from which the study data came. The obvious null hypothesis is that the proportions with the attribute are the same in the ``study'‘ and ``comparison'‘ populations. The PEEP study compared mortality rates in the two groups.6 The null hypothesis states that, in the underlying populations from which these study samples come, there is no difference in mortality rates.

For the comparison of two groups with the quantitative data, the null hypothesis is that the means in the underlying study and comparison populations are identical, or, alternatively, that the difference in population means is zero.

SIGNIFICANCE LEVEL

As we shall see, the test of significance involves making a decision under uncertainty and based on chance. Setting a significance level is the arbitrary selection of a small enough chance for making a choice. Convention in both medical and nonmedical research dictates 0.05 or 5% and 0.01 or 1% as the typical levels of significance. Choice of 5% indicates that an event occurring only 1 time in 20 or less is sufficiently rare to risk drawing a conclusion that excludes chance as a likely explanation of what was observed. Choice of 1% is more conservative and indicates that an event occurring only 1 time in 100 or less is sufficiently rare to risk drawing a conclusion that excludes chance as a likely explanation of what was observed.

How should we determine whether 5% or 1% or some other value is a proper level of significance? Because this is an arbitrary selection, a number of factors bear on this selection process. For instance, if a particular form of treatment carries a high risk of serious side-effects, we might choose a level of 1%. On the other hand, if the disease has an extremely high mortality rate, we might raise the level of significance to 5%, allowing a greater possibility for chance to have produced the results but not discarding a real result based on too strict a criterion. If the problem addressed is widespread and the proposed remedy simple and inexpensive, we might elect a higher level of significance, whereas an expensive, complicated, difficult-to-effect modality of therapy might more properly be judged by a lower significance level.

To refer again to the PEEP study, the chosen level of significance was 0.05. The level of significance was chosen because the mortality rate may have been considered high (favoring a higher level), the therapy was simple, and neither arm entailed higher charges or more complicated interventions.

ONE- OR TWO-SIDED TEST

This pertains to the nature of the alternatives one wishes to entertain in contrast to the null hypothesis. More specifically, if in the comparison of two groups one considers as alternatives to the null hypothesis only that the population mean or proportion in the study group may be higher than that in the comparison group, this is a one-sided test. If, however, one considers the possibility that the population mean or proportion in the study group may be either higher or lower than that in the comparison group, this is a two-sided test. In general, the two-sided test is the more conservative. In trials, although one may have every anticipation that the new therapy will perform better than the standard, there is often the possibility that it may perform worse. Hence, one would adapt the more conservative two-sided test in such a situation.

If we are increasing PEEP to attempt to increase arterial oxygen tension, we might analyze our results with a one-sided test. If we chose a two-sided test, we entertain the possibility that PEEP lowers the PO2. In practice, we chose the more conservative two-sided test to compare mortality rates in the two groups of the PEEP study.

4) RATIONALE FOR THE TEST OF SIGNIFICANCE

Central to the application of statistical methods is the notion that a study consists of a random sample from some underlying population. One calculates a descriptive statistic in the sample, a proportion (or rate or percentage) for quantal data, and a mean for quantitative data. The inference to be drawn concerns the respective statistic in the underlying population, that is, the population proportion or the population mean.

The situation we have prescribed then concerns two populations: a ``study'‘ and ``comparison'‘ and two samples, one from the study and one from the comparison population. We further presume that with quantal data we have calculated the sample proportion in each of our study and comparison samples. In the PEEP example, the sample proportions are 33.3% and 25.0%, respectively, in groups I and II. With quantitative data we presume calculation of the sample mean in each of our study and comparison samples.

The rationale for the test of significance is that we presume that the null hypothesis we have specified regarding the population(s) is true. We then determine, using methods based on the mathematical theory of probability, the chance that we would obtain results in our sample(s) that were as extreme as or more extreme than what we have actually observed. If this chance is sufficiently small, then we claim that the results obtained with our sample(s) are not compatible with the specified null hypothesis and our study has provided us evidence to refute the null hypothesis. This is the meaning of statistically significant.

If the chance we determine is not sufficiently small, then we claim that the results obtained with our sample(s) are indeed compatible with the specified null hypothesis and our study provides no evidence to refute the null hypothesis. This is the meaning of not statistically significant.

By ``sufficiently small,'‘ we mean the chance we had selected with our significance level specification, namely, the conventional 5% or 1%.

Finally, if we determine the chance of extremities in only one direction from the null hypothesis specification, we are performing a one-sided test. If we determine the chance of extremities in either direction from the null hypothesis specification, we are performing a two-sided test.

Stated simply, the test of significance is an indication of whether chance is a likely (not statistically significant) or an unlikely (statistically significant) explanation of the discrepancy between the stated null hypothesis and the observed results in the study sample(s).

5) DISTINCTION BETWEEN PAIRED AND INDEPENDENT SAMPLES FOR THE COMPARISON OF TWO GROUPS

We need to distinguish between two different forms of comparative studies: paired samples and independent samples. Paired samples occur when each observation in the study sample has, by the nature of the investigation, a matching or paired observation in the comparison sample. The most obvious paired situation is a ``before---after'‘ study or one in which the patient serves as his or her own control. Sometimes investigators individually match subjects on various characteristics (e.g., age, sex, race, socioeconomic status) and conduct a paired sample study.

When, as described in the study methodology, there is no evident individual matching of study with comparison sample observations, the design is obviously independent samples.

6) FOUR SITUATIONS FOR THE COMPARISON OF TWO GROUPS

With two types of data (quantal and quantitative) and two study designs (paired samples and independent samples), we have four situations for the comparison of two groups.

7) EXAMPLE

To illustrate the above points, consider the PEEP study discussed previously. Patients were randomized to one of two study groups, with 18 assigned to group I and 20 to group II. Among the various study outcomes investigated, Consider a test of significance comparing mortality rates in the two groups. A perusal of the Methods section of the paper indicates no individual matching or pairing of patients in the two groups; hence, we are dealing with independent samples. The data for this particular comparison, ``Died'‘ or ``Did Not Die,'‘ are clearly categorical. Hence, our interest is in comparison of proportions in independent samples, which, indicates chi-square as an appropriate method of statistical analysis.

Recall that we have stated previously that we are testing the null hypothesis of no difference in mortality in the underlying populations for the two interventions under study. When we perform our test of significance we are asking: if the null hypothesis is correct and there is no real difference, how often by chance alone would we obtain a mortality difference in samples of n = 18 in group I and n = 20 in group II of 33.3% 25.0% = 8.3% or something more extreme? When we specified a two-tailed test, we stated our interest in considering deviations as extreme or more extreme in both directions from the null hypothesis value of zero.

Following the calculation ritual for chi-square in independent samples, we obtain a chi-square value (with use of Yates’ correction and 1 df) of 0.04. From a table of the chi-square distribution, this yields a probability (p value) of 0.84 or 84%. Having chosen a significance level of 5%--that is, we considered 5% or 1 chance in 20 as sufficiently infrequent--clearly 84% far exceeds 5%. Hence, the test of significance concludes ``not statistically significant.’‘ In other words, chance or sampling variation does indeed constitute a plausible explanation for the observed difference in mortality rates for the two study groups. Thus, on the basis of the results of this study, we have no reason to doubt the null hypothesis. The observed difference in mortality rates between the two study groups is well within the possibility of chance variation.

The McMemar test is another type of Chi square analysis which one can use to compare proportions in a before and after determination on the same group of subjects, namely, a paired sample of categorical data.

In addition to mortality rate, the PEEP study examined other variables including number of days in the SICU, number of days intubated, number of arterial and venous blood gas determinations, number of inotropic or vasoactive drugs administered, and frequency with which more than 5 cm H2O PEEP was required to achieve the desired end points. The first five variables involve quantitative data, and, for each variable, the respective means and SDs for the 18 patients in group I and the 20 patients in group II were calculated. In each instance, a test of significance was conducted to compare the two group means. Since this is a comparison of two means in independent samples, a relevant calculation ritual is the independent samples t-test.

In each instance, the null hypothesis under test is that, in the underlying population from which these study data came, there is no difference in the means. Each test involved choice of a 5% significance level and use of a two-sided test. In each instance, the outcome of the test of significance was ``not significant’‘ (p 0.05).

Interpretation of these results is that in each of the five tests comparing means, the results observed in the study samples are indeed compatible with the null hypothesis statement that, in the respective populations, there is no difference in mean effects for each of these variables. In other words, if each of the null hypotheses of no difference in means is true, the results observed in the study sample, or something more extreme in either direction from the null, could well have occurred by chance or by sampling variation. Thus, on the basis of these study results, the two interventions do not differ with respect to mean effects for each of these variables.

The final data item analyzed gives the percentage of patients who required more than 5 cm H2O of PEEP to achieve the desired end points. As with the mortality data, comparison of the two groups on this variable entails a comparison of two proportions in independent samples and entails the chi-square test. Performing the same calculation ritual as with the mortality data leads to a chi-square (with use of Yates’ correction and 1 df) of 15.6. This yields a p value of 0.00008, or, more roughly, p 0.0001 (i.e., less than a 1 in 10,000 chance). Here, too, the null hypothesis is equal population frequency for the two interventions. The chosen significance level is 5% (or 1 chance in 20), and a two-tailed test was selected. Clearly, the p value we calculated, 1 in 10,000, is far below our chosen significance level, 1 in 20. Hence, our results are ``statistically significant’‘ (p 0.05). In other words, the sample results in this study are not compatible with the null hypothesis of no difference in the two interventions for this variable. The results we observed are unlikely to have occurred by chance or sampling variation alone. Thus, our study does provide sufficient evidence to doubt or refute the null hypothesis that these two interventions require the same frequency of use of more than 5 cm H2O of PEEP. We could state, ``Significantly more patients in the group II intervention compared with group I required more than 5 cm H2O of PEEP to achieve the desired end-points (p  0.05).’‘ Alternatively, the previous statement could appear with a parenthetical ``p 0.0001,’‘ or even with the actual p value calculated, ``p = 0.00008.’‘ In other words, this difference in sample rates, 27.8% in group I and 95% in group II, is most unlikely to have occurred purely by chance, and we can construe this as substantial evidence in this study to refute the null hypothesis.

INTERPRETATION OF STUDY RESULTS

(1) ``MULTIPLE PEEKS‘ PROBLEM: REPEAT ANALYSES OF ACCUMULATING DATA

Worshipping at the shrine of the 5% significance level, an interesting suggestion is to analyze accumulating data in a study after every few observations and to stop the study as soon as the cumulative data achieve statistical significance at the 5% level. Statistically, this is an entirely fallacious procedure. The actual significance level with this procedure will not be 5% but something much higher, thus increasing the likelihood that chance alone is responsible for the ``difference'' observed. Using conventional statistical tests such as t-tests and chi-square, the actual significance level will depend on the number of peeks or analyses performed with the accumulating data. In fact, it can be shown mathematically that even if the null hypothesis is exactly true, the investigator will ultimately find a statistically significant result although he or she may have to continue obtaining observations and analyzing the cumulative results for a long time.

If one wishes to take multiple peeks with repeat analyses of data as they accumulate, one must use special statistical techniques ($_{15}$) devised for this situation, namely, sequential methods (see also reference 7, Chap. 8).

(2) ``MULTIPLE COMPARISONS'' PROBLEM: MANY TESTS OF SIGNIFICANCE WITHIN A SINGLE STUDY

With the ability to collect many measurements on a wide variety of variables in a single study, and the availability of computers to grind out the calculation rituals for tests of significance, a study may involve a large number of tests of significance and the reporting of their respective p values.

With many variables tested, their corresponding p values require cautious interpretation. As we shall indicate a bit later, one way to view the chosen significance level is to consider it as the chance to make an incorrect decision and reject a null hypothesis that is true. Thus, specification of a 5% significance level means that we are willing to risk a 5% or 1 in 20 chance of coming to the wrong conclusion in our analysis and rejecting the null hypothesis when it is true. What this means is that among every 100 tests of significance we encounter in which the null hypothesis actually is true, five have resulted in the erroneous conclusion of statistical significance.

Thus, if a study involves 100 tests of significance with various data items, we would anticipate that 5% or five tests would produce statistically significant results by chance alone. The extreme in multiple statistical testing is to let the statistical analysis drive the investigation. An investigator may decide that rather than specify particular hypotheses for testing, he or she will conduct a study collecting information on as many variables as possible and then, from the tests of significance for each variable, choose and report those variables that turn out statistically significant at the 5% level. Not only is this improper use of statistical methods, but also it is poor science. It is often termed a ``fishing expedition,'' but in reality it would be a poor study design for a fishing trip as well--too much ocean and too few fish!

When a study does involve several statistical comparisons, specific techniques are available for dealing with this situation, namely, multiple comparisons procedures, which can preserve the predetermined significance level for the statistical testing. ($_{16}$,$_{17}$) Although, for illustrative purposes, we did not use any multiple comparisons procedure, certainly the PEEP study results would lend themselves to consideration of a multiple comparisons procedure.

Thus far we have limited discussion to comparison of two groups by means of t tests for quantitative data. If there are more than two groups, the appropriate test would be the analysis of variance, often abbreviated ANOVA. To determine statistical significance, this test analyzes the variance between groups and within groups. The null hypothesis in this instance states that there are no differences among any of the group means. However, if the null hypothesis is rejected, and analysis reveals that a difference is unlikely due to chance, ANOVA does not identify which group or groups are different. One must perform additional analysis. Again, most common statistical programs print the subsequent multiple comparisons procedures as part of the statistical results in the ANOVA package. Which test is appropriate for your study or determining whether the authors have used the appropriate test will require consultation with a statistician or further study.

Often times, clinical studies entail measurements of response of the variable under study over time in two or more different groups. Here, repeated measures ANOVA is the appropriate statistical test. For instance, we could use repeated measures ANOVA to compare the effect of PEEP on the 1st, 2nd and 3rd days of the study between the “adequate oxygenation” and “minimal shunt” groups.

(3) PLAYING BY THE RULES OF THE GAME

We realize that the structure of statistical tests of significance involves some rather arbitrary choices (such as choice of a significance level) and some perhaps unrealistic oversimplification in its dichotomization of the world (such as ``significant'' or ``not significant''). If one has chosen this hypothesis-testing framework for drawing inferences from study findings, one should report results properly, with correct terminology, according to the rules of the game and without editorializing. Thus, within the framework, after one has chosen a significance level, results are either ``significant’‘ or ``not significant’‘ at that chosen level. It is customary to indicate the smallest significance level at which that result would have been significant. For example, although the 5% significance level was chosen for the test, the calculated p value for comparison of percentage of patients requiring more than 5 cm H2O of PEEP was 0.0008. Hence, the results could be reported with one of the following three notations: p 0.05, p 0.0001, pl= 0.0008. What is irksome is when results are not significant at the chosen level, but authors often write, ``There was a clear trend toward significance’‘ or ``The findings suggest a difference, but they did not achieve statistical significance.’‘ Clearly, results are or are not significant. The qualifications of ``a clear trend’‘ or ``suggest a difference’‘ constitute unnecessary editorializing. Comments such as these, if made it all, should appear in the Discussion section of a paper and not in the Results. The astute reader is indeed wary of the arbitrariness and limitations of the process of statistical testing and does not need additional editorializing by the authors as he or she assesses a study’s findings.

Another perspective on this is that deployment of tests of significance according to the rules does mean that, say at a chosen 5% significance level, a calculated p value of 0.047 is ``significant’‘ whereas nearly the same sample result that leads to a calculated p value of 0.053 is ``not significant.’‘ The experienced researcher and reader are well aware of this limitation and apparent paradox with tests of significance. But, if one has chosen to play the game, one must play by the rules and not hedge or qualify the reporting of the findings.

The limitations and arbitrariness surrounding tests of significance are reasons why many investigators, and many biostatisticians, are shifting toward confidence limits for reporting research results. Although the same principles of statistical inference are involved in calculation and interpretation of a confidence interval, it provides more information than the mere reporting of a p value from a test of significance.

TYPE I AND TYPE II ERRORS

With the use of statistical tests of significance to test hypotheses, it is important to understand the two errors that arise in the conclusions drawn:

1) TYPE I OR A ERROR

When we choose the 5% significance, we, alternatively, have stated that we are willing to risk 5% chance of erroneously rejecting a true null hypothesis (i.e., claiming statistical significance when, in fact, the null hypothesis is true). This incorrect decision or error is called the type I error, a error, or error of the first kind.

(2) TYPE II OR ß ERROR

Obviously, the above definition implies that there is a corresponding type II error, ß error, or error of the second kind. In this situation, the null hypothesis is false, and some alternative hypothesis regarding the population values prevails. The incorrect conclusion we make is to fail to reject the null hypothesis when, in fact, the null hypothesis is false. In other words, ß or type II error is coming to the erroneous conclusion ``not statistically significant’‘ when, in fact, the null hypothesis is false and there really is a difference between the study and comparison groups.

We, as readers and authors, must make some decisions with respect to the possibility of introducing either type I or type II errors. For instance, if a particular form of treatment is hazardous and has life-threatening side-effects, we would want to be sure that there is a true difference before rejecting the null hypothesis. We may, therefore, choose a lower level of significance (p value = 0.01 rather than p value = 0.05) to make it less likely that the difference is a result of chance alone. We have diminished the likelihood of introducing a hazardous form of therapy based on results that, indeed, were the outcome of chance. On the other hand, if a safe form of treatment has a small therapeutic effect, we might not wish to fall into a type II error, in other words, coming to the erroneous conclusion that there was not a ``statistically significant’‘ difference between the two groups when, in fact, there was a true difference.

(3) POWER

Power of a statistical test is defined simply as the complement of the ß or type II error, that is,

Power = 1 ß

Thus, the power of a statistical test refers to the chance of correctly rejecting the null hypothesis when, in fact, the null hypothesis is false.

PLANNING THE SAMPLE SIZE OF A STUDY

With the above definitions we have three important quantities involved in a statistical test of significance:

1. the sample size, n

2. a error associated with the null hypothesis
3. ß error associated with an alternative hypothesis

Provision of any two of the above items will allow us to determine the value of the third. For example, with specification of (i) the sample size of the study and (ii) the a error chosen for the null hypothesis, we can determine, for any specific alternative hypothesis, the ß error, that is, the chance of erroneously concluding ``not statistically significant'‘ when, in fact, that specific alternative hypothesis is true.

Of more use, perhaps, is that with specification of (ii) the a error associated with the null hypothesis and (iii) the ß error for some specific alternative hypothesis, we can then determine n, the sample size needed for the study. For comparative investigations, this means that an investigator must provide an a error and a ß error for a specific difference (usually called a clinically meaningful difference). (Alternatively, rather than ß error, one could just as well frame the specification in terms of power, that is, the study is planned to have a certain specified power for a specific difference.)

The above is the rationale for the answer the statistician provides to the question, ``How big a sample do I need?'‘ The investigator must specify the a and ß errors, have some idea of the underlying variability in the measurements under consideration (expected SD), and select a value to represent a true difference between the experimental and control groups.

The equation to calculate the size of the sample contains the SD and the terms for the percentage point of a normal distribution for a and ß errors in the numerator and the true difference in the denominator. A higher specified level of significance for each will result in a lower numerator and a lower quotient or sample number. Thus, if we select an a or ß error of 0.05 compared with 0.01, we will need a smaller sample. On the other hand, if there is a greater variability in the measurements, a large sample size will be necessary. The sample size also depends on the value selected for a true difference. To determine whether mortality rate is reduced from 25% to 10% will take a smaller sample than an anticipated reduction from 25% to 20%. If the existing mortality rate is low (4%), therapy that would truly result in halving the rate to 2% would require a study involving a very large number of patients (for further details and methods of calculation, see reference 7, pp 142---146).

Consider the PEEP study we used earlier as an example to illustrate tests of significance. Consider the mortality data in which the mortality rate in both groups combined is roughly 30%. Suppose we plan some new intervention that we anticipate will reduce mortality. How large a comparative study do we need to test this new intervention (study group) against conventional therapy (comparison group)?

Clearly, the null hypothesis is that mortality in the study and that in the comparison groups are identical. When confronted with the statistical question regarding type I error, suppose we indicate that we plan to perform a two-sided test at the 5% level. We chose a two-sided test because even though we have every anticipation that our new intervention will reduce mortality, there is some possibility that it might result in an increase in mortality. If such an adverse consequence did occur, we certainly would want to be able to detect such a possibility in our analysis. Hence, we have opted for the more conservative two-sided test. When we chose the 5% significance level or a = 0.05, we have specified our type I error; that is, we have indicated that we wish a 5% or 1 in 20 chance of erroneously rejecting the null hypothesis, namely, erroneously claiming a ``statistically significant'‘ difference in our study when, in fact, there is no real difference in mortality.

We now have to specify our type II error. As a first step, we must choose what difference in mortality we consider clinically important. Suppose we choose a difference of 10%. This means that if, in fact, our new intervention reduces mortality by 10%, say, from 30% to 20%, we deem this as a clinically important effect that we wish to detect. In fact, when pressed, we might state that we wish to have a 10% type II error for this effect. In other words, we wish to have only a 10% or 1 in 10 chance of erroneously failing to reject the null hypothesis when this magnitude of effect exists, that is, if there is, in reality, an effect of reduction of mortality by 10% (i.e., absolute 10% difference), we wish our study to have only 1 chance in 10 of arriving at the erroneous conclusion of ``no significant difference'‘ in mortality.

We could, alternatively, have made the above specification in terms of statistical power rather than type II error. We would then say that if the real reduction in mortality is 10% (i.e., absolute 10% difference), we wish our study to have 90% power to detect such a difference. In other words, we wish to have a study such that there are 9 chances in 10 that our results will lead to the correct conclusion of ``statistically significant difference'‘ when, in fact, a true difference in mortality of this magnitude exists.

With these specifications, calculations reveal that we need a sample size of 420 patients in each of the study and comparison groups or a total of 840 patients in the investigation. If this number is beyond the resources available to us, we would have to relinquish something in our error specifications. Our sample size determination would yield a smaller number if we were to increase a error (e.g., from 5% to 10%) or increase our ß error (e.g., from 10% to 20%) or choose a larger clinically meaningful minimum difference in mortality rate (e.g., 15% instead of 10%).

Another approach to the collective bargaining between biostatistician and clinical investigator in determination of sample size is for the investigator to indicate to the biostatistician the maximum number of patients that he or she can anticipate for the study. The biostatistician can then determine, for various alternative choices of ``clinically meaningful differences,'' just what statistical power the study would have to determine such differences.

ANALYZING DIAGNOSTIC TESTS

Contempory medical literature abounds with descriptions of attributes of new diagnostic tests such as sensitivity and specificity. The definitions and calcuation of these terms can be derived from a 2 X 2 table.

ATTRIBUTE

Present Absent

TEST POSITIVE a b

TEST NEGATIVE c d

Sensitivity is the ability of a test to detect a disease and is calculated by dividing the true positives, the number of times the test is positive when the attribute is present, (a), by the same number plus the false negatives, the number of times the test is negative when the condition is present, (c). The forumla then is a/(a + c). Specificity refers to the test being negative when the disease is not present. It can be calculated by taking the true negatives, the number of times that the test is negative in the absence of disease, (d), and dividing it by the same quantity plus b or false positives, the test is positive although the attribute is absent. The formula for specificity thus is d/(d + b). Two other commonly used terms are positive and negative predictive values. The positive predictive value is the chance of having the attribute if the test is positive. This is calculated as the ratio of the true positives (a) in which the test is positive when the attribute is present and the sum of all situations in which the test is positive, which includes true positives (a) and false positives (b). The formula for positive predictive value then is a/(a + b). Negative predictive value states the accuracy of the test to exclude the attribute if the test is negative. It is calculated from the ratio of true negatives (d) and all negative test results, (d + c), both true and false negatives. The formula for negative predictive value is d/(d + c).

CONFIDENCE INTERVALS (LIMITS)

Confidence intervals are alternatives to tests of significance for drawing inferences regarding populations from observations in a sample. They are based on the same considerations of sampling variation as discussed with tests of significance. From results in a sample, a calculation ritual leads to determination of confidence limits and the confidence interval. The interval gives a range of values within which the true underlying population value lies. If a 95% confidence interval is calculated, the chance is 95% or 19 in 20 that the limits calculated embrace the true population value; the smaller the sample size, the wider the confidence interval.

For the comparison of two groups, we can calculate confidence intervals on the difference in percentages or in means. These, as indicated above, provide an interval within which the true difference in population proportions or means lies.

Again, consider the PEEP study as an illustration. With regard to mortality, the difference observed in the study was 33.3% 25.0% = 8.3%. Calculation of 95% confidence limits on this difference yields 20.6% to 37.2%. Thus, based on our sample results, we state with 95% confidence that in the population from which the study samples came, the difference in mortality rates is somewhere between 21% (i.e., an absolute difference in mortality rate 21% higher in group II than in group I) and + 37% (i.e., an absolute difference in mortality rate 37% higher in group I than in group II). The term ``95% confidence'' means that when we state that the true population difference is somewhere between 21% to + 37%, there is a 95% or 19 in 20 chance that these limits do embrace the true population difference; there is a 5% or 1 in 20 chance that the true population difference is outside these limits.

Clearly, the limits calculated above are wide due to the small sample sizes involved, namely, about 20 patients per group. Confidence limits are sensitive to sample size and shrink as sample size increases. For example, if these same sample results, namely, 33.3% and 25%, had arisen from a study quintupled in size, that is, with 100 patients in each of groups I and II, 95% confidence limits on the true population difference in percentage mortality would be 4.2% to + 20.9%.

Confidence limits are compatible with the results of tests of significance. For example, if 95% confidence limits on the difference in two proportions or means include zero, then a two-tailed test at the 5% level would yield the result ``not statistically significant.’‘ If the 95% confidence limits does not include zero, then the two-tailed test at the 5% level would yield the result ``statistically significant.’‘ In the PEEP study just mentioned, the 95% confidence intervals (21% to + 37%) included zero, which is compatible with the test result of ``no significant difference’‘ in mortality rates.

If, in a comparative study of proportions, one has interest in the ratio of the two proportions (i.e., relative risk), the confidence interval on the ratio can be calculated. If, for example, a 95% confidence interval of the ratio of two rates includes 1, there is no significant increase in relative risk, tested by a two-tailed test at the 5% level. If 1 is outside the 95% confidence interval, there is a significant increase in relative risk, with a two-tailed test at the 5% level.

One often encounters calculation of the odds ratio to give an estimate of relative risk, particularly in case-control studies. It can be calculated from a 2 X 2 contingency table. The ratio is calculated from the product of a (treatment used in cases) and d (treatment not used in controls) divided by b (treatment not used in cases) times c (treatment used in controls). The formula then is ad/bc.

Confidence intervals are gaining in popularity in clinical research. Particularly for epidemiologic studies, they provide more useful information than corresponding tests of significance.

However, often authors do not calculate confidence intervals. This is especially important when no occurrences of a particular outcome have occurred in a relatively small study group. It is erroneous to extrapolate this finding of zero responses to the general population. It has been shown that with sample sizes greater than 30, at a p = 0.05, ``... if none of n patients shows the event about which we are concerned, we can be 95% confident that the chance of this event is at most 3 in n.’‘ ([18]) For instance, in a study of in-hospital cardiopulmonary resuscitation, Taffet and colleagues ([19]) reported that none of the n = 68 patients older than age 70 who received CPR survived until discharge. Although the authors did not recommend setting age limits for CPR, one fears an inference of ``hopelessness’‘ from the study of of CPR in patients over the age of 70. Using the rule of 3/n, a one-sided upper 95% confidence limit is that we might reasonably expect up to 4.4% of patients over the age of 70 to survive. We should then discuss futility as a condition for withholding CPR with this latter figure in mind as a reasonable upper limit. It might well be that some might judge that, indeed, a less than 5% chance of survival justifies withholding CPR. However, this issue needs direct discussion, and we should clearly not draw the inference that that study proved that no patient over the age of 70 could possibly survive CPR.

WARNING

With confidence limits on a mean, it is important to note that these provide an interval within which the population mean likely lies. The confidence interval does not provide limits within which individual observations lie. It is entirely incorrect to interpret, for example, a 95% confidence interval on a mean as limits that encompass the values for 95% of individual subjects.

OTHER COMMON REGRESSION ANALYSES

In addition to linear regression, described previously (Y = a + bx), there are other mathematical relationships linking two variables including curvilinear (quadratic, cubic, etc.) or logarithmic functions. Once again all of the commonly used statistical packages provide these functions (perhaps too) easily and conveniently. There are also multiple regression techniques which can be used when there are several independent variables. The Harris-Benedict equation to predict energy expenditure is a commonly used multiple regression equation. The dependent variable predicted must be quantitative although both qualitative and quantitative variables can be used as predictors. Qualitative variables are described as present (using a coefficient of one) or absent (zero as coeffecient which cancels the term when multiplied). When the predicted or dependent variable is categorical, such as living or dead, the appropriate regression technique is logistic regression. The many severity of illness indices, such as APACHE and the Mortality Prediction Model, exemplify the use of logistic regression to calculate the risk of mortality.

CODA

In a way, we can compare our current medical journals with daily newspapers and television news programs. They report the most ``up-to-the-minute'‘ information; the reports may be incomplete, certain important details may be missing, and it may be difficult to fit the results reported into the existing framework of previous information. Sometimes this may indicate that the existing framework needs to be changed sharply; at other times, subsequent reports reveal flaws in the methodology or in the interpretation of the ``newest'‘ report.

Inertia may be interpreted in terms of the difficulty in moving a stationary object. Remember, too, inertia also refers to the difficulty in moving an object from the direction in which it is moving. When the preponderance of data points in one direction, we should not be too hasty in changing directions based on a single study that shows the opposite result. On the other hand, the purpose of this entire exercise, reading and contributing to the medical literature, is not only to add to existing knowledge or sharpen the focus, but also to change directions when necessary. A sense of proportion or balance is necessary: to choose a level of significance; to determine and accommodate types I and II errors; to distinguish chance from real effects; to separate statistical likelihood of difference from the importance of that difference; to give clinical dimensions to real experimental differences; to weigh costs and detrimental effects of therapy against the beneficial effects of improving outcome in devastating illness; or to improve the quality of life or diminish the costs of care when survival is not the only important determinant of outcome.

If the medical literature bears any resemblance to the news media in general, our tasks can only become more difficult in the future. As more and more information becomes more easily available with less and less validation, we may pass to a future that goes beyond Andy Warhol's dictum that everyone will be famous for 15 minutes. We may likely be heading for a future in which everything is true for only 15 minutes. Unfortunately, it will always take more than 15 minutes to analyze a medical report correctly. We will be faced with an overwhelming input of information. We must learn to reject quickly the flighty, flimsy, and faulty and to concentrate on science with substance.

References

1. Zeppa R: Commencement address, University of Miami School of Medicine, June 1987 (unpublished)
2. Feinstein AR: Biologic and Statistical Implications of Clinical Taxonomy. In Feinstein AR (ed): Clinical Judgment, p 209. Baltimore, Williams & Wilkins, 1967
3. Peter LJ, Dana B: The Laughter Prescription, p 131. New York, Ballantine Books, 1982
4. Kirby RR, Downs JB, Civetta JM, et al: High level positive end-expiratory pressure (PEEP) in acute respiratory insufficiency. Chest 1975; 67:156
5. Gallagher TJ, Civetta JM, Kirby RR: Terminology update: Optimal PEEP. Crit Care Med 1978; 6:323
6. Nelson LD, Civetta JM, Hudson-Civetta JA: Titrating positive end-expiratory pressure therapy in patients with early, moderate arterial hypoxemia. Crit Care Med 1987; 15:14
7. Colton T: Statistics in Medicine. Boston, Little, Brown, 1974
8. Hennekeus CH, Buring, J: Epidemiology in Medicine. Boston, Little, Brown, 1987.
9. Hudson-Civetta JA, Civetta JM, Martinez OV, et al: Risk and detection of pulmonary catheter-related infection in septic surgical patients. Crit Care Med 1987; 15:29
10. Applefeld JJ, Caruthers TE, Reno DJ, et al: Assessment of the sterility of the long term cardiac catheterization using the thermodilution Swan-Ganz catheter. Chest 1978; 74:377
11. Sackett DL: Biases in analytic research. J Chronic Dis 1979; 32:51
12. Real Paper, Boston, Massachusetts, 1978
13. Civetta JM: Critical illness--the non-steady state. Surg Forum 1972; 23:153
14. Angell M: Negative Studies. N Engl J Med 1989; 321:464
15. Armitage P: Sequential Medical Trials. Springfield, IL, Charles C Thomas, 1960
16. Bliss CI: Statistics in Biology, vol 1. New York, McGraw-Hill, 1967
17. Dixon WJ, Massey FJ Jr: Introduction to Statistical Analysis, 3rd ed. New York, McGraw-Hill, 1969
18. Hanley JA, Lippman-Hand A: If nothing goes wrong, is everything all right? JAMA 1983; 249:1743.
19. Taffet GE, Teasdale TA, Luchi RJ. In-hospital cardiopulmonary resuscitation. JAMA 1988; 260:2069

Author Information

Joseph M. Civetta, M.D.
Professor and Chairman, Department of Surgery, University of Connecticut Health Center