AJNR

The Value of a *P* Value

H.J. Cloft

AJNR Am J Neuroradiol 2006, 27 (7) 1389-1390
http://www.ajnr.org/content/27/7/1389

COMMENTARY

The Value of a *P* Value

"Nothing is impossible. Some things are just more unlikely than others."

—Jonathan Winters

When we physicians read the medical literature, we generally want to see a *P* value. We find it reassuring. There seems to be a pervasive perception that a *P* value can lend instant credibility to a research report. Readers can be seduced into thinking that a report has merit simply because the results yield a *P* value of less than 0.05. Investigators can become convinced that including a *P* value in their report will give them credibility and thus increase the likelihood that their work will be published. But how many of us look critically at each *P* value that we come across, and ask ourselves what each *P* value really means to practicing physicians and their patients? We should be careful to not always take a *P* value at its face value. When faced with a research result and an accompanying *P* value, asking yourself a few simple questions can help to clarify the true value of the result to you and your patients. These simple questions are: 1) *Why should I care?*; 2) *Is the result consistent with my experience?*; and 3) *Were the right tests and the right numbers used?* The first 2 questions do not require any formal statistical education, whereas the last question requires some basic statistical knowledge. The rationale for asking each of these questions is detailed below.

Why should I care? This question addresses the point that statistical significance is not equivalent to clinical significance. The *P* value must be interpreted in light of what practical meaning the result may have to a physician and/or a patient. By definition, the *P* value should test a hypothesis. Is the hypothesis something that I care about? A research publication may report abundant data that have statistical significance (based on an extremely low *P* value), and yet these may have absolutely no clinical significance. For example, a study of the use of a medical device in an animal model that reports a statistically significant ($P < .05$) treatment effect will have no clinical value if the animal model does not accurately model the human condition. Alternatively, a case report or small case series cannot have statistical significance, but can have substantial clinical significance. For example, a case report can alert a physician to a rare manifestation of disease that allows for a prompt diagnosis and thus averts serious morbidity or mortality in a number of patients in the future. The mere presence of statistical lingo might mislead a reader into believing that a useless report is useful. This point is illustrated by the publication of a report that actually found its way into the literature with the following clinically meaningless conclusion: "In this pilot study, the null hypothesis that both treatments will show equal results cannot be confirmed or rejected because of the small number of participants."[1]

Is the result consistent with my experience? This question addresses whether the result seems plausible when compared with everything else that you know. The *P* is an abbreviation of probability. It is the probability that an observed difference, or a difference more extreme, occurred purely by chance when, in fact, there is no true difference. An arbitrary value for *P* (usually 0.05 in medical studies) is then chosen to separate the probable from the improbable, or the significant from the insignificant. A *P* value of 0.05 suggests that, if the null hypothesis were true, there is a 5% chance of obtaining the observed results. This does not necessarily mean that there is only a 5% chance of being wrong in rejecting the null hypothesis. The chance of being wrong might be as high as 100%. The *P* value only evaluates the data from the study; it does not look at data from any other source. Consider a hypothetical study that concludes that findings on MR angiography of the circle of Willis predict what a patient ate for breakfast ($P < .05$). Your experience with MR physics and digestive physiology tells you that this conclusion is absurd and you conclude that, in your universe, there is still at least a 99.99% chance that the null hypothesis is correct; ie, findings on MR angiography of the circle of Willis are not related to what a patient ate for breakfast. This is an extreme example, but it illustrates that we must not interpret *P* values "in a vacuum," but rather we should interpret the *P* value in the light of our experience.

Were the right test and the right numbers used? Without formal training in statistical methods, this can be a difficult question to answer. Yet, it is an important consideration, because statistical methodology errors are quite common in the medical literature.[2–4] One reason for such a high prevalence of statistical methodology errors is that only a minority of journals have the statistical methods reported in manuscripts reviewed by statistical editors.[5] It is, therefore, important for readers of the medical literature to be wary of common statistical errors.

One such common statistical error is the use of the wrong statistical test. The choice of statistical test depends on the type of data accumulated. All data can be classified according to what are called scales of measurement. There are 4 different types of scales of measurement: nominal, ordinal, interval, and ratio. Nominal scale data consist of purely qualitative categories. Examples include sex, ethnic background, and blood type. Ordinal or rank scale data are numerical data, but the numbers represent position in an ordered series, and do not indicate how much difference exists between successive positions in the scale. Ordinal scale data can be hierarchically ordered, but do not have specific numerical values.[6] Examples of ordinal data include cancer stage, and Thrombolysis in Myocardial Infarction (TIMI) grade of arterial occlusion. An interval scale is a metric scale (ie, it has a fixed unit of measurement) with an arbitrary zero point; examples of interval scale data include Celsius temperature and Hounsfield Units. A ratio scale is a metric scale with an absolute zero that truly represents absence of the characteristic. Examples of a ratio scale include Kelvin temperature and cerebral blood flow.

Interval and ratio data are referred to as parametric data, and nominal and ordinal data are referred to as nonparametric data. Parametric means that the data follow a normal distribution or bell curve. Also, parametric means that the numbers can be added, subtracted, multiplied, and divided. Simple arithmetic should not be performed with ordinal variables because the data are not related in a linear manner. Because ordinal variables cannot be assigned to a linear numeric scale that makes sense, the computation of means and standard deviations for such data are not valid.

The scale of measurement limits the choice of statistical tests. For nonparametric data, statistical tests that rely on the computation of means and standard deviations, such as a *t* test, should not be applied. Unfortunately, incorrect application of parametric statistical tests to nonparametric variables can be found in abundance in the medical literature.[6] Tests designed for evaluation of nonparametric data, such as the Wilcoxon rank sum test and the Mantel-Haenszel test, should be used to analyze nonparametric data. Simply reading the names of statistical tests often elicits confusion, panic, or somnolence in physicians. Those physicans who do not have such a negative response and would like to understand these tests in detail are referred to statistical courses and textbooks. My goal is to simply clarify some basic statistical issues for those who would like to be more savvy consumers of the medical literature without formal study of statistics.

The practice of data mining can also lead to the use of inappropriate statistical methods. Data mining is the practice of retrospectively analyzing a collection of nonspecific data with numerous statistical tests. A variety of statistical tests are applied until a *P* value is obtained that suggests that some of the data have a statistically significant relationship. Data mining is not primarily driven by a hypothesis, but rather is driven by a search for a *P* value less than 0.05. The hypothesis of the work is then secondarily defined by the "significant" *P* value; ie, the work does not start with a hypothesis but rather ends with a hypothesis. The *P* value suggesting significance is reported and all others are discarded. Data mining is, at best, an unscientific practice, and it is an unethical practice in cases in which it is purposefully used to mislead the reader.

Publication bias drives the practice of data mining. Publication bias refers to the bias toward publication of "positive" or statistically significant results relative to "negative" or statistically insignificant results.[7] Authors of scientific reports generally know that their work is more likely to get published and be noticed if the data support a "positive" result, so they will be inclined to "mine" the data until a "positive" result is found. Because of publication bias, there are a lot of "negative" results (ie, $P > .05$) that will never see the light of day. This is unfortunate, because many "negative" results are clinically important to physicians. In addition to publication bias, the personal biases of the authors can have a substantial effect on the choice of statistical methods. The reader should give consideration to the author's potential to gain money, prestige, or something else of value, based on the results of their study. Unfortunately, such personal bias can both consciously and subconsciously affect the interpretation of results.

Few radiologists have had in-depth training in statistical methods. Rather, most radiologists, including myself, have learned statistics only on a "p.r.n." basis, studying the topic only when occasionally required as part of our training or employment. Such rudimentary training is often distilled down to a perception that one merely needs to look at the *P* value to decide whether a result is significant. Unfortunately, this is an oversimplification that can lead to numerous incorrect assumptions. It is possible, however, to improve upon our ability to understand the true clinical significance of a *P* value by simply clarifying a few basic concepts and by trying to critically evaluate the true value of each *P* value that we encounter.

References

1. Lamers HJ, Jamin RH, Zaat JO, et al. **Dietary advice for acute diarrhoea in general practice: a pilot study.** *Br J Gen Pract* 1998;48:1819–23
2. Gore SM, Jones IG, Rytter EC. **Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976.** *BMJ* 1977;1:85–87
3. Thorn MD, Pulliam CC, Symons MJ, et al. **Statistical and research quality of the medical and pharmacy literature.** *Am J Hosp Pharm* 1985;42:1077–82
4. Welch GE, Gabbe SG. **Statistics usage in the American Journal of Obstetrics and Gynecology: has anything changed?** *Am J Obstet Gynecol* 2002;186:584–86
5. Goodman SN, Altman DG, George SL. **Statistical reviewing policies of medical journals: caveat lector?** *J Gen Intern Med* 1998;13:753–56
6. Tello R, Ptak T. **Statistical methods for comparative qualitative analysis.** *Radiology* 1999;211:605–07
7. Easterbrook PJ, Berlin JA, Gopalan R, et al. **Publication bias in clinical research.** *Lancet* 1991;337:867–72

H.J. Cloft
Department of Radiology
Mayo Clinic
Rochester, Minn