

STATISTICS FOR CLINICIANS

P Values

Barry K. Moser, PhD
Associate Director
CALGB Statistical Center
Duke University Medical Center
Durham, NC

H&O How do you define *P* values?

BM *P* values are used in tests of statistical hypotheses. The *P* value is the probability of observing strictly by chance a test statistic that is at least as extreme as the value that was observed in the data, under the assumption that the null hypothesis is true. Regarding this definition, I should highlight two important aspects to keep in mind. First, it is important always to remember that the *P* value is calculated from the actual observed data, that is, it is a function of the observed data. Second, the *P* value is calculated assuming the null hypothesis is true.

H&O What is the null hypothesis?

BM The null hypothesis usually states that the factor being tested will have no influence on the outcome. For example, in a study trying to determine whether there is a difference in the outcomes from two treatment arms, the null hypothesis will generally state that no difference exists between the average outcomes of a population of patients assigned to the two arms. If a large value of the test statistic is observed from the data, the probability of observing strictly by chance a test statistic at least as extreme as the observed test statistic will be small under the null hypothesis. Therefore, the smaller the *P* value, the more difficult it is to believe that the null hypothesis is true, because if the null hypothesis were true it would be highly unlikely to observe a test statistic so extreme.

A *P* value is not, however, the probability that the null hypothesis is true. That is, a *P* value of .7 does not indicate a 70% chance that the null hypothesis is true. The *P* value is a probability calculated under the assump-

tion that the null hypothesis is true. It is the probability that a future observed statistic will be more extreme than the one observed, assuming the null hypothesis is true.

Similarly, it is incorrect to comment on the lack of statistical significance of a variable being tested. Rather, it is more appropriate to speak in terms of probabilities. For example, imagine a test of the effect of platelet count on relapse, with platelet count as the variable being tested and relapse as the outcome. If a large *P* value results from the test of whether platelet count has an influence on relapse, it would be incorrect to say that platelet count lacks statistical significance. It is more accurate to say that if platelet count has no effect on relapse, then there is a high probability of obtaining a test statistic more extreme than the observed one, and it is therefore easier to believe the null hypothesis is true.

H&O What is the difference between a *P* value and an observed significance level?

BM The terms *P* value and observed significance level (OSL) are synonymous, and I believe the OSL is the more descriptive term. Although OSL is not used very often in the medical field, it is used in other areas of science. Observed significance level is the more descriptive term because the *P* value does in fact give a measure of the significance level of the factor being tested; in addition, with OSL, we are stressing the fact that the significance is based on the observations.

H&O When is the *P* value a good measure from the data?

BM We established that the *P* value is based on the observed data, and therefore it has the potential to be a

valuable measure of significance of the variables or factors being tested. Ultimately the usefulness of the P value is dependent on the quality of the study design. Experiments are conceptualized by knowing what hypotheses are to be tested and then designed with those hypotheses in mind. Proper study design policies dictate that patients must be randomized correctly in order to minimize the possibility of bias. It is necessary to ensure that the appropriate numbers of patients are observed, or, in the case of survival analysis, the appropriate numbers of events occur, in order to perform a proper test. It is, in some cases, necessary to stratify samples in a specific way in order to prevent other variables from influencing the hypothesis being tested; with proper stratification it is possible to maintain balance in the data so that no spurious correlations affect the outcomes. These qualities plus the details of the statistical methods to be applied must be embedded in the study design from the outset. If that occurs, the P value is a very good measure of whether a given factor is significant or not. In the absence of these qualities, spurious biases can occur or influences (other than the factors of interest) can confound the conclusions.

H&O How can one discern if a P value is being misused?

BM One common misuse of P values is their calculation based on incomplete data. For example, if an experiment is designed to require 500 observations but a P value is calculated based on a selected subset of only 50 observations, it is more likely that spurious observations will lead to inflated significance levels indicated by small P values. However, if all the data were analyzed as designed, occasional spurious observations would have less effect on inflating the significance levels dictated by the P value.

The P value is a function of the observations, which in turn is related to how the data are collected and how the study is designed. Often a researcher will indicate that he or she has tested a factor and found a small (ie, significant) P value. Rather than accepting the result with enthusiasm, one should take a step back and ask some critical questions, such as: What was the original objective of the study? Does the P value address the original objective or something else? How were the data collected? How many observations were taken? How was the study randomized? How was the study stratified? Will any of these design criteria affect the P value generated? Did the researchers use all of the data or only a portion of the data when the P value was calculated?

Recently, I worked on a clinical trial to test the difference between two treatments on overall survival in 500 patients. As a side experiment, the researchers also

gathered tissue specimens from 130 of the patients. The researchers intended to produce a P value to check the relationship between gene expression/nonexpression and relapse. Therefore, the goal of the experiment shifted from checking overall survival for two treatment effects to checking whether a gene expression/nonexpression was related to relapse in a relatively small portion of the test sample. To make matters worse, the original experiment was designed to expect approximately 300 events to occur in 500 patients during the course of the study. Because the 130 patients had been observed for just a short period, only 25 events occurred. Analysis of this subset of data led to a P value of .03, from which the researchers hastily concluded that gene expression was related to relapse. However, these data were very preliminary, they included a selected 130 out of a potential 500 patients, and the P value was not associated with the hypothesis from the original design. As such, a prudent approach was for the researchers to wait for more data. In the next updated dataset, six more events were observed, for a total of 31 events, which changed the P value from the initial .03 to .20. This change was drastic because the small preliminary dataset was quite sensitive to the addition of a small number of events.

In this example, the study was not designed to answer the question regarding the relationship between relapse and gene expression, randomization was not performed to minimize any biases related to the question being addressed, and spurious correlations contributed to the sensitivity of the data. All of these influences contributed to the P value's lack of usefulness as a meaningful measure of significance. In isolation, a P value is not particularly useful; its worth is derived from its context. Without that context the reader must be cautious, as the P value can be misleading.

As a final comment on the misuse or misinterpretation of a P value, it should be noted that a large sample can produce a sufficiently small P value such that the associated factor is considered statistically significant while the estimated effect size may not be clinically meaningful. Therefore, interpretation of a P value should be performed simultaneously with the interpretation of the clinical meaningfulness of the estimated effect size.

H&O Do you believe it is therefore important always to detail the background when reporting a statistical value such as P ?

BM Absolutely. Rather than simply stating what the P value is, when reporting data, it is essential to contextualize it by describing the study design and the primary objectives of the study. In this context, the reader can

evaluate the P value measure. I believe researchers understand the importance of designing scientifically rigorous experiments, but I am less convinced that it is always understood that the P value is a function of the observed data and therefore strongly related to the study design, how the data are collected, the randomization and stratification of the patients, and whether bias and correlation have influenced the results.

H&O In data mining, many variables are often examined; can this practice lead to misuse of the P value?

BM Yes. It is relatively common for researchers to use data mining to search for any variables related to the outcome that interests them. Often, researchers will search through 20 or more variables until they find one that shows a small P value and then deem that variable significantly related to the outcome. The problem that arises in such cases is called multiplicity, meaning that the more variables checked, the more likely it is that, simply by chance, a variable with a small P value will arise regardless of whether or not it is significantly related to the outcome. It is important to be aware of how many variables have been tested before one variable is deemed significant. Writing a good protocol and/or statistical analysis plan that explicitly states what will be tested before the data are collected is thus essential. Examining many variables until one produces a small P value after the data have been collected is not the proper way to conduct research because the odds are that eventually a small P value will be found, but it may not have any clinical meaning on its own.

The previous remarks do not deny data mining's potential as a valuable analysis device. Data mining can supply valuable information provided analysis adjustments are made to account for multiplicity and that subsequent designed studies are performed to confirm the significance of the variables identified in the data mining process.

H&O What safeguards exist to prevent the misuse of P values?

BM Good practice dictates that a study protocol and/or a statistical analysis plan clearly specify the study objectives, the primary hypotheses being tested, the outcomes used to test the hypotheses, the statistical test procedures to be employed, the randomization and stratification schemes, the number of observations, and the expected power of the test. Additionally, the exact mechanism by which all the data are collected is described in detail and how the treatments are administered to the patients is specifically dictated. If all these criteria are met, then the P value is a valuable measure of the significance of the factors tested. Generally, problems arise when researchers begin to examine data after the fact, from a perspective that does not adhere to the original statistical analysis plan. In this way, the researchers are not constrained by the original study guidelines, which can lead to the misuse of P values.

Suggested Reading

Altman DG. *Practical Statistics for Medical Research*. London, England: Chapman and Hall; 1991.