Some Non-Random Views of Statistical Significance

Alfred O. Berg, MD
Seattle, Washington

Although research in family medicine is growing rapidly, few family physicians have had experience or training in statistical methods. Statistical significance and P values are often misunderstood and frequently misapplied. "Significance" is arbitrary; the actual P value is of greater interest than a significant/not significant statement; P values do not measure the strength of an association; statistical significance is not equivalent to "actual" significance; P values are largely dependent on sample size; and data "dredging" is guaranteed to yield spurious results. Competent statistical consultation, careful study planning, and recognition of statistical pitfalls are important to anyone who does research, and knowledge of these areas is useful as well to anyone reading the medical literature.

Research in family medicine is a growing field, yet few family physicians have had formal training in research methods. Discussions of P values dimly recall medical school lectures to the effect that "the smaller the better," but few have had occasion to calculate them or become familiar with their more subtle applications. This paper will point out some of the frequent errors and misapplications of tests of statistical significance, emphasizing use and interpretation of P values. After a definition, examples will be given drawn from the medical literature, illustrating some of the common problems, and concluding with recommendations and suggestions for those involved in research and for those reading the results.

Definition

No one knows who published the first P value, but it cannot have been before the 1930s, since it was only during that decade that Fisher developed the whole idea of statistical inference. Quite simply, any event has a probability of occurrence somewhere between zero and one. A P value is the probability that, if the experiment or study were exactly repeated, the size of the differences or changes demonstrated in the study could have been due to chance (random error) alone.

Examples

"...the difference was not significant at P=0.10."

"...the difference was significant at P=0.10."

From the Robert Wood Johnson Clinical Scholars Program, University of Washington, Seattle, Washington. Requests for reprints should be addressed to Dr. Alfred O. Berg, Robert Wood Johnson Clinical Scholars Program, University of Washington HQ-18, Seattle, WA 98195.

0094-3509/79/051011-04$01.00

These two examples illustrate that statistical significance is entirely arbitrary. In both cases, the probability that the observed difference could have been due to chance is one in ten, yet interpretations vary. There is no single level at which significance is guaranteed. Even a P value of less than 0.0001 does not assure that the event could not have been due to chance, only that it would be very unlikely.

A researcher was recently introduced to a game which had a 1 in 10,395 chance of being solved on the first try. It happened that he *did* solve it on the first try. The point is that rare things do happen.

The larger the number of comparisons in a research study, the more likely it becomes that some proportion of the findings will lead to spurious conclusions. If one sets a significance level of P=0.05, it is sobering to realize that five percent of all research findings published "significant" at that level are due to chance alone.

"...but the difference was not significant."

This is even worse—no P value at all. If the author set an arbitrary significance level at P=0.001, and rejected the finding because P was only 0.002, one might argue with the conclusions. It is always better to calculate the P values exactly and to report them. This allows the readers some freedom to make up their own minds about "significance."

"...Group I was older than Group II (P=0.005)."

In this particular case the readers were never told *how much* older Group I was than Group II. This illustrates that P values do not measure the strength of an association. One should always give a measure of the effect being tested—a difference, a percent change, or whatever.

A reviewer was recently given the preliminary findings of a large, well-designed study comparing patients in a family practice setting with those in an internist's practice, but the "findings" consisted only of pages and pages of P values, with no measures of how large the differences were. A P value without some measure of effect is of no value whatever.

"...the mean hematocrit in the first group was 38.6, in the second 37.5, and the difference was highly significant (P=0.003)."

This example illustrates the difference between statistical significance and actual significance. Statistical significance is largely a function of sample size, or how many subjects were studied. The larger the sample, the more likely it is that a difference observed will be statistically significant. In the above example, one could argue whether the authors proved anything, since a difference of less than three percent in hematocrit is unlikely to have clinical consequences, and certainly does not help one predict a given outcome.

A recent mail survey of physicians was designed to elicit attitudes on preventive medicine. On a five-point scale, five being most positive, internists average 4.1 and family physicians, 4.3. The difference was statistically significant (P=0.02) because there were large numbers of physicians in each group, but the finding is of very little value in predicting the attitude of a given internist or family physician, or in making a strong statement about true differences in attitudes, because the observed difference was so small.

"If you required treatment with an experimental drug, and drug A had just been shown effective over placebo on a sample of 500 patients (P=0.05), and drug B had just been shown effective over placebo on a sample of 20 patients (P=0.05), which drug would you choose?"

This is an example of the interaction between sample size and the "power" of a statistical test on the sample. The correct answer is to choose drug B, since it is very difficult to achieve statistical significance with such a small sample (hence the difference in *effect* must have been quite large), whereas it is fairly simple to achieve significance on a sample of 500, even if the difference in effect was very small. Calculating the power of a test is outside the scope of this paper, but, in general, the larger the study, the more likely one will be able to detect small differences between the study groups.

"The average numbers of clinic visits made by the experimental and control families were 6.4 and 7.8 per year, respectively, and the difference was not statistically significant (P=0.32) ...One can conclude that the educational program had no effect on the number of visits."

In this example, careful reading of the paper itself showed that only ten families were in each of the two groups, experimental and control. Using other data provided in the article, the chances of this study discovering a true difference of as much as 80 percent between the two groups was only fifty-fifty. This illustrates another aspect of the "power" problem mentioned in the previous example.

Frequently, a small study will fail to demonstrate statistically significant differences between the study groups, and the conclusion is made that no differences exist. When one calculates the power of the test, however, one finds that the results may be consistent with large differences in the overall population, but that the differences were missed because too few subjects were studied. In other words, watch out for "no significant difference" findings in small studies.

"...patients were compared on 78 characteristics, and only income and marital status differed between the two groups."

This sort of analysis is called "dredging" by some, a " fishing expedition" by others. The point is that the more tests one does, the more likely it is that significant differences will be found. In the above example, if the significance level was 0.05 (note that it is not given!), on the basis of chance alone one would have expected the authors to come up with three or four significant differences, not just two. There are some statistical ways around this problem, but very few researchers use them.

"The average decrease in serum cholesterol after changing diets was 22 mg percent, but the difference was not statistically significant (P=0.09)."

This example is more subtle than the preceding ones. Here the only problem is that the statistical test used is not identified. (Alert readers will note that the test names have not been stated in *any* of the previous examples). In this case, working backwards from the author's data, it is apparent that the unpaired t test was employed. Had the more appropriate paired t test been used, the observed difference would have been statistically significant at P=0.02. This illustrates the desirability of naming the statistical test used, in addition to giving the measure of effect and the exact P value. It is uncommon that only one statistical test might be appropriate in a given situation, and choosing between the several tests available is often a difficult decision. The reader should have the opportunity to judge the appropriateness of that decision, without having to laboriously retrace the author's calculations in order to guess at the statistical test employed.

Conclusions and Recommendations

No single discipline has a monopoly on confusion with P values. The examples chosen for this analysis were from journals in several medical specialties. A few summary comments are appropriate for those contemplating a study using statistical methods and for those who must read the results.

1. Obtain statistical consultation early, before the study begins. Few family physicians have the background or the interest for much statistical work, and many are uncomfortable with anything more complicated than a chi-square test. Find a consultant who is interested in your study and who is personally approachable. This is easier said than done. Even individuals with significant statistical training occasionally have difficulty obtaining the kind of statistical help they need.

2. Plan the analysis before the study begins. Know what kind of differences you would like to find, and how much money is available to spend on the study. It is possible to calculate the size of the study needed to answer your research question, and, in any event, you need to figure the power of the study to determine a specified difference once the size is established. You may find that the proposed study would be too expensive, but sometimes you may actually be able to trim costs by reducing the study size.

Plan the comparisons to be made in advance, and limit yourself to those. Avoid dredging the data—you are certain to "discover" meaningless relationships.

3. Always give some measure of the effect observed—a percent change, an absolute difference, a relative risk, a difference in means, or whatever. Only then should the statistical test be performed. It is occasionally a good idea to state the effect and calculate a confidence interval without *ever* giving a P value. For example, a difference of 15 percent, with a 95 percent confidence interval of 0 percent to 30 percent means that one is 95 percent certain that the "true" difference in

the population is somewhere between zero percent and 30 percent (in this case not a very impressive finding).

4. If you do choose to do significance tests, name the test used, and present the actual P values along with the effect measures. Allow the readers flexibility to make up their own minds about the appropriateness of the test, and whether or not the result is significant.

5. Distinguish between statistical significance and "actual" significance in the discussion. Very small P values do not mean that you have discovered something important.

6. Recognize that even with the best techniques and intentions, chance catches up with you. The whole idea of P values is based on the fact that random errors *do* exist. You may well discover some relationship which does not hold up under further testing. It is impossible to do an absolutely definitive study, as countless examples from medical history attest. Be comfortable with that.

7. If you wish to become heavily involved in research, or if competent statistical consultation is not available, further reading or special courses in the area may be necessary. Several excellent resource books are available.[2-5]

References

1. Fisher RA: The Design of Experiments. Edinburgh, Oliver and Boyd, 1935
2. Armitage P: Statistical Methods in Medical Research. New York, John Wiley,1971
3. Remington RD, Schork MA: Statistics with Applications to the Biological and Health Sciences. Englewood Cliffs, NJ, Prentice-Hall, 1970
4. Snedecor GW, Cochran WG: Statistical Methods, ed 6. Ames, Iowa, Iowa State University Press, 1967
5. Afifi AA, Azen SP: Statistical Analysis: A Computer-Oriented Approach. New York, Academic Press, 1972