Paper 1723-2014

P-values: Democrats or Dictators?

Brenda Beaty, MSPH, and Michelle Torok, PhD, University of Colorado School of Medicine, Aurora, CO

ABSTRACT

Part of being a good analyst and statistician is being able to understand the output of a statistical test in SAS®. *P*-values are ubiquitous in statistical output as well as medical literature and can be the deciding factor in whether a paper gets published. This shows a somewhat dictatorial side of them. But do we really know what they mean? In a democratic process, people vote for another person to represent them, their values, and their opinions. In this sense, the sample of research subjects, their characteristics, and their experience, are combined and "represented" to a certain degree by the *p*-value. This paper discusses misconceptions about and misinterpretations of the *p*-value, as well as how things can go awry in calculating a *p*-value. Alternatives to *p*-values are described, with advantages and disadvantages of each. Finally, some thoughts about *p*-value interpretation are given. To "disarm" the dictator, we need to understand what the democratic *p*-value can tell us about what it represents….and what it doesn't. This presentation is aimed at beginning to intermediate SAS statisticians and analysts working with SAS/STAT®.

INTRODUCTION

A t-shirt I once saw offered this bit of wisdom: "Being a statistician means never having to say you're sure." All of statistics is based on probability theory, and the '*p*' in '*p*-value' really stands for probability. There is always uncertainty in our results and one of the trickiest parts of statistics is quantifying that amount of uncertainty. *P*-values as measures of uncertainty are used extensively in scientific papers, and play a critical role in such things as career advancement, getting papers published, even approval of a new drug or device. Therefore, it has huge personal and economic repercussions.

WHERE DID IT COME FROM?

The modern use of p-values was popularized by Ronald Fisher (he of the "Fisher's exact test" and the "F-distribution" in the 1920's). According to Goodman, 2008:

"In Fisher's system, the *P* value was to be used as a rough numerical guide of the strength of evidence against the null hypothesis. There was no mention of "error rates" or hypothesis "rejection"; it was meant to be an evidential tool, to be used flexibly within the context of a given problem. Fisher proposed the use of the term "significant" to be attached to small *P* values, and the choice of that particular word was quite deliberate. The meaning he intended was quite close to that word's common language interpretation - something worthy of notice. In his enormously influential 1926 text, *Statistical Methods for Research Workers,* the first modern statistical handbook that guided generations of biomedical investigators, he said:

Personally, the writer prefers to set a low standard of significance at the 5 percent point . . . . A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.

In other words, the operational meaning of a *P* value less than .05 was merely that one should *repeat the experiment*. If subsequent studies also yielded significant *P* values, one could conclude that the observed effects were unlikely to be the result of chance alone. So "significance" is merely that: worthy of attention in the form of meriting more experimentation, but not proof in itself."

Statisticians in the early 1900's grappled mainly with industrial and agricultural research questions that required a decision for a limited number of courses of action. The dichotomous p-value may have made more sense in these quality control study contexts compared to its application to modern scientific research (Rothman and Greenland,1998).

WHAT *ISN'T* IT?

Do not put faith in what statistics say until you have carefully considered what they do not say.
-- William W. Watt (1913-1996; American Teacher, Author, and Poet)

The following **mis-**interpretations of *p*-values are adapted from Goodman, 2008.

"If *p*=.05, the null hypothesis has only a 5% chance of being true."

Why it's wrong: A *p*-value is calculated under the assumption that the null hypothesis *is* true. It cannot therefore measure the probability of the null hypothesis. Any number of hypotheses may be true, including the null hypothesis. The *p*-value isn't the probability that any hypothesis is true (Poole, 2001).

"A non-significant difference (e.g., $p > .05$) means there is no difference between groups."

Why it's wrong: This does not make the null hypothesis the most likely one; the range of effects within the confidence interval are also statistically consistent with the data.

"Studies with *p*-values on opposite sides of .05 are conflicting."

Why it's wrong: A study with a small sample size that yields imprecise estimates may not have statistically significant results, but another study that obtains the same effect estimate with a larger sample size could have statistically significant results.

A scientific conclusion or treatment policy should be based on whether or not the *p*-value is significant.

False: Large datasets, in particular, may produce statistically significant *p*-values, but the magnitude of the effect estimate may be small and clinically or substantively unimportant. Likewise, small studies may yield a clinically important effect estimate that is not statistically significant. The *p*-value confounds the precision and magnitude of the effect. It would not make sense to reach different conclusions about two studies on the same topic if one had $p=0.45$ and one had $p=0.51$? Furthermore, a body of knowledge in its entirety, not from a single study, should be used to guide decisions.

Also, avoid confusing Alpha and *p*-values

This is a common mistake that can be made clearer by reviewing the definitions of each term:

Alpha= prob (*p*-value < alpha | H_0)

p-value=prob (X >= x | H_0) where X is a random variable corresponding to some way of summarizing data (like means or proportions) and x is the observed result.

Alpha is the acceptable probability of a type 1 error: declaring significance when the null hypothesis is true. You set alpha in advance, it does not rely on your study data. The *p*-value is computed from the data.

If you conduct a null hypothesis test with alpha=0.05 and obtain $p=0.003$, the probability that you have made a type 1 error is not 0.003. It is 0.05.

WHAT *IS* IT?

The formal definition is as follows: In statistical significance testing, the *p*-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. A researcher will often "reject the null hypothesis" when the *p*-value turns out to be less than a certain significance level, often 0.05 or 0.01 (Goodman 1999).

WHAT DO I NEED TO CONSIDER?

Study size

"Studies with small sample sizes tend to miss important clinical differences in significance tests. In contrast, studies with large study sizes tend to produce significant findings that are clinically meaningless." (Stang, 2010)**.**

Statistical versus Clinical Significance

This is related to study size. Large sample sizes can give statistically significant effects that are clinically meaningless in practice. Conversely, sometimes a finding of no difference may be very important, especially if it disproves a common assumption. Statistical non-significance does not equal clinical unimportance. Remember, there are many studies demonstrating that a new drug does not provide benefits that are any different than a placebo. That's important information for a physician to know.

Multiple Comparisons Issues

P-values and multiple comparisons: If n independent associations are examined for statistical significance and all of the individual null hypotheses are true, the probability that at least one of them will be incorrectly found to be statistically significant is 1-(1-alpha) to the nth power. For example, *p*-values obtained from a genomics study in

which hundreds of comparisons are made should be viewed differently than *p*-values from an observational study in which a limited number of *a priori* comparisons made.

Underlying Assumptions and Considerations of Study Results

1) Using a *p*-value always means assuming that the null hypothesis is true. It is virtually impossible to **disprove** a specific hypothesis, therefore we posit no relationship and attempt to understand how the data match up with the null hypothesis.

1) . A *p*-value assumes randomly selected samples that represent the population completely and accurately. In reality, this is difficult to achieve. Different samples will lead to different study results.

2) As discussed above, the *p*-value is influenced by sample size. A small study may produce meaningful results with $p>0.05$ while a large study may yield clinically irrelevant estimates with $p<0.05$.

3) It is assumed that the correct causal model and appropriate variables in their appropriate forms are used. For example, including a mediating variable in a model or mis-specifying a continuous variable as linear when it should be quadratic can impact your results.

   It is assumed that the study is valid. Study validity is a function of many potential errors that are not due to chance, such as systematic bias and confounding.

4) It is expected that the assumptions for statistical models are met.

5) It is assumed that the correct statistical model is utilized. For example, choosing the wrong link function to describe the relationship of interest can influence results.

Given the assumptions and considerations above, it should be clear why the *p*-value should be seen as an estimate. With observational data, smaller sample sizes, non-normal data, etc. the *p*-value should be viewed as an even rougher estimation.

SUGGESTIONS

"Belief that “statistical significance” can alone discriminate between truth and falsehood borders on magical thinking." (Cohen, 2010)

1) Approach the analysis of epidemiologic data as an exercise in measurement, not decision making (like the earliest users of *p*-values!) (Rothman and Greenland, 1998).

   Measure and report precision and effect size separately (the *p*-value is a summary measure that mixes them):Present the magnitude of effect through the use of measures such as rates, risk differences, odds ratios. Report precision with standard errors or confidence intervals.

2) Emphasize confidence intervals, but be careful, as they can also be misinterpreted!

A 95% confidence interval does NOT mean that we are 95% certain that the true value lies within the confidence interval Instead, they mean that If the data were sampled and analyzed many times, AND if the variation between samples was due only to chance, the resulting set of confidence intervals would include the correct value for the point estimate at least 95% of the time.

A common interpretation of the p value is that the smaller it is, the less the result is influenced by chance. However, the estimates least influenced by chance and therefore more stable are those with narrow confidence intervals. As Charles Poole demonstrates in his paper “Low P-Values or Narrow Confidence Intervals: Which Are More Durable?” the results from one hypothetical study of a disease and exposure that yields a RR of 1.4, 95% CI 0.80-2.4, p=0.2 should be emphasized over another with RR=4.1, 95% CI: 1.2-14, p=0.02. Despite reaching statistical significance at the 0.05 level, the wide confidence intervals indicate that latter result is much more influenced by chance and is statistically unstable compared to the latter.

3) Perform Bayesian methods using prior probability distributions
   In scientific research, the data have been observed but the parameters are unknown. The Bayesian approach gives the probability of the parameter based on the data, whereas the frequentists calculate the probability of the data given the parameters. The Bayesian approach makes more sense in many situations, but most use the frequentist (traditional) approach (Lee, 2011).

CONCLUSION

P-values are not dictators, they are one of many tools we have as data analysts and researchers, As the editors of Epidemiology pointed out (2001) “"The question is not whether the p value is intrinsically bad, but whether it too easily substitutes for the thoughtful integration of evidence and reasoning." By considering what p-values are and are not, by remembering the many assumptions and limitations studies are subjected to, and by utilizing p-values in conjunction with other tools, you can disarm the dictator.

REFERENCES

Cohen HW. P values: use and misuse in medical literature. Am J Hypertens. 2011 Jan;24(1):18-23.

The Editors. The Value of P.[Editorial] Epidemiology. 12(3):286, May 2001.

Goodman S. A Dirty Dozen: Twelve P-value Misconceptions. Semin Hematol. 2008 Jul;45(3):135-40.

Goodman SN. P values, hypotheses tests and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 1993; 137:485-496.

Greenland S, Poole C. Living with P Values: Resurrecting a Bayesian Perspective on Frequentist Statistics. Epidemiology 2013;24:62-68.

Lee JJ. Demystify Statistical Significance-Time to move on From the P Value to Bayesian Analysis. JNCI 2011; 103(1).

Poole C. Low P-values or narrow confidence intervals: which are more durable? Epidemiology 2001 May;12(3):291-4.

Rothman KJ, Greenland S. Modern Epidemiology, 2nd Edition. Lippincott Williams &Wilkins, Philadelphia. 1998.

Rothman KJ. Six Persistent Research Misconceptions. J Gen Intern Med, January 2014.

Stang A,Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol. 2010 Apr;25(4):225-30.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Brenda Beaty, MSPH
Colorado Health Outcomes Program
University of Colorado Denver
Mail Stop F443
13199 E. Montview Ave., Suite 300
Aurora, CO 80045-0508
Work Phone: (303) 724-1076
E-mail: Brenda.Beaty@ucdenver.edu