Ecology, 95(3), 2014, pp. 627–630
© 2014 by the Ecological Society of America

P values are only an index to evidence: 20th- vs. 21st-century statistical science

K. P. Burnham[1] and D. R. Anderson

Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, Colorado 80523 USA

Overview Comments

We were surprised to see a paper defending *P* values and significance testing at this time in history. We respectfully disagree with most of what Murtaugh (2014) states. The subject of *P* values and null hypothesis significance tests is an old one and criticisms by statisticians began in the late 1930s and have been relentless (see *Commentaries on Significance Testing* for a partial impression of the large literature on the subject [*available online*]).[2] Oakes (1986) summed it up over 25 years ago, "It is extraordinarily difficult to find a statistician who argues explicitly in favor of the retention of significance tests ..."

For the most part, we do not comment point by point, instead we briefly contrast several historical and contemporary aspects of statistical science. The emphasis is on the information-theoretic (IT) approaches that permit computing several post-data quantities that are evidential, avoid conditioning on the null hypothesis, avoid *P* values, provide model likelihoods and evidence ratios, and allow formal inferences to be made based on all the models in an a priori set (multimodel inference).

Historical Statistics

Murtaugh (2014) reviews several of the historical methods for data analysis in simple situations; these methods focus on "testing" a null hypothesis by computing a "test statistic," assuming its asymptotic distribution, setting an arbitrary α level, and computing a *P* value. The *P* value usually leads to an arbitrary simplistic binary decision as to whether the result is "statistically significant" or not. In other cases, the *P* value is stated and interpreted as if it were evidential. The *P* value is defined as the pre-data probability: Prob{*a test statistic as large as, or larger, than that observed, given the null*}. That is, the anticipated data are being thought of as random variables.

Theory underlying these methods for statistical inference is thus based on pre-data probability statements, rather than on the exact achieved data, and reflects early approaches (e.g., Student's influential paper [Student 1908]). In general, these early methods are not useful for non-nested models, observational data, and large data sets involving dozens of models and unknown parameters. Step-up, step-down, and step-wise regression analyses represent perhaps the worst of these historical methods due partially to their reliance on a sequence of *P* values. There is a very large literature on problems and limitations of null hypothesis significance testing and it is not confined to ecology or biology.

At a deeper level, *P* values are not proper evidence as they violate the likelihood principle (Royall 1997). Another way to understand this is the "irrelevance of the sample space principle" where *P* values include probabilities of data never observed (Royall 1997). Royall (1997) gives a readable account of the logic and examples of why *P* values are flawed and not acceptable as properly quantifying evidence. *P* values are conditional on the null hypothesis being true when one would much prefer conditioning on the data. Virtually everyone uses *P* values as if they were evidential: they are not. *P* values are not an appropriate measure of strength of evidence (Royall 1997). Among other flaws, *P* values substantially exaggerate the "evidence" against the null hypothesis (Hubbard and Lindsay 2008); this can often be a serious problem. In controversial settings, such as many conservation biology issues, the null hypothesis testing paradigm, hence *P* values, put the "burden of proof" on the party holding the "null position" (e.g., state and federal agencies and conservation organizations).

In even fairly simple problems, one is faced with the "multiple testing problem" and corrections such as Bonferroni's are problematic when analyzing medium to large data sets. Anderson et al. (2000) provides a more detailed review of these and other technical issues. C. R. Rao, the well-known statistician and former Ph.D. student under R. A. Fisher (see Rao 1992), summarized the situation, "... in current practice of testing a null hypothesis, we are asking the wrong question and getting a confusing answer."

Manuscript received 6 June 2013; accepted 9 July 2013; final version received 17 July 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[1] E-mail: kenb@colostate.edu

[2] http://www.indiana.edu/~stigtsts

Statistical science has seen huge advances in the past 50–80 years, but the historical methods (e.g., *t* tests, ANOVA, step-wise regression, and chi-squared tests) are still being taught in applied statistics courses around the world. Nearly all applied statistics books cover only historical methods. There are perhaps two reasons for this: few rewards for updating course materials and lack of awareness of viable alternatives (e.g., IT and Bayesian). Students leave such classes thinking that "statistics" is no more than null hypotheses and *P* values and the arbitrary ruling of "statistical significance." Such courses are nearly always offered in a least squares setting, instead of the more general likelihood setting which would serve those wanting to understand generalized linear models and the Bayesian approaches. Murtaugh (2014) argues that *P* values and AIC differences are closely related (see his Fig. 2). However, the relationship holds only for the simplest case (i.e., comparison of two nested models differing by only one parameter). Thus, his "result" is *not* at all general. We believe that scientists require powerful modern methods to address the complex, real world issues facing us (e.g., global climate change, community dynamics, disease pandemics).

21st-Century Statistical Science

Methods based on Bayes theorem or Kullback-Leibler information (Kullback and Leibler 1951) theory allow advanced, modern approaches and, in this context, science is best served by moving forward from the historical methods (progress should not have to ride in a hearse). We will focus on the information-theoretic methods in the material to follow. Bayesian methods, and the many data resampling methods, are also useful and other approaches might also become important in the years ahead (e.g., machine learning, network theory). We will focus on the IT approaches as they are so compelling and easy to both compute and understand. We must assume the reader has a basic familiarity with IT methods (see Burnham and Anderson 2001, 2002, 2004, Anderson 2008).

Once data have been collected and are ready for analysis, the relevant interest changes to post-data probabilities, likelihood ratios, odds ratios, and likelihood intervals (Akaike 1973, 1974, 1983, Burnham and Anderson 2002, 2004, Burnham et al. 2009). An important point here is that the conditioning is on the data, not the null hypothesis, and the objective is inference about unknowns (parameters and models). Unlike significance testing, IT approaches are not "tests," are not about testing, and hence are free from arbitrary cutoff values (e.g., $\alpha = 0.05$).

Statisticians working in the early part of the 20th century understood likelihoods and likelihood ratios

$$\mathcal{L}(\theta_0)/\mathcal{L}(\hat{\theta}).$$

This is an evidence ratio about parameters, *given* the model and the data. It is the likelihood ratio that defines evidence (Royall 1997); however, Fisher and others, thinking of the data (to be collected) as random variables, then showed that the transformation

$$-2\log\left\{\mathcal{L}(\theta_0)/\mathcal{L}(\hat{\theta})\right\}$$

was distributed asymptotically as chi squared. Based on that result they could compute tail probabilities (i.e., *P* values) of that sampling distribution, given the null hypothesis. While useful for deriving and studying theoretical properties of "data" (as random variables) and planning studies, this transformation is unnecessary (and unfortunate) for data analysis. Inferential data analysis, given the data, should be based directly on the likelihood and evidence ratios, leaving *P* values as only an index to evidence. Such *P* values are flawed whereas likelihood ratios are evidential without the flaws of *P* values. Early statisticians (e.g., Fisher) had the correct approach to measuring formal evidence but then went too far by mapping the evidence into tail probabilities (*P* values). Likelihood ratios and *P* values are very different (see Burnham and Anderson 2002:337–339). Just because the two approaches can be applied to the same data should not, and does not, imply they are both useful, or somehow complementary. Inferentially they can behave quite differently.

The information-theoretic approaches allow a quantification of K-L information loss (Δ) and this leads to the likelihood of model *i*, given the data, $\mathcal{L}(g_i \,|\, \text{data})$, the probability of model *i*, given the data, $\text{Prob}\{g_i \,|\, \text{data}\}$, and evidence ratios about models. The probabilities of model *i* are critical in model averaging and unconditional estimates of precision that include model selection uncertainty. These fundamental quantities cannot be realized using the older approaches (e.g., *P* values).

Recent advances in statistical science are not always new concepts and methods, but sometimes an enlightened and extended understanding of methods with a long history of use (e.g., Fisher's likelihood theory). There is a close link between K-L information, Boltzmann's entropy ($H' =$ K-L), and the maximized log-likelihood. Akaike (1981, 1992) considered the information-theoretic methods to be extensions to Fisher's likelihood theory (Edwards 1992). In his later work, Akaike (1977, 1985) dealt more with maximizing entropy (H') rather than (the equivalent) minimizing K-L information. Entropy and information are negatives of each other (i.e., $-H' =$ information) and both are additive.

Twenty-first-century science is about making formal inference from all (or many of) the models in an a priori set (multimodel inference). Usually there is uncertainty about which model is actually "best." Information criteria allow an estimate of which model is best, based on an explicit, objective criterion of "best," and a quantitative measure of the uncertainty in this selection (termed "model selection uncertainty"). Estimates of precision, either for prediction or parameter estimation, include a component for model selection uncertainty, conditional on the model set.

Information-theoretic approaches are very different from historical methods that focus on *P* values. There is no need for a formal null hypothesis, no concept of the asymptotic distribution of a test statistic, no α level, no *P* value, and no ruling of "statistical significance." Furthermore, the "burden of proof" is the same across hypotheses/models when using an IT approach. Chamberlain's famous (1890) paper advocated hard thinking leading to multiple hypotheses that were thought to be plausible (most null hypotheses are false on a priori grounds). He wanted post-data probabilities of these alternatives. He must have been disappointed to see the field of statistics lean toward testing null hypotheses with little attention to evidence for or against a single alternative hypothesis, much less multiple alternative hypotheses.

Simple *P* values conditioned on the null hypothesis prevent several important approaches useful in empirical science: ways to rank models and the science hypotheses they represent, ways to deal with non-nested models (most model sets contain non-nested models), ways to incorporate model selection uncertainty into estimates of precision, ways to model average estimates of parameters or predictions, ways to reduce model selection bias in high dimensional problems (Lukacs et al. 2007, 2010), ways to assess the relative importance of predictor variables, ways to deal with large systems and data sets (e.g., 50–100 models, each with 10–300 parameters, where sample size might be in the thousands), ways to analyze data from observational studies (where the distribution of the test statistic is unknown).

The limitations of *P* values, as above, are very serious in our current world of complexity.

Comments on the "Scientific Method" and Statistical Science

While the exact definition of the so-called "scientific method" might be controversial, nearly everyone agrees that the concept of "falsifiability" is a central tenant of empirical science (Popper 1959). It is critical to understand that historical statistical approaches (i.e., *P* values) leave no way to "test" the alternative hypothesis. The alternative hypothesis is never tested, hence cannot be rejected or falsified! The breakdown continues when there are several alternative hypotheses (as in most real-world problems). The older methods lack ways to reject or falsify any of these alternative hypotheses. This is surely not what Popper (1959) or Platt (1964) wanted. "Support" for or against the alternative hypothesis is only by default when using *P* values. Surely this fact alone makes the use of significance tests and *P* values bogus. Lacking a valid methodology to reject/falsify the alternative science hypotheses seems almost a scandal.

It seems that Chamberlin's (1890) notion concerning alternative science hypotheses that are considered plausible should also be considered an integral part of the scientific method. Perhaps it is best if the "scientific method" embraced the concepts of formal evidence and likelihood in judging the relative value of alternative hypotheses because they provide a formal "strength of evidence."

Another serious limitation relates to the common case where the *P* value is "not quite" statistically significant (e.g., $P = 0.07$ when $\alpha = 0.05$). The investigator then concludes "no difference" and the null hypothesis prevails. Even worse, they often also conclude there is no evidence against the null hypothesis. Evidence ratios provide actual evidence in terms of odds, for example, at $P = 0.07$ (under normal theory and 1 df) the evidence is 5.2 to 1 against the null hypothesis. At $P = 0.05$, the evidence is 6.8 to 1 against the null. At $P = 0.096$ the evidence is 4 to 1 against the null, or equivalently, 4 to 1 in favor of the alternative. Depending on the context, even 3 to 1 odds might be useful or impressive; this is very different from concluding "no evidence against the null hypothesis." If the odds are, say, 224 to 1 (this is for $P = 0.001$), then the result must be considered as very convincing and evidence presented this way is much more understandable than saying $P = 0.001$. No automatic "cut-off" (e.g., $P = 0.05$) is relevant in an evidential paradigm such as IT. The interpretation of the evidence, being usually context specific, is left to the investigator: science is about evidence, not about sharp dichotomies or decisions.

Summary

Early statistical methods focused on pre-data probability statements (i.e., data as random variables) such as *P* values; these are not really inferences nor are *P* values evidential. Statistical science clung to these principles throughout much of the 20th century as a wide variety of methods were developed for special cases. Looking back, it is clear that the underlying paradigm (i.e., testing and *P* values) was weak. As Kuhn (1970) suggests, new paradigms have taken the place of earlier ones: this is a goal of good science. New methods have been developed and older methods extended and these allow proper measures of strength of evidence and multimodel inference. It is time to move forward with sound theory and practice for the difficult practical problems that lie ahead.

Given data the useful foundation shifts to post-data probability statements such as model probabilities (Akaike weights) or related quantities such as odds ratios and likelihood intervals. These new methods allow formal inference from multiple models in the a prior set. These quantities are properly evidential. The past century was aimed at finding the "best" model and making inferences from it. The goal in the 21st century is to base inference on all the models weighted by their model probabilities (model averaging). Estimates of precision can include model selection uncertainty leading to variances conditional on the model set. The 21st century will be about the quantification of information, proper measures of evidence, and multimodel inference. Nelder (1999:261) concludes, "The most important task before us in developing statistical science is to demolish the *P*-value culture, which has taken root to a frightening extent in many areas of both pure and applied science and technology."

Acknowledgments

The authors thank the Colorado Cooperative Fish and Wildlife Research Unit for support.

Literature Cited

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov and F. Csaki, editors. Second International Symposium on Information Theory. Akademiai Kiado, Budapest, Hungary.

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control AC 19:716–723.

Akaike, H. 1977. On entropy maximization principle. Pages 27–41 *in* P. R. Krishnaiah, editor. Applications of statistics. North-Holland, Amsterdam, The Netherlands.

Akaike, H. 1981. Likelihood of a model and information criteria. Journal of Econometrics 16:3–14.

Akaike, H. 1983. Information measures and model selection. International Statistical Institute 44:277–291.

Akaike, H. 1985. Prediction and entropy. Pages 1–24 *in* A. C. Atkinson and S. E. Fienberg, editors. A celebration of statistics. Springer, New York, New York, USA.

Akaike, H. 1992. Information theory and an extension of the maximum likelihood principle. Pages 610–624 *in* S. Kotz and N. L. Johnson, editors. Breakthroughs in statistics. Volume 1. Springer-Verlag, London, UK.

Anderson, D. R. 2008. Model based inference in the life sciences: a primer on evidence. Springer, New York, New York, USA.

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64:912–923.

Burnham, K. P., and D. R. Anderson. 2001. Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife Research 28:111–119.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer, New York, New York, USA.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociological Methods and Research 33:261–304.

Burnham, K. P., D. R. Anderson, and K. P. Huyvaert. 2009. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behavioral Ecology and Sociobiology 65:223–235.

Chamberlin, T. C. 1890. The method of multiple working hypotheses. Science 15:92–96. (Reprinted 1965, Science 148:754–759.)

Edwards, A. W. F. 1992. Likelihood: expanded edition. Johns Hopkins University Press, Baltimore, Maryland, USA.

Hubbard, R., and R. M. Lindsay. 2008. Why *P* values are not a useful measure of evidence in statistical significance testing. Theory Psychology 18:69–88.

Kuhn, T. S. 1970. The structure of scientific revolutions. Second edition, University of Chicago Press, Chicago, Illinois, USA.

Kullback, S., and R. A. Leibler. 1951. On information and sufficiency. Annals of Mathematical Statistics 22:79–86.

Lukacs, P. M., K. P. Burnham, and D. R. Anderson. 2010. Model selection bias and Freedman's paradox. Annals of the Institute of Statistical Mathematics 62:117–125.

Lukacs, P. M., W. L. Thompson, W. L. Kendal, W. R. Gould, W. R. Doherty, K. P. Burnham, and D. R. Anderson. 2007. Comments regarding a call for pluralism of information theory and hypothesis testing. Journal of Animal Ecology 44:456–460.

Murtaugh, P. A. 2014. In defense of *P* values. Ecology 95:611–617.

Nelder, J. A. 1999. Statistics for the millennium. Statistician 48:257–269.

Oakes, M. 1986. Statistical inference: a commentary for the social and behavioral sciences. Wiley, New York, New York, USA.

Platt, J. R. 1964. Strong inference. Science 146:347–353.

Popper, K. R. 1959. The logic of scientific discovery. Harper and Row, New York, New York, USA.

Rao, C. R. 1992. R. A. Fisher: The founder of modern statistics. Statistical Science 7:34–48.

Royall, R. M. 1997. Statistical evidence: a likelihood paradigm. Chapman and Hall, London, UK.

Student. 1908. The probable error of a mean. Biometrika 6:1–25.