Mundry, R. 2011. Issues in information theory-based statistical inference—a commentary from a frequentist's perspective. Behavioral Ecology and Sociobiology 65:57–68.
Nakagawa, S., and I. C. Cuthill. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews 82:591–605.
Quinn, J. F., and A. E. Dunham. 1983. On hypothesis testing in ecology and evolution. American Naturalist 122:602–617.
Ramsey, F., and D. Schafer. 2002. The statistical sleuth: a course in methods of data analysis. Second edition. Duxbury Press, Belmont, California, USA.
Rinella, M. J., and J. J. James. 2010. Invasive plant researchers should calculate effect sizes, not *P*-values. Invasive Plant Science and Management 3:106–112.
Royall, R. M. 1997. Statistical evidence: a likelihood paradigm. Chapman and Hall, New York, New York, USA.
Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. Akaike information criterion statistics. D. Reidel Publishing, Hingham, Massachusetts, USA.
Stephens, P. A., S. W. Buskirk, G. D. Hayward, and C. Martinez del Rio. 2005. Information theory and hypothesis testing: a call for pluralism. Journal of Applied Ecology 42:4–12.
Strong, D. R. 1980. Null hypotheses in ecology. Synthese 43:271–285.
Yoccoz, N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bulletin of the Ecological Society of America 72:106–111.

Ecology, 95(3), 2014, pp. 617–621

The common sense of *P* values

PERRY DE VALPINE[1]

Department of Environmental Science, Policy and Management, 130 Mulford Hall #3114, University of California, Berkeley, California 94720-3114 USA

When perplexed graduate students ask me about the anti-*P*-value arguments they've heard, I offer them many of the same responses as Murtaugh (2014), and some others as well. Now I can start by having them read his paper. In this comment, I will support his basic message but dig more deeply into some of the issues.

What are *P* values for? The purpose of *P* values is to convince a skeptic that a pattern in data is real. Or, when you are the skeptic, the purpose of *P* values is to convince others that a pattern in data could plausibly have arisen by chance alone. When there is a scientific need for skeptical reasoning with noisy data, the logic of *P* values is inevitable.

Say there is concern that the chemical *gobsmackene* is toxic to frogs, but *gobsmackene* is an effective insecticide. The proponents of *gobsmackene* are vehement skeptics of its toxicity to frogs. You run an experiment, and the resulting *P* value is 0.001 against the null hypothesis that *gobsmackene* has no effect on frogs. As the expert witness in a trial, you explain to the judge that, if *gobsmackene* is not toxic to frogs, the chances of obtaining data at least as extreme as yours just by a fluke are tiny: just one in a thousand. The judge interprets this probabilistic statement about your evidence and bans *gobsmackene*.

Manuscript received 2 July 2013; revised 10 September 2013; accepted 10 September 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[1] E-mail: pdevalpine@berkeley.edu

Now take an example from the other side, where you are the skeptic. Suppose someone claims to have a treatment that neutralizes a soil contaminant. She presents experimental data from 20 control and 20 treated plots, and the treated plots have 30% less of the contaminant than the control plots. You examine the data and determine that the variation between replicates is so large that, even if the treatment really has no effect, there would be a 20% chance of reporting an effect at least that big, i.e., $P = 0.20$. Since that is more likely than having three children turn out to be all boys, you are not convinced that their treatment is really effective. Of course, one doesn't need good-guy/bad-guy cartoons to imagine the kind of serious skepticism for which *P* value reasoning is useful.

These uses of *P* value reasoning seem like common sense. Why, then, is there so much controversy about such reasoning? I agree with Murtaugh (2014) that many anti-*P*-value arguments boil down to frustrations with practice rather than principle. For example, the arbitrariness of the conventional 0.05 threshold for significance is an example of the "fallacy of the beard." How many whiskers does it take to make a beard? Because it is impossible to give a precise answer that doesn't admit exceptions, should you choose never to discuss beards? Similarly, the arbitrariness of 0.05 is unavoidable, but that doesn't mean we shouldn't consider *P* values as one way to interpret evidence against a null hypothesis. And if a null hypothesis is silly, there will be no skeptics of the alternative, so *P* values are unnecessary.

However, other anti-*P*-value arguments are worth taking more seriously, and Murtaugh (2014) does not do them justice. Specifically, it is worth examining what people mean when they argue that *P* values do not measure evidence. Such arguments are the starting point for dismissing the use of *P* values, which motivates the use of information theoretic and Bayesian methods. An excellent, ecologically oriented volume on scientific evidence was edited by Taper and Lele (2004).

I will argue that while *P* values are not a general measure of evidence, they can provide valuable interpretations of evidence. A useful set of distinctions that emerges from the debates about statistical philosophy is among "model fit," "evidence," and "error probabilities" (Lele 2004, Mayo 2004). For purposes here, likelihoods measure "model fit," likelihood ratios compare "evidence" between two models, and *P* values are "error probabilities." Thus, likelihood ratios (to which *F* ratios are closely related) are the central quantity for comparing models, and *P* values are one way to interpret them. Separating these ideas goes a long way towards clarifying a healthy role for *P* values as one type of reasoning about hypotheses. In this light, commonly used *P* values such as for linear models, analysis of variance, generalized linear (and mixed) models, and other likelihood ratio tests all make sense and are philosophically sound, even if they are not the tool one needs for every analysis.

Sample space arguments against P *values*

What are the philosophical arguments against *P* values? Many dismissals of *P* values are built on claims that sample spaces must by force of logic be irrelevant to measuring evidence, from which it follows that P values cannot measure evidence. For example, Murtaugh (2014) quotes Hobbs and Hilborn (2006) on this point, who in turn cited Royall (1997). The sample space of a probability model is the mathematical "space" of all possible data sets. A *P* value is the probability of observing evidence at least as strong against the null hypothesis as the actual evidence. Therefore, a *P* value is a summation (or integration) over probabilities of data that could have been observed but weren't, i.e., a summation over the sample space. If sample space probabilities can have no role in reasoning about hypotheses, then *P* values are useless. This is considered to be an implication of the "likelihood principle," which states that "all the information about [parameters] ... is contained in the likelihood function" (Berger and Wolpert 1988:19). Under the likelihood principle, likelihood ratios alone—but not the corresponding *P* value—compare hypotheses.

Before considering the arguments against sample space probabilities, it is important to see why the *P* value concept *alone* is considered inadequate as a general measure of evidence for comparing hypotheses (Berger and Wolpert 1988, Royall 1997, Lele 2004). Hypothetically, one could construct *P* values (or confidence intervals) that are technically valid but bizarre. For example, one could make a procedure where the confidence interval is calculated in randomly chosen ways yet still rejects the null at the nominal probability level (Royall 1997). Such examples have almost no bearing on statistical practice, but they do establish that the *P* value concept alone is not enough. We need *P* values based on a good measure of evidence (or test statistic), generally likelihood ratios, which illustrates why it is useful to separate "evidence" (likelihood ratio) from "error probabilities" (e.g., *P* values).

Now we are ready to consider the arguments against sample space probabilities. Two common types of hypothetical examples are those for stopping rules and those for multiple testing (Berger and Wolpert 1988, Royall 1997). In a stopping-rule example, one considers data collected under two protocols. Either the sample size is pre-determined, or intermediate analyses of some of the data may be used to decide whether to continue the study, i.e., there is a "stopping rule." Now suppose that at the end of each study, each protocol happens to have generated exactly the same data. Then, it is asserted, it makes no sense to reach different conclusions about various hypotheses.

However, if one uses *P* values, the second study must be viewed as a stochastic process in which the sample space includes different times at which the study might have been stopped based on intermediate results. Therefore, the probability of obtaining evidence at least as strong against a null hypothesis involves a different sample space probability for the two studies. This violates the likelihood principle by considering not just the final data but also the decisions used to obtain the data. To be sure, some thought experiments are less blatant, or more bizarre, than this one, but they all involve the same data generated by different protocols (Royall 1997).

The other type of argument against sample-space probabilities involves multiple testing. Again one compares two people who obtain the same data in different ways (Royall 1997). Say both are studying human traits associated with myopia (near-sightedness), and both use a sample of 100 people. In addition to measuring myopia, one person measures only birth month and the other measures birth month and 999 other variables. If both obtain the same data for birth month and myopia, they should have the same evidence about that specific relationship. It should not matter that the second person measured other variables, but that is exactly what a multiple-testing correction (e.g., Bonferroni) would enforce.

The flaw in both types of examples is that they dismiss the possibility that how a study is conducted really can impact the probability of obtaining spurious evidence. In the stopping-rule example, if you tell experimentalists they are allowed to check their data at every step and stop when they have a result they like, some really would

do so, and that protocol really would shape the probabilities of spurious outcomes. In the multiple-testing example, if you tell someone they can inspect 1000 relationships separately and then write headline news about whichever one has the strongest evidence while ignoring the fact that they started with 1000 variables, they have a much higher chance of reporting a spurious result than does someone who looked at only one relationship. Mayo (2004) argues that this disconnect between philosophy and practice arose because many philosophers of science begin their reasoning once the data are in hand, and the process of obtaining data may seem irrelevant.

The situation is not helped by the rhetoric of irritation: It seems plainly ridiculous that if you "peek" at your data, you are automatically invalidating your results, or that if you decide to measure some extra variables in addition to your primary variable (e.g., birth month) you must change you interpretations about birth month (Royall 1997). But those misconstrue the logic of *P* values: if you merely peek and then proceed, or if you treat birth month as an a priori hypothesis no matter what else you measure, then you have caused no harm. But if you stop when you decide your data look good, or if you study many variables and report whichever has the strongest evidence, you are de facto changing the probability of obtaining a strong result. In summary, the anti-sample-space arguments against *P* values ignore the fact that how data are collected can influence the probabilities of the final data in hand.

Another problem with the multiple-testing argument is that it pretends one must control the family-wise error rate (e.g., Bonferroni correction). However, alternatives include presenting each result separately and letting the reader evaluate them or presenting a false discovery rate. In other words, disliking Bonferroni corrections is unrelated to whether *P* value logic is sound.

Model fit, evidence, and error probabilities

Making the distinctions between model fit, evidence, and error probabilities can unwind a lot of tension in the above debates, or at least focus them more narrowly. In both cases above, if the likelihood ratio (comparison of model fits) represents "evidence," then the likelihood principle is satisfied. This relies upon all the statistical foundations about why likelihoods are so useful (Lele 2004). Then, as an aid to deciding what to believe from the evidence, the error probabilities represented by *P* values are one useful type of reasoning. Indeed, even Berger and Wolpert (1988), in their seminal anti-*P*-value work, state that "most classical procedures work very well much of the time," even if they would disagree about why. Such classical procedures would seem to include *P* values and confidence intervals based on likelihood ratios or the closely related *F* ratios, i.e., the vast majority of *P* values that ecologists generate.

Using these distinctions, the two investigators in the examples above might have the same evidence, but different error probabilities. One might have obtained the evidence by a procedure more likely than the other to generate spurious results, such as a stopping rule or multiple testing. In these distinctions, the usage of "evidence" is construed in a narrow sense to mean "likelihood ratio." It may be confusing that a broad dictionary definition of "evidence" would include anything used to decide what to believe, which could encompass *P* values based on likelihood ratios.

These distinctions also clarify that likelihoods are more fundamental, and *P* values are one way to use likelihoods. Indeed, Murtaugh (2014) is implicitly talking not about any kind of hypothetical *P* values but rather about *P* values *from likelihood ratios* (or *F* ratios). In this case, the two have a monotonic relationship, and one can see why he chooses to speak of *P* values directly as "evidence." However, after so much philosophical debate has gone into separating the two concepts, it strikes me as confusing to try to put them back together. In essence, Murtaugh (2014) is using the broad sense of "evidence," but statisticians have gone to great lengths to posit the narrow sense of "evidence" to mean likelihood ratios. By discussing *P* values based on likelihood ratios, Murtaugh has blurred this distinction, although his fundamental points remain sound.

Unfortunately an occasional rhetorical follow-on to the dismissal of sample space probabilities is to describe them derisively as probabilities of unobserved data. This makes *P* values sound contrary to common sense because rather than focusing on the data actually observed, they consider possible unobserved data, which sounds foolish. This is misleading. If one is using a probability model for the data, which is the basis for likelihoods in the first place, then part and parcel of that model are the probabilities associated with unobserved data. A likelihood calculation *would not exist* if the model didn't describe a distribution for other hypothetical data. Using sample space probabilities is not on its face ridiculous.

Another problem with the arguments against *P* values is to treat their validity as akin to a mathematical theorem: it must be universally either true or false. But there is no reason to dismiss a principle that works in some situations and not others; doing so should violate ecologists' healthy sense of pragmatism. The types of hypotheticals used for these philosophical debates typically have no bearing on, say, finding the *P* value of one parameter in a model for simply collected data. Indeed, most anti-sample-space thought experiments do not consider the common situation of nested models, and hence don't address the fact that a "bigger" model will always fit the data better and so we often need to apply skeptical reasoning (Forster and Sober 2004).

A final argument against *P* values as evidence is tied to the relationship between *P* values and accept/reject hypothesis testing. In a nutshell, the statement "we found no evidence ($P = 0.06$)" appears to offend common sense. If $P = 0.06$, there is certainly evidence

against the null, and to state otherwise sounds Orwellian. This is not a flaw with *P* values, but rather with presentation customs. I interpret "we found no evidence ($P = 0.06$)" to be shorthand for a more elaborate explanation that the evidence was not strong enough to convince a skeptic and hence no claim about it will be attempted. Unfortunately this has created a feeling that *P* value *practices* enforce ridiculous statements even though the principles are sound. Again, distinguishing the concepts of evidence and error probabilities clarifies the issue.

In summary, the philosophical arguments to dismiss *P* value logic fail to appreciate its role in skeptical reasoning about serious hypotheses. Skepticism is fundamental to science, and *P* values are fundamental to skeptical reasoning.

Bayesian methods and P *values*

Murtaugh (2014) does not go far into Bayesian arguments against *P* values, but there is much to say there. It is common for a pro-Bayesian argument to begin with a rejection of frequentism as represented by sample space probabilities and *P* values. However, there are also "Bayesian *P* values" and credible intervals (as Murtaugh points out, confidence intervals come from a continuous application of *P* values, so without *P* values there would be no confidence intervals), so one must ask if these are really more meaningful than their frequentist counterparts. For example, a common pro-Bayesian argument is that frequentist confidence intervals are approximate while Bayesian credible intervals are exact. I will argue next that this is a misleading comparison.

The reason a confidence interval is approximate is that the procedure is trying to do something objective but hard: cover the correct parameter value in 95% of data sets. In a Bayesian analysis, a 95% credible interval is defined with "degree of belief" probability, so it is meaningless to call it "accurate" vs. "approximate." Sure, it is "exact" in the sense of a correct execution of obtaining the posterior distribution, but we could just as well say the frequentist confidence interval is exact because we obtained it with mathematically correct profile likelihood calculations. So the pro-Bayesian argument amounts to saying "we prefer an exact calculation of something subjective to an approximation of something objective."

Let us make a more stark example. Suppose a weather forecaster predicts a 10% probability of rain tomorrow. What should that mean? The frequentist answer is that 10% means "1 out of 10 times on average": out of many "10%" predictions, it should have actually rained 10% of the time. This could be tested with data. The Bayesian definition of 10% *probability* of rain tomorrow is 10% *degree-of-belief* in rain, which can mean whatever the forecaster wants it to. If you collect data and determine that "10% degree-of-belief" corresponds to rain 20% of the time, you would have done nothing to change what the forecaster wants 10% to mean, nor could you. Are you happier with an "exact" 10% degree-of-belief that can mean anything or an "approximate" 10% frequentist probability that aims for objectivity?

If this example makes you uncomfortable, you are in good company. In a seminal paper, Rubin (1984) argued that to be scientists, we must have objective criteria that allow the possibility of rejecting a model (or hypothesis) from data. Since a pure Bayesian outlook does not provide this possibility, we should seek Bayesian results that are "calibrated" to frequentist probability. In other words, Bayesians still need frequentism. In practice, many Bayesian results will be *approximately* calibrated to frequentist interpretations because of asymptotic likelihood theory. In summary, the need for skeptical reasoning in science leads to *P* values, and the objectivity desired in such reasoning should make one uncomfortable with a pure Bayesian stance.

Model selection and P *values*

Murtaugh's (2014) emphasis on the relationship between AIC model selection and *P* values is good medicine for the way they are often viewed in opposition to each other. However, they are derived to solve different problems: AIC is for finding the best model for predicting new data, and for all one knew it might have led somewhere unrelated to *P* values. They turn out to be closely related (for nested models), which is what Murtaugh emphasizes, but the different problems they solve make the use of one vs. the other more than a "stylistic" difference. A more subtle point is that the derivation of AIC is valid even when the models are not nested and do not include the "correct" model (Burnham and Anderson 2002), while likelihood ratio tests rely upon nested, well-chosen models.

Murtaugh highlighted that some conventions for interpreting AIC differences, such as thresholds of 4 and 7, would be conservative by hypothesis-testing standards. On the other hand, I think there is a temptation to over-interpret the AIC winner between two models, which is a liberal threshold by hypothesis testing standards. If model F has one more parameter than model R, it must have a maximum log likelihood more than 1.0 higher than R in order to have a better AIC (Murtaugh 2014: Eq. 4). The *P* value from the corresponding likelihood ratio test is 0.16 (Murtaugh 2014: Eq. 5; $\Delta\text{AIC} = 0$ in Murtaugh 2014: Fig. 2), liberal rather than conservative. An amazing outcome of the AIC derivation is that it actually puts meaning on this particular *P* value threshold. (There is a deeper reason to be careful mixing hypothesis tests with model selection, which is that a test comparing the top two models from a larger set does not incorporate the stochasticity of the AIC ranks.).

History, headaches, and P *values*

I suspect that many ecologists are happy to see *P* values get philosophically whacked because they are

such a bloody pain. Long ago, ecology was a science of storytelling. It was an uphill battle to infuse hypothesis testing into the science, but eventually it became the gold standard for presenting results. Unfortunately, that meant that authors and journals over-emphasized *P* values and under-emphasized effect size, biological significance, and statistical prediction. Against that background, those who pushed forward model selection, model averaging, and Bayesian methods faced another uphill battle to broaden the scope of ecological statistics beyond *P* values. As a result, *P* values have sometimes been bashed rather than put in healthy perspective.

In complicated situations, simply obtaining a valid *P* value (or confidence interval) can be so difficult that sometimes practice is confused with principle. It would be nice if we could dismiss them as unimportant when they are impractical. Ironically, however, some of the methods that arose behind anti-*P*-value arguments are useful for obtaining better *P* values for difficult problems. For example, one use of model-averaging with AIC weights is to obtain more accurate *P* values and confidence intervals by incorporating uncertainty due to model selection. Similarly, there are claims that sometimes Bayesian procedures can provide more accurate frequentist coverage.

Suppose you are backed up against a wall by a gang of wild raving pure scientists, and they are forcing you to *decide what you believe*. Under such duress, objective statements about error probabilities can provide one useful line of reasoning for Bayesians and frequentists alike. What would science be if one can't argue objectively that someone else's claims are statistically spurious?

Acknowledgments

I thank Daniel Turek and Subhash Lele for helpful comments.

Literature Cited

Berger, J. O., and R. L. Wolpert. 1988. The likelihood principle: a review, generalizations, and statistical implications. Second edition. IMS lecture notes. Monograph series, volume 6. Institute of Mathematical Statistics, Hayward, California, USA.

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.

Forster, M., and E. Sober. 2004. Why likelihood? Pages 153–165 *in* M. L. Taper and S. R. Lele, editors. The nature of scientific evidence. University of Chicago Press, Chicago, Illinois, USA.

Hobbs, N. T., and R. Hilborn. 2006. Alternatives to statistical hypothesis testing in ecology. Ecological Applications 16:5–19.

Lele, S. R. 2004. Evidence functions and the optimality of the law of likelihood. Pages 191–216 *in* M. L. Taper and S. R. Lele, editors. The nature of scientific evidence. University of Chicago Press, Chicago, Illinois, USA.

Mayo, D. G. 2004. An error-statistical philosophy of evidence. Pages 79–97 *in* M. L. Taper and S. R. Lele, editors. The nature of scientific evidence. University of Chicago Press, Chicago, Illinois, USA.

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

Royall, R. M. 1997. Statistical evidence: a likelihood paradigm. Chapman and Hall, New York, New York, USA.

Rubin, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Annals of Statistics 12(4):1151–1172.

Taper, M. L., and S. R. Lele. 2004. The nature of scientific evidence. University of Chicago Press, Chicago, Illinois, USA.

Ecology, 95(3), 2014, pp. 621–626

To *P* or not to *P*?

Jarrett J. Barber[1,3] and Kiona Ogle[2]

[1]*School of Mathematical and Statistical Sciences, Arizona State University, Tempe, Arizona 85287-1804 USA*
[2]*School of Life Sciences, Arizona State University, Tempe, Arizona 85287-6505 USA*

Introduction

We appreciate Murtaugh's (2014) very readable defense of *P* values. Murtaugh argues that most of the criticisms of *P* values arise more from misunderstanding or misinterpretation than from intrinsic shortcomings of *P* values. After an introductory musing on a familiar definition of the *P* value, we discuss what appears to be an "intrinsic shortcoming" of the *P* value, rather than a misunderstanding or misinterpretation; the *P* value lacks what might be considered a very reasonable property to be desired in measures of evidence of hypotheses (Schervish 1996). Then, we attempt to provide a sense of the misunderstanding of *P* values as posterior probabilities of hypotheses (Murtaugh's fifth criticism) or as error probabilities; *P* values can often be much

Manuscript received 22 July 2013; revised 23 August 2013; accepted 26 August 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[3] E-mail: Jarrett.Barber@asu.edu