The problem with p-values is how they're used*

Andrew Gelman†

5 July 2013

I agree with Murtaugh (and also with Greenland and Poole 2013, who make similar points from a Bayesian perspective) that with simple inference for linear models, p-values are mathematically equivalent to confidence intervals and other data reductions, there should be no strong reason to prefer one method to another. In that sense, my problem is not with p-values but in how they are used and interpreted.

Based on my own readings and experiences (not in ecology but in a range of social and environmental sciences), I feel that p-values and hypothesis testing have led to much scientific confusion by researchers treating non-significant results as zero and significant results as real. In many settings I have found estimation rather than testing to be more direct. For example, when modeling home radon levels (Lin et al. 1999), we constructed our inferences by combining direct radon measurements with geographic and geological information. This approach of modeling and estimation worked better than a series of hypothesis tests that would, for example, reject the assumption that radon levels are independent of geologic characteristics.

I have, on occasion, successfully used p-values and hypothesis testing in my own work, and in other settings I have reported p-values (or, equivalently, confidence intervals) in ways that I believe have done no harm, as a way to convey uncertainty about an estimate (Gelman 2013). In many other cases, however, I believe that null hypothesis testing has led to the publication of serious mistakes, perhaps most notoriously in the paper by Bem (2011), who claimed evidence for extra-sensory perception (ESP) based on a series of statistically significant results. The ESP example was widely recognized to indicate a crisis in psychology research, not because of the substance of Bem's implausible and unreplicated claims, but

*Discussion of "In defense of P-values," by Paul Murtaugh, for *Ecology*. We thank two reviewers for helpful comments and the National Science Foundation for partial support of this work.

†Department of Statistics, Columbia University, New York, N.Y.

because the research methods used to purportedly demonstrate the truth of these claims were nothing but the standard null-hypothesis significance tests that are standard in so many fields.

As researchers in medicine and psychology such as Ioannidis (2005), Simmons, Nelson, and Simonsohn (2011), Yarkoni (2011), and Francis (2013) have discussed, the problem is not merely with claims that are ridiculous on scientific grounds, but more broadly that many statistically significant claims will be in error. Gelman and Weakliem (2009) discuss the "statistical significance filter": results that succeed in having a low p-value will inherently yield overestimates of the magnitude of effects and comparisons ("Type M," or magnitude, errors) and are also likely to go in the wrong direction ("Type S," or sign, errors).

The article under discussion reveals a perspective on statistics which, by focusing on static data, is much different from mine. Murtaugh writes:

> Data analysis can be always be redone with different statistical tools. The suitability of the data for answering a particular scientific question, however, cannot be improved upon once a study is completed. In my opinion, it would benefit the science if more time and effort were spent on designing effective studies with adequate replication, and less on advocacy for particular tools to be used in summarizing the data.

I do not completely agree with this quotation, nor do I entirely agree with its implications. First, the data in any scientific analysis are typically not set in stone, independent of the statistical tools used in the analysis. Often I have found that the most important benefit derived from a new statistical method is that it allows the inclusion of more data in drawing scientific inferences. Here are some quick examples:

- Meta-analysis and hierarchical models allow partial pooling.
- Multinomial discrete-data regression models allow researchers to make fuller use of their measurements, going beyond the simple binary thresholding required for basic

logistic regression.

- Multivariate methods such as factor analysis allow the use of multiple correlated measurements.

- Regularized regression methods such as lasso make it possible to include large numbers of predictors in regression models, much more than is possible using least squares methods for variable selection.

My second point of disagreement with the quotation above is in the implication that too much time is spent on considering how to perform statistical inference. (Murtaugh writes of "advocacy" but this seems to me to be a loaded term.) It is a well-accepted principle of the planning of research that the design of data collection is best chosen with reference to the analysis that will later be performed. We cannot always follow this guideline—once data have been collected, they will ideally be made available for any number of analyses by later researchers—but it still suggests that concerns of statistical methods are relevant to design. Beyond this, as noted above, the choice of statistical method is not just about deciding how to summarize "the data" but also influences what data are included in the analysis.

In conclusion, I share the long-term concern (see Krantz 1999, for a review) that the use of p-values encourages and facilitates a sort of binary thinking in which effects and comparisons are either treated as zero or are treated as real, and also an old-fashioned statistical perspective under which it is difficult to combine information from different sources. The article under discussion makes a useful contribution by emphasizing that problems in research behavior will not automatically be changed by changes in data reductions. The mistakes that people make with p-values, could also be made using confidence intervals and AIC comparisons, and I think it would be good for statistical practice to move forward from the paradigm of yes/no decisions drawn from stand-alone experiments.

Hypothesis testing and p-values are so compelling in that they fit in so well with the Popperian model in which science advances via refutation of hypotheses. For both theoretical

and practical reasons I am supportive of a (modified) Popperian philosophy of science in which models are advanced and then refuted (Gelman and Shalizi 2013). But a necessary part of falsificationism is that the models being rejected are worthy of consideration. If a group of researchers in some scientific field develops an interesting scientific model with predictive power, then I think it very appropriate to use this model for inference and to check it rigorously, eventually abandoning it and replacing it with something better if it fails to make accurate predictions in a definitive series of experiments. This is the form of hypothesis testing and falsification that is valuable to me. In common practice, however, the "null hypothesis" is a straw man that exists only to be rejected. In this case, I am typically much more interested in the size of the effect, its persistence, and how it varies across different situations. I would like to reserve hypothesis testing for the exploration of serious hypotheses and not as in indirect form of statistical inference that typically has the effect of reducing scientific explorations to yes/no conclusions.

References

Bem, D. J. 2011. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. Journal of Personality and Social Psychology 100:407–25.

Francis, G. 2013. Replication, statistical consistency, and publication bias (with discussion). Journal of Mathematical Psychology.

Gelman, A. 2013. P values and statistical practice. Epidemiology 24:69–72.

Gelman, A., and Shalizi, C. 2013. Philosophy and the practice of Bayesian statistics (with discussion). British Journal of Mathematical and Statistical Psychology 66:8–18.

Gelman A., and Weakliem, D. 2009. Of beauty, sex, and power: statistical challenges in estimating small effects. American Scientist 97:310316.

Greenland S., and Poole, C. 2013. Living with P-values: resurrecting a Bayesian perspective on frequentist statistics. Epidemiology 24:62–68.

Ioannidis, J. 2005. Why most published research findings are false. PLOS Medicine 2(8):e124.

Krantz, D. H. 1999. The null hypothesis testing controversy in psychology. Journal of the American Statistical Association 44:1372–1381.

Lin, C. Y., Gelman, A., Price, P. N., and Krantz, D. H. 1999. Analysis of local decisions using hierarchical modeling, applied to home radon measurement and remediation (with discussion). Statistical Science 14:305–337.

Simmons J., Nelson L., and Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. Psychological Science 22:1359–1366.

Yarkoni, T. (2011. The psychology of parapsychology, or why good researchers publishing good articles in good journals can still get it totally wrong. Citation Needed blog, 10 Jan. http://www.talyarkoni.org/blog/2011/01/10/the-psychology-of-parapsychology-or-why-good-researchers-publishing-good-articles-in-good-journals-can-still-get-it-totally-wrong/