# *P* Values and Statistical Practice

*Andrew Gelman*

Sander Greenland and Charles Poole[1] accept that *P* values are here to stay but recognize that some of their most common interpretations have problems. The casual view of the *P* value as posterior probability of the truth of the null hypothesis is false and not even close to valid under any reasonable model, yet this misunderstanding persists even in high-stakes settings (as discussed, for example, by Greenland in 2011).[2] The formal view of the *P* value as a probability conditional on the null is mathematically correct but typically irrelevant to research goals (hence, the popularity of alternative—if wrong—interpretations). A Bayesian interpretation based on a spike-and-slab model makes little sense in applied contexts in epidemiology, political science, and other fields in which true effects are typically nonzero and bounded (thus violating both the "spike" and the "slab" parts of the model).

I find Greenland and Poole's[1] perspective to be valuable: it is important to go beyond criticism and to understand what information is actually contained in a *P* value. These authors discuss some connections between *P* values and Bayesian posterior probabilities. I am not so optimistic about the practical value of these connections. Conditional on the continuing omnipresence of *P* values in applications, however, these are important results that should be generally understood.

Greenland and Poole[1] make two points. First, they describe how *P* values approximate posterior probabilities under prior distributions that contain little information relative to the data:

> This misuse [of *P* values] may be lessened by recognizing correct Bayesian interpretations. For example, under weak priors, 95% confidence intervals approximate 95% posterior probability intervals, one-sided *P* values approximate directional posterior probabilities, and point estimates approximate posterior medians.

I used to think this way, too (see many examples in our books), but in recent years have moved to the position that I do not trust such direct posterior probabilities. Unfortunately, I think we cannot avoid informative priors if we wish to make reasonable unconditional probability statements. To put it another way, I agree with the mathematical truth of the quotation above, but I think it can mislead in practice because of serious problems with apparently noninformative or weak priors.

Second, the main proposal made by Greenland and Poole is to interpret *P* values as bounds on posterior probabilities:

> [U]nder certain conditions, a one-sided *P* value for a prior median provides an approximate lower bound on the posterior probability that the point estimate is on the wrong side of that median.

This is fine, but when sample sizes are moderate or small (as is common in epidemiology and social science), posterior probabilities will depend strongly on the prior distribution. Although I do not see much direct value in a lower bound, I am intrigued by Greenland and Poole's[1] point that "if one uses an informative prior to derive the posterior probability of

Partially supported by the Institute for Education Sciences (grant ED-GRANTS-032309-005) and the U.S. National Science Foundation (grant SES-1023189).
From the Departments of Statistics and Political Science, Columbia University, New York, NY.
***Editors' note:*** *Related articles appear on pages 62 and 73.*
Correspondence: Andrew Gelman, Departments of Statistics and Political Science, Columbia University, New York, NY 10027. E-mail: gelman@stat.columbia.edu.

ISSN: 1044-3983/13/2401-0069
DOI: 10.1097/EDE.0b013e31827886f7

the point estimate being in the wrong direction, $P_0/2$ provides a reference point indicating how much the prior information influenced that posterior probability." This connection could be useful to researchers working in an environment in which *P* values are central to communication of statistical results.

In presenting my view of the limitations of Greenland and Poole's[1] points, I am leaning heavily on their own work, in particular on their emphasis that, in real problems, prior information is always available and is often strong enough to have an appreciable impact on inferences.

Before explaining my position, I will briefly summarize how I view classical *P* values and my experiences. For more background, I recommend the discussion by Krantz[3] of null hypothesis testing in psychology research.

## WHAT IS A *P* VALUE IN PRACTICE?

The *P* value is a measure of discrepancy of the fit of a model or "null hypothesis" *H* to data *y*. Mathematically, it is defined as $\Pr(T(y^{rep})>T(y)|H)$, where $y^{rep}$ represents a hypothetical replication under the null hypothesis and *T* is a test statistic (ie, a summary of the data, perhaps tailored to be sensitive to departures of interest from the model). In a model with free parameters (a "composite null hypothesis"), the *P* value can depend on these parameters, and there are various ways to get around this, by plugging in point estimates, averaging over a posterior distribution, or adjusting for the estimation process. I do not go into these complexities further, bringing them up here only to make the point that the construction of *P* values is not always a simple or direct process. (Even something as simple as the classical chi-square test has complexities to be discovered; see the article by Perkins et al[4]).

In theory, the *P* value is a continuous measure of evidence, but in practice it is typically trichotomized approximately into *strong evidence*, *weak evidence*, and *no evidence* (these can also be labeled highly significant, marginally significant, and not statistically significant at conventional levels), with cutoffs roughly at $P = 0.01$ and 0.10.

One big practical problem with *P* values is that they cannot easily be compared. The difference between a highly significant *P* value and a clearly nonsignificant *P* value is itself not necessarily statistically significant. (Here, I am using "significant" to refer to the 5% level that is standard in statistical practice in much of biostatistics, epidemiology, social science, and many other areas of application.) Consider a simple example of two independent experiments with estimates (standard error) of 25 (10) and 10 (10). The first experiment is highly statistically significant (two and a half standard errors away from zero, corresponding to a normal-theory *P* value of about 0.01) while the second is not significant at all. Most disturbingly here, the difference is 15 (14), which is not close to significant. The naive (and common) approach of summarizing an experiment by a *P* value and then contrasting results based on significance levels, fails here, in implicitly giving the imprimatur of statistical significance on a comparison that could easily be explained by chance alone. As discussed by Gelman and Stern,[5] this is not simply the well-known problem of arbitrary thresholds, the idea that a sharp cutoff at a 5% level, for example, misleadingly separates the $P = 0.051$ cases from $P = 0.049$. This is a more serious problem: even an apparently huge difference between clearly significant and clearly nonsignificant is not itself statistically significant.

In short, the *P* value is itself a statistic and can be a noisy measure of evidence. This is a problem not just with *P* values but with any mathematically equivalent procedure, such as summarizing results by whether the 95% confidence interval includes zero.

## GOOD, MEDIOCRE, AND BAD *P* VALUES

For all their problems, *P* values sometimes "work" to convey an important aspect of the relation of data to model. Other times, a *P* value sends a reasonable message but does not add anything beyond a simple confidence interval. In yet other situations, a *P* value can actively mislead. Before going on, I will give examples of each of these three scenarios.

### A *P* Value that Worked

Several years ago, I was contacted by a person who suspected fraud in a local election.[6] Partial counts had been released throughout the voting process and he thought the proportions for the various candidates looked suspiciously stable, as if they had been rigged to aim for a particular result. Excited to possibly be at the center of an explosive news story, I took a look at the data right away. After some preliminary graphs—which indeed showed stability of the vote proportions as they evolved during election day—I set up a hypothesis test comparing the variation in the data to what would be expected from independent binomial sampling. When applied to the entire data set (27 candidates running for six offices), the result was not statistically significant: there was no less (and, in fact, no more) variance than would be expected by chance alone. In addition, an analysis of the 27 separate chi-square statistics revealed no particular patterns. I was left to conclude that the election results were consistent with random voting (even though, in reality, voting was certainly not random—for example, married couples are likely to vote at the same time, and the sorts of people who vote in the middle of the day will differ from those who cast their ballots in the early morning or evening). I regretfully told my correspondent that he had no case.

In this example, we cannot interpret a nonsignificant result as a claim that the null hypothesis was true or even as a claimed probability of its truth. Rather, nonsignificance revealed the data to be compatible with the null hypothesis; thus, my correspondent could not argue that the data indicated fraud.

### A *P* Value that Was Reasonable but Unnecessary

It is common for a research project to culminate in the estimation of one or two parameters, with publication turning

on a *P* value being less than a conventional level of significance. For example, in our study of the effects of redistricting in state legislatures (Gelman and King),[7] the key parameters were interactions in regression models for partisan bias and electoral responsiveness. Although we did not actually report *P* values, we could have: what made our article complete was that our findings of interest were more than two standard errors from zero, thus reaching the $P < 0.05$ level. Had our significance level been much greater (eg, estimates that were four or more standard errors from zero), we would doubtless have broken up our analysis (eg, studying Democrats and Republicans separately) to broaden the set of claims that we could confidently assert. Conversely, had our regressions not reached statistical significance at the conventional level, we would have performed some sort of pooling or constraining of our model to arrive at some weaker assertion that reached the 5% level. (Just to be clear: we are not saying that we would have performed data dredging, fishing for significance; rather, we accept that sample size dictates how much we can learn with confidence; when data are weaker, it can be possible to find reliable patterns by averaging.)

In any case, my point is that in this example it would have been just fine to summarize our results in this example via *P* values even though we did not happen to use that formulation.

### A Misleading *P* Value

Finally, in many scenarios *P* values can distract or even mislead, either a nonsignificant result wrongly interpreted as a confidence statement in support of the null hypothesis or a significant *P* value that is taken as proof of an effect. A notorious example of the latter is the recent article by Bem,[8] which reported statistically significant results from several experiments on extrasensory perception (ESP). At brief glance, it seems impressive to see multiple independent findings that are statistically significant (and combining the *P* values using classical rules would yield an even stronger result), but with enough effort it is possible to find statistical significance anywhere (see the report by Simmons et al[9]).

The focus on *P* values seems to have both weakened that study (by encouraging the researcher to present only some of his data so as to draw attention away from nonsignificant results) and to have led reviewers to inappropriately view a low *P* value (indicating a misfit of the null hypothesis to data) as strong evidence in favor of a specific alternative hypothesis (ESP) rather than other, perhaps more scientifically plausible, alternatives such as measurement error and selection bias.

## PRIORS, POSTERIORS, AND *P* VALUES

Now that I have established my credentials as a pragmatist who finds *P* values useful in some settings but not others, I want to discuss Greenland and Poole's proposal to either interpret one-sided *P* values as probability statements under uniform priors (an idea they trace back to Gossett)[10] or else to use one-sided *P* values as bounds on posterior probabilities (a result they trace back to Casella and Berger).[11]

The general problem I have with noninformatively derived Bayesian probabilities is that they tend to be too strong. At first, this may sound paradoxical, that a noninformative or weakly informative prior yields posteriors that are too forceful—and let me deepen the paradox by stating that a stronger, more informative prior will tend to yield weaker, more plausible posterior statements.

How can it be that adding prior information weakens the posterior? It has to do with the sort of probability statements we are often interested in making. Here is an example from Gelman and Weakliem.[12] A sociologist examining a publicly available survey discovered a pattern relating attractiveness of parents to the sexes of their children. He found that 56% of the children of the most attractive parents were girls, when compared with 48% of the children of the other parents, and the difference was statistically significant at $P < 0.02$. The assessments of attractiveness had been performed many years before these people had children, so the researcher felt he had support for a claim of an underlying biological connection between attractiveness and sex ratio.

The original analysis by Kanazawa[13] had multiple-comparisons issues, and after performing a regression analysis rather than selecting the most significant comparison, we get a *P* value closer to 0.2 rather than the stated 0.02. For the purposes of our present discussion, though, in which we are evaluating the connection between *P* values and posterior probabilities, it will not matter much which number we use. We shall go with $P = 0.2$ because it seems like a more reasonable analysis given the data.

Let $\theta$ be the true (population) difference in sex ratios of attractive and less attractive parents. Then the data under discussion (with a two-sided *P* value of 0.2), combined with a uniform prior on $\theta$, yield a 90% posterior probability that $\theta$ is positive. Do I believe this? No. Do I even consider this a reasonable data summary? No again. We can derive these "No" responses in three different ways: first, by looking directly at the evidence; second, by considering the prior; and third, by considering the implications for statistical practice if this sort of probability statement were computed routinely.

First, a claimed 90% probability that $\theta > 0$ seems too strong. Given that the *P* value (adjusted for multiple comparisons) was only 0.2—that is, a result that strong would occur a full 20% of the time just by chance alone, even with no true difference—it seems absurd to assign a 90% belief to the conclusion. I am not prepared to offer 9-to-1 odds on the basis of a pattern someone happened to see that could plausibly have occurred by chance alone, nor for that matter would I offer 99-to-1 odds based on the original claim of the 2% significance level.

Second, the prior uniform distribution on $\theta$ seems much too weak. There is a large literature on sex ratios, with factors such as ethnicity, maternal age, and season of birth

corresponding to difference in probability of girl birth of <0.5 percentage points. It is a priori implausible that sex-ratio differences corresponding to attractiveness are larger than for these other factors. Assigning an informative prior centered on zero shrinks the posterior toward zero, and the resulting posterior probability that $\theta > 0$ moves to a more plausible value in the range of 60%, corresponding to the idea that the result is suggestive but not close to convincing.

Third, consider what would happen if we routinely interpreted one-sided *P* values as posterior probabilities. In that case, an experimental result that is 1 standard error from zero—that is, exactly what one might expect from chance alone—would imply an 83% posterior probability that the true effect in the population has the same direction as the observed pattern in the data at hand. It does not make sense to me to claim 83% certainty—5-to-1 odds—based on data that not only could occur by chance alone but in fact represent an expected level of discrepancy. This system-level analysis accords with my criticism of the flat prior: as Greenland and Poole[1] note in their article, the effects being studied in epidemiology are typically range from −1 to 1 on the logit scale; hence, analyses assuming broader priors will systematically overstate the probabilities of very large effects and will overstate the probability that an estimate from a small sample will agree in sign with the corresponding population quantity.

Rather than relying on noninformative priors, I prefer the suggestion of Greenland and Poole[1] to bound posterior probabilities using real prior information. I would prefer to perform my Bayesian inferences directly without using *P* values as in intermediate step, but given the ubiquity of *P* values in much applied work, I can see that it can be helpful for researchers to understand their connection to posterior probabilities under informative priors.

## SUMMARY

Like many Bayesians, I have often represented classical confidence intervals as posterior probability intervals and interpreted one-sided *P* values as the posterior probability of a positive effect. These are valid conditional on the assumed noninformative prior but typically do not make sense as unconditional probability statements. As Sander Greenland has discussed in much of his work over the years, epidemiologists and applied scientists in general have knowledge of the sizes of plausible effects and biases. I believe that a direct interpretation of *P* values as posterior probabilities can be a useful start—if we recognize that such summaries systematically overestimate the strength of claims from any particular dataset. In this way, I am in agreement with Greenland and Poole's interpretation of the one-sided *P* value as a lower bound of a posterior probability, although I am less convinced of the practical utility of this bound, given that the closeness of the bound depends on a combination of sample size and prior distribution.

The default conclusion from a noninformative prior analysis will almost invariably put too much probability on extreme values. A vague prior distribution assigns much of its probability on values that are never going to be plausible, and this disturbs the posterior probabilities more than we tend to expect—something that we probably do not think about enough in our routine applications of standard statistical methods. Greenland and Poole[1] perform a valuable service by opening up these calculations and placing them in an applied context.

## REFERENCES


1. Greenland S, Poole C. Living with *P*-values: resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*. 2013;24:62–68.
2. Greenland S. Null misinterpretation in statistical testing and its impact on health risk assessment. *Prev Med*. 2011;53:225–228.
3. Krantz DH. The null hypothesis testing controversy in psychology. *J Am Stat Assoc*. 1999;94:1372–1381.
4. Perkins W, Tygert M, Ward R. Computing the confidence levels for a root-mean- square test of goodness-of-fit. *Appl Math Comput*. 2011;217:9072–9084.
5. Gelman A, Stern HS. The difference between "significant" and "not significant" is not itself statistically significant. *Am Stat*. 2006;60:328–331.
6. Gelman A. 55,000 residents desperately need your help! *Chance.* 2004;17:28–31.
7. Gelman A, King G. Enhancing democracy through legislative redistricting. *Am Polit Sci Rev*. 1994;88:541–559.
8. Bem DJ. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *J Pers Soc Psychol*. 2011;100:407–425.
9. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci*. 2011;22:1359–1366.
10. Gossett, WS. The probable error of the man. *Biometrika.* 1908;6:1–25.
11. Casella G, Berger RL. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J Am Stat Assoc*. 1987;82:106–111.
12. Gelman A, Weakliem D. Of beauty, sex, and power: statistical challenges in estimating small effects. *Am Sci*. 2009;97:310–316.
13. Kanazawa S. Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis (gTWH). *J Theor Biol*. 2007;244:133–140.