Low P-Values or Narrow Confidence Intervals: Which Are More Durable?

Charles Poole

What should be the role of P-values and confidence intervals in the interpretation of scientific results? This question is not new and our field of epidemiology is far from alone in struggling with it.\(^1\) I have four suggestions for authors and readers. The first is quite broad, so I offer that one before describing current practices. I then turn to the other three. My remarks are confined to settings in which P-values and confidence intervals accompany estimates of effect measures, such as the relative risk.

Briefly, here are my suggestions. One, we should work harder than ever to avoid strict or exact interpretations of P-values and confidence intervals in observational research, where these statistics lack a theoretical basis. Two, we should stop interpreting P-values and confidence intervals as though they measure the probability of hypotheses. Three, when we want to know the probability of hypotheses, we should use Bayesian methods, which are designed expressly for that purpose. Four, we should get serious about precision and look for narrow confidence intervals instead of low P-values to identify results that are least influenced by random error.

Real Life Is Not Randomized

When treatment or exposure is randomized, we have a solid theoretical basis, testable in simulations, for the probability models from which P-values, confidence intervals, and likelihoods are deduced. In observational research, all we can do is hope that the social, behavioral, and physical processes by which people become exposed to risk factors in the unrandomized real world do not differ too greatly from randomization.\(^4\) Unfortunately, each time we find that risk factors are associated with each other in observational studies, we find evidence against that hope. We cannot remind ourselves too often of this fundamental problem. At the very least, it should cause us to avoid hairsplitting interpretations of probabilistic statistics in observational research, where they are intrinsically fuzzy.

Contemporary Uses of P-Values and Confidence Intervals

Significance testing unquestionably dominates epidemiology today. In attempting to refrain from this practice over the past 17 years,\(^5\) I have often been expected, assumed, encouraged, and sometimes even forced to engage in it by editors, reviewers, colleagues, professors, students, funding sources, regulators, attorneys, and journalists. It is not easy to be a non-tester in a testing world.

After Rothman's highly influential 1978 essay, "A Show of Confidence," an immense and easily documented shift in reporting style took place.\(^7\) Whereas P-values or "S" (significant) and "NS" (not significant) once were reported exclusively, the reporting of confidence intervals has now become accepted practice, with or without P-value accompaniment. Confidence intervals have a survival advantage for the tiny non-testing minority to which I belong. They enable us to gauge the precision of estimates easily, but without depriving the established majority of its beloved tests.

Epidemiologists who see no purpose to a confidence interval other than its use in significance testing sometimes wonder why this shift in reporting practice has occurred. The P-value provides the information they desire more efficiently and exactly. Some are vaguely aware that confidence intervals supposedly convey information that P-values do not, but are unsure what that extra information is and even less sure how it might be useful. The word "precision" seems to be used with increasing regularity nowadays, and confidence intervals are occasionally described as "wide," but "wide" and "imprecise" often seem nothing more than code words for "includes the null value" and hence for "not statistically significant."

Improbable Observations Do Not Imply Improbable Hypotheses

When we estimate a parameter such as the relative risk, each possible value of that parameter is the expected value under some hypothesis, and each hypothesis has a P-value.\(^8,9\) What we call the P-value for the null hypothesis. Approximately, each P-value is the probability of obtaining an estimate at least as far from a specified value as the estimate we have
obtained, if that specified value were the true value. It follows that no P-value, for the null hypothesis or any other, is the probability that the specified hypothesis is true. As an obvious example, the hypothesis corresponding to the point estimate has a (two-sided) P-value of 1.0. However, we do not treat our point estimates as absolutely certain to be true. Neither is the point estimate, in general, the most probable value.

For a given estimate, the 95% confidence interval is the set of all parameter values for which $P \geq 0.05$. For the value at each limit of a 95% confidence interval, $P = 0.05$ (two-sided). Thus, if either of the 95% confidence limits for a relative risk estimate equals 1.0 (the null value of this parameter), we can infer that the null P-value is 0.05. From this link between confidence intervals and P-values, it follows that a 95% confidence interval is not a range of values within which the unknown true value lies with 95% probability.

The well-known “coverage probability” of confidence intervals pertains to a parameter value that is known to be true and the probability that an as yet unknown confidence interval will contain it. Coverage probability does not pertain to a known confidence interval and an unknown true value. To interpret a given 95% confidence interval as having a 95% probability of including the unknown true value is to mistake a frequentist confidence interval for a Bayesian probability interval. This error is merely an extension of the logical fallacy of mine, but our biostatistical teachers give us the hypotheses, which require Bayesian methods to determine whether the specified hypothesis is true.

Why do we turn probability logic on its head in this way? We very much want to know the probabilities of the null hypothesis or any other hypothesis true? These are the probabilities of ultimate concern in significance testing – the probability of declaring “significance” when the null hypothesis is true. One minus the Type II error rate is the test’s power or sensitivity: the probability of declaring “significance” when the alternative hypothesis is true. No informed patient would want to know which values are more and less likely, more and less plausible, etc., we must specify prior probabilities for those values and use Bayes’s theorem to update those probabilities when new data are in hand.

It has become increasingly clear that the null P-value (hereafter called “the” P-value) does not do a very good job of the task for which it was originally intended: to quantify the statistical evidence against the null hypothesis. The reason is simple. The familiar Type I and Type II error rates upon which Neyman and Pearson taught us to focus beg vitally important questions.

One minus the Type I error rate is the specificity of a significance test: the probability of not declaring “significance” when the null hypothesis is true. One minus the Type II error rate is the test’s power or sensitivity: the probability of declaring “significance” when the alternative hypothesis is true. No informed patient would want to know which values are more and less likely, more and less plausible, etc., we must specify prior probabilities for those values and use Bayes’s theorem to update those probabilities when new data are in hand.

Significance tests are no different. In the same frequency terms that Neyman and Pearson used, the researcher who wishes to be fully informed should be interested in questions such as the following: How often is the null hypothesis true when we fail to reject it? When we do reject the null hypothesis, how often is the alternative hypothesis true? These are the probabilities of ultimate concern in significance testing – the predictive values of “NS” and “S.” There is no way to determine them without postulating (stated again in frequency terms) how often the null and alternative hypotheses are true.
The interest many epidemiologists express in how low the \( P \)-value is, if it is lower than 0.05,\(^{19-26}\) raises still other questions. How much evidence against the null hypothesis do we have when \( P = 0.04 \), or when \( P = 0.001 \)? To answer these questions, we need to consider the probabilities under the null and alternative hypotheses of obtaining these particular \( P \)-values, not just the probabilities of obtaining \( P < 0.05 \).

Statisticians who have examined these questions in detail\(^{19-26}\) have found, under widely ranging conditions, that \( P \)-values on the order of 0.05, 0.01, and even lower provide much less evidence against the null hypothesis than they appear to provide at face value. As a general matter, \( P \)-values in the vicinity of 0.05 provide almost no evidence against the null hypothesis at all. \( P = 0.04 \), for instance, is typically found to be almost equally probable under the null and alternative hypotheses.

One upshot of this work has been a statistical research program devoted to calibrating, standardizing, conditioning, or adjusting low \( P \)-values to make them higher, so that they reflect more realistically the limited statistical evidence they provide against the null hypothesis.\(^{19-26}\) Now that Bayesian methods are computationally feasible, one wonders whether these efforts to patch up \( P \)-values will ultimately be viewed a transitional stopgap.

**Taking Precision Seriously**

Transitional stopgaps should not be dismissed lightly, especially when the transitions in question take decades to unfold. Stopgaps can be particularly valuable when it seems that the only alternative is to cry in the (frequentist) wilderness for a (Bayesian) revolution. In epidemiology, the advent of confidence intervals creates an opportunity to take another small step toward more widespread use of Bayesian methods, while at the same time improving overall interpretation. This step is merely to take precision seriously.

Epidemiologists have many reasons to emphasize certain results over others. Some results may pertain to particularly topical research questions. Some may be more valid than others. And some may be less influenced by random error. This last consideration seems to be an important one to many epidemiologists, who regularly use \( P \)-values to determine the degree to which chance influences their results. They believe that the lower the \( P \)-value, the less the influence of chance. Unfortunately, this extremely common use of the \( P \)-value is a misuse and an abuse of that statistic. The estimates least influenced by chance are not those with low \( P \)-values, but those with narrow confidence intervals.

Consider the four hypothetical relative risk estimates in Table 1. The ratio of the upper to lower 95% confidence limits (CLR) is a handy measure of confidence interval width, and thus of precision. (For a difference measure such as the risk difference, the difference between the upper and lower confidence limits would serve the same purpose.) The example was devised to dramatize four clear-cut combinations of statistical “significance” and precision.

<table>
<thead>
<tr>
<th>Exposure or Disease</th>
<th>RR (95% CI)</th>
<th>( P )</th>
<th>95% CLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.5 (0.80–8.0)</td>
<td>0.1</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>1.7 (1.2–2.4)</td>
<td>0.003</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>4.1 (1.2–14)</td>
<td>0.02</td>
<td>12</td>
</tr>
<tr>
<td>D</td>
<td>1.4 (0.80–2.4)</td>
<td>0.2</td>
<td>3</td>
</tr>
</tbody>
</table>

To the extent that the role of chance would be taken into account in deciding which of these results to emphasize, the conventional choices would be the statistically “significant” estimates B and C. These would be the “associations unlikely to be due to chance alone.” But one of them, estimate C, is very unstable. That estimate is influenced much more by random error, and from that standpoint is much less dependable, than estimate B.

Of equal importance, when C is compared with D, estimate C is influenced much more by chance and in that regard is much less trustworthy, even though estimate C is statistically “significant” and estimate D is not. Estimates B and D – not B and C – are this study’s most precise estimates. Estimates B and D stand the best chance of holding up, conditional on their validity, in the context of existing and future research. Estimates B and D would weigh more heavily into meta-analyses and would exert stronger influences on probability distributions in properly conducted Bayesian analyses.

It is sometimes said that confidence intervals are especially valuable, and that increases in sample size and statistical efficiency are particularly needed, when statistical “significance” has not been attained. To the contrary, an estimate that has a wide confidence interval is imprecise and unstable no matter how low its \( P \)-value. Based solely on the results in Table 1, larger sample sizes, special study populations and statistically more efficient designs would be particularly desirable for A and C, regardless of the fact that one of these estimates is statistically “significant” and the other is not.

Some epidemiologists wonder what all the fuss over \( P \)-values and confidence intervals is about. This hypothetical example shows how an emphasis on precision rather than statistical “significance” can affect which results we may choose to highlight. I invite the reader to examine published research reports in which the estimates with the lowest \( P \)-values have been singled out for emphasis, and to imagine how differently those papers would read if the estimates with the narrowest confidence intervals had been highlighted instead.

**CONCLUSION**

Our results that deserve the greatest reliance are those that are most stable and trustworthy. With regard to random error, a very poor way of identifying dependable results is to select associations with impressively low
P-values. Inference and decision-making would be far better served by choosing estimates with narrow confidence intervals, which are least vulnerable to the play of chance. These are the results for which, by virtue of intentional or accidental features of our research methods, our studies provide the most evidence (as distinguished from the most valid evidence).

By taking precision seriously, we can easily identify those research questions on which our studies provide the greatest quantity of statistical evidence, and those questions for which larger and more statistically efficient studies are needed. In terms of resistance to random error, our most durable results are our most precise estimates - however unspectacular, unsensational, and “non-significant” many of those estimates might be.

References