In Praise of Confidence Intervals

By David Romer*

There has been a revolution in econometric practice in recent decades. The intense focus on identification, the emphasis on economic as well as statistical significance, the attention to ensuring that standard errors are correct, and the development of new techniques to deal with dynamics, heterogeneity, and nonlinearities are enormous advances over the empirical research of previous generations.

The thesis of this paper is that there is at least one dimension on which there is still considerable room for improvement. As I document, most modern empirical papers concentrate on two characteristics of their findings: whether the point estimates are statistically significantly different from zero, and the economic interpretation of the point estimates. But in almost all applications, there are potential values of the parameters other than the point estimates and zero that are of interest. Focusing on point estimates and statistical significance obscures the implications of the findings for those values. In addition, as discussed below, this focus also leaves out important information even about the strength of the evidence against a parameter value of zero.

This problem can be solved easily. Reporting confidence intervals for key estimates and discussing their economic interpretation allows a researcher to communicate the essence of the missing information.1

I. Some Evidence

I suspect most economists will find the statement that current empirical papers generally emphasize point estimates and statistical significance uncontroversial. Nonetheless, to provide systematic evidence on this issue, I examine empirical papers published in 2019 in the three leading nonspecialized general interest journals in economics: the American Economic Review (AER), the Journal of Political Economy (JPE), and the Quarterly Journal of Economics (QJE). Specifically, I start by examining all papers in the QJE in 2019 that estimate parameters statistically (rather than being entirely theoretical or reporting only descriptive statistics). There are 35 such papers. I then consider the first 35 papers in 2019 that estimate parameters statistically in each of the AER and JPE, for a total of 105 papers.

I classify the papers into five groups: (1) ones that discuss confidence intervals prominently (for example, by reporting a confidence interval in the abstract, introduction, or conclusion, or reporting multiple confidence intervals elsewhere in the text), or otherwise discuss the implications of the results for parameter values other than point estimates or zero; (2) ones that report standard errors (or t-statistics) prominently in the text but do not discuss confidence intervals; (3) ones that mention confidence intervals in passing (for example, a single brief mention of a confidence interval in the text); (4) ones that mention standard errors or t-statistics in passing but do not mention confidence intervals; (5) ones that mention neither confidence intervals nor standard errors in the text.2

---

*University of California, Berkeley (email: dromer@econ.berkeley.edu). I thank Amy Finkelstein, Guido Imbens, Patrick Kline, and Christina Romer for valuable comments and discussions.

1Not surprisingly, versions of all the ideas in this paper have been advanced in various forms before. For a few examples, see Manski (forthcoming) on the temptations of overstating the degree of certainty associated with empirical findings and the dangers of such unwarranted certitude; Abadie (forthcoming) on the drawbacks to a single-minded focus on a null hypothesis of zero and the benefits of reporting results in ways that convey information about the full range of possible parameter values; and the special issue of The American Statistician, Statistical Inference in the 21st Century: A World Beyond p < 0.05 (Jeske 2019).

2There are two reasons for not considering information reported only in tables and figures in classifying papers. First, and most importantly, what authors put in the text indicates what they consider important and shapes how their findings

---

1Go to https://doi.org/10.1257/pandp.20201059 to visit the article page for additional materials and author disclosure statement.
Table 1 reports the results. Only 14 percent of the papers (15 out of 105) report confidence intervals at all prominently. An additional 10 percent discuss standard errors somewhat prominently. Six percent mention confidence intervals in passing, and 7 percent mention standard errors in passing.

The remaining 64 percent of the papers make no mention of confidence intervals or standard errors in the text. Some of these papers report only that the estimates are statistically significant; some provide qualitative information about statistical significance (such as “strongly” or “moderately”); others report information on p-values (such as “at the 5 percent level” or “p < 0.01,” or sometimes, actual p-values); and some do not even say whether the estimates are statistically significant. In these papers, it is usually clear from the text that the standard errors on the focal estimates are less than half the size of the point estimates, but there is little information beyond that. The tone is often that once it is known that estimates are significantly different from zero, the only aspect of the results that matters is the point estimates—almost as though when an estimate is significantly different from zero, it can be treated as exact.

The sample size in Table 1 is large enough that the results are unlikely to be greatly affected by the random variation in which empirical papers are in the sample. For example, the upper end of the 2 standard error confidence interval for the fraction of papers that report confidence intervals prominently is 21 percent, and the lower end of the 2 standard error confidence interval for the fraction that do not mention confidence intervals or standard errors at all is 54 percent.

<table>
<thead>
<tr>
<th>Information Reported in the Text of Empirical Papers in Three Leading Journals in 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussed prominently</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Mentioned in passing</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Notes: Standard errors are in parentheses. See text for details.

II. Why Do We Care?

Knowing only a point estimate and that it is significantly different from zero is not enough to know what values of the parameter other than zero the data provide strong evidence against, and what values they provide little reason to object to. Since in almost all applications there are hypotheses of interest other than that a parameter is zero, this is an important omission.

To see this concretely, consider two classic examples, one from microeconomics and one from macroeconomics. The microeconomic example is the rate of return to education—the percent increase in earnings from an additional year of schooling. In most contexts, the possibility that this rate of return is zero is not of great interest: it would make little sense for individuals to devote so much of their lives to education if it had no economic benefit, and the vast previous work in this area already provides overwhelming evidence that the rate of return is positive.

Thus if a new dataset or empirical strategy provides new evidence on the rate of return, most readers’ interest will be in how consistent or inconsistent that evidence is with various values of the return other than zero. For example, they might want to know whether it casts strong doubt on the view that the return is 9.4 percent, which is the median estimate from the numerous instrumental variables estimates using US data surveyed by Card (2001, table 2). Or they might want to know what it tells us about the values of 6 percent for men and 7 percent
for women suggested by Trostel, Walker, and Woolley (2002, p. 15) based on an analysis of
data from 28 countries. Individuals considering
more schooling and policymakers considering
new projects to promote education might be
interested in the implications of the new evi-
dence for whether the return is above whatever
they view as the cutoff for moving forward. And
so on. Knowing the point estimate from a new
study and that it is statistically significantly
different from zero provides little information
about the implications for these various can-
didate values of the rate of return.

The second example is the fiscal multiplier—
the short-run output effect of a one-unit increase
in government purchases with the economy in
a specific set of circumstances (for example,
with output below normal and with a particular
assumption about the behavior of monetary pol-
icy). There are some models where the multi-
plier is zero (notably, flexible price models with
inelastic supply), so here the null hypothesis of
zero is an interesting one. But other values are
also potentially important. One focal value is a
multiplier of one, which is both the value pre-
picted by some models under certain conditions
(Woodford 2011) and the boundary between
stimulus increasing or decreasing private eco-
nomic activity. A policymaker designing a stim-
ulus package might be interested in comparisons
with results from prior work about multipliers
for various types of tax cuts. And, as with the
return to education, values obtained in previous
work, such as the figure of 1.8 suggested in a
recent survey of cross-sectional research by
Chodorow-Reich (2019), are also of interest.
Again, a focus on the point estimate and whether
it is statistically significantly different from zero
is misplaced if a reader’s interest is in knowing
what the evidence tells us about any of these var-
ious other possible values.5

3 Although a central argument of this paper is that taking
zero as the focal null hypothesis is generally not appropriate
in most economic applications, there are other contexts in
which it is appropriate. Indeed, one reason for the common
emphasis on statistical significance and point estimates may
be that this focus made considerable sense in some of the
settings where statistical inference was developed. Consider,
for example, a simplified view of old-fashioned drug devel-

dment: a company synthesizes chemical compounds more
or less at random and then tests their efficacy against a pre-
viously untreated disease. In that setting, there are strong
reasons to expect most of the compounds to have essentially

III. When a Value of Zero Is the Relevant Null

One can take the argument a step further. Even when the issue is whether a parameter is
zero, focusing on point estimates and statistical significance is problematic. Consider a paper
that reports only that an estimate is significantly
different from zero or that a p-value is less than
some focal level, such as 0.05 or 0.01. Such an
approach conveys very little about whether the
t-statistic is, on the one hand, only slightly or
moderately above 2 (for example, 2.1, which
corresponds to \( p = 0.036 \), or 2.6, which cor-
responds to \( p = 0.009 \)), or, on the other, far
above 2 (for example, 4.0, which corresponds
to \( p = 0.0001 \)). But these two cases have quite
different implications concerning the hypothesis
that the parameter is zero. There are several
reasons.

First, absent any concerns with the estimates
and standard errors, a t-statistic of 2.1 or 2.6
is merely strong evidence against a value of
zero, while a t-statistic of 4.0 is overwhelming
evidence.

Second, although the smaller t-statistics pro-
vide strong evidence against a value of zero,
they provide only moderate evidence against
many values that are much closer to zero than
to the point estimate. In many cases, this means
that the smaller t-statistics provide only moder-
ate evidence against values of the parameter that
are economically small.

Third, and crucially, it is rare for there to be
no issues or concerns with the estimates
and standard errors. There may be grounds for think-
ing that there is some bias, so that an unbiased
estimate is somewhat closer than the point esti-
mate to zero. There may be reasons to fear that
the standard errors are somewhat understated.
There may be concerns about external validity;
that is, there may be factors suggesting that the
situation being studied is unusual, so that again
the best estimate in most relevant circumstances
is somewhat closer than the point estimate to
zero. And there is the possibility of reporting
and publication bias, so that the results are more

4 Although a central argument of this paper is that taking
zero as the focal null hypothesis is generally not appropriate
in most economic applications, there are other contexts in
which it is appropriate. Indeed, one reason for the common
emphasis on statistical significance and point estimates may
be that this focus made considerable sense in some of the
settings where statistical inference was developed. Consider,
for example, a simplified view of old-fashioned drug devel-

dment: a company synthesizes chemical compounds more
or less at random and then tests their efficacy against a pre-
viously untreated disease. In that setting, there are strong
reasons to expect most of the compounds to have essentially

no effect on the disease. Further, a key issue is whether the
compound has any effectiveness; and once there is strong
evidence that a compound is effective (which, given the
large number of compounds tested, would require a p-value
far below 0.05), one of the most important aspects of the
results is the point estimate.
extreme than a typical examination of the issue would find. Collectively, these concerns can easily change the apparently strong evidence against the null of zero provided by a $t$-statistic somewhat above 2 into only moderate evidence. But in the absence of some large identifiable problem with a study’s econometrics, they are unlikely to turn the seemingly overwhelming evidence against the null of zero provided by a $t$-statistic far above 2 into anything less than very strong evidence.

IV. Possible Alternatives

The previous sections argue that a common approach to discussing empirical results often leaves out important information about the results’ implications. At a general level, addressing this omission is straightforward. What is needed is information about the implications of the results for hypotheses of interest. And although there are various ways of providing this information, a natural one is to report and discuss a confidence interval. In contrast to reporting a point estimate and whether it is statistically significantly different from zero, reporting a confidence interval provides information about the full range of possible values of the parameter.

Although the key point is that reporting and discussing confidence intervals would often convey much more about the implications of papers’ findings than the usual approach, this leaves the smaller issue of which confidence interval is the best choice. Importantly, it is not obvious that it is the usual 2 standard error interval. A common shortcut way of interpreting a confidence interval is that the results provide strong evidence against parameter values outside the interval and are essentially equally supportive of all values inside it. But when that shortcut interpretation is applied to a 2 standard error confidence interval (or, almost equivalently, the 95 percent interval), it leads to an exaggerated sense of the uncertainty associated with the results. For example, suppose initially researchers viewed the point estimate and a value at the boundary of the 2 standard error confidence interval as equally likely. Ex post (assuming normality for simplicity), they should view the point estimate as roughly seven times as likely as the value at the boundary. That is, even though both values are in the confidence interval, the results are considerably more supportive of the point estimate than the value at the boundary.

The fact that the standard shortcut interpretation applied to the usual confidence interval produces an overstated sense of uncertainty may be a reason researchers tend not to report confidence intervals. Consider, for example, a paper where the $t$-statistic on the focal estimate is 2.5. The authors may be reluctant to report the 2 standard error confidence interval because that may cause readers to think the results provide “no evidence” against a value only one-fifth the size of the point estimate.

One way to obtain the advantages of reporting confidence intervals without the disadvantages caused by the shortcut interpretation would be to report somewhat narrower intervals than the usual 2 standard error bands. Since some papers already report 90 percent (1.645 standard error) bands, they appear to be the most natural alternative to the traditional 2 standard error ones. With a 90 percent interval, for a researcher who initially viewed the point estimate and the value at the boundary of the interval as equally likely, ex post the point estimate would be roughly four times (rather than seven times) as likely as the value at the boundary.

Even better would be to report both 1 standard error and 2 standard error bands for papers’ key estimates (something that is now sometimes done in figures). Ex post, a researcher who started with flat priors would view the point estimate as only moderately more likely than the other values in the 1 standard error band (concretely, considerably less than twice as likely). Thus in this case, the natural, and roughly correct, shortcut interpretation would be that the results provide little information about the relative merits of different values within the 1 standard error interval, moderate evidence against values in the 2 standard error but not the 1 standard error interval relative to the point estimate, and strong evidence against values outside the 2 standard error band relative to the point estimate.⁴

⁴Of course, the factors that may cause stated confidence intervals to be narrower than the true intervals mean that reporting a 90 percent confidence interval or discussing both the 1 standard error and 2 standard error intervals may understate the amount of uncertainty associated with the estimates. But addressing this by focusing on 95 percent confidence intervals would amount to trying to correct one
V. Conclusion

Consider two possible papers estimating the multiplier for government purchases. In both, the point estimate is 3.0. In one, the standard error is 1.3, while in the other it is 0.7. With the usual current approach to discussing empirical results, the two papers would describe their findings in similar terms. Both would observe that the estimate is statistically significant and would focus on the economic interpretation of a multiplier of 3.

In fact, however, the two results would have very different implications for most questions about the multiplier economists are interested in. With an emphasis on confidence intervals, the papers’ discussions would reflect those differences. The one that obtained a standard error of 1.3 (implying a 1 standard error confidence interval of (1.7, 4.3) and a 2 standard error interval of (0.4, 5.6)) would, as before, observe that the hypothesis of a multiplier of zero is rejected. But it would go on to emphasize that the estimate was not very precise: the results provide little evidence against more conventional values of the multiplier such as 1.8, only moderate evidence against a multiplier of 1 or slightly below, and very strong but not overwhelming evidence against a multiplier of zero. The paper with a standard error of 0.7 (and thus a 1 standard error confidence interval of (2.3, 3.7) and a 2 standard error interval of (1.6, 4.4)), in contrast, would observe that a multiplier of zero was not merely rejected but rejected overwhelmingly. And it would proceed to point out that the results provide quite strong evidence against not just zero but values of 1 and below, and that they even provide moderately strong evidence against recent estimates in the vicinity of 1.8.

All this could be accomplished with the addition of a sentence or two reporting confidence intervals, a few adjectives (“imprecise,” “overwhelming,” “moderate,” and so on), and a few sentences describing the confidence intervals’ implications for key candidate values of the multiplier. Expanding the papers by a very small amount so that two sets of results that imply very different conclusions are presented differently would surely pass a cost-benefit test overwhelmingly.

The thesis of this paper is that these issues are not specific to this pair of possible papers, but common. Because a parameter value of zero is rarely the only one of interest, focusing on statistical significance and point estimates usually leaves out important information about the implications of papers’ findings. In some other fields, such as medicine, reporting and discussing confidence intervals is routine.\(^5\) Doing the same in economics could have considerable value: a practice of reporting confidence intervals for papers’ focal estimates and briefly discussing their key implications would add only trivially to papers’ length, and would often greatly increase the amount of information conveyed about the economic implications of the findings.

REFERENCES


Jeske, Dan, ed. 2019. The American Statistician: Statistical Inference in the 21st Century: A World beyond \(p < 0.05\), Vol. 73, Supplement 1.


\(^5\)For example, the submission guidelines for The New England Journal of Medicine (https://www.nejm.org/author-center/new-manuscripts) state, “Significance tests should be accompanied by confidence intervals for estimated effect sizes, measures of association, or other parameters of interest.” I am grateful to Amy Finkelstein for this point.