thebmj

BMJ 2019;366:l5381 doi: 10.1136/bmj.l5381 (Published 10 September 2019)

HEAD TO HEAD

HEAD TO HEAD

Are confidence intervals better termed "uncertainty intervals"?

Debate abounds about how to describe weaknesses in statistics. **Andrew Gelman** has no confidence in the term "confidence interval," but **Sander Greenland** doesn't find "uncertainty interval" any better and argues instead for "compatibility interval"

Andrew Gelman *professor of statistics and political science*[1], Sander Greenland *professor of epidemiology and statistics*[2]

[1]Columbia University, New York, USA; [2]Department of Epidemiology and Department of Statistics, University of California, Los Angeles, USA

Yes—Andrew Gelman

Science reformers are targeting P values and statistical significance, and rightly so.[1-3] It's wrong to take P≤0.05 as indicating that an effect is real, and it's wrong to take P>0.05 as a reason to act as though an effect is zero.

One proposed reform is to replace statistical significance with confidence intervals: instead of simply reporting whether the 95% interval contains zero or reporting a P value, report the entire interval. But this approach has problems too,[4] in that there can be good reasons in some cases to think that the true effect is likely to be outside the interval entirely. Confidence intervals excluding the true value can result from failures in model assumptions (as we've found when assessing US election polls[5]) or from analysts seeking out statistically significant comparisons to report, thus inducing selection bias.[6]

Confidence intervals can be a useful summary in model based inference. But the term should be "uncertainty interval," not "confidence interval," for four key reasons.

Difficulties in interpretation

My first concern with the term "confidence interval" is the well known confusion in interpretation. Officially, all that can be interpreted are the long term average properties of the procedure that's used to construct the interval, but people tend to interpret each interval implicitly in a bayesian way—that is, by acting as though there's a 95% probability that any given interval contains the true value. For example, I recently reviewed a popular science book that went to the trouble of defining confidence intervals but unfortunately said that they indicated "the level of confidence in the results." The confidence interval shouldn't inspire confidence: it's a measure of uncertainty.

Secondly, in statistics we use models to make predictions, which can propagate uncertainty in parameters to uncertainty in predictions by using predictive simulation. In linear regression we can obtain an uncertainty interval for each coefficient a and b and can predict ranges for future observations, where $y=a+bx+$error. Similar approaches for propagating uncertainty can occur in more complicated models using bayesian simulation or cross validation. Using any of these methods, uncertainty is a unifying principle regulating inferences about parameters and forecasts about the future.

Intuitive sense

My third concern is the awkwardness of explaining that confidence intervals are big in noisy situations where you have less confidence and are small when you have more confidence. The rule that bigger "uncertainty intervals" correspond to more uncertainty makes intuitive sense.

Finally, expressing uncertainty is the most legitimate goal of estimating intervals, in my view. Using confidence intervals to rule out zero (or other parameter values) involves all of the well known problems of significance testing. So, rather than constructing this convoluted thing called a confidence procedure, which is defined to have certain properties on average but can't generally be interpreted for individual cases, I prefer to aim for an uncertainty interval, using the most appropriate statistical methods to get there. In some cases a reasonable answer can be obtained by using classical confidence interval procedures and simply relabeling; in other settings we might prefer bayesian or machine learning methods, but the goal of assessing uncertainty is the same.

Let's use the term "uncertainty interval" instead of "confidence interval." The uncertainty interval tells us how much uncertainty

Correspondence to: A Gelman gelman@stat.columbia.edu, S Greenland lesdomes@ucla.edu

we have. As of this writing, "confidence interval" outnumbers "uncertainty interval" in an online search by the huge margin of 26 million to 138 000. "Uncertainty interval" doesn't even have a Wikipedia page. So, we have some way to go.

No—Sander Greenland

Astute writers have long complained that statistical significance is poorly correlated with practical significance. This, in turn, has led to advice to drop "significance" as a shorthand for P≤0.05 and instead present actual P values for observations.[7-10] For example, the wording "The difference was significant at the 0.05 level," in reference to a P value of 0.024, would become "The difference had P=0.024." Similarly, "There was no significant difference," in reference to a P value of 0.31, would become "The difference had P=0.31."

It's also been suggested that an association with P>0.05 could be described as highly compatible with the data and background assumptions (the statistical model) used to compute the P value.[8 10] This change avoids confusing P>0.05 with any unwarranted implications that the observed difference is not significant practically, averting the fallacy of claiming that "no association was observed" when the observed association is of important magnitude despite having P>0.05.[8 10]

"Compatibility" also emphasizes the dependence of the P value on the assumptions as well as on the data, recognizing that P<0.05 can arise from assumption violations even if the effect under study is null.

Confidence tricks and uncertainty laundering

Criticisms of "significance" translate immediately into criticisms of "confidence" jargon. The label "95% confidence interval" evokes the idea that we should invest the interval with 95/5 (19:1) betting odds that the observed interval contains the true value (which would make the confidence interval a 95% bayesian posterior interval[11]). This view may be harmless in a perfect randomized experiment with no background information to inform the bet (the original setting for the "confidence" concept); more often, however, the 95% is overconfident because it takes no account of procedural problems and model uncertainties that should reduce confidence in statistical results.[12] Those possibilities include uncontrolled confounding, selection bias, measurement error, unaccounted-for model selection, and outright data corruption. They afflict not just single studies but meta-analyses as well.[13 14]

The label "confidence interval" thus provides no basis for confidence as we commonly understand it. In fact, "confidence intervals" were initially deemed a "confidence trick."[15] One response is to relabel confidence intervals as "uncertainty intervals"—but, like "confidence," the word "uncertainty" gives the illusion that the interval properly accounts for all important uncertainties. In reality, conventional statistics can't begin to capture uncertainties about study limitations, and valid uncertainty intervals require much more thorough modeling of uncertainty sources.[12 14] The term "uncertainty interval" is thus a glaring example of uncertainty laundering[16]—misrepresenting uncertainty as if it were a known quantity.

Choose compatibility, not confidence

As noted, no conventional interval adequately accounts for procedural problems that afflict data generation or for uncertainties about the statistical assumptions. Such inadequacies invalidate claims that confidence intervals cover the true value 95% of the time, and they render false the implication (fostered by the label "uncertainty interval") that a conventional interval captures the key uncertainties about the results.

Nonetheless, all values in a conventional 95% interval can be described as highly compatible with data under the background statistical assumptions, in the very narrow sense of having P>0.05 under those assumptions. In equivalent terms: given any value in the interval and the background assumptions, the data should not seem very surprising.[10] This leads to the intentionally modest term "compatibility interval" as a replacement for "confidence interval." The "compatibility" label offers no false confidence and no implication of complete uncertainty accounting; instead, it treats the interval as nothing more than an exhibit of relations between the data and various possibilities under the analysis assumptions.

In summary, both "confidence interval" and "uncertainty interval" are deceptive terms, for they insinuate that we have achieved valid quantification of confidence or uncertainty despite omitting important uncertainty sources. Such labels misrepresent deep knowledge gaps as if they were mere random errors, fully accounted for by the intervals.

Replacing "significance" and "confidence" labels with "compatibility" is a simple step to encourage honest reporting of how little we can confidently conclude from our data. This, in turn, can mitigate the overconfidence seen in many (now discredited) treatment recommendations, while encouraging new recommendations to remain cautious and compatible with what has been observed.

AG thanks the US Office of Naval Research for partial support of this work.

Competing interests: Both authors have read and understood BMJ policy on declaration of interests and have no relevant interests to declare.

Provenance and peer review: Commissioned; not externally peer reviewed.

1 Greenland S. The need for cognitive science in methodology. *Am J Epidemiol* 2017;186:639-45. 10.1093/aje/kwx259 28938712
2 McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. *Am Stat* 2019;73:235-4510.1080/00031305.2018.1527253.
3 Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;567:305-7. 10.1038/d41586-019-00857-9 30894741
4 Carlin JB, Gelman A. Some natural solutions to the p-value communication problem—and why they won't work. *J Am Stat Assoc* 2017;112:899-90110.1080/01621459.2017.1311263.
5 Shirani-Mehr H, Rothschild D, Goel S, Gelman A. Disentangling bias and variance in election polls. *J Am Stat Assoc* 2018;113:607-1410.1080/01621459.2018.1448823.
6 Vasishth S, Mertzen D, Jäger LA, Gelman A. The statistical significance filter leads to overconfident expectations of replicability. *J Mem Lang* 2018;103:151-7510.1016/j.jml.2018.07.004.
7 Wasserstein R, Lazar N, Schirm A. Editorial: Moving to a world beyond p<0.05. *Am Stat* 2019;73:1-19. https://tandfonline.com/doi/full/10.1080/00031305.2019.158391310.1080/00031305.2019.1583913.
8 Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;567:305-7. https://www.nature.com/articles/d41586-019-00857-9. 10.1038/d41586-019-00857-9 30894741
9 McShane B, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. *Am Stat* 2019;73:235-45. https://www.tandfonline.com/doi/full/10.1080/00031305.2018.152725310.1080/00031305.2018.1527253.
10 Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *Am Stat* 2019;73:262-70. www.tandfonline.com/doi/pdf/10.1080/00031305.2018.154313710.1080/00031305.2018.1543137.
11 Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian data analysis*. 3rd ed. Chapman and Hall/CRC, 2014.
12 Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014;43:1969-85. 10.1093/ije/dyu149 25080530
13 Garmendia CA, Nassar Gorra L, Rodriguez AL, Trepka MJ, Veledar E, Madhivanan P. Evaluation of the inclusion of studies identified by the FDA as having falsified data in the results of meta-analyses: the example of the apixaban trials. *JAMA Intern Med* 2019 (published online 4 Mar). 10.1001/jamainternmed.2018.7661.
14 Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JB. Models for potentially biased evidence in meta-analysis using empirically based priors. *J R Stat Soc* 2009;172:119-3610.1111/j.1467-985X.2008.00548.x.
15 Bowley AL. Discussion on Dr Neyman's paper. *J R Stat Soc* 1934;97:607-10.
16 Gelman A. The problems with p-values are not just with p-values. *Am Stat* 2016;70:S1 (comment 10). http://www.stat.columbia.edu/~gelman/research/published/asa_pvalues.pdf.