American Journal of Epidemiology

Vol. 189, No. 7
DOI: 10.1093/aje/kwaa004
Advance Access publication:
January 29, 2020

The *AJE* Classroom

Can Confidence Intervals Be Interpreted?

Ashley I. Naimi* **and Brian W. Whitcomb**

* Correspondence to Dr. Ashley I. Naimi, Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, 130 DeSoto Street, Parran 503, Pittsburgh, PA 15261 (e-mail: ashley.naimi@pitt.edu).

Initially submitted December 12, 2019; accepted for publication January 7, 2020.

The confidence interval is a fundamental tool for quantifying uncertainty due to sampling associated with a point estimate. However, the manner in which that uncertainty is quantified often results in confusion (1). There has long been criticism of null hypothesis significance testing via *P* values, and growing advocacy for the use of confidence intervals instead (2). As a result, it is important to address points of confusion regarding the interpretation of confidence intervals to be clear about what they do and do not mean. Here, we demonstrate key properties of frequentist confidence intervals, clarify their interpretation, and explain common misunderstandings.

Suppose we are interested in estimating the association between an exposure (X) and outcome (Y), adjusting for a set of 5 confounders (C_1 to C_5) in a sample of 100 observations. In a simulated data set (Web Appendix, available at https://academic.oup.com/aje), we use logistic regression and obtain an estimated regression coefficient (log of the odds ratio) of 0.15 (corresponding to an odds ratio of 1.20) and a model-based standard error of 0.06. Using the standard Wald equation (estimate $\pm$ 1.96 $\pm$ standard error), we calculate 95% confidence limits of 1.02 and 1.31.

CONFIDENCE INTERVALS

The primary question we address here is: what is the English language interpretation of the numerical interval (1.02, 1.31)? To start, we begin with the formal definition of a confidence interval:

> Over infinite repeated sampling, and in the absence of selection, information, and confounding bias, the α-level confidence interval will include the true value in α% of the samples for which it is calculated.

This definition describes an expectation observable over a large number of repeated samples. But, unfortunately, for any single study, we cannot know if the estimated 95% confidence interval is among the 95% that actually included the true value, or among the 5% that did not include the true value. The intervals from any given study will either include or exclude the true value (i.e., the probability that the estimated bound will contain the truth will be either 0 or 1). This property is illustrated in Figure 1, showing data from the scenario used for our simulation, repeated 100 times to represent 100 samples.

It would be incorrect to state there is a 95% probability that the true odds ratio lies between 1.02 and 1.31 in this example. Furthermore, it would be misleading to state we are "95% confident" the true values lies between 1.02 and 1.31. In this setting, the word "confident" does nothing to express what is precisely conveyed by the technical notion of a confidence interval. Indeed, attempts to draw inferences from the confidence limits of 1.02 and 1.31 from our single sample in any frequency-based probabilistic framework can lead to interpretation problems because these numbers provide no information on the actual (probabilistic) degree to which the true odds ratio is captured by the estimated interval values. In contrast to *P* values, the confidence interval provides a range of parameter estimates with units that may be compatible with the data-generating mechanism. However, the upper and lower bounds of the interval, per se, do not provide information regarding confidence, as is sometimes misunderstood.

ESTIMANDS, ESTIMATORS, AND ESTIMATES

Correctly interpreting confidence intervals is easier when the distinction among an estimand, estimator, and estimate is clear. These concepts, often encountered in statistics and causal inference, can provide clarity on why estimated confidence interval values can be easily misunderstood. For confidence intervals, the *estimand* is the parameter(s) of interest—specifically, upper and lower bound values with the properties defined in the previous section: over repeated sampling, bounds that will include the true value at the chosen α-level when there is no confounding, selection, or information bias. This confidence interval estimand can be

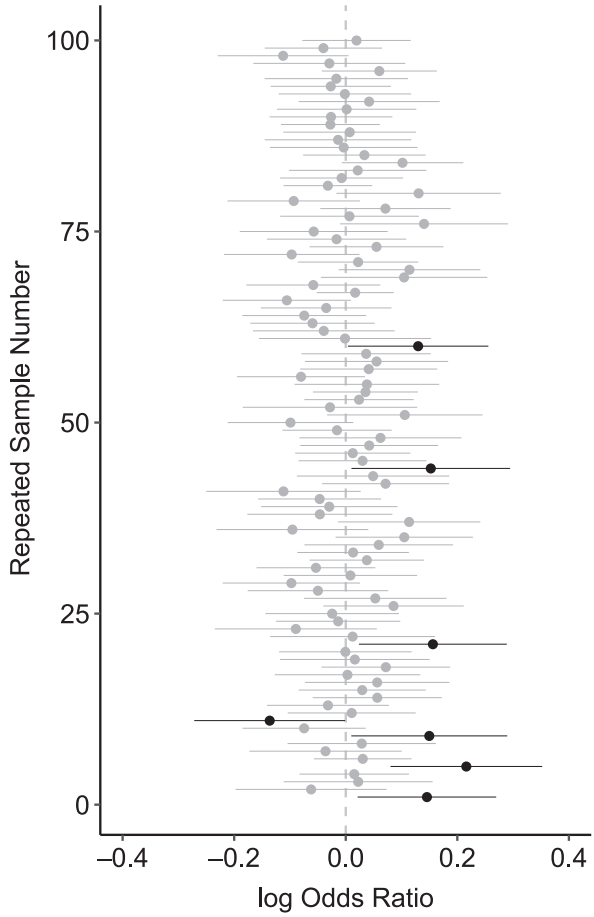

Figure 1. Demonstration of the properties of a confidence interval estimator versus its realized values. Points and lines represent odds ratios and 95% confidence intervals from repeated samples (presented on the log scale). Over 100 repeated samples, 93 of the estimated values cover the true log of the odds ratio = 0 (gray points and lines), whereas 7 do not (black points and lines). For a given realized value, which is what would arise in an actual study, one will never know whether the confidence interval is one that includes or excludes the truth.

targeted using a variety of approaches, or *estimators*, in practice. These estimators include the Wald estimator with model-based standard errors (which we used in the beginning of this article), bootstrap estimators, and many other approaches. Depending on the setting (e.g., small vs. large samples; regular vs. nonregular point estimators; asymptotically linear vs. nonlinear point estimators), some confidence interval estimators will provide better nominal coverage than others. Finally, applying a given confidence interval estimator to data, one obtains a confidence interval *estimate.*

Critically, the definition of the confidence interval applies to the estimand but does not apply to the estimate. That is, one cannot interpret the realized values of a confidence interval estimator (in our case, 1.02 and 1.31) by describing them in terms of the confidence interval estimand. These realized values result from an estimation process to which the notion of a common understanding of a confidence interval applies; however, the 95% confidence interval estimate cannot be characterized by how well it covers the true value. By definition, any given 95% confidence interval estimate will either include or exclude the truth with 100% probability.

INTERPRETING CONFIDENCE INTERVAL ESTIMATES

For all these reasons, we cannot interpret confidence interval estimates by invoking properties that apply to the estimand. Once this feature of confidence intervals is understood, 1 question inevitably arises: how can we interpret confidence interval estimates (in our case, the values of 1.02 and 1.31)? If we cannot gain any information about the extent to which the true parameter value may be contained within the bounds of the confidence interval estimates, what information can be gleaned from these values?

One feature of the estimated confidence interval provides important information about the extent to which the estimate is subject to variability: the estimated confidence interval width (3). As a direct function of the standard error, the confidence interval width measures the degree of precision characterizing the point estimate of interest. Very wide estimated confidence interval limits suggest the results are subject to a high degree of variability or a low degree of precision. In contrast, narrow interval bounds suggest the study results are not subject to a high degree of random variation. One may even compute the confidence limit difference or ratio to obtain a summary measure of the degree of random variation in the estimated results. Finally, upper and lower bounds can (and should) be interpreted with subject matter knowledge in mind, such as the outcome under study, and the effect being estimated. Together with a given point estimate, we may take a sense of the location (value) of an estimate as well as its precision (confidence interval width).

Summarizing the impact of random variation on a study's results is an essential step in conducting any scientific study. Different approaches can be used to do this, including Bayesian credible intervals (4), the surprisal index (5), and confidence intervals. However, the confidence interval estimates should not be interpreted as the confidence interval estimand. Because they reflect the precision of an estimate and a range of values, in contrast to *P* values, which combine both these pieces of information into a single number, there are advantages to interval estimation; however, for confidence intervals, just as with *P* values, correct interpretation of the statistic is key to drawing appropriate inferences from research findings.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, University of Pittsburgh, Pittsburgh, Pennsylvania (Ashley I. Naimi); and Department of Biostatistics and Epidemiology, University of Massachusetts Amherst, Amherst, Massachusetts (Brian W. Whitcomb).

Conflict of interest: none declared.

REFERENCES

1. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337–350.
2. Rothman KJ. A show of confidence. *N Engl J Med*. 1978; 299(24):1362–1363.
3. Poole C. Low *P* values or narrow confidence intervals: which are more durable? *Epidemiology*. 2001;12(3): 291–294.
4. Greenland S. Bayesian perspectivesd for epidemiological research: I. foundations and basic methods. *Int J Epidemiol*. 2006;35(3):765–775.
5. Greenland S. Valid *P* values behave exactly as they should: some misleading criticisms of *P* values and their resolution with *S*-values. *Am Stat*. 2019;73: 106–114.