INVITED FEATURE: STATISTICS PRIMER

Understanding confidence intervals and why they are so important

Claire Cameron, Robin Turner, Ari Samaranayaka

Introduction

Back in 1986, a paper came out in the British Medical Journal (BMJ) called "Confidence intervals rather than P values: estimation rather than hypothesis testing."[1] That was 35 years ago, and we are (unfortunately) still continuing the same discussion about P values and confidence intervals today. We touched on this topic and presented the American Statistical Association (ASA) statement on P values in a previous article.[2]

Now we want to take the opportunity to showcase confidence intervals and shed some light on their meaning and interpretation. In the 1986 paper, the authors say that looking at study findings and categorising them as statistically significant or not "is not helpful and encourages lazy thinking."[1] This is because the purpose of most research investigations in medicine is to determine the size of some measure(s) of interest. A P value from a hypothesis-test simply attempts to determine whether the size of this measure is statistically significant. Let us explain.

What is a confidence interval?

Most of what we do in biostatistics involves answering a research question using a sample. A confidence interval gives a measure of uncertainty or error around an estimated statistic. It is important to note that the interval can only reflect the uncertainty that arises from taking a sample from the underlying population (that is, sampling variability). Non-sampling issues such as bias, accuracy of the measures, or generalisability of the result cannot be inferred from a confidence interval.[3]

Generally, the confidence interval takes the form:

statistic ± *multiplier* × *standard error of the statistic.*

In this formula, the "statistic" is the estimate from our sample; this is our best guess at the true value in the population. This statistic could be a mean, a proportion, a difference in means, a regression coefficient, a relative risk, a hazard ratio, and so on. The "multiplier" is a value from a particular theoretical distribution (e.g. normal distribution, t distribution); the applicable distribution depends on the type of statistic. The value selected from that distribution reflects the "confidence" that the unknown parameter will lie in the confidence interval. Most commonly, we select a value corresponding to 95% confidence (we will come back to what we mean by this). The "standard error of the statistic" represents the variability due to sampling. As shown by the formula, these intervals are usually symmetrical around the statistic. However, confidence intervals for some measures (e.g. relative risk, odds ratio, hazard ratio) are actually estimated on a log scale; therefore, confidence intervals for these measures will not be symmetrical. We are interested in what values are contained in the interval and how wide the interval is. Alongside that, we are interested in the magnitude of the statistic (often called the effect size).

Interpretation

The correct interpretation of a 95% confidence interval is that we are 95% confident that the true value (also known as the population parameter) is contained within the interval. What we mean by this is that if we repeat the sampling in an identical way many times and produce confidence intervals using each sample, 95% of these intervals would include the true (unknown) population parameter.

It is *not correct* to say that there is a 95% chance that an interval will contain the population parameter.[4]

Answering questions with confidence intervals

As mentioned earlier, a confidence interval tells us the size and precision of the estimate, which is more useful than a P value. The questions we could answer using confidence intervals include: What is the difference (in proportions) between these two groups? what is the mean for this population and how does it compare with previously reported means? what is the estimated relative risk? what is the size of the association between this risk factor and the outcome?" Confidence intervals can answer all these questions. Under the hypothesis testing framework, these questions would become: Is there a difference (in proportions) between these two groups? is the mean for this population different to previously reported means? is the relative risk different from 1? is the risk factor "significantly" related to the outcome? These latter questions may feel more comfortable, as we have lived under the P value way of thinking for a long time, however, we would suggest that these are less useful questions to answer than the first set of questions.

As an example, in our P value paper,[2] we presented an estimate of the height difference for men and women. When we had 1000 people per group, the difference in mean height (men minus women) was 2.10 cm, with a 95% confidence interval of 1.23 cm to 2.98 cm. This suggests that men in the sample are, on average, taller than women, because the interval is entirely positive. If we follow the hypothesis testing framework, our finding would be "men and women do not have the same mean height." This latter finding is less informative than the former. When we reduced the sample size to 40 per group, keeping the same difference (2.10 cm), the confidence interval ranged from -2.01 cm (women taller than men) to 4.11 cm (men taller than women). This interval tells us that our data supports a true difference of between -2.01 cm and 4.11 cm. Women may be taller than men (negative differences) by up to 2.01 cm, or men may be taller than women (positive differences) by up to 4.11 cm. We can't rule out the possibility that men and women, on average, are the same height (because the interval contains 0). If we take a closer look at this interval, it is not symmetrical as we would expect; clearly, we have made an arithmetic error. We wish we could say we planned it to demonstrate a point for this paper, but it was a genuine mistake. The actual interval should go from -2.01 cm to 6.22 cm. We wanted to highlight this here

to show how easy it is for mistakes to slip into publications and how, by thinking about them carefully, you can detect these mistakes.

To further demonstrate the utility of confidence intervals, we turn to an example from a recent meta-analysis which reports estimates and confidence intervals from a number of studies.[5] This meta-analysis focused on the mean difference in sodium intake using 24-hour diet recall compared to 24-hour urine collection. For the sake of this example, let us assume that a difference of 500 mg/day of sodium between the 24-hour diet recall and the 24-hour urine measurements is the minimum important difference, i.e. differences less than this are clinically unimportant. Figure 1 shows the results of a subset of six of the 28 studies that were included in the meta-analysis.

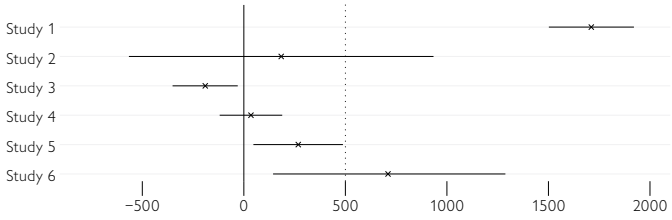

Figure 1. Mean difference in sodium intake and 95% confidence intervals for six selected studies from McLean et. al.[5] The estimated difference was calculated as the sodium intake using 24-hour urine minus sodium intake using 24-hour diet recall. The dashed line shows the minimum difference that is considered important, i.e. differences below this are considered clinically unimportant.

In light of what we know about confidence intervals, how should we interpret the results from these six examples? Study 1 clearly shows a large, clinically important difference between the measures. The confidence interval is well above both zero and 500 mg/day. Study 2 has an interval that includes zero and the clinically important difference, so the interval suggests that the measures could be the same (zero difference) or very different (over 500 mg/day). This interval has provided an inconclusive result. Study 3 has an entirely negative and quite narrow interval, suggesting that the urine measure was slightly less than the 24-hour diet recall measure on average. Whilst the interval does not include zero, indicating that there is a difference in the measures, the interval in its entirety is smaller than 500 mg/day, so the small observed difference is unimportant. This result is similar to that of Study 5, except that the confidence interval of that study is entirely positive. Those two studies have quite narrow intervals, indicating the estimates are reasonably precise. Study 4 has a narrow interval that includes zero. This would suggest that there is no evidence of a difference between the two measures in this study. Study 6 shows a wide interval. The interval suggests the true difference may be greater than zero (and very small) or greater than 500 mg/day (so clinically important). It is difficult to interpret the importance of the effect, because the interval is so wide.

Confidence intervals and hypothesis testing are closely related. For example, if the 95% confidence interval for a difference includes zero, we know that the P value must be greater than 0.05. This is where "lazy thinking" can creep in. We must resist the temptation to treat confidence intervals like hypothesis tests. If we use the interval to say "the interval does not contain zero, therefore the result is significant" and vice versa, we feed into the issue of using P values as a yes or no answer. We then totally overlook the extra information contained in the confidence interval, including thinking about the importance of the results clinically. We must take the opportunity to look at the information in more depth. We should be challenging ourselves to write and interpret results from studies without using the words "statistically" or "significant."

In most studies about human health, investigators are interested in, for example, determining the size of a difference in outcome in the population between groups, rather than a simple indication of whether or not they are different. A confidence interval comprises a range of values, derived using a single sample of data, which tells us the reasonable range for the unknown population value that our sample supports. We find ourselves, 35 years on, still advocating for the use of confidence intervals over hypothesis testing, and advocating for the removal of the term "statistically significant" from the statistical lexicon. Understanding the meaning of confidence intervals is incredibly useful, both for reporting our own research and for reading the literature with a critical eye. We hope the readers of this journal will champion this approach alongside us into the future.

References

1. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J. 1986 Mar;292(6522):746-50.

2. Cameron C, Samaranayaka A, Turner R. Features Articles: P-values: What are their significance? New Zealand Medical Student Journal. 2020 Apr;30:48-9.

3. Altman DG, Bland JM. Uncertainty beyond sampling error. BMJ. 2014 Nov 25;349:g7065.

4. Spiegelhalter D. The art of statistics: learning from data. United Kingdom: Penguin UK; 2019.

5. McLean R, Cameron C, Butcher E, Cook NR, Woodward M, Campbell NRC. Comparison of 24-hour urine and 24-hour diet recall for estimating dietary sodium intake in populations: a systematic review and meta-analysis. The J Clin Hypertens. 2019 Dec;21(12):1753-62.

About the authors

› Dr Claire Cameron, BSc(Hons), DipGrad, MSc, PhD, is a Senior Research Fellow and Biostatistician in the Biostatistics Centre, Division of Health Sciences, University of Otago.
› Associate Professor Robin Turner, Bsc(Hons), MBiostat, PhD, is the Director of the Biostatistics Centre, Division of Health Sciences, University of Otago.
› Dr Ari Samaranayaka, BSc, MPhil, PhD, is a Senior Research Fellow and Biostatistician in the Biostatistics Centre, Division of Health Sciences, University of Otago.

Correspondence

Dr Claire Cameron: claire.cameron@otago.ac.nz