BMJ 2014;349:g7064 doi: 10.1136/bmj.g7064 (Published 25 November 2014)

RESEARCH METHODS & REPORTING

STATISTICS NOTES

Uncertainty and sampling error

Douglas G Altman *professor of statistics in medicine*[1], J Martin Bland *professor of health statistics*[2]

[1]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, UK; [2]Department of Health Sciences, University of York, York YO10 5DD, UK

Medical research is conducted to help to reduce uncertainty. For example, randomised controlled trials aim to answer questions relating to treatment choices for a particular group of patients. Rarely, however, does a single study remove uncertainty. There are two reasons for this: sampling error and other (non-sampling) sources of uncertainty. The word "error" comes from a Latin root meaning "to wander," and we use it in its statistical sense of meaning variation from the average, not "mistake." Sampling error arises because any sample may not behave quite the same as the larger population from which it was drawn. Non-sampling error arises from the many ways a research study may deviate from addressing the question that the researcher wants to answer.

Sampling error is very much the concern of the statistician, who imagines that the group of people in the study is just one of the many possible samples from the population of interest. Despite it being widely condemned,[1] the dominant way of summarising the evidence from a research study is by the P value. It should be obvious that the evidence from a research study cannot reasonably be summarised as just a single number, but the use of P values remains unshakeable. Further, the practice of labelling P values as significant or not significant leads not only to dichotomous decisions but often also to the belief that the research question has been answered.

P values represent the probability that the observed data (or a more extreme result) could have arisen when the true effect of interest is zero—for example, the true treatment effect in a randomised trial. It is common to interpret P<0.05 ("significant") as clear evidence that there is a real effect, and P>0.05 ("not significant") as evidence that there is no effect. However, the former interpretation may be unwise, and the latter is wrong. Although 0.05 is the conventional decision point, P<0.05 is far from representing certainty. One in 20 studies could have a difference of the observed size if there were really no difference in the population. "Not significant" indicates that we found insufficient evidence to conclude that there is a real effect, not that we have shown that there is not one.[2] Referring to results as *statistically* significant, or not, only helps a bit.

Interpretation of a study's results should be primarily based on the estimated effect and a measure of its uncertainty. In mainstream statistics, the uncertainty of estimates is indicated by the use of confidence intervals. Before the mid-1980s, confidence intervals were rarely seen in clinical research articles. Around 1986 things changed,[3] and these days almost all clinical research articles in major journals include confidence intervals. The confidence interval is a range of uncertainty around the estimate of interest, such as the treatment effect in a controlled trial.

So, for example, in a study of the impact of a mental health worker on the management of depression in primary care, it was reported that "After adjustment for baseline depression, mean depression score was 1.33 PHQ-9 points lower (95% confidence interval 0.35 to 2.31, P=0.009) in participants receiving collaborative care than in those receiving usual care at four months."[4] This means that we estimate that, in the population which these trial participants represent, the average difference in mean depression score if all were offered collaborative care would be between 0.35 and 2.31 scale points less than if all were treated in the usual way. It is only an estimate. For 2.5% of studies the confidence interval will be entirely below the true population difference, and 2.5% will have the interval entirely above it. We don't think "P=0.009" adds much to this, but researchers can seldom bear to do without it. The inevitable uncertainty from sampling error can be reduced by increasing the sample size, but usually only modestly. To halve the width of the confidence interval we would need to quadruple the sample size.

A common mistake is to believe that the confidence interval expresses all the uncertainty. Rather, the confidence interval expressed uncertainty from just one cause—namely the uncertainty due to having taken a sample from the population defined by the inclusion criteria. Often there are other sources of uncertainty that may be even more important to consider, in particular relating to possibly biased results. We address these in our linked statistics note.[5]

Contributors: DGA and JMB jointly wrote and agreed the text.

Correspondence to: D G Altman doug.altman@csm.ox.ac.uk

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; not externally peer reviewed.

1 Sterne JA, Davey Smith G. Sifting the evidence—what's wrong with significance tests? *BMJ* 2001;322:226-31.
2 Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311:485.
3 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ* 1986;292:746-50.
4 Richards DA, Hill JJ, Gask L, Lovell K, Chew-Graham C, Bower P, et al. Clinical effectiveness of collaborative care for depression in UK primary care (CADET): cluster randomised controlled trial. *BMJ* 2013;347:f4913.
5 Altman DG, Bland JM. Uncertainty beyond sampling error. *BMJ* 2014;349:g7065.

Cite this as: *BMJ* 2014;349:g7064