Statistics
with confidence

2nd edition

Edited by

Douglas G Altman
David Machin
Trevor N Bryant
Martin J Gardner

Contents

[4 Means and their differences](#)

[5 Medians and their differences](#)

[6 Proportions and their differences](#)

[7 Epidemiological studies](#)

18 Tables for the calculation of confidence intervals

Index

*This book is dedicated to the memory of
Martin and Linda Gardner.*

STATISTICS WITH CONFIDENCE

Confidence intervals and statistical guidelines

Second edition

Edited by

DOUGLAS G ALTMAN
Director, Imperial Cancer Research Fund Medical Statistics Group and Centre for Statistics in Medicine, Oxford

DAVID MACHIN
Director, National Medical Research Council, Clinical Trials and Epidemiology Research Unit, Singapore

TREVOR N BRYANT
Senior Lecturer in Biocomputation, University of Southampton

and

MARTIN J GARDNER
Former Professor of Medical Statistics, MRC Environmental Epidemiology Unit, University of Southampton

Contributors

Douglas G Altman, *Director, Imperial Cancer Research Fund Medical Statistics Group and Centre for Statistics in Medicine, Institute of Health Sciences, Oxford*

Trevor N Bryant, *Senior Lecturer in Bio computation, Medical Statistics and Computing* (*University of Southampton*), *Southampton General Hospital, Southampton*

Michael J Campbell, *Professor of Medical Statistics, Institute of Primary Care, University of Sheffield, Northern General Hospital, Sheffield*

Leslie E Daly, *Associate Professor of Public Health Medicine and Epidemiology, University College Dublin, Ireland*

Martin J Gardner, *former Professor of Medical Statistics, MRC Environmental Epidemiology Unit* (*University of Southampton*), *Southampton General Hospital, Southampton*

Sheila M Gore, *Senior Medical Statistician, MRC Biostatistics Unit, Cambridge*

David Machin,* *Director, National Medical Research Council Clinical Trials and Epidemiology Research Unit, Ministry of Health, Singapore*

Julie A Morris, *Medical Statistician, Department of Medical Statistics, Withington Hospital, West Didsbury, Manchester*

Robert G Newcombe, *Senior Lecturer in Medical Statistics, University of Wales College of Medicine, Cardiff*

Stuart J Pocock, *Professsor of Medical Statistics, London School of Hygiene and Tropical Medicine, London*

* Now *Professor of Clinical Trials Research, University of Sheffield*

Source of contents

INTRODUCTION Specially written for second edition

PART I Estimation and confidence intervals

1 Gardner MJ, Altman DG. Estimating with confidence. *BMJ* 1988;**296**:1210–1 (revised)

2 Specially written for second edition

3 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ* 1986;**292**:746–50 (revised)

4 From appendix 2 of source reference to chapter 3 (revised and expanded)

5 Campbell MJ, Gardner MJ. Calculating confidence intervals for some non-parametric analyses. *BMJ* 1988;**296**:1454–6 (revised and expanded)

6 Specially written for second edition (replaces chapter 4 of first edition)

7 Morris JA, Gardner MJ. Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *BMJ* 1988;**296**:1313–6 (revised and expanded)

8 Altman DG, Gardner MJ. Calculating confidence intervals for regression and correlation. *BMJ* 1988;**296**:1238–42 (revised and expanded)

9 Machin D, Gardner MJ. Calculating confidence intervals for survival time analyses. *BMJ* 1988;**296**:1369–71 (revised and expanded)

10 Specially written for second edition

11 Specially written for second edition. Includes material from Altman DG. Confidence intervals for the number needed to treat. *BMJ* 1998;**317**:1309–12.

12 Daly LE. Confidence intervals and sample sizes: don't throw out all your old sample size tables. *BMJ* 1991;**302**:333–6 (revised)

13 Specially written for second edition

PART II Statistical guidelines and checklists

14 Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *BMJ* 1983;**286**:1489–93 (revised)

15 Gardner MJ, Machin D, Campbell MJ. Use of checklists in assessing the statistical content of medical studies. *BMJ* 1986;**292**:810–2 (revised and expanded)

PART III Notation, software and tables

16 Specially written for this book. Minor revisions for second edition

17 Specially written for second edition

18 Specially prepared for this book. Minor revisions for second edition

Introduction

DOUGLAS G ALTMAN, DAVID MACHIN, TREVOR N BRYANT

In preparing a new edition of a book, the editors are usually happy in the knowledge that the first edition has been a success. In the current circumstances, this satisfaction is tinged with deep personal regret that Martin Gardner, the originator of the idea for *Statistics with Confidence*, died in 1993 aged just 52. His achievements in a prematurely shortened career were outlined in his obituary in the *BMJ*.[1]

The first edition of *Statistics with Confidence* (1989) was essentially a collection of expository articles concerned with confidence intervals and statistical guidelines that had been published in the *BMJ* over the period 1986 to 1988. All were coauthored by Martin. The other contributors were Douglas Altman, Michael Campbell, Sheila Gore, David Machin, Julie Morris and Stuart Pocock. The whole book was translated into Italian[2] and the statistical guidelines have also appeared in Spanish.[3]

As may be expected, several developments have occurred since the publication of the first edition and Martin had discussed and agreed some of the changes that we have now introduced into this new and expanded edition. Notably, this second edition includes new chapters on Diagnostic tests (chapter 10); Clinical trials and meta-analyses (chapter 11); Confidence intervals and sample sizes (chapter 12); and Special topics (substitution method, exact and mid-P confidence intervals, bootstrap confidence intervals, and multiple comparisons) (chapter 13). There is also a review of the impact of confidence intervals in the

medical literature over the ten years or so since the first edition (chapter 2). All the chapters from the first edition have been revised, some extensively, and one (chapter 6 on proportions) has been completely rewritten. The list of contributors has been extended to include Leslie Daly and Robert Newcombe. We are grateful to readers of the first edition for constructive comments which have assisted us in preparing this revision.

Alongside the first edition of *Statistics with Confidence*, a computer program, Confidence Interval Analysis (CIA), was available. This program, which could carry out the calculations described in the book, had been written by Martin, his son Stephen Gardner and Paul Winter. An entirely new Windows version of CIA has been written by Trevor Bryant to accompany the book, and is packaged with this second edition. It is outlined in chapter 17. The program reflects the changes made for this edition of the book and has been influenced by suggestions from users.

Despite the enhanced coverage we would reiterate the comment in the introduction to the first edition, that this book is not intended as a comprehensive statistical textbook. For further details of statistical methods the reader is referred to other sources.[4–7]

We were all privileged to be colleagues of Martin Gardner. We hope that he would have approved of this new edition of *Statistics with Confidence* and would be pleased to know that he is still associated with it. In 1995 the Royal Statistical Society posthumously awarded Martin the inaugural Bradford Hill medal for his important contributions to medical statistics. The medal was accepted by his widow Linda. As we were completing this second edition in October 1999 we were greatly saddened to learn that Linda too had died from cancer, far too young. We dedicate this book to the memory of both Martin and Linda Gardner.

1 Obituary of MJ Gardner. *BMJ* 1993;**306**:387.

2 Gardner MJ, Altman DG (eds) *Gli intervalli di confidenza. Oltre la significatività statistica.* Rome: II Pensiero Scientifico Editore, 1990.

3 Altman DG, Gore SM, Gardner MJ, Pocock SJ. Normas estadisticas para los colaboradores de revistas de medicina. *Archivos de Bronconeumologia* 1988; **24**:48–56.

4 Altman DG. *Practical statistics for medical research.* London: Chapman & Hall, 1991.

5 Armitage P, Berry G. *Statistical methods in medical research.* 3rd edn. Oxford: Blackwell Science, 1994.

6 Bland M. *An introduction to medical statistics.* 3rd edn. Oxford: Oxford University Press, 2000.

7 Campbell MJ, Machin D. *Medical statistics. A commonsense approach.* 3rd edn. Chichester: John Wiley, 1999.

Part I

Estimation and confidence intervals

1

Estimating with confidence

MARTIN J GARDNER, DOUGLAS G ALTMAN

*Editors' note: this chapter is reproduced from the first edition (*with minor adjustments*). It was closely based on an editorial published in 1988 in the* British Medical Journal. *Chapter 2 describes developments in the use of confidence intervals in the medical literature since 1988.*

Statistical analysis of medical studies is based on the key idea that we make observations on a sample of subjects and then draw inferences about the population of all such subjects from which the sample is drawn. If the study sample is not representative of the population we may well be misled and statistical procedures cannot help. But even a well-designed study can give only an idea of the answer sought because of random variation in the sample. Thus results from a single sample are subject to statistical uncertainty, which is strongly related to the size of the sample. Examples of the statistical analysis of sample data would be calculating the difference between the proportions of patients improving on two treatment regimens or the slope of the regression line relating two variables. These quantities will be imprecise estimates of the values in the overall population, but fortunately the imprecision can itself be estimated and incorporated into the presentation of findings. Presenting study findings directly on the scale of original measurement, together with information on the inherent imprecision due to sampling variability, has distinct advantages over just giving P values usually dichotomised

into "significant" or "non-significant". This is the rationale for using confidence intervals.

The main purpose of confidence intervals is to indicate the (im)precision of the sample study estimates as population values. Consider the following points for example: a difference of 20% between the percentages improving in two groups of 80 patients having treatments A and B was reported, with a 95% confidence interval of 6% to 34% (see chapter 5). Firstly, a possible difference in treatment effectiveness of less than 6% or of more than 34% is not excluded by such values being outside the confidence interval—they are simply less likely than those inside the confidence interval. Secondly, the middle half of the 95% confidence interval (from 13% to 27%) is more likely to contain the population value than the extreme two quarters (6% to 13% and 27% to 34%)—in fact the middle half forms a 67% confidence interval. Thirdly, regardless of the width of the confidence interval, the sample estimate is the best indicator of the population value—in this case a 20% difference in treatment response.

The *British Medical Journal* now expects scientific papers submitted to it to contain confidence intervals when appropriate.[1] It also wants a reduced emphasis on the presentation of P values from hypothesis testing (see chapter 3). *The Lancet*,[3] the *Medical Journal of Australia*,[4] the *American Journal of Public Health*,[5] and the *British Heart Journal*,[6] have implemented the same policy, and it has been endorsed by the International Committee of Medical Journal Editors.[7] One of the blocks to implementing the policy had been that the methods needed to calculate confidence intervals are not readily available in most statistical textbooks. The chapters that follow present appropriate techniques for most common situations. Further articles in the *American Journal of Public Health* and the

Annals of Internal Medicine have debated the uses of confidence intervals and hypothesis tests and discussed the interpretation of confidence intervals.[8–14]

So when should confidence intervals be calculated and presented? Essentially confidence intervals become relevant whenever an inference is to be made from the study results to the wider world. Such an inference will relate to summary, not individual, characteristics—for example, rates, differences in medians, regression coefficients, etc. The calculated interval will give us a range of values within which we can have a chosen confidence of it containing the population value. The most usual degree of confidence presented is 95%, but any suggestion to standardise on 95%[15]

Thus, a single study usually gives an imprecise sample estimate of the overall population value in which we are interested. This imprecision is indicated by the width of the confidence interval: the wider the interval the less the precision. The width depends essentially on three factors. Firstly, the sample size: larger sample sizes will give more precise results with narrower confidence intervals (see chapter 3). In particular, wide confidence intervals emphasise the unreliability of conclusions based on small samples. Secondly, the variability of the characteristic being studied: the less variable it is (between subjects, within subjects, from measurement error, and from other sources) the more precise the sample estimate and the narrower the confidence interval. Thirdly, the degree of confidence required: the more confidence the wider the interval.

1 Langman MJS. Towards estimation and confidence intervals. *BMJ* 1986;**292**:716.

2 Anonymous. Report with confidence [Editorial]. *Lancet* 1987;**i**:488.

3 Bulpitt CJ. Confidence intervals. *Lancet* 1987;**i**:494–7.

4 Berry G. Statistical significance and confidence intervals. *Med J Aust* 1986;**144**:618–19

5 Rothman KJ, Yankauer A. Confidence intervals vs significance tests: quantitative interpretation (Editors' note). *AmJ Public Health* 1986;**76**:587–8.

6 Evans SJW, Mills P, Dawson J. The end of the P value? *By Heart J* 1988;**60**:177–80.

7 International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *BMJ* 1988;**296**:401–5.

8 DeRouen TA, Lachenbruch PA, Clark VA, *et al*. Four comments received on statistical testing and confidence intervals. *Am J Public Health* 1987;**77**:237–8.

9 Anonymous. Four comments received on statistical testing and confidence intervals. *Am J Public Health* 1987;**77**:238.

10 Thompson WD. Statistical criteria in the interpretation of epidemiological data. *Am J Public Health* 1987;**77**:191–4.

11 Thompson WD. On the comparison of effects. *Am J Public Health* 1987;**77**:491–2.

12 Poole C. Beyond the confidence interval. *AmJ Public Health* 1987;**77**:195–9.

13 Poole C. Confidence intervals exclude nothing. *Am J Public Health* 1987;**77**:492–3.

14 Braitman, LE. Confidence intervals extract clinically useful information from data. *Ann Intern Med* 1988;**108**:296–8.

15 Gardner MJ, Altman DG. Using confidence intervals. *Lancet* 1987;**i**:746.

2

Confidence intervals in practice

DOUGLAS G ALTMAN

As noted in chapter 1, confidence intervals are not a modern device, yet their use in medicine (and indeed other scientific areas) was quite unusual until the second half of the 1980s. For some reason in the mid-1980s there was a spate of interest in the topic, with many journals publishing editorials and expository articles (see chapter 1). It seems that several such articles in leading medical journals were particularly influential. Since the first edition of this book there have been many further such publications, often contrasting confidence intervals and significance tests. There has been a continuing increase in the use of confidence intervals in medical research papers, although some medical specialties seem somewhat slower to move in this direction. This chapter briefly summarises some of this literature.

Surveys of the use of confidence intervals in medical journals

There is a long tradition of reviewing the statistical content of medical journals, and several recent reviews have included the use of confidence intervals. Of particular interest is a review of the use of statistics in papers in the *British Medical Journal* in 1977 and 1994, before and after it adopted its policy of requiring authors to use confidence intervals.[1] One of the most marked increases was in the use of confidence intervals, which had risen from 4% to 62% of papers using some statistical technique, a large increase but still well short of that required. Similarly, between 1980 and 1990 the use of confidence intervals in the *American Journal of Epidemiology* approximately doubled to 70%, and it was around 90% in the subset of papers related to cancer, [2] despite a lack of editorial directive.[3] This review also illustrated a wider phenomenon, that the increased use of confidence intervals was not so much instead of P values but as a supplement to them.[2]

The uptake of confidence intervals has not been equal throughout medicine. A review of papers published in the *American Journal of Physiology* in 1996 found that out of 370 papers only one reported confidence intervals![4] They were presented in just 16% of 100 papers in two radiology journals in 1993 compared with 52% of 50 concurrent papers in the *British Medical Journal*.[5]

Confidence intervals may also be uncommon in certain contexts. For example, they were used in only 2 of 112 articles in anaesthesia journals (in 1991–92) in conjunction with analyses of data from visual analogue scales.[6]

Editorials and expository articles

Editorials[7-19] and expository articles[20-31] related to confidence intervals have continued to appear in medical journals, some being quite lengthy and detailed. In effect, the authors have almost all favoured greater use of confidence intervals and reduced use of P values (a few exceptions are discussed below). Many of these papers have contrasted estimation and confidence intervals with significance tests and P values.

Such articles seem to have become rarer in the second half of the 1990s, which may indicate that confidence intervals are now routinely included in introductory statistics courses, that there is a wide belief that this particular battle has been won, or that their use is so widespread that researchers use them to conform. Probably all of these are true to some degree.

Medical journal policy

As noted in chapter 1, when the first edition of this book was published in 1989, a few medical journals had begun to include some mention of confidence intervals in their instructions to authors. In 1988 the influential 'Vancouver guidelines'[32] (originally published in 1979) included the following passage:

> Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results. When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as the use of P values, which fails to convey important quantitative information.

This passage has survived intact to May 1999 apart from one trivial rewording.[33] The comment on confidence

intervals is, however, very brief and rather nebulous. In 1988 Bailar and Mosteller published a helpful amplification of the Vancouver section,[34] but this article is not cited in recent versions of the guidelines. Over 500 medical journals have agreed to use the Vancouver requirements in their instructions to authors.[33]

Despite the continuing flow of editorials in medical journals in favour of greater use of confidence intervals,[7–19] it is clear that the uptake of this advice has been patchy, as illustrated by reviews of published papers and also journals' instructions to authors. In 1993, I reviewed the 'Instructions to Authors' of 135 journals, chosen to have high impact factors within their specialties. Only 19 (14%) mentioned confidence intervals explicitly in their instructions for authors, although about half made some mention of the Vancouver guidelines. Journals' instructions to authors change frequently, and not necessarily in the anticipated direction. Statistical guidelines published (anonymously) in 1993 in *Diabetic Medicine* included the following: 'Confidence intervals should be used to indicate the precision of estimated effects and differences'.[35] At the same time they published an editorial stating *'Diabetic Medicine* is now requesting the use of confidence intervals wherever possible'.[14] These two publications are not referenced in the 1999 guidelines, however, and there is no explicit mention of confidence intervals, although there is a reference to the Vancouver guidelines.[36]

Kenneth Rothman was an early advocate of confidence intervals in medical papers.[37] In 1986 he wrote: 'Testing for significance continues today not on its merits as a methodological tool but on the momentum of tradition. Rather than serving as a thinker's tool, it has become for some a clumsy substitute for thought, subverting what

should be a contemplative exercise into an algorithm prone to error.'[38] Subsequently, as editor of *Epidemiology*, he has gone further:[39]

> When writing for *Epidemiology*, you can also enhance your prospects if you omit tests of statistical significance. Despite a widespread belief that many journals require significance tests for publication, the Uniform Requirements for Manuscripts Submitted to Biomedical Journals discourages them, and every worthwhile journal will accept papers that omit them entirely. In *Epidemiology*, we do not publish them at all. Not only do we eschew publishing claims of the presence or absence of statistical significance, we discourage the use of this type of thinking in the data analysis, such as in the use of stepwise regression.

Curiously, this information is not given in the journal's 'Guidelines for Contributors' (http://www.epidem.com/), perhaps reflecting the slightly softer position of a 1997 editorial: 'it would be too dogmatic simply to ban the reporting of all P-values from *Epidemiology*.'[40] Despite widespread encouragement to include confidence intervals, I am unaware of any other medical journal which has taken such a strong stance against P values.

A relevant issue is the inclusion of confidence intervals in abstracts of papers. Many commentators have noted that the abstract is the most read part of a paper,[41] yet it is clear that it is the part that receives the least attention by authors, and perhaps also by editors. A few journals explicitly state in their instructions that abstracts should include confidence intervals. However, confidence intervals are often not included in the abstracts of papers even in journals which have signed up to guidelines requiring such presentation.[42,43]

Misuse of confidence intervals

The most obvious example of the misuse of confidence intervals is the presentation in a comparative study of separate confidence intervals for each group rather than a confidence interval for the contrast, as is recommended (chapter 14). This practice leads to inferences based on whether the two separate confidence intervals, such as for the means in each group, overlap or not. This is not the appropriate comparison and may mislead (see chapters 3 and 11). Of 100 consecutive papers (excluding randomised trials) that I refereed for the *British Medical Journal*, 8 papers out of the 59 (14%) which used confidence intervals used them inappropriately.[44]

The use for small samples of statistical methods intended for large samples can cause problems. In particular, confidence intervals for quantities constrained between limits should not include values outside the range of possible values for the quantities concerned. For example, the confidence interval for a proportion should not go outside the range 0 to 1 (or 0% to 100%) (see chapters 6 and 10). Quoted confidence intervals which include impossible values – such as the sensitivity of a diagnostic test greater than 100%, the area under the ROC curve greater than 1, and negative values of the odds ratio – should not be accepted by journals.[45,46]

One criticism of confidence intervals as used is that many researchers seem concerned only with whether the confidence interval includes the 'null' value representing no difference between the groups. Confidence intervals wholly to one side of the no effect point are deemed to indicate a significant result. This practice, which is based on a correct link between confidence interval and the P value, is indeed

common. But even if the author of a paper acts in this way, by presenting the confidence interval they give readers the opportunity to take a different and more informative interpretation. When results are presented simply as P values, this option is unavailable.

Dissenting voices

It is clear that there is a considerable consensus among statisticians that confidence intervals represent a far better approach to the presentation and interpretation of results than significance tests and P values. Apart from those, mostly statisticians, who criticise all frequentist approaches to statistical inference (usually in favour of Bayesian methods), there seem to have been very few who have spoken out against the general view that confidence intervals are a much better way to present results than P values.

In a short editorial in the *Journal of Obstetrics and Gynecology*, the editor attacked several targets including confidence intervals.[47] He expressed the unshakeable view that only positive results (P < 0.05) indicate important findings, and suggested that 'The adoption of the [confidence interval] approach has already enabled the publication in full of many large but inconclusive studies … '

Charlton[48] argued that confidence intervals do not provide information of any value to clinicians. In fact, he criticised confidence intervals for not doing something which they do not purport to do, namely indicate the variation in response for individual patients.

Hilden[49] cautioned that confidence intervals should not be presented 'when there are major threats to accuracy besides sampling error; or when a characteristic is too local

and study-dependent to be generalizable'. Hall[50] took this line of reasoning further, arguing that confidence intervals 'should be used sparingly, if at all' when presenting the results of clinical trials. He also argued, contrary to the common view, that they might be particularly misleading 'when a clinical trial has failed to produce anticipated results'. His reasoning was that patients in a trial are not a random sample and thus the results cannot be generalised, and also that 'a clinical trial is designed to confirm expectation of treatment efficacy by rejecting the null hypothesis that differences are due to chance'. He went further, and suggested that 'there are few, if any, situations in which a confidence interval proves useful'. This line of reasoning has a rational basis, but he has taken it to unreasonable extremes. Other articles in the same journal issue[51,52] presented a more mainstream view.

It is interesting that there is no consensus among this small group of critics about what are the failings of confidence intervals. It is right to observe that we should always think carefully about the appropriate use and interpretation of *all* statistics, but it is wrong to suggest that all confidence intervals are meaningless or misleading.

Comment

Like many innovations, it is hard now to imagine the medical literature without confidence intervals. Overall, this is surely a development of great value, not least for the associated downplaying (but by no means elimination) of the wide use of $P < 0.05$ or $P > 0.05$ as a rule for interpreting study findings. However, as noted, confidence intervals can be both misused and overused and there are arguments in favour of other approaches to statistical inference. Also, despite a large increase in the use of confidence intervals,

even in those journals which require confidence intervals – such as the *British Medical Journal* – their use is not widespread, and in some fields, such as physiology and psychology, their use remains uncommon.

Confidence intervals are especially valuable to aid the interpretation of clinical trials and meta-analyses[53] (see chapter 11). In cases where the estimated treatment effect is small the confidence interval indicates where clinically valuable treatment benefit remains plausible in the light of the data, and may help to avoid mistaking lack of evidence of effectiveness with evidence of lack of effectiveness.[54] The CONSORT statement[43] for reporting randomised trials requires confidence intervals, as does the QUOROM statement[55] for reporting systematic reviews and meta-analyses (see chapters 11 and 15).

None of this is meant to imply that confidence intervals offer a cure for all the problems associated with significance testing and P values, as several observers have noted.[56,57] We should certainly expect continuing developments in thinking about statistical inference.[58–61]

1 Seldrup J. Whatever happened to the *t*-test? *Drug Inf J* 1997;**31**:745–50.

2 Savitz DA, Tolo K-A, Poole C. Statistical significance testing in the *American Journal of Epidemiology*, 1970–1990. *AmJ Epidemiol* 1994;**139**:1047–52.

3 Walter SD. Methods of reporting statistical results from medical research studies. *AmJ Epidemiol* 1995;**141**:896–908.

4 Curran-Everett D, Taylor S, Kafadar K. Fundamental concepts in statistics: elucidation and illustration. *J Appl Physiol* 1998;**85**:775–86.

5 Cozens NJA. Should we have confidence intervals in radiology papers? *Clin Radiol* 1994;**49**:199–201.