



## Confidence intervals, missing data and imputation: a salutary illustration

Stephen Gorard

School of Education, Durham University

s.a.c.gorard@durham.ac.uk

### ABSTRACT

This paper confirms that confidence intervals are not a generally useful measure or estimate of anything in practice. CIs are recursive in definition and reversed in logic, meaning that they are widely misunderstood. Perhaps most importantly, they should not be used with cases that do not form a complete and true random sample from a known population – the latter is a key premise underlying their calculation. This means that, whatever their merits, CIs should not be used in the vast majority of real-life social science analyses. The second part of the paper illustrates the dangers of ignoring this premise, perhaps on some purported pragmatic grounds. Using 100 simulations of a sample of 100 integers from a uniform population with members in the range 0 to 9, it shows that CIs are very misleading as soon as there is deviation from randomness. For example, when 5% of the cases in each sample are deleted a reported 95% CI would be no better than a 66% CI in reality. If 10% of the lowest score cases are replaced with the achieved mean for the sample, then a reported 95% CI would be more like a 43% CI in reality. In addition, the simulation shows that the mean and standard deviation for any sample are correlated (an issue of linked scale). This illustrates that using the sample standard deviation as an estimate for the SD of the sampling distribution in order to try and assess whether the sample mean is close to the mean of the sampling distribution will simply make matters worse. The best and only available estimate of the sampling distribution mean, in practice, is the sample mean.

### Keywords

Confidence intervals, credible intervals, attrition, missing data

### Academic sub-disciplines

Education evaluation, statistical analysis, formatting of results

### Subject classification

Evaluation

### Type

Simulation, innovation, critique

# Council for Innovative Research

Peer Review Research Publishing System

**Journal:** INTERNATIONAL JOURNAL OF RESEARCH IN EDUCATION  
METHODOLOGY

Vol 5, No.3

[www.cirworld.org/journals](http://www.cirworld.org/journals), [ijremeditor@gmail.com](mailto:ijremeditor@gmail.com)



## INTRODUCTION

“A confidence interval calculated for a measure of treatment effect shows the range within which the true treatment effect is likely to lie” according to [http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/what\\_are\\_conf\\_inter.pdf](http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/what_are_conf_inter.pdf), accessed 28/3/14

Richard Feynman is supposed to have joked ‘if you think you understand quantum mechanics, you don’t...’. Something similar, although at a lower level of comprehension, could be said about confidence intervals. Those who appear most certain about the interpretation of confidence intervals (CIs) are generally those who have the weakest (or indeed no) explanation of it [1]. The difficulties stem from CIs being recursive in the sense that the definition includes itself, and because the logic is *modus tollens* or the reverse of the approach expected in everyday discourse [2]. These and other difficulties of comprehension then lead to the common errors of using CIs as though they could handle non-random cases, and of interpreting a CI as a range of likelihood within which a desired parameter will fall. Neither is true [3].

The paper starts by describing what a confidence interval is, and what it is not. It reminds reader why a CI is not a valid measure of uncertainty when the cases involved are not randomly sampled, and why the utility of a CI is to be doubted even when the cases are randomly sampled. The paper then describes the methods and findings of a simple simulation, before summing up what can be learnt about the dangers of using confidence intervals when any cases are missing or biased.

## WHAT IS A CONFIDENCE INTERVAL?

A confidence interval is widely used as a measure of uncertainty around an estimate for a population parameter based on a sample measurement. For example, a researcher may collect measurements from a random sample of cases and report their mean score as 5.3 with a 95% CI of 4.5 to 6.1. What does this signify and how useful is it?

Given a known set of measurements for a population, it is possible to draw a random sample of cases from that population. The mean and standard deviation of the measurements for the random sample may or may not be similar to those in the set of population measurements from which they have been drawn. A simple comparison of the means and standard deviations for the population and sample would reveal how close they are.

If a large number of samples of the same fixed size were taken from the same population, there would be a sampling distribution. The means of the numerous samples would be approximately normally distributed around the overall mean of this sampling distribution. As the number of samples approaches infinity, the distribution of their means would tend towards perfectly normal, and the overall mean of the sampling distribution would tend towards the mean of the population from which this huge number of samples has been taken. It is known that 95% of the area under a normal curve lies within 1.96 standard deviations (either way) from its mean. This implies that 95% of the repeated random sample means will lie within 1.96 standard errors of the sampling distribution mean, where the standard error is the standard deviation of the population divided by the square root of the fixed number of cases in each random sample. It then follows that 95% of the time the mean of the sampling distribution will lie within 1.96 standard errors (for the sampling distribution) of the mean of any of the fixed size samples. Each sample mean can then be said to have a 95% ‘confidence interval’ of itself plus or minus 1.96 of the standard error for the sampling distribution.

In practice an analyst will have only one sample mean available. The 95% is not the likelihood of the population mean being within the 95% CI. It is the likelihood of generating a CI containing the mean based on numerous repeated samples. Nevertheless, if the mean of the sampling distribution were identical to the mean of the population this might give consumers of the result some (rather strange because recursive and logically backwards) idea of the quality of the sample mean. In practice however, it is very rare to have a true and complete random sample. When working with population data, convenience and snowball samples, CIs are irrelevant (because there would be no standard error, of course). The same is true when working with incomplete random samples – i.e. samples that were designed to be random but which are no longer random because there has been at least one case of non-response or dropout. Even a full random sample in social science is most unlikely to be a true random one since cases are rarely replaced after selection, and so the artificial lack of any possibility of duplication of cases distorts the probabilities then cited. So, none of the above explanation of CIs is relevant for most pieces of real-life research.

Anyway and in practice, even for an analyst in the unlikely position of working with a real random sample, the calculation above is still impossible. The mean of the sampling distribution is not necessarily identical to the population mean unless the number of samples taken is infinite, which is literally impossible. There would be no point in calculating a CI for a sample mean when the population mean, or even the sampling distribution mean, was already known. This is because if it was known then the precise distance from the sample mean to the population could be calculated by simple arithmetic. A CI would be absurd in such a context. If the population (or the sampling distribution) mean is not known then it is not possible to know the standard deviation of either the population or the sampling distribution, since calculation of the standard deviation involves summing the squares of the deviation of each population measure from that population mean. It then follows that the standard error for the sampling distribution can never be calculated in real-life since this involves dividing the population standard deviation by the square root of the fixed number of cases in each sample. Without knowing the population parameters, or at least those for the sampling distribution, there is no way of knowing the standard error. But whenever we can know the actual standard error we do not need it, by definition.



In practice, the mean and standard deviation of the population are not known, and there is no known sampling distribution. Only the mean and standard deviation for the sample are really available. Common practice is that the standard deviation around the sample mean is used instead of that for the population in order to estimate the standard error of the unknown sampling distribution. In practice therefore, a CI is calculated from the sample mean and the sample standard deviation *only*. This does not mean that the true population is near the centre of the CI (or even in it). It is only the sample mean that defines the centre of the CI. A CI is merely the sample mean and sample standard deviation aggregated in a complex way, and then writ large. This has the perverse result that a CI calculated for a sample from one population would be the same as that for the same sample drawn from a completely different population. For example, the CI around the probability of drawing 10 red balls at random from a bag containing 50 red and 50 blue balls is exactly the same as the CI around the probability of drawing 10 red balls at random from a bag containing 90 red and 10 blue balls. The CI says nothing about the actual population from which the sample was drawn, and therefore gives no idea of uncertainty in the sampling (other than the scale). It is, for all practical purposes for which it is intended, useless.

The true meaning of a sample CI as calculated in practice is that if many other samples of the same size had been created many times and their CIs calculated then around 95% of these other samples would include the mean of the sampling distribution (and so perhaps the mean of the population). This is a recursive definition because it uses a very large number of CIs to define any one CI. It is also reverse logic since it does not say that the distribution mean is 95% likely to be within the achieved sample CI. A CI is a statement about repeated experiments that have never occurred and presumably never will (and the statement is thereby argued to contravene the likelihood principle [4]). A CI is a result generated by a procedure that will give valid CIs around 95% of the time. This is not a great help when faced with only one sample CI. This one sample CI either does or does not include the unknown population mean [5]. But then the same tautology could be said about any interval at all – it either does or does not include the value of interest.

An analogy might be being faced with a biased coin that gives more of one result when tossed (either head or tail) and less of the other result (either head or tail). This much is known, but it is not known in which direction the coin is biased. An analyst is faced with only one trial and has one result or sample – perhaps it is a head. This analyst would be best advised to simply use the achieved result as the best estimate of the overall bias, and declare that the coin is biased towards heads. If the one sample result (heads) is the right one, then the conclusion (that the coin is biased towards heads) will be correct. If the one sample result is not the common result (the coin is really biased towards tails) then the analyst will be wrong. In the same way, given a one sample mean this will be the best available estimate of the population mean. If the analyst assumes that the sample mean is a good estimate, they may be right or wrong. Creating a 'confidence' interval around the mean, as is common practice, changes nothing in that assumption. The interval is not somehow a probability that the analyst is right (or wrong). It cannot be because it is based entirely on the sample. It is not a Bayesian 'credible' interval, for example. Credible intervals generally do not coincide with so-called confidence intervals because the former uses prior knowledge, and problem-specific contextual information, of the kind that mainstream statistics simply ignores [6].

The widespread misunderstanding created by CIs is a standard logical fallacy usually illustrated as – all Greeks are human therefore all humans are Greeks – which only works if Greeks are the only humans (i.e. a tautology). So the 'logic' says that in order to estimate how close the sample mean is to the unknown population mean we must assume that the sample standard deviation is close to the unknown population standard deviation. But if the sample mean can be a poor estimate of the population mean then the sample standard deviation can also be a poor estimate of the population standard deviation. And if the sample standard deviation is a poor estimate, whether the sample mean is a itself poor estimate or not, then it follows that any CI based on the sample standard deviation will be a poor estimate of whether the sample mean is a poor estimate of the population mean. And even if the sample standard deviation is a reasonable (order of magnitude) estimate of the population standard deviation, it is not true that exactly 95% of the area of the population will lie within 1.96 sample standard deviations of the population mean (nor even of the sample mean since the sample itself may not be normally distributed). This means that the 95% figure will, in practice, always be quoted incorrectly, and that the accuracy of the CI limits (or interval) will always be spurious unless quoted to so few significant figures as to be practically worthless anyway.

CIs are a way of re-writing significance tests and p-values [7]. And significance testing and p-values are easily misunderstood, give misleading results about the substantive nature of results, and are 'best avoided' [8]. CIs share the logical difficulties of p-values, by referring to the probability of the data achieved given certain assumptions about the population parameters. What analysts actually want to know is the probability of the assumption about the population parameter(s) being true given the data achieved. To create the latter from the former via Bayes' Theorem would require knowledge of the parameters for the sampling distribution (or ideally the actual population). But, as already explained, this knowledge does not exist in practice. So, calculating the CI for any one sample does not provide a likelihood for whether that one sample mean is, or is not, near the population mean. CIs also share with the flawed approach of significance testing the problem that they do not and cannot address the most important analytical questions about any sample results. Systematic biases in design, measurement or sampling are generally more substantively important than random variation in deciding on the trustworthiness of any research results. For example, a confidence interval cannot distinguish between a sample of 100 cases with a 50% response rate and a random sample of 100 cases with a 100% response rate. Both will be assessed by this peculiar technique as providing equal 'confidence'. Clearly though the latter is far superior in quality and therefore the level of trust that can be placed in it (a true concept of confidence). Similarly, all other things being equal, the CI for a sample of 200 cases with a response rate of 20% (i.e. 1,000 cases were approached) will be reported as superior to the CI for a random sample of 100 cases with a 100% response rate. In both examples, CIs should not be used since at least one of the samples is no longer random at all. The non-response cannot be assumed to be random, and careful research has shown that non-responding cases are not a random sub-set of all others. Their mere existence





and occurrence creates bias. But the kind of error, misunderstanding and mis-representation that uses CIs to estimate uncertainty in findings is widespread. It is also dangerous [9, 10]. The next section shows why the use of CIs with incomplete data is so misleading.

## METHODS USED IN SIMULATION/ILLUSTRATION

The illustration used in this paper is based on 100 simulations run in Excel. Each simulation involved creating a sample of 100 random integers in the range 0 to 9, with a uniform distribution (see Figure 1).

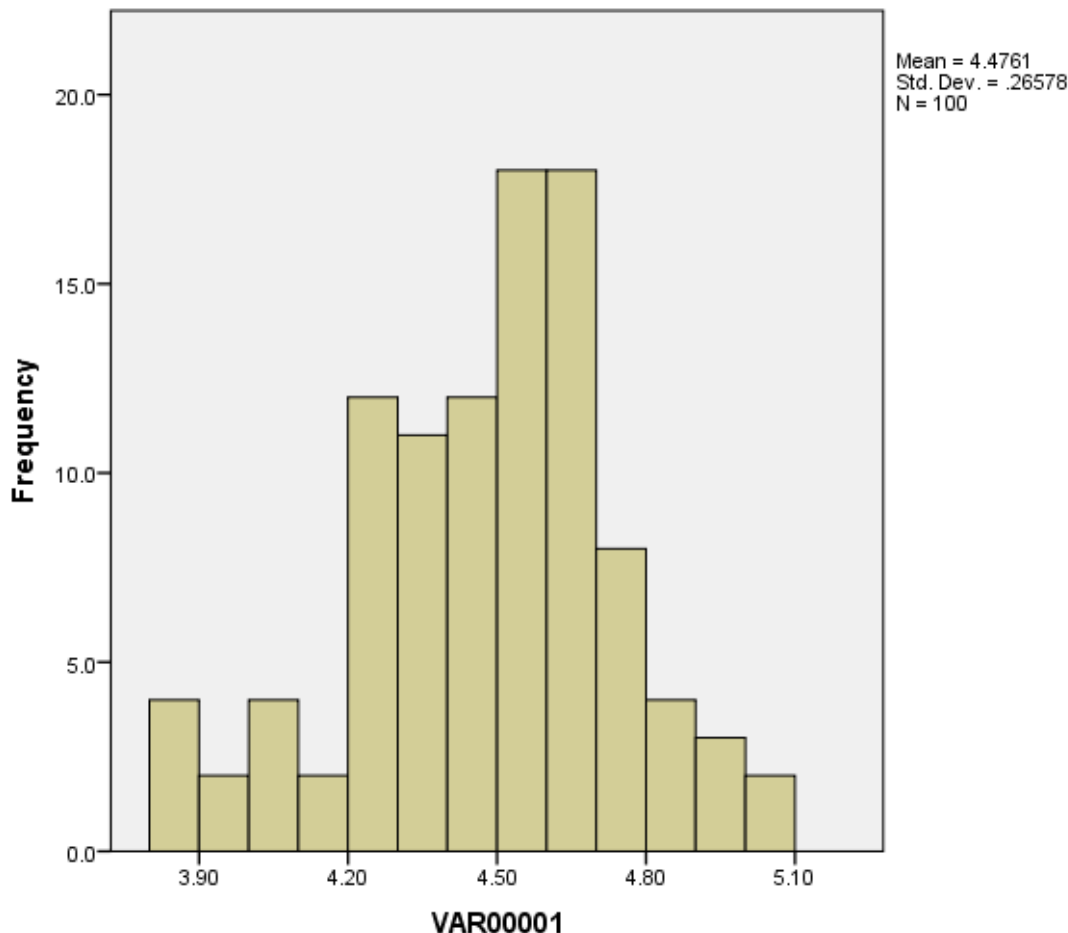


Figure 1 – Distribution of means of 100 samples of 100 random integers (0-9)

The means, standard deviations, and 95% confidence intervals for the means for each sample, and the Pearson's R correlation between means and standard deviations for all samples, were calculated. Then, the 5% lowest scores in all samples were simply deleted and the CIs recalculated, then the same 5% lowest scores were replaced by the achieved mean in each sample (with the mean calculated after the lowest scores had been deleted). Finally the same things were done with the 10% lowest scores in all samples. This was to simulate the likely impact of biased missing data. At each stage CIs were recalculated.

## THE IMPACT OF 'MISSING' DATA

The 100 sample means varied from 3.86 to 5.3, with standard deviations ranging from 2.84 to 3.14. The first finding of note is the level of correlation between the samples means and their standard deviations ( $R=+0.2$ ). This implies that when the sample mean is not close to the true population mean then there is a tendency for the sample standard deviation also to be further from the true population standard deviation. And this matters because, as explained above, neither the true mean nor the true standard deviation is known when calculating a CI in practice. The sample standard deviation is therefore used to form an estimate for the standard error of the sampling distribution, in order to help assess how good an estimate the sample mean is for the mean of the sampling distribution. But if the sample mean and SD are correlated then a poor estimate of the population mean is in more danger of being condoned by the relatedly poor estimate of the population SD for the same sample. But all of this will be invisible in practice when an analyst is faced with only one sample, and wants to use the sample standard deviation to help assess how good the mean is as an estimate for the population. It cannot be done.



The 100 calculated confidence intervals ranged from 3.27-4.45 to 4.81-5.79 (neither of which extreme CIs contained the true population mean of 4.5). In total six of the 100 samples had CIs not containing the population mean. This is close to 95%, but is of course based on ideal random numbers rather than the more realistic datasets with more extreme scores found in practice [11, 12].

When 5% of the lowest integers (552 random zeroes out of 10,000 cases) were deleted from all samples, 24 of the resultant CIs did not contain the true mean. This shows how ineffective and misleading CIs would be in practice, if used incorrectly with what are now no longer random samples. What would be reported by any analyst making this kind of mistake as a 95% CI could be a 76% CI or worse. When 5% of the lowest integers were replaced with the achieved sample mean for each sample, 34 of the resultant CIs did not contain the true mean. This again shows how ineffective CIs would be, and how far they are from anything like 95% accurate in real-life. This result also shows that the common practice of trying to replace missing data on the basis of the data that was obtained is worse than useless. This practice generally exacerbates the bias caused by missing data, and should cease [13].

When 10% of the original lowest integers (985 zeroes out of 10,000 cases) were deleted from all samples, the situation was even worse of course. Now, 41 of the recalculated CIs did not contain the true mean. And, when the 10% missing data was replaced by the achieved mean for the sample, 57 of the purported 95% confidence intervals did not include the mean of the sampling distribution. This means that an analyst reporting a 95% CI for a sample with 90% response rate could actually be citing a 43% CI or worse. Confidence intervals take no account of bias or missing data. And this bias or potential bias is a far bigger threat to the security of findings than random sampling variation (which is only relevant when a full and true random sample is available – in practice, never).

## CONCLUSIONS

In general, a larger sample will tend to produce a more accurate estimate for the mean than a smaller one. Other than this, an analyst faced with deciding whether a sample mean is a good representation of the population mean only has the sample mean and sample standard deviation as a guide. These could well be correlated meaning that using the sample standard deviation to try and assess the quality of the sample mean becomes even more problematic. There is no way of using these to assess how close the one sample mean is to the unknown population mean. This is why confidence intervals are confusing. They appear to offer something that they do not and cannot.

Even if they were useful and well-understood, confidence intervals were only ever intended to be used with complete random samples – a very rare phenomenon. Any deviation from randomness, such as the usual levels of attrition in social science, makes even the valid initial mathematical argument about CIs for populations irrelevant. CIs cannot address bias, systematic measurement error, attrition or any of the other manifold threats to the validity of a study. Yet they are being routinely used in just this way with non-random cases, and as though they covered issues like attrition and sample quality. This is leading to widespread errors in analysis and reporting. One example of the problem lies in the traditional use of forest plots for research syntheses [14]. An up-to date description is:

“A vertical line representing no effect is also plotted. If the confidence intervals for individual studies overlap with this line, it demonstrates that at the given level of confidence their effect sizes do not differ from no effect for the individual study. The same applies for the meta-analysed measure of effect: if the points of the diamond overlap the line of no effect the overall meta-analysed result cannot be said to differ from no effect at the given level of confidence.”

[http://en.wikipedia.org/wiki/Forest\\_plot](http://en.wikipedia.org/wiki/Forest_plot), accessed 28/3/14

There are the same recurring two problems with this account and usage. The one trial CIs do not provide a precise level of confidence that each trial has a result different from zero. Perhaps most importantly, these CIs do not take account of sample quality. Some readers may think that the 5% and 10% attrition rates used in the simulation for this paper are rather large. However, 90% and 95% completion rates are actually very impressive and rather rare in social science. Papers using forest plots often portray studies with completion rates as low as 75% or even 70%. The What Works Clearinghouse ([http://ies.ed.gov/ncee/wwc/pdf/reference\\_resources/wwc\\_procedures\\_v3\\_0\\_draft\\_standards\\_handbook.pdf](http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_draft_standards_handbook.pdf), accessed 28/3/14) describes up to 29% post-allocation dropout in any sample as ‘low’ attrition. Therefore, as little as 71% can be high completion. The WWC standard is merely concerned with the disparity in rates of attrition between groups within the overall sample. The assumption is that if this attrition is reasonably even in quantity then no bias has occurred. This assumption is clearly wrong, and 29% is an absurdly high level of attrition to tolerate. It is easy to envisage dropout that was equal in scale between two groups but very different in character. For example, a rather low-level education intervention may appear patronising to high attaining students. Thus, dropout in the treatment group of a randomised controlled trial could be higher for high attainers. This is the kind of bias that the simulation was trying to assess in terms of CIs. The dropout from the control group might be equal in number but either neutral in terms of attainment, or the lowest attainers might be the more demoralised by not receiving the treatment after allocation becomes clear. Either way the dropout will be biased. In either situation, as well as being hard to interpret and technically unjustified, a confidence interval will be entirely inaccurate. This paper has shown that as little as 10% biased dropout means that over half of all possible CIs will be entirely misleading. It is difficult to imagine how misleading a CI would be when a sample had 29% dropout. The practice of using CIs in this way should cease.



## REFERENCES

1. Gorard, S. (2014) The widespread abuse of statistics by researchers: what is the problem and what is the ethical way forward?, *Psychology of Education Review*, 38, 1, 3-10
2. Gorard, S. (2010) All evidence is equal: the flaw in statistical reasoning, *Oxford Review of Education*, 36, 1, 63-77
3. Watts, D. (1991) Why is introductory statistics difficult to learn?, *The American Statistician*, 45, 4, 290-291
4. Lindley, D. and Phillips, L. (1976) Inference for a Bernoulli process, *The American Statistician*, 30, 3, 112-119
5. Jaynes, E. (1976) Confidence intervals vs Bayesian intervals, in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, The University of Western Ontario Series in Philosophy of Science Volume 6b, pp. 175-257
6. Jaynes, E. (1976) Confidence intervals vs Bayesian intervals, in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, The University of Western Ontario Series in Philosophy of Science Volume 6b, pp. 175-257
7. Coe, R. (2002) It's *the effect size, stupid*, presentation to British Annual Research Association Conference, Exeter, September 2002
8. Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K. and Busick, M. (2012) *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*, Washington DC: Institute of Education Sciences, p.13
9. Carver, R. (1978) The case against statistical significance testing, *Harvard Educational Review*, 48, 378-399
10. Matthews, R. (1998) *Bayesian Critique of Statistics in Health: The great health hoax*, <http://www2.isye.gatech.edu/~brani/isyebayes/bank/pvalue.pdf>
11. Barnett, V. and Lewis, T. (1978) *Outliers in statistical data*, Chichester: John Wiley and Sons
12. Huber, P. (1981) *Robust Statistics*, New York: John Wiley and Sons
13. Gorard, S. (2013) *Research Design: Robust approaches for the social sciences*, London: SAGE
14. Lewis, S. and Clarke, M. (2001) Forest plots: trying to see the wood and the trees, *BMJ* 2001; 322 doi: <http://dx.doi.org/10.1136/bmj.322.7300.1479>



## AUTHOR BIOGRAPHY

**Stephen Gorard** is Professor of Education and Public Policy, and Fellow of the Wolfson Research Institute, at Durham University. He is a Methods Expert for the US government Institute of Education Science, member of the ESRC Grants Awarding Panel, and Academician of the Academy of Social Sciences. He is author of nearly 1,000 books and papers.