CONFIDENCE INTERVALS AND REPLICATION: THE REALITY

Ian R. Gordon

Statistical Consulting Centre and School of Mathematics and Statistics
The University of Melbourne, Victoria, Australia 3010
irg@unimelb.edu.au

In the climate of 'replication crises' across many disciplines, but particularly psychology, there has been a focus on the reproducibility of statistical inferences in various forms. Confidence intervals and the perspective of estimation have long been regarded as more insightful than P-values and hypothesis testing, for representing simple statistical inferences. It has been argued that P-values should be abandoned altogether. Two recent arguments in favour of this proposition are that confidence intervals have (1) superior "replication" to, and (2) less variation than, P-values. I discuss these arguments and show they are unsound; a different statistical perspective is presented.

BACKGROUND

Since the 1970s at least, there has been a call for the use of confidence intervals in preference to P-values, when representing statistical inference on a single parameter. This began in the medical and epidemiological field (for example, Rothman (1978), Berry (1986), Gardner and Altman (1986)). Journals took action along these lines: editorial advice to authors of the *Medical Journal of Australia* strongly encouraged authors "… to express their main conclusions in confidence interval form" (Berry, 1986), and in the 1990s *Epidemiology* discouraged the reporting of P-values, without banning them altogether (Lang et al., 1998). The result of this decades-long campaign was assessed by Stang et al. (2017), who concluded that in major medical journals there has been an observable trend towards confidence intervals and away from hypothesis testing.

This debate and recommendations about practice will be familiar to anyone who has engaged even slightly in matters of basic statistical education, and, in particular, the communication of statistical inferences. It is an important issue because of the ubiquitous need to present inferences, across all areas of quantitative research.

Many contributions have been made to the consideration of this question. Some focus on statistical aspects; there has also been attention given to how different approaches are understood. One way that the competition between P-values and confidence intervals has been framed is to examine the properties of both, with a view to demonstrating that one is better than the other, because of features taken to be important and desirable. In several publications, Geoff Cumming has argued for the abandonment of hypothesis testing and P-values. His argument is multi-faceted, and several elements of the debate have been studied; see Cumming et al. (2004); Cumming and Maillardet (2006); Cumming (2008, 2012) and Cumming and Calin-Jageman (2017).

This is a substantial body of work. I agree with much of the general thrust of it, and it is not the purpose of this paper to discuss and review it all; rather, two of Cumming's main ideas are assessed and questioned. This should not be taken – at all – as a defense of P-values and a case for their use in favour of confidence intervals. However, it is argued here that two key arguments advanced by Cumming are unsound. This matters: if readers become aware that unsound reasons are being used to support a recommended practice, they may legitimately doubt the force of the recommendation. Further, the two matters considered here involve the fundamentals of inference, and if misleading and wrong arguments in this area are promoted, there is an undesirable risk of misinformation and confusion, something that statistical educators care about deeply.

The two key concepts discussed here, which frame the debate, are *replication* and *variation*. The best encapsulation of Cumming's position is this statement:

"I am especially interested in replication … One of the many reasons that confidence intervals are better than P-values is that confidence intervals give quite good information about what is likely to

happen in replication of an experiment, whereas a P-value gives almost no information about replication. The dance of the P-values illustrates how P-values vary enormously with replication, thus indicating how terribly uninformative they are." (Cumming, 2017b).

REPLICATION FOR P-VALUES AND CONFIDENCE INTERVALS

While the concept of replication has received attention in many contexts more generally, Cumming introduced attempts to consider the "replication" of P-values and confidence intervals in a more specific way, in Cumming et al. (2004) and Cumming and Maillardet (2006).

The general context of the paper by Cumming (2008) is a debate concerning the merits of different representations of statistical inference; in particular, inferences about a single parameter. Under repeated sampling using the same statistical framework (sample size and so on), he considers the likely pattern of P-values from subsequent studies, following an initial study, and defines this to be the "replication" of the P-values. He introduces a P interval and asserts that it is "surprisingly" wide. Most of the paper is devoted to this point, but a key element of Cumming's (2008) argument is that – in contrast to the claimed poor "replication" performance of P-values – confidence intervals perform better.

In this often-heated debate, sound statistical arguments are needed, and here it is suggested that Cumming's (2008) article contains a fallacious argument regarding the "replication" of confidence intervals. The point is asserted in the title of the paper: ". . .P-values predict the future only vaguely, but confidence intervals do much better". Cumming's argument is that "confidence intervals, by contrast [to P-values] give useful information about replication. There is an 83% chance that a replication gives a mean that falls with the 95% confidence interval from the initial experiment." This is said to be "superior information about replication". There are a number of arguments against this conclusion.

1. In Cumming (2008), no clear basis for comparison is offered. If we want to assert that one method is better than the other, we need a sensible scale to use for comparison, but none is provided.
2. The comparison implicit in Cumming (2008) is subjective. This shows in the language used; for example, under Figure 1: "Note the extremely large variation in p, from < .001 to 0.759." But one could equally note, in a similarly subjective way, the variation in the limits of the 95% confidence intervals, and claim that this was 'extremely large'; after all, in a context in which the true difference in means is 10 units, the lower limits of the 25 95% confidence intervals range from about −8 to just over 10. It is asserted that "p gives dramatically uncertain information about replication", while "confidence intervals give useful information about replication". These are subjective assertions.
3. A technical argument is offered to support the effective "replication" performance of confidence intervals: the probability of a "replication mean" falling within the previous 95% confidence interval is 0.83. Formally, this result says that for random samples on a normal distribution, the probability that one sample mean falls within the 95% confidence interval for μ, based on another sample mean, is 0.83. Presumably (although implicitly) 0.83 is a suitably large probability.

When replication is being considered, we ought to envisage a process in which 'everything non-random is the same'. Cumming calls this the framework of 'exact replication'. But in using the probability of 0.83 to support the replication of confidence intervals, this principle is not followed. The very essence of replication is abandoned. When addressing the question of whether the confidence interval is replicated, Cumming examines not the next confidence interval, but the next point estimate. He compares a point with an interval.

Secondly – and this criticism is associated with the first – this means that the measurement of replication, the assessment of whether the first interval is replicated or not, is reduced to a binary outcome: either the next point estimate is in the first confidence interval (replicated) or it is not in the first confidence interval (not replicated).

This binary decision parallels the use of a threshold in hypothesis testing, and hence deciding, in a binary way, whether a test is statistically significant. Many authors have been critical of such an impoverished approach. Gardner and Altman (1986) refer to the “arbitrary convention of using the 5% level of statistical significance to define two alternative outcomes — significant or not significant — which is not helpful and encourages lazy thinking”.

The same principle applies here: one ought not consider “replication” of confidence intervals in a binary, either/or fashion, for much the same reason: “replication” (in this sense) is a matter of degree and extent and can be measured and reported accordingly; when this is done, the conclusion is different from Cumming’s.

REPLICATION FOR CONFIDENCE INTERVALS: A BETTER APPROACH

Is it feasible to define replication for a confidence interval in a way that respects its nature, and that preserves the integrity of the idea of replication? A simple approach to do so is suggested here. Rather than considering a confidence interval to be replicated or not according to whether the next point estimate is in the original confidence interval, the proposal is to compare the next interval with the original confidence interval, and hence define *R*, the *extent* of confidence interval replication, to be the fraction of the second interval that is contained within the first one. Defined in this way, *R* is a measure which ranges between zero (there is no overlap between the intervals) and one (the two intervals exactly coincide, or the second interval is shorter than, and entirely contained within, the first interval). Now replication is not binary, changing abruptly and therefore unrealistically from “no” to “yes”, but subtle, gradual and nuanced, properly reflecting the expression of precision in an interval. In Cumming’s version of “replication” of confidence intervals, the interval moves disjointedly from being replicated (next estimate is just inside the first interval) to not replicated (next estimate is just outside the first interval) for a trivial change in the estimate. On the other hand, the measure defined here moves smoothly between zero and one, according to the position of the next interval.

For the case of inference on the mean of a normal distribution with known variance, from a sample of size *n*, the distribution of *R* can be easily derived and the cumulative distribution function for R is shown in Figure 1.

The distribution has a discrete probability at r=0: Pr(R = 0) = 0.00557. This is the probability that the two confidence intervals do not overlap. The mean value of *R* is 0.71, the median is 0.76 and the lower quartile is 0.58. This distribution reflects the replication of a confidence interval in a more realistic and authentic way than the single probability of 0.83, based on comparing a point with an interval.

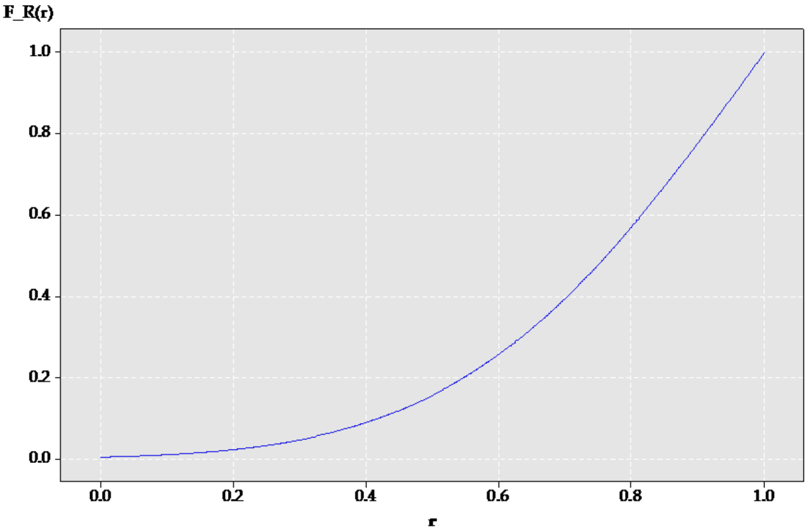

Figure 1: Cumulative distribution function for *R*, the extent of replication of a 95% confidence interval for the population mean, when the population standard deviation is known.

It is possible to examine the case when the population standard deviation is unknown and must be estimated; in this case the distribution of *R* depends on the sample size. The cumulative distribution function is shown for a few values of *n* in Figure 2.

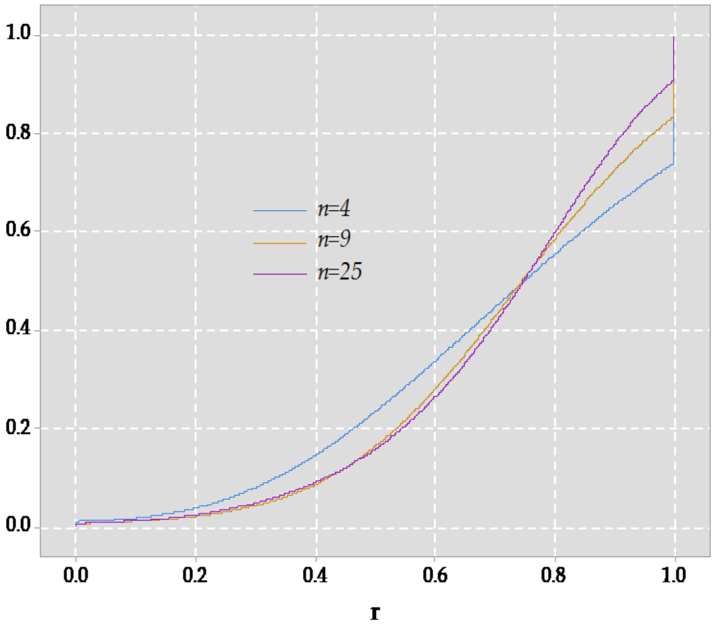

Figure 2: Cumulative distribution function for *R*, the extent of replication of the 95% confidence interval for a population mean, normal sample, unknown variance, three values of *n*.

The extent of replication of a confidence interval varies, in a smooth way, from 0 to 1, and is not well represented by quoting a single figure of 83%.

In any case, it would be better to emphasise the true nature of replication in a confidence interval, which is that centred around the true parameter value, rather than around one estimate of this parameter.

The question of ‘what you expect next time, given what you saw this time’ changes every time there is a new experiment. The “replication” framework used by Cumming to make his point uses a different framework from that in which we understand frequentist inference, leading to further cognitive difficulties in understanding and interpretation.

VARIATION OF P-VALUES AND CONFIDENCE INTERVALS

Cumming has a second key argument for the superiority of confidence intervals. The claim is that is that P-values vary a lot, and more than confidence intervals. This argument is framed, often, against presumptions or expectations in a general, psychological sense, of what is going to occur. But is the issue the variation, or the expectations? Where do these expectations come from?

The “dance of the P-values”, an evocative phrase, is presented in Cumming (2012, p. 130) and also in a recent video clip (Cumming, 2017a), where the P-value is said to be “extremely unreliable, enormously variable” and “not to be trusted”.

The sampling variability of the P-value under the null hypothesis is, of course, uniform. Further, this is a good feature, in that it correctly expresses the “lack of evidence” conclusion that we want when the null is true. For any departure from the null, the distribution of the P-value departs from uniformity and becomes skewed, increasingly concentrated near 0 as the power of the test increases. There is nothing strange or untoward or unexpected about these properties of its distribution. They are inevitable features of the construction of the P-value.

So why is this variation said to be “astonishing”, “enormous”, “extremely large” or “immense”? These are all descriptors used by Cumming.

Confidence intervals “dance” too, so it is not surprising that Cumming also uses the phrase “the dance of the confidence intervals” (Cumming, 2012, p. 78) and also in the video clip (Cumming, 2017a). In that clip, he shows this dance and comments (at 4:17) on “these intervals dancing around *exactly as we should expect*” (my emphasis). I suggest that P-values also dance around “exactly as we should expect”. It is question of whether our expectations are aligned with reality.

Further, and importantly, confidence intervals and P-values dance “in step”. P-values and (corresponding) confidence intervals can be derived from the other, as is familiar to data analysts, who sometimes need to obtain a confidence interval from an estimate and a P-value, or a P-value from a confidence interval.

Assume a generic context for a parameter θ and a null hypothesis H_0: $\theta = \theta_0$. Assume that inference based on normal theory is approximately valid. Then if we have a 100(1-α)% confidence interval for θ, (L, U), the two-sided P-value can be obtained, since the point estimate is the average of L and U and the standard error of the estimate is a simple function of L and U. For example, for a 95% confidence interval, the standard error of the estimate is equal to (U – L)/(2 × 1.96).

Conversely, if we are given the P-value and the point estimate, we can obtain the 100(1-α)% confidence interval. The point estimate is needed in order to resolve which of the two possible confidence intervals apply, according to the whether the point estimate is less than or greater than θ_0. This is realistic; it is uncommon to know the P-value but not the point estimate.

This one-to-one correspondence tells us that however a P-value is dancing, the corresponding confidence interval is dancing in parallel. Neither dance is more exotic, volatile or variable than the other.

CONCLUSION

There are good reasons for data analysts to prefer confidence intervals over P-values. However, two suggested reasons promoted by Cumming, namely, superior “replication” and less variation for confidence intervals, are not sound, and their promulgation has the potential to add to confusion about the meaning and properties of these common representations of inference.

REFERENCES

Berry, G. (1986). Statistical significance and confidence intervals. *Medical Journal of Australia, 144,* 618–619.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3,* 286–300.

Cumming, G. (2012). *Understanding the New Statistics; Effect Sizes, Confidence Intervals, and Meta-Analysis,* New York, NY: Routledge.

Cumming, G. (2017a). Significance Roulette 1. Retrieved from https://www.youtube.com/watch?v=OcJImS16jR4 .

Cumming, G. (2017b). Staff profile: Professor Geoff Cumming. http://www.latrobe.edu.au/psychology/staff/profile?uname=GDCumming.

Cumming, G. and Calin-Jageman, R. (2017), *Introduction to the New Statistics,* New York, NY: Routledge.

Cumming, G. and Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods, 11,* 217–227.

Cumming, G., Williams, J., and Fidler, F. (2004). Replication, and researchers’ understanding of confidence intervals and standard error bars. *Understanding Statistics, 3,* 299–311.

Gardner, M. J. and Altman, D. G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal (Clinical Research Ed.), 292,* 746–750.

Lang, J., Rothman, K. J., and Cann, C. (1998), “That confounded P-value,” Epidemiology, 9, 7–8.

Rothman, K. J. (1978). A show of confidence. New England Journal of Medicine, *299,* 1362–1363.

Stang, A., Deckert, M., Poole, C., & Rothman, K. (2017). Statistical inference in abstracts of major medical and epidemiology journals 1975-2014: a systematic review. European Journal of Epidemioly, *32*, 21–29.