

Running head: CONFIDENCE INTERVALS AND REPLICATION

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, 11, xxx-yyy.

© American Psychological Association. Journal website: <http://www.apa.org/journals/met/>

This article may not exactly replicate the final version published in the journal. It is not the copy of record.

Confidence Intervals and Replication: Where Will the Next Mean Fall?

Geoff Cumming¹ and Robert Maillardet²

¹La Trobe University, Melbourne, Victoria, Australia

²University of Melbourne, Victoria, Australia

Confidence intervals (CIs) give information about replication, but many researchers have misconceptions about this information. One problem is that the percentage of future replication means captured by a particular CI varies markedly, depending on where in relation to the population mean that CI falls. We investigated the distribution of this percentage, for σ known and unknown, for various sample sizes, and for robust CIs. The distribution has strong negative skew: Most 95% CIs will capture around 90% or more of replication means, but some will capture a much lower proportion. On average, a 95% CI will include just 83.4% of future replication means. We present figures designed to assist understanding of what CIs say about replication, and also extend the discussion to explain how p values give information about replication.

Confirmation comes from repetition.... Repetition is the basis for judging... significance and confidence. (Tukey, 1969, pp. 84-85)

A phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result. (Fisher, 1966, p. 14)

Considering whether an effect is replicable is at the heart of drawing inferences from data. A confidence interval (CI) gives inferential information, and so one important way to think about CIs is to consider what they tell us about replication. Given a sample mean and 95% CI, what is the probability that a repetition of the experiment, with an independent sample of participants, would give a mean that falls within the original CI? The correct answer is not .95. In this article we consider this and related questions.

First, consider Figure 1, which shows the mean and 95% CI for each of a sequence of 20 independent samples from a normal population with fixed mean $\mu = 50$, that is, a sequence of 20 replications of an experiment. The population SD σ is assumed known, so all CIs have the same width. In the long run we expect 5% of the CIs will not include the population mean μ ; here just one does not include μ and its mean is shown as an open circle. Inspect each mean in turn and note where it falls in relation to the preceding CI: In three cases the mean is not captured by the preceding CI, and these means are marked with a short horizontal line. We will show that in the long run 16.6% of CIs do not capture the next replication mean. Figure 1 illustrates the contrast

between CIs capturing the population parameter, and capturing the next replication mean. Capture of the next sample mean is less frequent because it is influenced by two sources of variability: the variation of the original mean about μ , and the variation of the replication mean about μ . The relevant variable is the difference between two successive means, and this random variable is important in our analysis below.

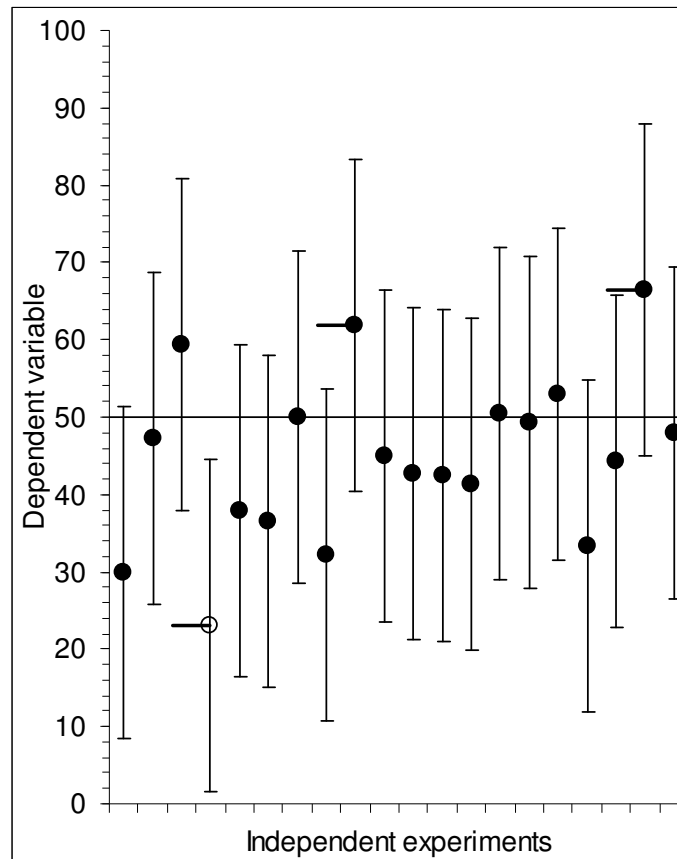


Figure 1. Means with 95% confidence intervals (CIs) for 20 independent samples (i.e., 20 replications) from a normal population. The CIs all have the same width because σ is assumed known. The population mean $\mu=50$ is marked by the horizontal line. In cases where the CI does not include μ , the sample mean is shown as an open circle: There is one case here, and in the long run 5% of samples will give CIs that do not include μ . In cases where a mean does not fall within the previous CI, the sample mean is marked with a short horizontal line: There are 3 cases here, and in the long run 16.6% of means will fall outside the previous CI.

Let us define *capture percentage* (CP) to be the percentage of replication means that will fall within a given original CI. We assume that replications come from the same population, so variation from sample to sample is caused only by sampling variability, and not by any other changes in population or experimental characteristics. Our simulated replication is thus different from most real-life replication, in which some details of procedure or population vary from experiment to experiment. Figure 2 illustrates how, for a given fixed μ , different original CIs have different CP values. The left panel shows an original CI for which $CP=83$, which we will show to be the average value of CP. The middle panel illustrates the only case in which $CP=95$,

and this is the maximum value CP can take when σ is known and so CI width is fixed, without scope to vary from sample to sample. The right panel shows one of the 5% of original CIs for which $CP < 50$ because the original CI does not include μ . These images reinforce the message that there are two sources of variability: The first source is the variability of the original mean about μ , and the second is the variability of replication means around μ . When we consider where a replication mean will fall, in relation to an original sample mean, these two sources of variability are compounded, and so it should not be surprising that some replications may seem quite far from an initial result, and CP values can be low.

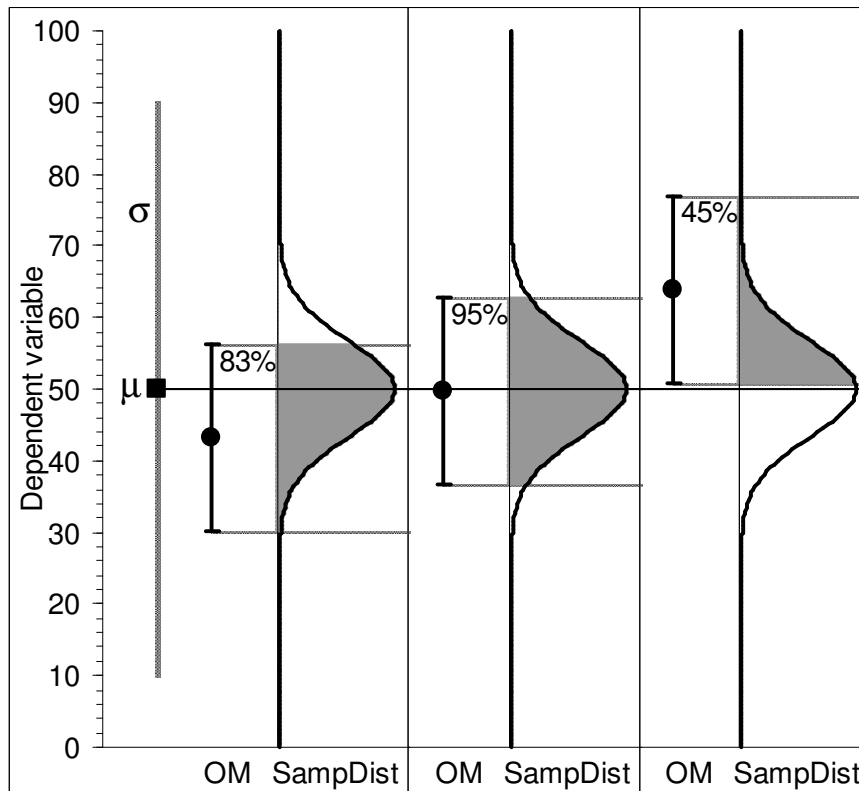


Figure 2. Three examples of a 95% CI and the sampling distribution of replication means, for a normal population and $\sigma=40$ assumed known. At the far left is the mean μ and SD σ of the population, then in each panel is shown the original mean (OM) and 95% CI for our original sample, of size $n=36$. To the right in each panel is shown the sampling distribution (SampDist) for all possible replication means. The dotted horizontal lines and shading indicate CP, the percentage of replication means that would fall within each original CI, and the CP values are shown. In the left panel 83% of replication means would fall within the original CI; in the middle panel the sample mean happens to fall very close to μ , and so $CP=95$; and in the right panel the CI does not include μ , so $CP < 50$. The value of CP is influenced both by the departure of the original mean from μ , and by the variation of replication means about μ . These illustrations are of simulations, in which μ is known and so it, and the sampling distribution of replication means can be displayed. In research practice, of course, μ is not known and only the original mean and its CI can be displayed.

This brief discussion, and Figure 2, illustrate that capture percentages vary, 95 is the maximum (for 95% CIs, with σ known), most are somewhat below 95, and a small proportion are much lower. The mean CP must therefore be distinctly below 95, and the distribution of CP values is strongly negatively skewed. We believe that understanding capture percentages and replication can assist researchers to understand CIs more fully, and thus have a better basis for using CIs to interpret research findings.

It is timely to consider how researchers think about CIs because psychologists are now being urged to make wider use of CIs to improve research communication (American Psychological Association, 2001, p. 22). One advantage of CIs is that they support inference without any need for specification of a null hypothesis. This and other advantages of CIs were discussed by Cumming and Finch (2001, 2005). However, Belia, Fidler, Williams and Cumming (2005), and Cumming, Williams and Fidler (2004) found that leading researchers in psychology, behavioral neuroscience, and medicine have a range of severe misconceptions about CIs. The five CI misconceptions we identified can be compared with the extensive lists of misconceptions about statistical significance testing and p values that have been described by many writers, including Carver (1978) and Kline (2004, chapter 3).

Cumming et al. (2004) investigated researchers' beliefs about the chance that a replication mean would fall within an original CI. They found that a majority of researchers in the three disciplines hold the confidence level misconception (CLM), which is the erroneous belief that about $C\%$ of future replication means will fall within an original $C\%$ CI, where C is the *confidence level* of a CI. A holder of the CLM believes that, on average, a 95% CI will include about 95% of future means. Following Estes (1997), Cumming et al. explained that for confidence level $C=95$ the correct percentage is 83.4: On average, 83.4% of replication means, just 5 out of 6, will fall within an original 95% CI. There is some analogy between the CLM and Carver's (1978) replicability fallacy, which is the erroneous belief a result that is statistically significant at $p < .05$ will be similarly statistically significant in 95% of replications.

In this article we investigate the average of CP, and the probability distribution of CP over different original CIs. We first assume σ known, then drop that assumption. Our main discussion assumes a normal population, and considers the sample mean, and CIs based on normal or t distributions. We then consider other situations, including robust statistics, and sampling from populations that are not normal. We extend the consideration of CIs and replication to the relation between p values and replication. We present images and suggest guidelines that we hope can help researchers and students overcome the CLM, and gain a more accurate appreciation of capture percentages and what CIs tell us about replication. We hope this will assist researchers to realize the advantages that CIs have for improved research communication.

Replication Means, σ Known

When the population is assumed normal and σ is known, a critical value of the normal distribution is used to calculate the CI, and the CI width is the same for all samples of size n , for given σ and confidence level C . Different original CIs have the same width, but will capture different percentages of replication means as a function of where the original mean falls in relation to μ . The CP attains its maximum value of C (in this case, 95) only when the original mean falls exactly at μ , and CP drops as the original mean falls further from μ .

We follow Estes (1997) to find the average value of CP when σ is known. Let M_1 be our original sample mean, and M_2 a replication mean. Let $CP(M_1)$ be the capture percentage for

original mean M_1 . Both M_1 and M_2 are distributed as $N(\mu, \sigma^2/n)$, meaning Normal with mean μ and variance σ^2/n . Then $M_2 - M_1$ is the difference between the means of two independent samples, each of size n , from the same population. The sampling distribution of this difference is $N(0, 2\sigma^2/n)$ because variance is additive when independent variables (the two sample means) are subtracted. Clearly M_2 is only in the original CI, centered on M_1 , if the absolute value of $(M_2 - M_1)$ is less than the *margin of error* of that original CI, where the margin of error is half the total width of the CI.

The CI on M_1 has margin of error $w = z_C \times \sigma / \sqrt{n}$, where z_C is the standard normal critical value for a $C\%$ CI (i.e., $P[-z_C < N(0, 1) < z_C] = C/100$, and so $z_C = 1.96$ for $C=95$). Therefore the area we seek lies between $-w$ and $+w$, under the normal distribution with variance $2\sigma^2/n$, which is the area under the standard normal distribution between $-z_C/\sqrt{2}$ and $+z_C/\sqrt{2}$. This may be expressed as $2[\Phi(z_C/\sqrt{2}) - 0.5]$, where Φ is the cumulative distribution function of the standard normal.

If $C=95$, then $z_C=1.960$, $z_C/\sqrt{2}=1.386$, and so the probability is .834. In other words, the average CP is 83.4, and the left panel in Figure 2 illustrates an original CI that has this CP. The Microsoft Excel statement for the average CP is

$$200*(\text{NORMSDIST}(\text{NORMSINV}(C/200+0.5)/\text{SQRT}(2))-0.5). \quad (1)$$

The heavy line in Figure 3 shows how the average CP varies with C . For σ known or n not small, error bars that are ± 1 SE correspond to a CI with $C=68.3$, and Figure 3 shows that for this case the average CP is just 52.1. We refer to such error bars as *SE bars*, and for these bars there is barely a 50-50 chance that on average a replication mean will fall within the original error bars.

The derivation above also indicates that an interval that is $\sqrt{2}$ times as wide as the 95% CI will have an average CP of 95%. This interval could be referred to as the *95% prediction interval* (PI; Cumming, 2005; Killeen, 2005), and for this interval $C=99.44$, a value that could also be read from the heavy curve in Figure 3, for $CP=95$. If in Figure 2 an interval were plotted around each of the three original sample means that is $\sqrt{2}$ times as wide as the intervals shown in the figure, these would be 99.44% CIs or, equivalently, 95% PIs. These three intervals would capture, respectively, 96%, 99%, and 75% of replication means, and the average CP over all such intervals would be 95%.

Equation 1 gives the average percentage of future means that will fall within the original CI, that is, within $z_C \times \sigma / \sqrt{n}$ of M_1 . A simple extension of the argument gives the average percentage of future cases in which the CI around M_2 will have some overlap with the original CI around M_1 . This will be the average percentage of future means that will fall within $2z_C \times \sigma / \sqrt{n}$ of M_1 , which is

$$200*(\text{NORMSDIST}(2*\text{NORMSINV}(C/200+0.5)/\text{SQRT}(2))-0.5).$$

For $C=95$, this is 99.44, meaning that, on average, in 0.56% of cases (roughly 1 in 180) the next 95% CI will not overlap with the original 95% CI. (In Figure 1 there are no examples of successive CIs not overlapping, the closest being the third and fourth CIs.) For lower values of C this average percentage increases notably. For $C=90, 80$, and 50 , the average percentages of future CIs that do not overlap the original interval are 2.0, 7.0, and 34.0 respectively. For a 50%

CI, on average there is fully a one in three chance that the 50% CI for a replication will have no overlap with the original interval. If SE bars are shown, on average 15.7% of replication SE bars will not overlap with the original interval.

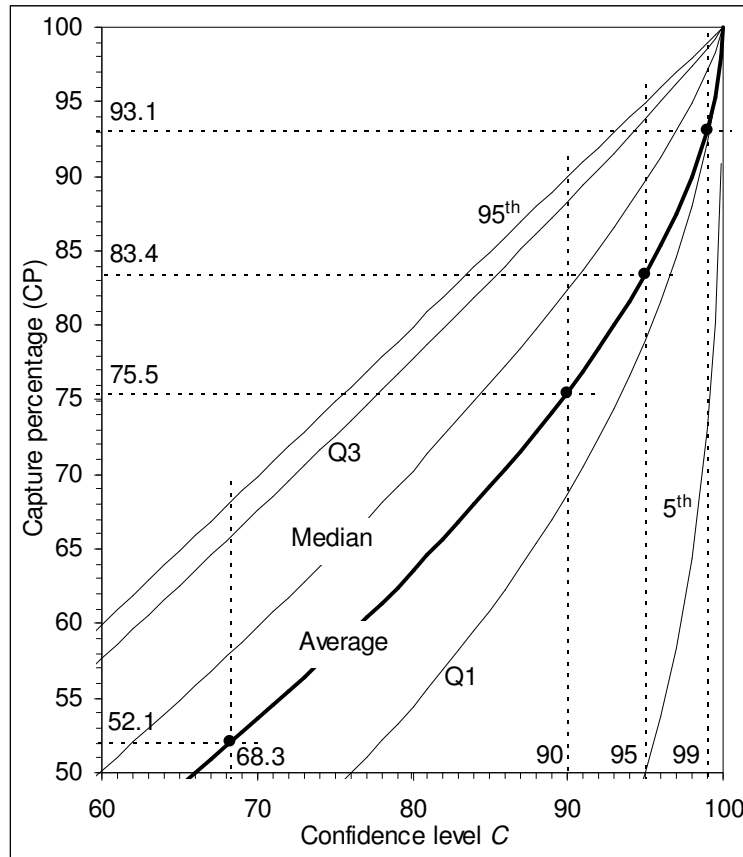


Figure 3. The heavy curve shows average CP as a function of confidence level C , assuming a normal population and that σ is known. (CP is the percentage of replication means that will fall within an original CI.) Average CP values are shown next to the vertical axis for 99%, 95%, 90%, and 68.3% CIs. (68.3% CIs correspond to error bars that are ± 1 SE, referred to in the text as SE bars.) The light curves mark percentiles of the CP probability distribution: The three quartiles are marked Q1, Median, and Q3, and the 5th and 95th percentiles are also shown. Note for example where the dotted vertical at $C=95$ intersects the curves: For 95% CIs, the 5th percentile of the CP distribution is 50, Q1 is 79.0, the median is 89.6, Q3 is 93.8, and the 95th percentile is 94.95.

Note the distinction between the overlap of CIs on two replication means, as we consider above, and the overlap of CIs on two independent means, which may represent two different treatments in an experiment. Cumming and Finch (2005) discussed the latter case of overlap, and described an approximate mapping from the extent of overlap to the p value for the comparison of the two means. Overlap that is half the average of the two margins of error is, under broad conditions, equivalent to a p value of approximately .05; and zero overlap—the two CIs just touch end-to-end—corresponds approximately to $p = .01$. In the case of CIs on two replication means the population mean is assumed fixed, meaning that it makes little sense to think of the p

value for a comparison of the two means, because the null hypothesis for that comparison is assumed true.

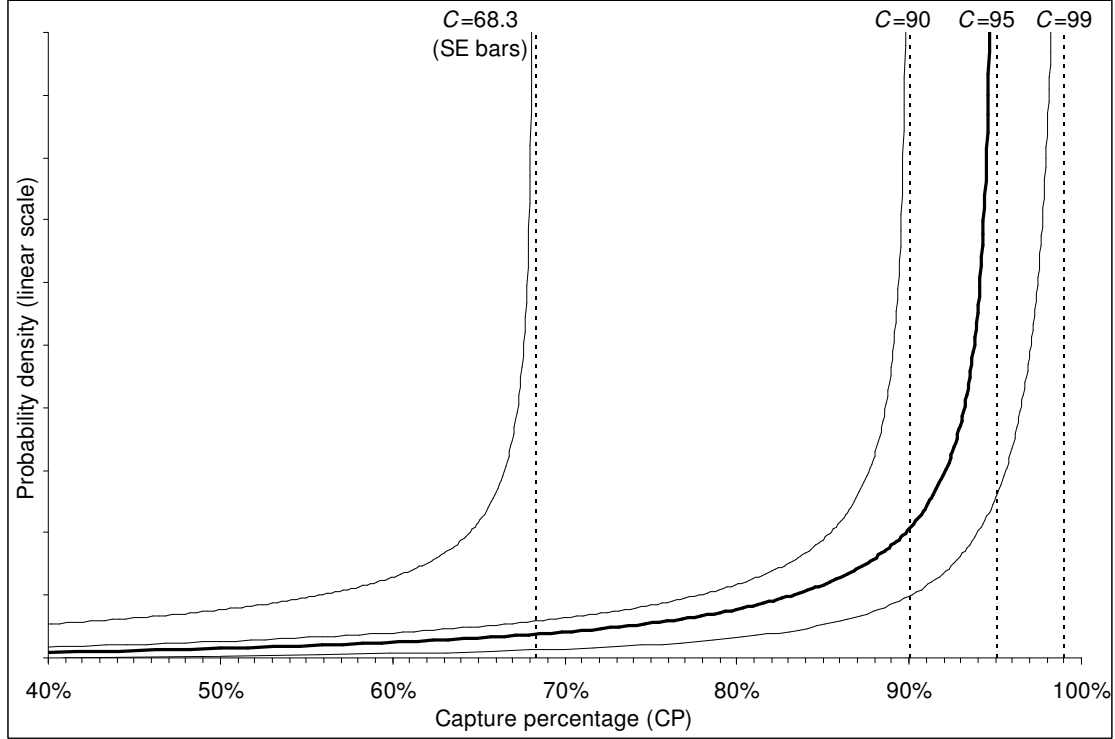


Figure 4. The probability distribution of CP, for four values of C , for a normal population and σ known. The heavy curve is for $C=95$. Vertical dotted lines at C are asymptotes. The curves are independent of n , and strongly negatively skewed.

While discussing the comparison of the means of two different treatments, it is worth considering the difference between those means, which is often the research focus. Replication and the difference between means is best thought of in terms of the single CI on that difference. Cumming and Finch (2005) discussed that CI, and noted that in the independent means case it is almost certainly wider than the CI on either individual mean, whereas in the case of paired data it is usually narrower than either, but can have any width from zero to considerably larger than either individual CI, depending on the correlation between the two measures. In either case, focusing on the single CI on the difference allows our discussion of replication to be applied.

We now consider the capture percentage $CP(M_1)$ further, still assuming σ known. We can write:

$$\begin{aligned}
 CP(M_1) &= 100 \times P\left[M_2 \in \left\{M_1 \pm z_c \left(\frac{\sigma}{\sqrt{n}}\right)\right\}\right] \\
 &= 100 \times P\left[\frac{M_2 - \mu}{\sigma/\sqrt{n}} \in \left\{\frac{M_1 - \mu}{\sigma/\sqrt{n}} \pm z_c\right\}\right] \\
 &= 100 \times [\Phi(Z_1 + z_c) - \Phi(Z_1 - z_c)], \tag{2}
 \end{aligned}$$

where $Z_1 = \frac{M_1 - \mu}{\sigma/\sqrt{n}}$ is $N(0, 1)$. Equation 2 is consistent with the observation that $CP(M_1)$ is

symmetric, and decreases monotonically either side of its maximum value at $M_1 = \mu$.

In each panel in Figure 2, the shaded area under the normally distributed sampling distribution of replication means is $CP/100$, and this area can be seen as the difference between two left-hand tail areas under that distribution. Equation 2 expresses that difference, and is the basis for a simple relationship between the quantiles of $CP(M_1)$ and those of Z_1 . If c_q is the q^{th} quantile of $CP(M_1)$, this relationship can be expressed as

$$c_q = \Phi(z_{100(1-q)} + z_C) - \Phi(z_{100(1-q)} - z_C), \text{ for } 0 \leq q \leq 1,$$

where again z_C is the critical value for a $C\%$ CI, so $z_C = \Phi^{-1}\left(\frac{1}{2}\left(1 + \frac{C}{100}\right)\right)$ and

$$z_{100(1-q)} = \Phi^{-1}\left(1 - \frac{q}{2}\right).$$

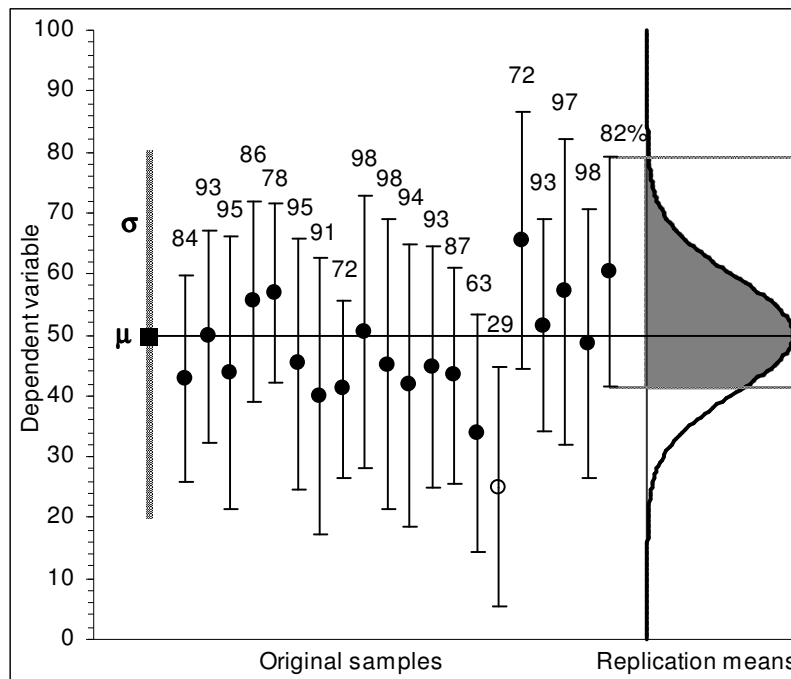


Figure 5. Means and 95% CIs for 20 original samples, with $n=10$, a normal population and σ not known. The population μ and σ are shown at left, but are not used to calculate CIs. The sampling distribution of replication means is shown at right, with the area corresponding to CP for the most recent CI shown shaded. Capture percentages (the numbers above the CIs) vary because (i) sample means fall at various distances from μ , and (ii) CI width varies. In one case (open dot) the CI does not include μ , and CP is low. Some CPs are higher than 95.

Figure 3 shows five such quantiles plotted as a function of C . Figure 4 is also based on this relationship, and shows the probability density function of CP, for σ known, for four values of C . The heavy curve is for $C=95$. The figure shows the large negative skew: Many CIs have CP values close to the maximum C , and a few have low CPs. Just $(100-C)\%$ of CIs will have CP

less than 50. Figure 3 shows that, as would be expected for highly skewed curves, average CP values (heavy line) are considerably lower than the medians. The skew is also indicated by the closeness of the 75th (Q3) and 95th percentile lines in Figure 3, and the very wide spacing of the 25th (Q1) and 5th percentile lines. Note that all curves in Figures 3 and 4 are independent of n .

Replication Means, σ not Known

When the population is assumed normal and σ is not known, sample SD is used as an estimate of σ and t is used to calculate CI width, which varies over samples. Figure 5 shows that, for σ not known, CP varies widely, influenced by variation in both (i) where the original mean falls in relation to μ , and (ii) the sample SD and thus the width of the original CI. Some CIs will have a large SD, and may have a CP greater than C .

We seek an expression for average CP, with σ not known. Let s_1 be the SD for the original sample. The original CI is $M_1 \pm t_c (s_1/\sqrt{n})$, where t_c is the critical value of t corresponding to C . The average CP is

$$\begin{aligned} & 100 \times P[M_2 \in \{M_1 \pm t_c (s_1/\sqrt{n})\}] \\ &= 100 \times P\left[\frac{M_2 - M_1}{\sqrt{2\sigma^2/n}} \in \left\{\pm t_c / \sqrt{2}\right\} (s_1/\sigma)\right]. \end{aligned} \quad (3)$$

As before, $(M_2 - M_1)$ is $N(0, 2\sigma^2/n)$, and so $((M_2 - M_1)/\sqrt{2\sigma^2/n})$ is $N(0, 1)$. Because the population is assumed normal, M_1 and s_1^2 are independent, and $(n-1)s_1^2/\sigma^2$ has a chi-square distribution with $df=n-1$. Consider

$$T_{n-1} = \frac{((M_2 - M_1)/\sqrt{2\sigma^2/n})}{\sqrt{(n-1)s_1^2/\sigma^2(n-1)}},$$

which is the ratio of a standard normal to the square root of a chi square distribution divided by its df , and thus has a t distribution, with $df=n-1$.

Substituting into Equation 3, we are seeking

$$P[T_{n-1} \in \{\pm t_c / \sqrt{2}\}],$$

which is readily found from the t distribution, with $df=n-1$. The Microsoft Excel statement for the average CP is

$$100*(1-TDIST(TINV(1-C/100,N-1)/SQRT(2),N-1,2)). \quad (4)$$

Figure 6 shows how average CP varies with C . The heavy line shows the relation when σ is known, and the light lines the relation for various values of n when σ is not known. The figure shows that average CP values are little influenced by n , unless n is very small, for example less than 10.

Figure 7 shows the distribution of CP when σ is not known, for $C=95$, for various values of n . For each n we took 200,000 independent random samples. For this and all subsequent simulations we used the pseudorandom number generator of Wichman and Hill (1982, 1987), as implemented in Microsoft Excel 2003. Figure 7 also reports means and medians for the distributions pictured. The curves for σ unknown differ substantially from the curve for σ known, even for very large n , especially in the region of $CP=95$ and above, although the mean and median values vary little for any n greater than about 20.

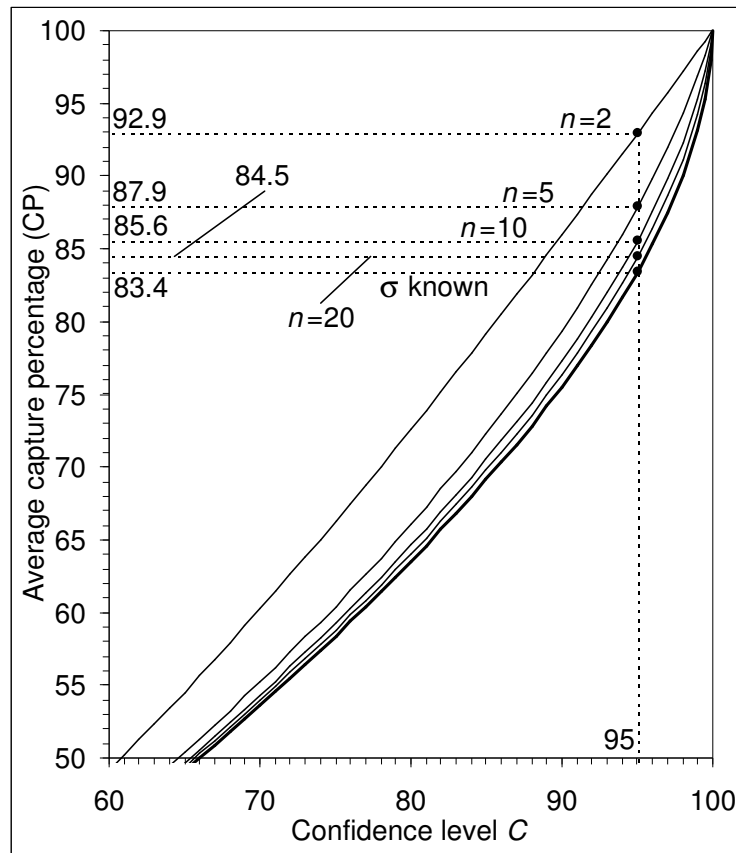


Figure 6. Average CP as a function of C . The heavy curve shows the relation for a normal population and σ known, and is the same as the Average curve of Figure 2. The light curves show the relation for marked values of n , for σ not known. Values of average CP are marked for $C=95$, for the various values of n . Unless n is very small, less than 10, it has little effect.

Robust Measures

As a first step of generalization we made an initial exploration of CPs for robust CIs. We used 20% trimmed means, and the 20% Winsorized variance to calculate robust 95% CIs, using the method of Tukey and McLaughlin (1963; Wilcox, 2003, p. 128). We used $n=10$ and 20, and the following population distributions: normal, mixed normal, uniform, exponential, and lognormal. The mixed normal was a mixture of $N(0, 1)$ with $p=.9$, and $N(0, 100)$ with $p=.1$. These distributions represent a range of symmetric and skew departures from normality. For each combination of n and population distribution we took 20,000 independent random samples. Table 1 reports percentiles of the observed distributions of CP values, the average CP, and also the observed coverage of the CIs: the percentage of samples in which the robust CI included the population trimmed mean. As a baseline for comparison, the first three rows of Table 1 show percentiles and the average CP for the cases investigated earlier: normal population, with σ known; then σ not known, and $n=10$ and 20. The percentiles and averages shown in italics in these rows of Table 1 were calculated using the formulas presented earlier.

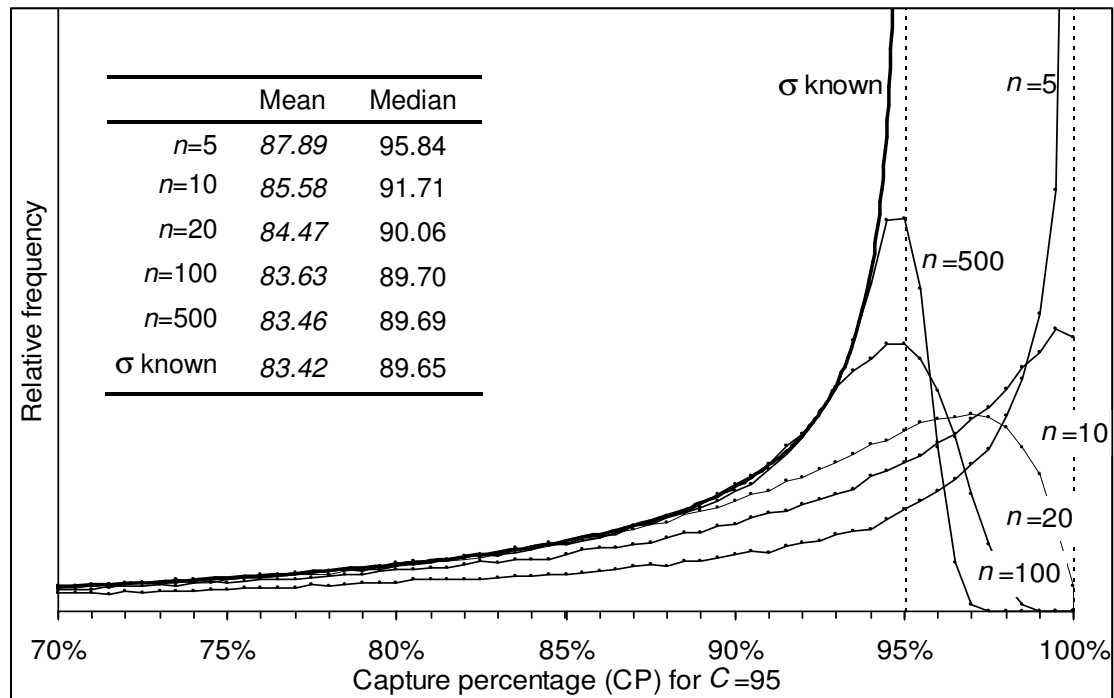


Figure 7. Relative frequency distribution of the capture percentage (CP) generated by simulation, assuming a normal population and σ not known, for various values of n , with $C=95$. For σ known, the heavy curve and tabulated values are calculated, as are the tabulated means for each n . (Values in the table that are calculated are shown in italics.) For the various values of n , the fine curves and tabulated medians are based on simulations of 200,000 samples at each value of n . For $n=5$, and other very small values of n , the curve is high near 100%, but there is no vertical asymptote.

The rightmost column of Table 1 indicates that, as expected, the robust CIs achieve coverage fairly close to the nominal 95%, although in some cases less so for $n=10$. Overall, the percentiles and average CP values reported in Table 1 indicate that the basic robust techniques of trimming and Winsorizing give robust CIs with CP values that are quite similar over situation and sample size, at least down to $n=10$, and quite similar to those for CIs for a normal population calculated using conventional methods. It is striking that robust CIs give such similar capture probabilities, despite the quite severe departures from normality that were used in the simulations. Our results do not of course imply that the assumption of normality is unimportant when using conventional CIs, but rather that robust CIs give stable CP results for a wide variety of population distributions.

CIs for Proportions

We made one attempt to study a discrete distribution, by investigating 95% CIs for proportions. We took samples of n Bernoulli trials, each having probability of success π . To calculate CIs we used the method recommended by Newcombe and Altman (2000), based on Newcombe's (1998) comparative evaluations of seven approaches. It is the score method of Wilson (1927), which yields well-located CIs that are appropriately asymmetric, and which may

be used even for small n and proportions near 0 or 1. For the discrete binomial distributions, the probability distributions of CP are not continuous, as in Figure 4, but have nonzero probabilities at only some $n+1$ values, fewer if $\pi=.5$. For each of $n=10$ and 20, we took 20,000 samples at each of the ten π values from .50 to .95 in steps of .05. At each combination of n and π we noted percentiles of the frequency distribution of CP, the average CP, and the obtained coverage of the nominally 95% CIs. However, observed percentile values varied haphazardly with π , especially for $n=10$, because of the discrete nature of the CP distribution and the small number of values that CP can take. For example, with $n=10$ the 5th percentile was 26 for $\pi=.65$, 5.8 for $\pi=.75$, and 60 for $\pi=.95$. The 25th percentile was 76 for $\pi=.65$, 53 for $\pi=.75$ but again 60 for $\pi=.95$. This wide and haphazard variation merely reflected the small number of discrete values that CP could take, and the arbitrary percentiles we chose to examine, and precludes any simple quantitative summary of the type we could make for continuous distributions, as shown in Table 1.

Table 1. Percentiles, and the average, of the probability distribution of the capture percentage (CP) of 95% confidence intervals (CIs) for various measures, sample sizes, and population distributions.

Population		Percentiles of distribution of CP								CI
distribution	n	Measure	Method	5%	25%	50%	75%	95%	Average	Coverage ^a
Normal	any	Mean	σ known	<i>50^b</i>	79.0	89.6	93.8	95.0	83.4 ^b	95 ^c
Normal	10	Mean	σ not known	49.7	80.5	91.7	96.9	99.5	85.6 ^d	95 ^c
Normal	20	Mean	σ not known	49.7	79.6	90.1	95.1	98.3	84.5 ^d	95 ^c
Normal	10	TM20 ^e	Simulation	61.0	81.3	89.3	92.6	93.6	84.8	93.5
Normal	20	TM20	Simulation	55.1	80.4	89.5	93.1	94.1	84.1	94.1
Mixed normal	10	TM20	Simulation	57.4	81.1	89.5	93.0	94.0	84.4	94.0
Mixed normal	20	TM20	Simulation	52.6	79.8	89.4	93.3	94.4	83.7	94.4
Uniform	10	TM20	Simulation	64.9	82.0	88.6	91.2	91.9	84.8	92.0
Uniform	20	TM20	Simulation	58.5	80.6	89.2	92.7	93.8	84.4	93.8
Exponential	10	TM20	Simulation	53.6	79.8	89.2	92.5	93.5	83.5	91.7
Exponential	20	TM20	Simulation	53.3	79.5	89.1	93.1	94.3	83.6	93.2
Lognormal	10	TM20	Simulation	48.1	77.7	88.7	93.1	94.3	82.4	91.2
Lognormal	20	TM20	Simulation	49.2	78.5	89.4	93.5	94.7	83.0	92.8

Notes ^aPercentage of CIs that include the population parameter

^bFor σ known, all percentiles and the average are calculated, as indicated by italics.

^cCoverage is exact for normal population, σ known or unknown.

^dCalculated value, as indicated by italics.

^eTM20 is the trimmed mean, with .2 taken from each tail. The .2 Winsorized sample SD is used to calculate the robust 95% CI (see text).

CI, p Values and Replication

Our analysis of CIs and replication can be extended to give a relation between p values and replication. CIs support inference without any need for specification of a null hypothesis, but a p value is, of course, relative to a chosen point null hypothesis μ_0 . Figure 8 shows sample mean M and CIs for a single sample of size $n=20$, from a normal population, for which $t=1.55$ ($df=19$), with respect to a null hypothesized value μ_0 , and the one-tailed p value is .069. For a CI, it is natural to ask what percentage of replication means will, on average, fall within the interval. For a p value, however, there is no directly corresponding question that can be asked about replication, but a reasonable question may be to ask what is the average percentage of replication means that will fall on the same side of μ_0 as the original sample mean. We refer to this as the average percentage of replication (APR) of a p value. When divided by 100, this is the same as the p_{rep} of Killeen (2005a) and the average probability of replication of Cumming (2005).

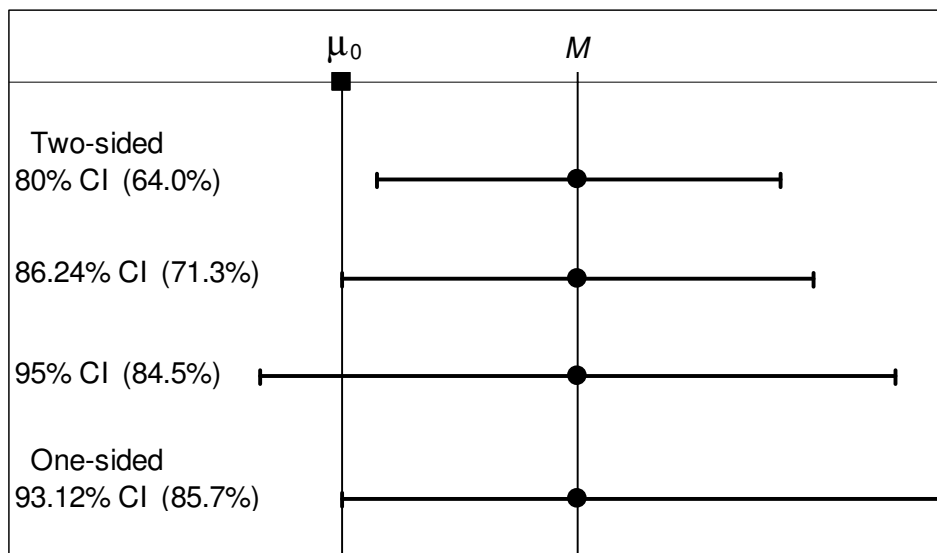


Figure 8. Means and CIs for a single-sample example with $n=20$ for which the t value relative to the null hypothesized value μ_0 is 1.55 (with $df=19$), and one-tailed $p=.069$. The population is assumed normal. The filled dots are the sample mean M . Four CIs are shown. The first is an 80% CI, which does not contain μ_0 ; the second is an 86.24% CI, whose lower limit is exactly μ_0 ; and the third is a 95% CI, which contains μ_0 . For each CI, C is shown, and the percentage in parentheses is the average CP for that interval, calculated using Equation 4. The average CP for the one-sided interval is also, by definition, the average percentage of replication means (APR, see text) that will fall on the same side of μ_0 as the original sample mean M .

Figure 8 illustrates how our earlier analysis of the average CP of a CI, for the normal, σ not known case, can give expressions for APR. The average CP for the 95% and 80% CIs illustrated can be calculated from Equation 4, giving values of 64.0% and 84.5% respectively, as indicated in Figure 8. Now consider the two-sided CI whose lower limit equals μ_0 (second CI from the top), for which $C = 100 \times (1 - 2 \times .069) = 86.2$, because the one-tailed p value is .069 (Cumming & Finch, 2005). Equation 4 gives the average CP for this CI as 71.3%. Now consider the one-sided CI whose lower limit equals μ_0 (the bottom CI in Figure 8): It follows from symmetry that the average CP for this interval is $50+71.3/2=85.7\%$. Now, by the definition of

APR, the average CP for such a one-sided CI that ends at μ_0 is the same as APR, and so Equations 1 and 4 can be used to calculate APR for any one-tailed p value. Table 2 shows APR values for selected p values, for σ not known and $n=10$ and 20 using Equation 4, and for σ known using Equation 1. Note that APR varies relatively little with n , at least down to $n=10$, except possibly for very small p values. Our analysis here extends that of Cumming (2005) to the σ not known case, as well as bringing together the discussion of replication in relation to CIs and p values.

The average CP for a CI is an average over a wide range of CP values. Similarly, the percentages in Table 2 are averages over a wide range of values. Consider a case in which the observed sample mean is greater than μ_0 and we calculate one-tailed $p=.025$. Table 2 gives APR=92%. This does not of course mean that 92% of replications will give means that are statistically significantly larger than μ_0 . However, nor does it mean that, if we repeated our experiment—with our particular value of μ —many times, 92% of the replications would give effects in the same direction as our original experiment. In our particular situation (our particular μ) more, or less—perhaps markedly less—than 92% of replication effects would be in the original direction, as is illustrated for the σ known case by the simulation described by Cumming (2005). What it does mean is that, averaged over all situations (all possible μ values) that give $p=.025$, 92% of replication means would be greater than μ_0 . Alternatively, we could say that, over a lifetime of research in which $p=.025$ is obtained in a variety of situations, averaged over all those situations (all those μ values) 92% of replication effects would be of the same sign as the original effect. The relationship between p values and average replication percentages needs to be interpreted with care, but it does give a quantitative response to the question of whether p values give information about replication (Sohn, 1998).

Table 2. Average percentage of replication means (APR) that will fall on the same side of μ_0 for selected values of one-tailed p .

One-tailed p	.5	.25	.1	.05	.025	.01	.005	.0025	.001
APR, $n=10^a$	50	68.4	82.3	88.6	92.8	96.1	97.6	98.6	99.3
APR, $n=20^a$	50	68.4	82.0	88.2	92.2	95.6	97.1	98.2	99.0
APR, σ known ^b	50	68.3	81.8	87.8	91.7	95.0	96.6	97.6	98.6

Notes ^aAssuming a normal population, σ not known, and using Equation 4 as the basis for APR calculation.

^bAssuming a normal population, and using Equation 1 as the basis for APR calculation.

Figure 2 and our discussion of the average CP for CIs emphasized that different values of $(M_1-\mu)$, the distance between the original sample mean and the population mean, give different values of CP. We spoke of different values of $(M_1-\mu)$ in terms of different original CIs (different M_1 values) for a particular μ . We can, however, equally speak, at least for the σ known case, of various μ values that give a particular sample mean M_1 , and then the average CP is the average over all μ values that can give our particular M_1 value; the average is weighted by the likelihood that a μ value gives M_1 . In the case of average CP of a CI we have these two ways to speak of

the variation of $(M_1 - \mu)$: Regard it as variation in M_1 for an assumed fixed μ , or as reflecting numerous possible μ values for a given M_1 . By contrast, when considering the APR for a p value we have only the second, possibly less natural, option: The average percentage of replication outcomes in the original direction (i.e., APR) is an average over all possible μ values, the weights for the average being the likelihoods that the various μ values give the originally observed p value (Cumming, 2005). It is not surprising, therefore, if understanding the APR for a p value is more challenging than appreciating the average CP of a CI, even if Figure 8 shows how the two can be related.

Discussion

The CLM is the misconception that a 95% CI will capture 95% of replication means. In fact on average only 83.4%, just 5 out of 6, of replication means will fall within an original 95% CI, but the distribution of CP values has strong negative skew. We offer five images that may assist understanding. Figure 1 shows 95% CIs on the means of 20 independent samples and contrasts capture by a CI of μ , and the generally less frequent capture by the CI of the next replication mean. Figure 2 shows three example CIs to illustrate how two sources of variability contribute to the way replication means fall in relation to a CI: the variation of the original mean about μ and, independently, the variation of the replication mean about μ . Figure 2 thus illustrates why average CP is less than C and why CP varies so widely over different CIs. Figure 5 shows 95% CIs on 20 replication means, for samples with $n=10$, and marks the CP value for each interval. The variability in these CP values is especially marked when σ is unknown, as here, and n is small, because CIs vary in width from sample to sample. Figure 4 displays the probability distribution of CP and shows its marked negative skew. Figure 7 displays this distribution for σ unknown, and shows that in this case it generally has especially large variance, even for large values of n .

Our simulation results suggest that the above conclusions may hold quite generally, including down to $n=10$ for σ not known, and for robust CIs with a range of strongly nonnormal populations. To make the implications concrete, suppose we assume a normal population and calculate from a dataset without assuming σ known that the 95% CI is (27.5, 44.1). The CP for this interval comes from one of the distributions in Figure 7, depending on n . It is most likely our CI will capture around 85 to 95% of future replication means, but there is about a 1 in 4 chance that its CP is less than 80%, and a 1 in 20 chance it is an extreme interval with CP less than 50%. For any particular interval, such as (27.5, 44.1), we never know CP—unless we can actually repeat the experiment many times—but we do know that over a lifetime of calculating 95% CIs in many situations the above probabilities will hold.

Cumming and Finch (2005) followed Cleveland (1994) in preferring CIs to SE bars, but SE bars are often shown in figures, so need to be considered. As mentioned earlier, for n at least 10, SE bars can be regarded as giving, approximately, a 68% CI. The average CP is 52%, but in a particular case there is about a 1 in 4 chance that the CP is less than 40%, and a 1 in 20 chance it is an extreme interval with CP less than 17% (a chance of only 1 in 6 of capturing a replication mean). Also, there is on average a .16 chance that the SE bars for a replication will not overlap the original SE bars.

Froman and Shneyderman (2004) also explored the range of likely outcomes when an experiment is replicated, and concluded that these “span an impractically large range, rendering the post hoc analysis of replicability without merit” (p. 365). We agree that variability over replication is large and that a single study often gives insufficient information to warrant any

highly confident conclusion, but we suggest that understanding replication may have merit by offering a valuable perspective on the extent and implications of sampling variability. Replication is central to the scientific method, and we should not turn a blind eye to it just because it makes salient the inherent uncertainty of a single study.

CIs do indeed give information about replication, as do p values, but the value of average CP or APR must be used with care, because they are averages of distributions that have considerable variance and are highly skewed. Our results emphasize that the extent of variation from replication to replication is high—perhaps for many people surprisingly so—and that for a minority of CIs the percentage of replication means captured will be especially low. This has to be our main conclusion: Sampling variability is so large and influential that repeating an experiment can easily give results that differ considerably from the initial results.

Note also that we have studied only the random sampling component of the variation over replication. In any practical situation, it is almost inevitable that variation from experiment to experiment will be even greater, because there are likely to be some changes in experimental context, fine details of the procedure, or the population from which participants are sampled.

We suspect that further exploration of the average CP of CIs, and especially of the APR corresponding to p values, is needed. There are likely to be interesting links to a Bayesian analysis, as Killeen (2005a, b) claimed, and two of the comments on his article debated (Doros & Geier, 2005; Macdonald, 2005). Additional challenges are to find better ways to illustrate and explain the concepts and, most fundamentally, to assess the extent to which they may give greater insight into the drawing of inferences from data. Given the well-known shortcomings of the dichotomous decision making of statistical significance testing, can approaches based on replication give inference practices that will be better understood and enable better research decision making (Killeen, in press)?

References

- American Psychological Association. (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and other error bars. *Psychological Methods, 10*, 389-396.
- Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.
- Cleveland, W. S. (1994). *The elements of graphing data*. Summit, NJ: Hobart Press.
- Cumming, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science, 16*, 1002-1004.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 530-572.
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170-180.
- Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*, 299-311.
- Doros, G., & Geier, A. B. (2005). Probability of replication revisited. Comment on "An alternative to null-hypothesis significance tests". *Psychological Science, 16*, 1005-1006.
- Estes, W. K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review, 4*, 330-341.

- Fisher, R. A. (1966). *The design of experiments* (8th ed.). Edinburgh: Oliver & Boyd.
- Froman, T., & Shneyderman, A. (2004). Replicability reconsidered: An excessive range of possibilities. *Understanding Statistics*, 3, 365-373.
- Killeen, P. R. (2005a). An alternative to null hypothesis significance tests. *Psychological Science*, 16, 345-353.
- Killeen, P. R. (2005b). Replicability, confidence, and priors. *Psychological Science*, 16, 1009-1012.
- Killeen, P. R. (in press). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review*.
- Kline, R. B. (2004). *Beyond significance testing. Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.
- Macdonald, R. R. (2005). Why replication probabilities depend on prior probability distributions. A rejoinder to Killeen (2005). *Psychological Science*, 16, 1007-1008.
- Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, 17, 857-872.
- Newcombe, R. G., & Altman, D. G. (2000). Proportions and their differences. In D. G. Altman, D. Machin, T. N. Bryant, & M. J. Gardner (2000). *Statistics with confidence: Confidence intervals and statistical guidelines* (2nd ed.) (pp. 45-56). London: British Medical Journal Books.
- Sohn, D. (1998). Statistical significance and replicability. Why the former does not presage the latter. *Theory & Psychology*, 8, 291-311.
- Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, 24, 83-91.
- Tukey, J. W., & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample. Trimming/Winsorization 1. *Sankhya A*, 25, 331-352.
- Wichman, B. A., & Hill, J. D. (1982). Algorithm AS183: An efficient and portable pseudo-random number generator. *Applied Statistics*, 31, 188-190.
- Wichman, B. A., & Hill, J. D. (1987, March). Building a random-number generator. *BYTE*, pp. 127-128.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, 209-212.

Author Note

Geoff Cumming, School of Psychological Science, La Trobe University. Robert Maillardet, Department of Mathematics and Statistics, The University of Melbourne.

This research was supported by the Australian Research Council. We thank Fiona Fidler for comments on a draft.

Figures 1, 2, 3 and 5 are derived from a component of ESCI (“ess-key”; Exploratory Software for Confidence Intervals), which runs under Microsoft Excel. This component of ESCI may be downloaded, for personal use without cost, from www.latrobe.edu.au/psy/esci

Correspondence about this article may be addressed to Geoff Cumming, School of Psychological Science, La Trobe University, Melbourne, Australia 3086, or Robert Maillardet, Department of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia 3010. Email: G.Cumming@latrobe.edu.au, or rjmail@unimelb.edu.au.