UNDERSTANDING REPLICATION: CONFIDENCE INTERVALS, *p* VALUES, AND WHAT'S LIKELY TO HAPPEN NEXT TIME

Geoff Cumming
La Trobe University, Australia
g.cumming@latrobe.edu.au

Science loves replication: We conclude an effect is real if we believe replications would also show the effect. It is therefore crucial to understand replication. However, there is strong evidence of severe, widespread misconception about p *values and confidence intervals, two of the main statistical tools that guide us in deciding whether an observed effect is real. I propose we teach about replication directly. I describe three approaches: Via confidence intervals (What is the chance the original confidence interval will capture the mean of a repeat of the experiment?); Via* p *values (Given an initial* p *value, what is the distribution of* p *values for replications of the experiment?): and via Peter Killeen's '*p_{rep}*,' which is the average probability that a replication will give a result in the same direction. In each case I will demonstrate an interactive graphical simulation designed to make the tricky ideas of replication vividly accessible.*

> "Confirmation comes from repetition.… Repetition is the basis for judging… significance and confidence." (Tukey, 1969, pp. 84-85)
> "Given the problems of statistical induction, we must finally rely, as have the older sciences, on replication." (Cohen, 1994, p. 1002)

REPLICATION IS AT THE HEART OF SCIENCE; WE SHOULD TEACH IT EXPLICITLY

Considering whether an effect is replicable is at the heart of drawing inferences from data. In science, one of the strongest reasons for belief an effect is real is evidence that it can be observed again and again. There is, however, strong evidence (Tversky ad Kahneman, 1971) of widespread misconception about replication, especially the severe under-estimation of the extent and consequences of sampling variability. I propose we should teach explicitly about replication and, in particular, help students gain accurate intuitions about the extent of sampling variability.

There are two main reasons why replication is valuable. First, it reduces the chance an observed effect was a fluke, caused merely by sampling variability. Second, and importantly, any replication is a little different from the original experiment: It occurs at a different time, and may involve different researchers in a different place, perhaps with small differences of procedure and context. If an effect replicates, these differences provide some evidence of its robustness and generality. In this article I consider only the first—sampling variability—aspect of replication; in practice, therefore, variation over replication is likely to be even greater than in my analyses. I assume all replications are exact repetitions of the original experiment, with a different random sample from the same population, which I assume to be normal.

I describe three ways to picture and think about replication. The first considers what confidence intervals (CIs, Cumming and Finch, 2005) say about a repeat of the experiment. The second considers what a *p* value indicates about a subsequent experiment, and the third is Peter Killeen's (2005) p_{rep}, or average probability of replication. It would also be natural to consider replication in relation to resampling techniques, but I leave that for another day.

GIVEN A CONFIDENCE INTERVAL, WHERE WILL THE NEXT MEAN FALL?

Given a sample mean and 95% CI, what is the probability that a replication gives a mean that falls within the original CI? The correct answer is not .95. Consider Figure 1, which is the familiar picture of intervals given by a sequence of random samples from a single population, mean μ. Figure 1 thus shows a sequence of 20 replications of an experiment. In practice, of course, we never know μ, marked with the horizontal line, and we have only a single mean and CI. We expect in the long run 5% of the CIs not to include μ, and in the figure just one misses μ, the tenth CI, marked with the open symbol. Examine the results one by one and look to see, in each case, whether the CI captures the following mean. In three cases, marked by triangles, the mean falls outside the preceding CI.

The key realisation is that 95% refers to how the intervals fall in relation to the fixed μ, whereas the replication capture probability depends on variation in *both* the original CIs and the replication means. Because replications are independent, the SD of the distribution of differences between successive means is √2 times the SD of individual means. Using this fact, Cumming, Williams and Fidler (2004) showed that a 95% CI captures, on average, 83.4% of replication means: Just 5 out of 6 replication means will, on average, fall within an original 95% CI.

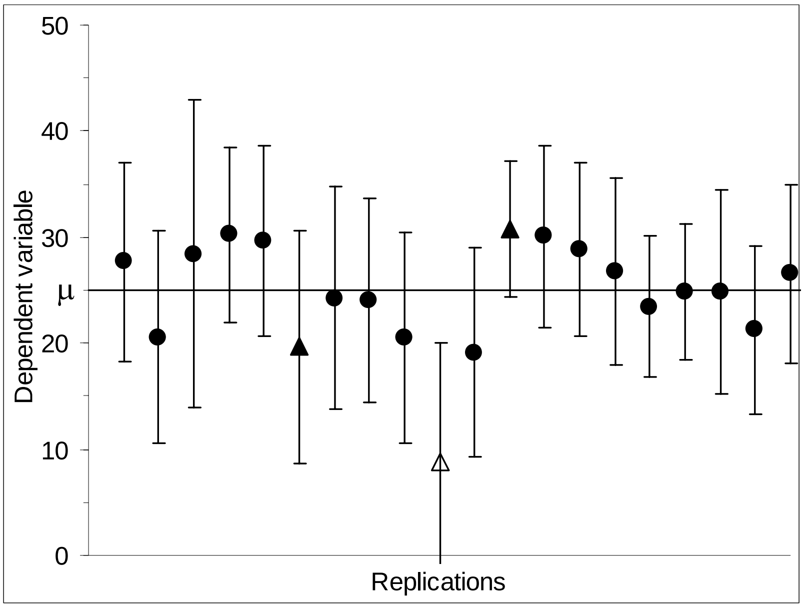

Figure 1: Mean and confidence interval (CI) for 20 replications, based on independent samples of size *n*=20 from a population with mean μ. Means whose CI does not contain μ are shown with an open symbol; means that do not lie within the preceding CI are marked with a triangle

Cumming, Williams and Fidler (2004) reported an internet study in which we obtained responses from authors of articles in leading journals in psychology, behavioural neuroscience, and medicine. Respondents worked with an applet that showed a single mean and CI, and clicked to mark where they thought 10 replication means could plausibly fall. The results, and the answers to questions about how the respondents thought about and performed the task, suggested that a majority—probably a large majority—of leading researchers in those three disciplines hold the *Confidence Level Misconception* (CLM). The CLM is the (erroneous) belief that a *C*% CI will, on average, capture *C*% of replication means. Most researchers seem to believe that a 95% CI will, on average, include about 95% of future means; the correct value is 83.4%. In other words, on average 16.6% of replication means will fall outside a 95% CI, but researchers underestimate this by more than a factor of 3: They believe it is 5%. This misconception can be regarded as yet one more manifestation of the severe under-estimation of the extent and consequences of sampling variability described by Tversky and Kahneman (1971).

Figure 2 shows a simulation revealing the distribution of replication capture percentages. The most recent replication, at the right, captures 96% of means, as indicated by the shaded area under the sampling distribution of means. Note that some means (open circles) do not include μ, and so capture less than 50% of means. Note also that CI width varies because it is calculated from sample SD, not σ. Therefore some CIs, that happen to be long, capture more than 95% of means. The lower panel is a dot plot of the capture percentages for a sequence of 500 replications. This distribution is highly skewed: Most original CIs will capture 90 to 95%, or even more, of replication means. Occasional CIs will capture many fewer, and these are the CIs on original means that happen to fall some way from μ. We know that 5% of original CIs will not include μ because they lie entirely above or below μ, so these CIs will capture less than 50% of replication means. The mean of the 500 values is 86.1 and the median 91.7, values consistent with the theoretical analysis by Cumming and Maillardet (2005). Cumming and Maillardet described this distribution, for σ known and not known, and investigated further aspects of CI capture percentages. For σ known, mean capture percentage is 83.4 and the median is 89.6. For σ not known, and n=20, the mean is 84.5 and median 90.1. My simulation *ESCI Ustanding Stats*, which runs under Microsoft Excel, illustrates CIs, replication and capture probabilities, and is available from www.latrobe.edu.au/psy/esci.

Unless n is very small, SE bars (error bars that are ±1 SE) give approximately a 68% CI, with average capture percentage of 52.1, meaning that on average the next mean has only about a coin toss chance of landing within the original SE bars. Further, there is on average a .157 chance (about 1 in 6) that the SE bars on the next mean will have no overlap with the original SE bars.

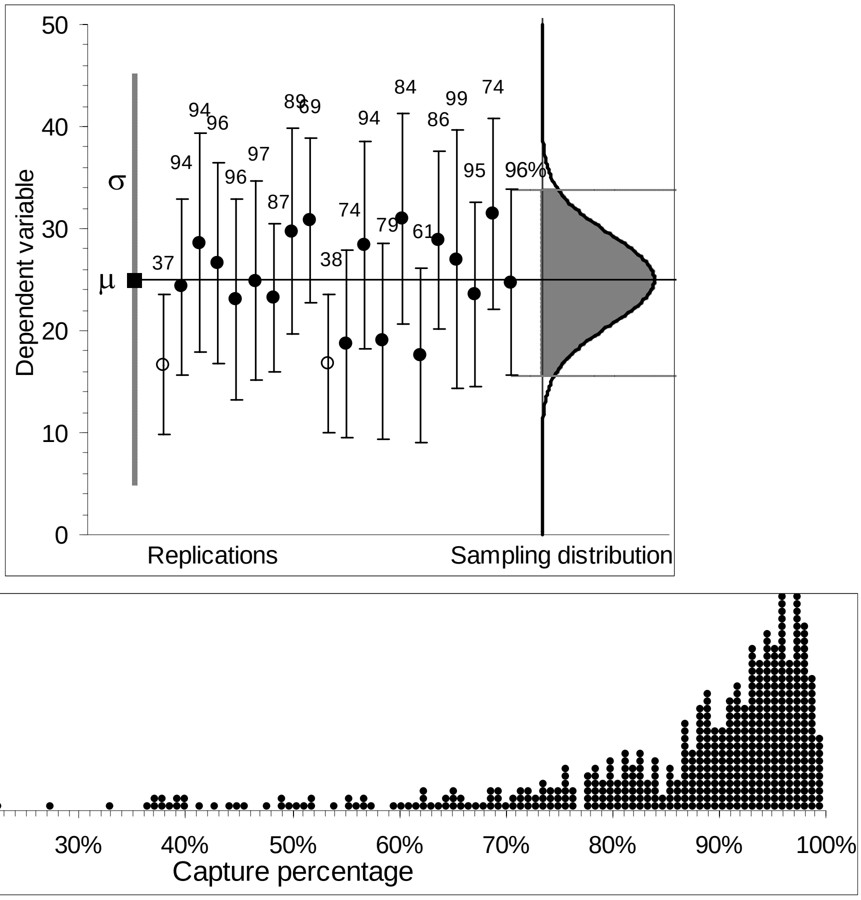

Figure 2: Replication capture percentages

GIVEN A p VALUE, WHAT p VALUE IS LIKELY NEXT TIME?

What does a p value say about replication? What p value is a replication likely to give? Sackrowitz and Samuel-Cahn (1999) investigated the probability distribution of the p value, for a given population effect of δ. (Effect size δ is Cohen's d, measured in population SD units. An effect size of .5 is often regarded as 'medium,' although interpretation of effect size should only be made in a particular context.) I am using the Sackrowitz *et al.* analysis to develop simulations for exploration of p values and, especially, how widely they vary over replication. Figure 3 shows graphs of the probability density function for one-tailed p, single samples of size n=25, σ known, for population effect sizes δ=0 (H_0 true), and δ=.1, .3, and .5. Dotted verticals mark for each δ value the upper 95th percentile of the p value. These curves can be regarded as a generalisation of statistical power functions. The key conclusion is that p varies very widely and so, on replication, p is likely to be quite different from the value initially obtained.

If we assume population effect size δ equals the observed point estimate *d*, we can derive prediction intervals (PIs) for future *p* values. For n=25 and observed d=.3, we calculate one-tailed p=.067 for our sample. If we assume δ=.3 also, then assuming σ known a 50% PI for one-tailed *p* is (.015, .20) and a 95% PI is (0, .56). These intervals may appear astonishingly wide! If we obtain a *p* value of .067, and make the assumption that our point estimate accurately informs us of the population parameter, about all we can be reasonably confident of is that a replication is likely to give a *p* value not greater than .56! Further, it turns out that for σ known these intervals are independent of *n*: However large our experiment, if we calculate from our data that p=.067 (and assume our point estimate is accurate) then the PIs for replication *p* values are as broad as stated.

Once again analysis and simulation demonstrate that we are likely to under-estimate, perhaps drastically, the extent of variation on replication. The conclusion may be that researchers place too much weight on any obtained single *p*, given that an experiment could very easily have given a quite different value. The solution may be to turn to meta-analysis (Cumming, 2006).

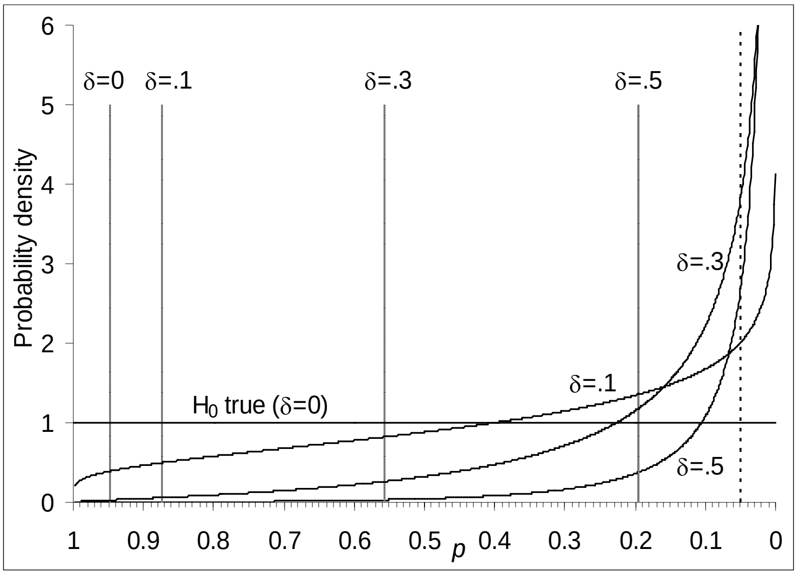

Figure 3: The probability density function of the one-tailed *p* value

PETER KILLEEN'S PROBABILITY OF REPLICATION, p_{rep}

Peter Killeen (2005), in a lead article in *Psychological Science*, argued that *p* values should be replaced as the basis of statistical inference by p_{rep}, the probability that a replication will give a result in the same direction as the original experiment. Killeen's proposal has led James Cutting, the editor of *Psychological Science*, to request all authors to report and use p_{rep} as their basis for inference. The proposal is, not surprisingly, controversial, and three replies, including mine (Cumming, 2005) were published in December 2005.

Suppose an experiment finds an effect of size d_1>0, in population SD units. If we make an assumption about the true population effect δ, it is easy to calculate the probability that d_2, the result of a replication, is also greater than zero. That is the probability of replication (PR) for a population with that δ. For example we could assume, as I did in the preceding section on *p* values, that δ= d_1. Killeen's ingenuity was to find an expression for the replication probability that does not involve δ. The basis for Killeen's analysis was in fact a rediscovery of an analysis by Fisher (1959, p. 114). Killeen's p_{rep} is the average of PR over all values of δ, weighted by the likelihood that each value would have given d_1.

I felt that Killeen did not sufficiently explain the averaging that underlies p_{rep}, and my comment (Cumming, 2005) presents a simulation to illustrate that averaging. It is *ESCI APR Figures*, and is available from www.latrobe.edu.au/psy/esci. It is shown in Figure 4. The starting point is the conventional assumption that d_1 is drawn randomly from a sampling distribution with fixed but unknown mean δ. Focus on $\Delta = d_1 - \delta$, the sampling error of d_1, whose distribution is shown in Figure 4a. Taking a value Δ_K from this distribution yields a value δ_K for the mean of a population that could have given d_1. The sampling distribution for this population is shown in Figure 4b and the shaded area is the PR for that population. (PR is the chance that $d_2>0$ if sampling from that population.) The simulation takes successive values of Δ_K, each of which defines a population that could have given d_1, and each of which gives a PR value. The most recent three cases are shown in Figures 4b, c, and d; the PR values are shown at the right. All PR values are also shown in the dot plot (Figure 4f); the most recent, PR=.718, is the heavy dot.

Killeen's p_{rep} is the average of all possible PR values, and so I refer to it as the *average probability of replication* (APR). Figure 4e is the distribution of d_2 given d_1, based on Killeen's analysis, which gives APR directly as the shaded area under the curve. The mean of the 104 PR values given so far by the simulation is 0.836, not far from the calculated value of 0.892. Note that Killeen's average is based on a highly skewed distribution.

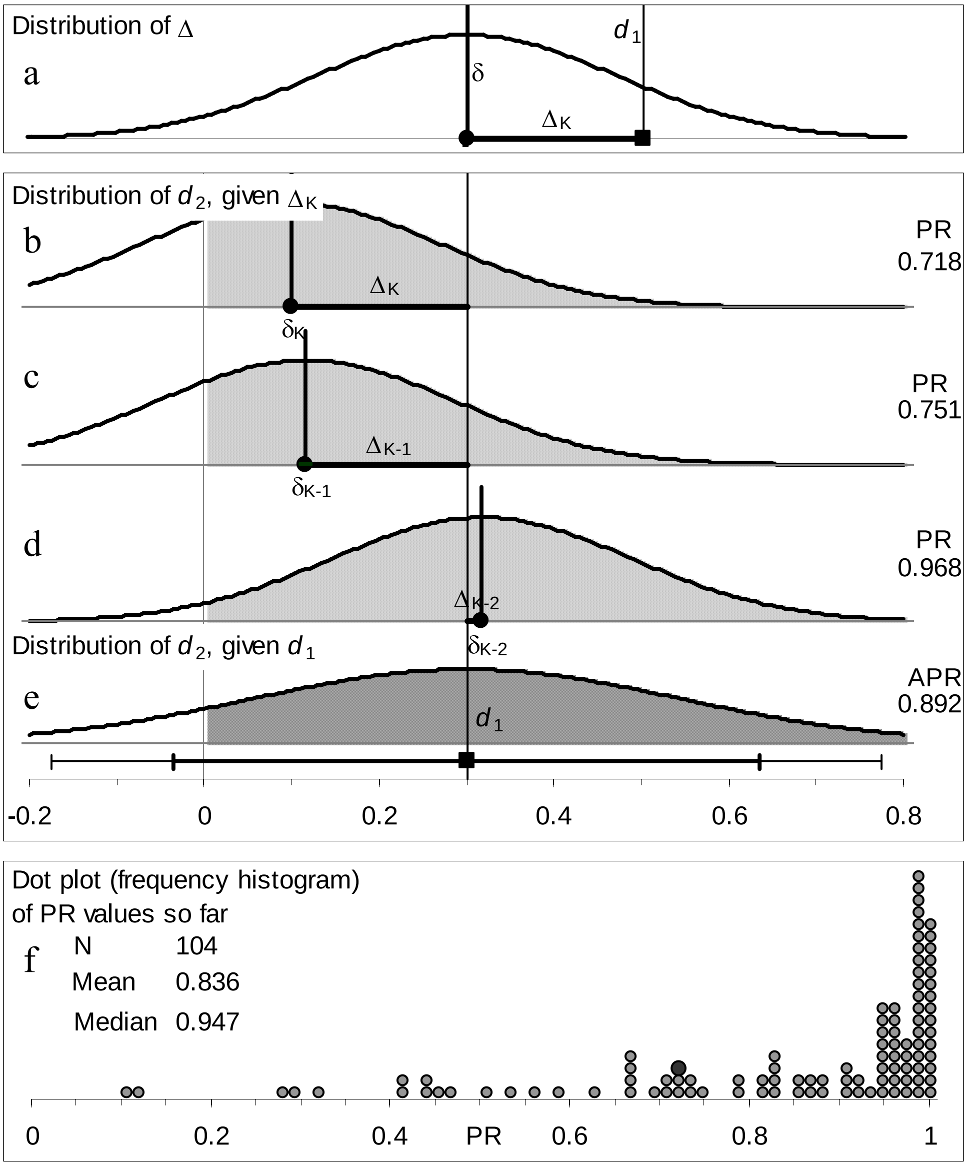

Figure 4: Simulation to illustrate the averaging underlying Killeen's (2005) p_{rep}

In our particular research situation that yielded our observed d_1 (and that has a fixed but unknown δ) we never know the probability of replication (of obtaining a same-sign result next time) unless we can repeat our experiment many times. That is, we never in practice know PR. We can however use Killeen's method to calculate p_{rep} (same as my APR), which is the chance of getting a same-sign result next time 'averaged' over all worlds likely to have given d_1. In most cases (most worlds, most δ_K values) PR will be a little greater than the calculated p_{rep}, but occasionally it will be much less, as the highly skewed dot plot in Figure 4f illustrates.

Killeen's proposal is intriguing, and its implications are yet to be worked out, but it may offer an approach to inference that is natural and intuitive. 'The probability an effect will, on average, be seen again next time' (that's p_{rep}) may be more readily understood and used than (here comes the *p* value) 'the probability that, if there were no effect, we would have obtained this result or something more extreme'!

TEACHING AND UNDERSTANDING REPLICATION

Researchers and students severely underestimate sampling variability, and consequently overestimate power and underestimate the extent a replication result is likely to differ from the original. These misconceptions need to be overcome: I suggest we approach the problem by teaching about replication, a concept central to inference, and to science. My experience explaining CIs and replication encourages me that Figures 1 and 2, and their simulations, will prove useful in teaching. I am planning evaluations of these, and also further development of the simulations of Figures 3 and 4 in preparation for trials with students.

There is a relation between p_{rep} and CI capture of replication means (Cumming, 2005), as similarities between the dot plots in Figures 2 and 4 might suggest. There is also a relation between p_{rep} and *p* values. Sohn (1998) sparked lively debate by arguing that *p* values do not give information about replicability. It is now clear, however, that the three inter-related approaches I have outlined, including *p* values, all give information about replication. Indeed, given an experimental result, we can calculate any or all of a CI, a *p* value, and a p_{rep} value, and each of these can be used to make quantitative statements about what is likely to happen on replication. The challenge now is to find how best to use these developments to enhance statistics education.

ACKNOWLEDGEMENT

I thank Fiona Fidler for comments on a draft.

REFERENCES

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist,* 49, 997-1003

Cumming, G. (2006). Meta-analysis: Pictures that explain how experimental findings can be integrated. In A. Rossman and B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics,* Salvador, Brazil. Voorburg: The Netherlands: International Statistical Institute.

Cumming, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*, 16, 1002-1004.

Cumming, G. and Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist,* 60, 170-180.

Cumming, G. and Maillardet, R. (2005). Confidence intervals and replication: Where will the next mean fall? (Submitted).

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

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference* (2nd edition). Edinburgh: Oliver and Boyd.

Killeen, P. R. (2005). An alternative to null hypothesis significance tests. *Psychological Science,* 16, 345-353.

Sackrowitz, H. and Samuel-Cahn, E. (1999). *P* values as random variables. *The American Statistician,* 53, 326-331.

Sohn, D. (1998). Statistical significance and replicability. Why the former does not presage the latter. *Theory and Psychology,* 8, 291-311.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist,* 24, 83-91

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin,* 92, 105-110.