Published in final edited form as:
Psychol Sci. 2005 December ; 16(12): 1009–1012. doi:10.1111/j.1467-9280.2005.01653.x.

Replicability, Confidence, and Priors

Peter R. Killeen
Arizona State University

Abstract

All commentaries concern priors. In this issue of *Psychological Science*, Cumming graphically demonstrates the implications of our ignorance of δ. Doros and Geier found mistakes in my argument and provide the Bayesian account. Macdonald notes that my program is like Fisher's, Fisher's is like the Bayesians', and the Bayesians' is incoherent. These Commentaries strengthen the foundation while leaving all conclusions intact.

REPLICATING p_{rep}

Cumming reminds us that p_{rep} is an estimate of the probability that a replication with the same power will support the original finding—that it will give an effect of the same sign. The histogram of probabilities of replication (PRs) at the bottom of his Figure 1 is therefore reassuring: All but 6 of the 139 cases have PRs greater than .5: More than 95% of the cases therefore support the original finding. Indeed, because the distribution of PR is negatively skewed, we can generally expect the typical (median) replicability to be better than claimed, as was the case in Cumming's example. In that sense, p_{rep} is a conservative estimate of replicability.

Cumming's real concern is not that a few replications may be victims of sampling error, but that the original experiment might have been a victim. Again, there is consolation to be found in his histogram: By current standards (corresponding to $p_{rep} = .9$), for none of his 139 cases did Δ go far enough in the wrong direction to have supported a decision to publish an unreplicable finding (i.e., in no case was PR < .1). Define *strong evidence* as a p_{rep} greater than p_s. If we set p_s to a relatively liberal .8, the probability that replication of an experiment that provided strong evidence in the first place will provide strong contradictory evidence (the replication's own p_{rep} is greater than *p*s, but the effect is in the wrong direction) is less than .05.[1] Given Cumming's original p_{rep} of .89, approximately 3 of Cumming's cases should have strongly contradicted the original; 2 did so.

Neither p_{rep} nor any other statistic can overcome the probabilistic nature of the relation between evidence and inference. There is no surety, but only the relative safety of numbers, good experimental design, and empirical replication.

I was edified by Cumming's explanation of replication intervals in terms of confidence intervals (CIs). Yet, although everyone agrees on the importance of reporting some measure of effect size, CIs are less than ideal: First, most researchers do not understand what a CI

Address correspondence to Peter Killeen, Department of Psychology, Box 1104, McAllister St., Arizona State University, Tempe, AZ 85287-1104; e-mail: killeen@asu.edu..

[1]The probability of a replicate speaking strongly against an original, where *strong* means the replicate has a p_{rep} of its own of p_S, is 1 −NORMSDIST(NORMSINV(p_S) + NORMSINV(p_{rep})), where NORMSDIST is the standardized normal distribution, and NORMSINV is its inverse. The probability of a replication speaking strongly for an original is 1 − NORMSDIST(NORMSINV(p_S) − NORMSINV(p_{rep})). If we would call the top quartile of *p*reps *supportive*, the bottom quartile *contradictory*, and the rest *ambiguous*, then set p_S as equal to .75.

means (Cumming, Williams, & Fidler, 2004, p. 299). The problem is not confined to psychologists: "A confidence interval is an assertion that an unknown parameter lies in a computed range, with a specified probability [sic]" (Rinaman, Heil, Strauss, Mascagni, & Sousa, 1996, p. 608). Such misunderstanding may be part of the reason why "of the 15 measurements of the Astronomical Unit that [Youden, 1972] presented, not a single one fell within the range of the possible values given by its immediate predecessor" (Stigler, 1996, p. 780)—or at least may be a reason for the bemusement that attends such observations. Second, as Fidler, Thomason, Cumming, Finch, and Leeman (2005) noted, "what to construct CIs around—and how to display them—remain issues for debate" (p. 495). Third, CIs are an impure measure of effect size, because they invoke a sampling distribution to set the relation between level and interval (May, 2003), and that is an easily avoided source of error: Just use *d* or *r*.

If there is "still much to learn about confidence intervals" (Fidler et al., 2005), there is fortunately much less to learn about replication intervals: Calculate the standard error, center it over the statistic, and the long-run probability of a replication falling within those limits (Cumming's average probability of capture, or APC) is approximately 50%. Perhaps it is time to start explaining the complicated in terms of the simple.

Cumming's table, figures, and Web site should help readers to understand this alternative to null-hypothesis significance testing, as his insightful and encouraging comments helped me to understand it in the first place.

ERROR AND CORRECTION

I arrived at p_{rep} by conditioning on the unknown δand integrating it out, assuming flat priors. This is also how Cumming simulated his PRs. I recognized this to be tantamount to a convolution and took the variables I was differencing to be the sampling errors of the original and replicate. But as Doros and Geier show, my reduction of the argument to $d'_2 = d'_1 - \Delta 1 + \Delta 2$, although correct in any particular case, does not give the expected value of d'_2 Their fourth proposal (B2) provides the Bayesian route to my result. Treat σ^2_δ as a prior and divide the numerator and denominator of their equation leading to Equation 4 by σ^2_δ If knowledge of μ*c* is vague or *n* is large, then $\sigma^2_\delta >> \sigma^2_d$, whereupon their Equation 4 reduces to $P(d'_2>0|d'_1)=\int_{-\infty}^{\mu_c/\sigma_c} N(0,1)$, with $\mu_c \approx d'_1$ (its maximum likelihood estimate), and $\sigma_c=\sigma_{d_R} \approx \sqrt{2}\sigma_d$, just as in my original report (Killeen, 2005).

I did not use σ^2_δ as a prior but as the variance of the hyper-parameter δ_j for the reference population of experiments *j*, and should have subscripted it as $\sigma^2_{\delta_j}$ (see the appendix for errata and further discussion of priors). Then my Equation 7, written as $\sigma_{d_R} = \sqrt{2\left(\sigma^2_{d_i}+\sigma^2_{\delta_j}\right)}$, is correct. As a realization variance, $\sigma^2_{\delta_j}$ represents the divergence of different populations of subjects, measurements, or operations, and approaches zero only for identical replications.

PRIOR IGNORANCE

Macdonald argues that the distribution of replicate effect sizes may be derived either from Fisher's fiducial arguments or from Bayesian analyses, but that the former are invalid and the latter incoherent. Viable interpretations of Fisher's arguments reduce to a Bayesian model, such Doros and Geier's, with uniform priors on the location parameter δ_j. Seidenfeld (1979, p. 131) blamed Fisher's failure on the difficulty in formulating uninformative priors that were invariant over arbitrary transformations of the variables. But such invariance is a

useless luxury for scientists. Most of the inferential statistics we use depend on the additivity of random variables, and those remain additive only under linear transformations. If simple reaction times are normally distributed on *log*(*t*), then *log*(*t*), not *t*, is the scale on which to express priors. Such measurement constraints,[2] long dismissed by statisticians (Hand, 2004), de-mark the boundaries within which Fisher's fiducial probabilities and Bayesian inferences are both valid and coherent.[3] Statistics lose their authority to the extent that the variables and their transformations depart from linear comparability; their justification then must be found in their less principled, but often considerable, pragmatic utility.

Priors

Statistics can address three different types of questions (Royall, 1997):

- What should I believe?
- What should I do?
- How should I evaluate this evidence?

The first question requires Bayesian updating of priors to incorporate new data. If the priors are subjective, Bayesian analysis is "a code of consistency for the person applying it, not a system of predictions about the world around him" (Savage, 1972, p. 59, who nonetheless took personal probability as "the only probability concept essential to science," p. 56). If the priors are objective, Bayesian updating is the tool of choice for secondary meta-analysis, and provides the machinery for a cumulative science. Had the astrophysicists cited by Youden (1972) incorporated priors in their final parameter estimates, there would have been less humor and more truth in the title of his article. Scientists wanting to know what to believe about claims—their own or others'—should respect prior information (Field, 2003). After Bayesian updating, p_{rep} provides an excellent prognostic.

Neyman and Pearson avoided the Bayesian implications of the first question by skipping to the second, asserting that a counsel to action carries no implications for belief (Neyman, 1960, p. 290). But an answer to the second question requires both efficient use of the data—not possible in their schema—and a payoff matrix. By providing the first, p_{rep} lays the groundwork of a decision theory for scientific inference.

The standard answer to the third question is that results should be evaluated by classifying them as either significant or nonsignificant. But this approach "is an impoverished, potentially misleading way to describe evidence" (Dixon, 2003, p. 200; J.E. Hunter, 1997). Given the typical case of a composite alternative hypothesis (e.g., "not the null"), p_{rep} predicts the probability that replications will provide evidence supporting the original effect. Given well-defined alternative hypotheses, likelihood analysis (Royall, 1997), corrected for bias (Forster & Sober, 2004), estimates the strength of evidence favoring the alternatives. If additional statistical evaluation is wanted, randomization of the constituent log likelihoods will provide empirical sampling distributions from which p_{rep} may be inferred. In either case, priors "can obfuscate formal tests by including information not specifically contained within the experiment itself" (Maurer, 2004, p. 17); they flavor the evidence with the idiosyncratic taste of the evaluator. Flat (uninformative) priors provide the level playing field necessary for unbiased evaluation. After evidence passes a filter such as p_{rep}, it may be

[2]Seidenfeld's (1979) "smoothly invertible canonical pivotal variables" concisely embody the necessary constraints, but he considered those too restricting. The issues are subtle; consult Macdonald's references in this issue of *Psychological Science* and Seidenfeld's (1979) book. Note, however, that Seidenfeld's paragon estimation of the volume of a cube from weights of capacity and side does not survive dimensional analysis; the weight of his ruler is a volumetric measure and should be added, not cubed.

[3]Transformation techniques permit nonlinear transforms by appropriately warping one of the scales; for statistical utility, the scale on which the central-limit theorem holds should be treated as privileged.

weighted and added to the canon. Belief is best constructed from independently established facts, composed with an eye to their cumulating effect.

Supernatural Paradoxes

If we knew that $\delta = 0$, as in Macdonald's example, then the probability of a positive effect in replication would be .50, no matter what p_{rep} predicts. But Macdonald assumes supernatural knowledge; p_{rep} does not. Individual experiments do not establish parameters; meta-analyses converge on parameters. To know what to believe, enter all relevant information into that inferential engine. To know what research to advise students to undertake, attend to priors. To evaluate experimental results, however, use p_{rep}, unflavored. It comes with the proviso of *ceteris paribus*, and its doubled variance allows for sampling error in the original and the replicate.

Doros and Geier conclude that, because p_{rep} can be calculated[4] from p, it inherits the shortcomings of null-hypothesis significance testing. Wrong. These statistics, although informationally equivalent, are distinguished by the inferences they warrant; p_{rep} is a valid posterior predictive probability, p is not. That is precisely why Fisher pursued the fiducial argument, which, absent measurements on interval scales, is unattainable. With linearity, "selection of an 'ignorance' prior can be made without fear of violating the probability calculus" (Seidenfeld, 1979, p. 133).

THE REFERENCE SET FOR p_{rep}

Much of my discussion thus far is, in the end, irrelevant to most readers of this article. Virtually all psychological data are observational or are drawn from convenience samples, subsets of which are randomly assigned to control or experimental conditions. These standard empirical procedures are incompatible with the normal statistical models, which assume random sampling from a reference set or population (Lunneborg, 2000). Randomization tests emulate our experimental operations (Byrne, 1993), do not depend on priors, do not depend on the form of the populations sampled, and permit fiducial inference (Pitman, 1937). Their logic is straightforward; M.A. Hunter and May (2003) have provided a clear overview and useful references. The p value from such a test gives the proportion of occasions on which the data would have segregated into such disparate groups (or have been so correlated with a predictor) by chance.[5] The corresponding p_{rep} estimates the probability of replication in samples from the same data set (cf. Pitman's w statistic). It also predicts replicability in general, with its accuracy depending on the similarity of the subjects and procedures in the original and replicate. Permutation tests and p_{rep} respect what we do and tell us what we need to know. They are the right analytic tools for most of our primary research questions.

Acknowledgments

National Science Foundation Grant IBN 0236821 and National Institute of Mental Health Grant 1R01MH066860 supported this work.

References

Bernardo, J.M. (in press). Reference analysis. In D. Dey & C.R. Rao (Eds.), *Handbook of statistics* (*Vol 25*). Amsterdam: Elsevier.

[4]The calculation is as follows: p_{rep} = NORMSDIST((NORMSINV(1 −p))/SQRT(2)).

[5]Permutation tests evaluate any difference in samples. They may be modified to test differences of means (Efron & Tibshirani, 1993).

Byrne, M.D. (1993). A better tool for the Cognitive Scientist's toolbox: Randomization statistics. In W. Kintsch (Ed.), *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 289–293). Mawah, NJ: Erlbaum (Available from http://chil.rice.edu/byrne/pubs.html)

Cumming G. Understanding the average probability of replication: Comment on Killeen (2005). Psychological Science. 2005; 16:1002–1004. [PubMed: 16313666]

Cumming G, Williams J, Fidler F. Replication and researchers' understanding of confidence intervals and standard error bars. Understanding Statistics. 2004; 3:299–311.

Dixon P. The *p*-value fallacy and how to avoid it. Canadian Journal of Experimental Psychology. 2003; 57:189–202. [PubMed: 14596477]

Doros G, Geier AB. Probability of replication revisited: Comment on "An alternative to null-hypothesis significance tests. Psychological Science. 2005; 16:1005–1006. [PubMed: 16313667]

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap* London: Chapman & Hall.

Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Still much to learn about confidence intervals. Psychological Science. 2005; 16:494–495. [PubMed: 15943677]

Field AP. The problems in using fixed-effects models of meta-analysis on real-world data. Understanding Statistics. 2003; 2:105–124.

Forster, M., & Sober, E. (2004). Why likelihood? In M.L. Taper & S.R. Lele (Eds.), *The nature of scientific evidence: Statistical, philosophical, and empirical considerations* (pp. 153–190). Chicago: University of Chicago Press.

Hand, D.J. (2004). *Measurement theory and practice* New York: Oxford University Press.

Hunter JE. Needed: A ban on the significance test. Psychological Science. 1997; 8:3–7.

Hunter MA, May RB. Statistical testing and null distributions: What to do when samples are not random. Canadian Journal of Experimental Psychology. 2003; 57:176–188. [PubMed: 14596476]

Killeen PR. An alternative to null-hypothesis significance tests. Psychological Science. 2005; 16:345–353. [PubMed: 15869691]

Lee MD, Wagenmakers EJ. Bayesian statistical inference in psychology: Comment on Trafimow (2003). Psychological Review. 2005; 112:662–668. [PubMed: 16060758]

Lunneborg, C.E. (2000). *Data analysis by resampling: Concepts and applications* Pacific Grove, CA: Brooks/Cole/Duxbury.

Macdonald RR. Why replication probabilities depend on prior probability distributions: A rejoinder to Killeen (2005). Psychological Science. 2005; 16:1007–1008. [PubMed: 16313668]

Maurer, B.A. (2004). Models of scientific inquiry and statistical practice: Implications for the structure of scientific knowledge. In M.L. Taper & S.R. Lele (Eds.), *The nature of scientific evidence: Statistical, philosophical, and empirical considerations* (pp. 17–50). Chicago: University of Chicago Press.

May K. A note on the use of confidence intervals. Understanding Statistics. 2003; 2:133–135.

Neyman, J. (1960). *First course in probability and statistics* New York: Holt, Rinehart and Winston.

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics: Vol. 2B: Bayesian inference* (2nd ed.). New York: Oxford University Press.

Pitman EJG. Significance tests which may be applied to samples from any populations. Supplement to the Journal of the Royal Statistical Society. 1937; 4:119–130.

Rinaman, W.C., Heil, C., Strauss, M.T., Mascagni, M., & Sousa, M. (1996). Probability and statistics. In D. Zwillinger (Ed.), *CRC standard mathematical tables and formulae* (30th ed., pp. 569–668). Boca Raton, FL: CRC Press.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm* London: Chapman & Hall.

Savage, L.J. (1972). *The foundations of statistics* (2nd ed.). New York: Dover.

Seidenfeld, T. (1979). *Philosophical problems of statistical inference: Learning from R. A. Fisher* London: D. Reidel.

Stigler SM. Statistics and the question of standards. Journal of Research of the National Institute of Standards and Technology. 1996; 101:779–789.

Youden WJ. Enduring values. Technometrics. 1972; 14:1–11.

APPENDIX

Errata

S. Sirois (personal communication, May 10, 2005) noticed that the standard error of replication on p. 347 in my original article should have been $\sigma_{d_R} = \sqrt{2}\sigma_d$. The second variance under the radical in Equation 7 should have been $\sigma^2_{\delta_j}$. An unembellished *d*, as used by Cumming, simplifies notation.

Flat Priors

Bayesians recommend either Jeffrey's priors (Lee & Wagenmakers, 2005, have provided an excellent Bayesian tutorial) or reference priors (Bernardo, in press). The Jeffrey's prior for the mean of normally distributed data is uniform. Alas, over an infinite range, that leaves any particular prior equaling an unproductive zero. But this is not a problem if the range is merely huge (e.g., spread with $\sigma^2 \approx 10^{10}$), as the prior's influence will then fall below the measurement error of rational data. "If prior information is genuinely weak relative to the data, the posterior distribution should be robust to any *reasonable* choice of prior distribution [including improper priors]" (O'Hagan & Forster, 2004, p. 107).

Priors that are flat for *d* cannot also be flat for r^2 (Macdonald, this issue). Ignorance has structure. Reference priors cash out that structure against the models tested. The reference prior $\pi(d, \sigma) = (\sigma\sqrt{1+d^2/2})^{-1}$ is relatively flat for the effect sizes and variances involved whenever statistical analysis is deemed necessary. For the range of effect sizes that concern psychologists, whether they use Jeffrey's priors or reference priors, *d* or r^2, it is all pretty much Kansas.