Replication probabilities and prep. Online Supplement 3 to Serious stats: A guide to advanced statistics for the...

Chapter · January 2012

CITATIONS
0

READS
25

1 author:

Thom S Baguley
Nottingham Trent University
64 PUBLICATIONS **889** CITATIONS

SEE PROFILE

Online Supplement 3
Replication probabilities and p_{rep}

This supplement draws primarily on Chapters 2, 3, 4 and 11.

OS3.1 Replication probabilities

A number of researchers have attempted to estimate the probability of replicating an effect. Greenwald *et al*. (1996) tried to estimate the probability that an identical replication of a statistically significant finding would be statistically significant. Their approach begins with the assumption that the original effect is indicative of the true effect. This assumption is hard to justify and, as Greenwald *et al*. (*ibid*.) note, leads to overestimates of replicability. Attempts to extend this approach have attracted heavy criticism (Macdonald, 2005; Froman and Schneyderman, 2004). Froman and Schneyderman's analysis is particularly insightful because they show that replication probability (in the sense of obtaining statistical significance in an identical study) is a relabeling of *post hoc power*. Treating the sample statistics as if they are population parameters (rather than estimates of population parameters) inevitably ignores some of the uncertainty in the sample and leads to spurious certainty in the estimates of replication probabilities. Froman and Schneyderman (2004) argue that this makes replication probabilities unusable in practice.

Recent work by Killeen (2005) attempts to avoid some of these problems. His p_{rep} statistic has been proposed as an alternative to a conventional p value. A central feature of Killeen's approach is to adopt a definition of replication that involves obtaining an effect in the same direction as the original study. Thus p_{rep} is the probability that an identical replication obtains an effect with the same sign. Figure OS3.1 shows the sampling distribution of a non-zero standardized mean difference (labeled δ_1) in relation to its expected value under a null hypothesis of no effect.

Killeen (*ibid*.) proposed that the probability of an identical replication equates to the shaded area in this figure. A new effect sampled from the same population as δ_1 (i.e., an identical

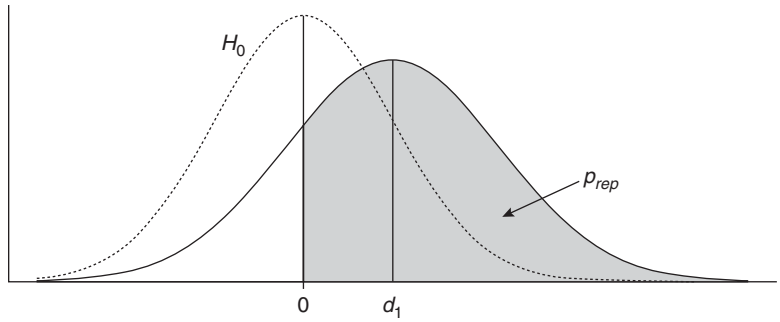

Figure OS3.1 Replication probability (p_{rep}) in relation to the sampling distribution of the null hypothesis ($\delta = 0$) and the true effect ($\delta = \delta_1$)

replication of the original study) should have the same sign as the original effect with probability equal to the area of this region.

The precise value of the replication probability depends on δ_1 (which is unknown). Following Greenwald *et al.*, Killeen argued that it can be estimated if the sampling distribution of δ_1 is normal. The crucial quantity required to estimate p_{rep} is the variance of this sampling distribution. Killeen employs an approximation for the sampling variance of a standardized mean difference derived by Hedges and Olkin (1985):

$$\hat{\sigma}_{\delta}^2 = \frac{N^2}{n_1 n_2 (N-4)} \qquad \text{Equation OS3.1}$$

This approximation works well for effects sampled from a normal distribution provided δ is not too large (e.g., its absolute value is ≤ 1). Killeen (2005) argues that because the sampling distribution of δ_1 combines variability from both the observed effect and the replication attempt, it can be estimated as twice this value (i.e., $\hat{\sigma}_{rep}^2 \approx 2\hat{\sigma}_{\delta}^2$). Determining p_{rep} involves calculating the area of the shaded region of the δ_1 curve in Figure OS3.1. As δ_1 is a standardized mean difference (estimated from $\hat{\delta}$ or g) this can be accomplished using a standard normal distribution where $z = \hat{\delta}/\hat{\sigma}_{rep}^2$.

Because the smallest probability of obtaining the same sign in a replication (if there is no observed effect whatsoever) is one half, it follows that $.5 \leq p_{rep} \leq 1$. Unlike p, it only makes sense to calculate p_{rep} for effects with one degree of freedom (df), and which can therefore be considered directional (e.g., t tests and correlations). Nevertheless, p_{rep} is intimately related to a conventional p value (and can be obtained from it). At one level, p_{rep} is merely a monotonic transformation of p. The argument in favor of p_{rep} is that the transformation produces a statistic that is more meaningful. This conclusion is hotly disputed.

Several methods exist for calculating p_{rep}, however some are problematic (Iverson *et al.*, 2009). Care must be taken even with software that calculates p_{rep}, because it is easy to obtain output less than .5 or equal to one. For values less than .5 (indicating that p_{rep} has been calculated for the wrong tail of the distribution) the correct value should be one minus the calculated value.

Iverson *et al*. (*ibid*.), although highly critical of p_{rep}, provide one of the clearest explanations of its correct calculation. The following formulas take one-sided p values as input. The Iverson *et al*. formula is:

$$p_{rep} = \Phi\left[\frac{\Phi^{-1}\left(\max\{p, 1-p\}\right)}{\sqrt{2}}\right] \qquad \text{Equation OS3.2}$$

The quantity $\max\{p, 1-p\}$ is the larger of the values p or $1 - p$, where p is the one-sided p from a t test, correlation or other 1 *df* test. The function Φ is the familiar cumulative distribution function (*cdf*) of the z distribution. Its inverse Φ_{-1} is the quantile function for z. The formula therefore takes the larger of p or $1 - p$ and uses it to obtain the z score that cuts off the required proportion of the standard normal distribution. This value is divided by $\sqrt{2}$. This last step comes into play because the distribution of the replications incorporates a double dose of uncertainty (from both the original study and its replication). The z distribution has $\sigma = 1$, and doubling the variance will make $\sigma = \sqrt{2}$. Diving by root 2 simply corrects the z score for this extra variability. The final step is to calculate the cumulative probability associated with the corrected z score.

Example OS3.1 A published experiment might report an independent t test as $t(50) = 2.68$, $p = .01$. The corresponding one-sided p value is therefore approximately .005. Subtracting .005 from one gives .995. As $1 - p$ is larger than p, it is entered into Equation OS3.2. The corresponding z score is roughly 2.58 and so:

$$p_{rep} = \Phi\left[\frac{\Phi^{-1}(.995)}{\sqrt{2}}\right] = \Phi\left[\frac{2.58}{\sqrt{2}}\right] = \Phi(1.82) = .966$$

This matches the results of the `p.rep()` function in the R `psych` package, which returns $p_{rep} = .965726$ for one-sided $p = .005$. Care needs to be taken using this function (e.g., it will return 'impossible' p_{rep} values (< .5) if one-sided $p > .5$).

Calculating p_{rep} this way is instructive. First, the calculations are relatively easy (but fiddly). Second, p_{rep} values for statistically significant effects (using the conventional criterion of $\alpha = .05$) range from about .917 to .999. (p_{rep} is a monotonic function inversely related to p if $p < .05$.) This is quite a narrow range. When p_{rep} statistics were widely reported in place of or in addition to p values (e.g., for a few years in the journal *Psychological Science*), values in the range .85 to one were considered evidence of replicability. Its practical impact was largely to replace $\alpha = .05$ with a somewhat looser threshold (as $p_{rep} = .85$ roughly equates to one-sided $p = .075$).

OS3.2 Criticisms of p_{rep}

There are good reasons to be cautious of using p_{rep}. Some of these arise from the practical difficulty of interpreting the statistic. Unlike p, its interpretation is restricted to effects that are directional. This is not a major obstacle, because there are reasons to prefer directional tests and 1 *df* effects. More problematic is the narrow range of p_{rep} values. The narrow range

makes it appear that $p_{rep} = .90$ is not dissimilar to $p_{rep} = .98$. It would be better to represent the 'replicability' in terms of the odds of obtaining an effect with the same sign:

$$O_{rep} = \frac{p_{rep}}{1 - p_{rep}} \qquad \text{Equation OS3.3}$$

Switching to odds makes it clearer that $p_{rep} = .90$ and $p_{rep} = .98$ are not that similar. The corresponding odds of the same sign in an identical replication would be 9 and 49. Odds would also make p_{rep} values close to .5 appear less impressive (e.g., for $p_{rep} = .59$, $O_{rep} = 1.4$).

Another difficulty is the variability of p_{rep} scores. Both those critical of (e.g., Iverson *et al.*, 2009) and those generally supportive of p_{rep} (Cumming and Fidler, 2009) have pointed out that it is, at best, a very imprecise estimate of the true replicability of the direction of an effect. Cumming and Fidler (*ibid.*) demonstrate this by comparing the variability of p_{rep} with that of p and the width of a CI (see Figure 11.2). Furthermore, p_{rep} is closely related to the controversial concept of *post hoc* power (see Maraun and Gabriel, 2010).

However, the strongest attacks on p_{rep} are on theoretical grounds (see Macdonald, 2005; Iverson *et al.*, 2009; Iverson *et al.*, 2010). One criticism in particular is fundamental. Several commentators (e.g., Macdonald, 2005; Iverson *et al.*, 2009) have noted that Killeen's estimate of p_{rep} implicitly makes the assumption that the true size of the effect could take any value. This criticism is related to Bayesian criticisms of NHST reviewed later in the chapter. For most research questions, extremely large true effects (in either direction) are very unlikely.[1] For example, standardized effect sizes such as $\delta = .8$ or 1.5 occur infrequently, but effects such as $\delta = 10$ or $\delta = 100$ are implausible (most published findings probably being in the range $-1 < \hat{\delta} < 1$). In addition, true effects close to $\delta = 0$ may be particularly common (e.g., in experimental research where H_0 is plausible).

The practical impact of this assumption is that the probability of obtaining an effect in the same direction as the original effect is overstated. This is explained in detail by Iverson *et al.* (2009), and happens because p_{rep} under-weights the chance of small effect sizes. It therefore implicitly assumes the true effect is quite a bit larger than it probably is. Assuming the effect is quite large increases the estimate of the probability that a future study will find an effect in the same direction. Iverson *et al.* (2009) argue that p_{rep} provides an estimate of the upper limit of the probability of a same-sign replication. The true value could be substantially lower.

Although p_{rep} is an interesting idea, it requires additional assumptions about the distributions of true effects to get accurate replication probabilities. Even if these assumptions can be justified, the variability of p_{rep} as a statistic (see Figure 11.2) and the compressed range of values it can take make it impractical as a decision tool. Although using replication odds might solve the latter problem, the former problem is inherent in the formulation of the statistic. Recent criticisms of p_{rep} have explored confusion about its precise definition and suggest that it is not an accurate estimate of the true replication probability (Maraun and Gabriel, 2010; Trafimow *et al.*, 2010). There appears to be no simple short cut to obtaining the probability of a replication. Nor is the probability of a replication, whether accurate or inaccurate, an adequate substitute for replicating an effect in a new study (Maraun and Gabriel, 2010; Serlin, 2010).

OS3.3 R code for Online Supplement 3

OS3.3.1 Calculating p_{rep} (Example OS3.1)

The exact one-sided *p* value for a *t* statistic of 2.68 with 50 *df* can be obtained directly from software output or from the `pt()` function:

```
p.obs <- pt(2.68, 50, lower.tail = FALSE)
p.obs
```

This returns the value .004971346. To get p_{rep} from the observed one-sided *p* value it is possible to use the `pnorm()` and `qnorm()` functions:

```
pnorm(qnorm((1 - p.obs))/sqrt(2))
```

The `psych` package also includes a `p.rep()` function. Care needs to be taken, as it will happily return p_{rep} values less than .5. Rounding the *p* value to .005 has little impact on the result:

```
library(psych)
p.rep(p.obs)
p.rep(.005)
```

Proofing the statistic against *p* values greater than .5 is also possible:

```
p.obs <- .62
if(p.obs > .5) p.new <- 1 - p.obs
p.rep(p.new)
```

OS3.3.2 R packages

Revelle, W. (2011) *psych*: Procedures for psychological, psychometric, and personality research. R package version 1.0-95.

OS3.4 Note

1. This assumption of p_{rep} is by no means obvious. At first glance you might think that it assumes a normal distribution of true effect sizes (and thus that effects close to zero are more likely). In fact, it assumes only that the sampling distributions of the observed effect and the replication (reflecting sampling error) are normal. The distribution of true effect sizes is assumed to be uniform, extending from $-\infty$ to ∞.

OS3.5 References

Cumming, G., and Fidler, F. (2009) Confidence Intervals Better Answers to Better Questions. *Zeitschrift für Psychologie*, 217, 15–26.

Iverson, G. J., Lee, M. D., Zhang, S., and Wagenmakers, E.-J. (2009) p_{rep}: An Agony in Five Fits. *Journal of Mathematical Psychology*, 53, 195–202.

Iverson, G. J., Wagenmakers, E.-J., and Lee, M. D. (2010) A Model Averaging Approach to Replication: The Case of p_{rep}. *Psychological Methods*, 15, 172–81.

Froman, T., and Schneyderman, A. (2004) Replicability Reconsidered: An Excessive Range of Possibilities. *Understanding Statistics*, 3, 365–73.

Greenwald, A. G., Gonzalez, R., Harris, R. J., and Guthrie, D. (1996) Effect Sizes and p Values: What should be the Reported and what should be Replicated? *Psychophysiology*, 33, 175–83.

Hedges, L. V., and Olkin, I. (1985) *Statistical Methods for Meta-analysis.* Orlando, FL: Academic Press.

Killeen, P. R. (2005) An Alternative to Null Hypothesis Statistical Tests. *Psychological Science*, 16, 345–53.

Macdonald, R. R. (2005) Why Replication Probabilities Depend on Prior Probability Distributions. *Psychological Science*, 16, 1007–8.

Maraun, M., and Gabriel, S. (2010) Killeen's (2005) p_{rep} Coefficient: Logical and Mathematical Problems. *Psychological Methods*, 15, 182–91.

Serlin, R. C. (2010) Regarding p_{rep}: Comment Prompted by Iverson, Wagenmakers, and Lee (2010); Lecoutre, Lecoutre, and Poitevineau (2010); and Maraun and Gabriel (2010). *Psychological Methods*, 15, 203–8.

Trafimow, D., MacDonald, J., Rice, S., and Clason, D. L. (2010) How Often is Prep Close to the True Replication Probability?. *Psychological Methods*, 15, 300–307.