Leonhard Held is professor of biostatistics at the University of Zurich (UZH) and director of the Center for Reproducible Science at UZH.

Replication power and regression to the mean

If a scientific study reports a discovery with a *p*-value at or around 0.05, how credible is it? And what are the chances that a replication of this study will produce a similarly "significant" finding? **Leonhard Held**, **Samuel Pawel** and **Simon Schwab**'s answers may surprise you

Statistical significance is frequently used to support a claim of scientific discovery. However, the fact that a study yields a statistically significant result does not mean that it found a genuine effect, nor does it mean that repeating the study will necessarily lead to significance again. This is obvious to a statistician, but the fragility of a statistically significant result is still not as widely recognised as it should be.

Consider, for example, the highly cited study by Ackerman *et al.*, which was published in 2010 in *Science*.[1] In one of the reported experiments, participants had to rate curricula vitae (CVs or résumés) of individuals, which led to a remarkable outcome; CVs attached to a relatively heavy clipboard were regarded more favourably than those on a lighter clipboard. This intriguing outcome, moreover, carried a so-called *p*-value of 0.049. By convention, this made the finding (just) statistically significant. A few years later, a large collaboration of researchers led by Camerer attempted to replicate important findings from the social sciences, including this study.[2] Based on the same study design but involving almost 600 participants – 11 times the original – this replication found a far smaller effect size; so small, indeed, that it failed to reach statistical significance ($p = 0.13$).

One may wonder how such a large discrepancy between the original and the replication study could have occurred. One helpful approach to answer this question is the concept of replication power.

A central issue

Having sufficient power to obtain a significant result is a central issue in the design and planning of scientific studies. Power depends on the effect size to be detected, the sample size, and the required significance level (see "*p*-values and significance tests defined"). There are numerous formulas and software packages for calculating the power for a given sample size or the required sample size to achieve a certain power, with power often set to at least 80%.

Suppose an original study provides an estimate (assumed to be normally distributed) of the effect size together with a confidence interval from which the *p*-value can be derived. It is then natural to ask what the power of an identically designed replication study would be if we assume that the effect size found by the original study equals the (unknown) true effect size. This is known as replication power or replication probability. There are in fact two forms of such power: *conditional*, meaning that the statistical uncertainty of the original effect estimate, as represented by its confidence interval, is ignored; and *predictive*, meaning that the uncertainty is incorporated. Either way, the replication power depends essentially only on the original *p*-value and the sample size of the replication study relative to the original study.

Let us take the simple case of a replication that uses the same sample size as the original study.[3] If the original *p*-value was close to the significance level, say 0.05 (or 5%), then both conditional and predictive power turn out to be only 50%. This often shocks many non-statisticians: it means that if the same sample size is used, a borderline statistically significant result is no more likely to be replicated than a coin-toss! To gain some intuition why this is so, think of the *p*-value as a summary measure calculated from the data. If two sets of data have been generated independently but in exactly the same way, the corresponding *p*-values have the same distribution. The probability that one is larger than the other is then 50%. The original *p*-value needs to be as small as 0.005 to have a conditional replication power around the usual standard of 80%. This is shown in Figure 1. But matters are even worse if we take into account standard scientific practice.

Publication bias

The assumption that the original effect estimate is the true effect size is crucial in the calculation of replication power, but often some scepticism is appropriate. If we assume that most effects studied by researchers are small and that the results of studies are more likely to be published if they are statistically significant ("publication bias"), then the statistical phenomenon known as "regression to the mean" predicts that we should actually expect the replication effect estimate to be, on average, smaller than its original counterpart. Regression to the mean is a simple consequence of the fact that extreme observations of random quantities tend to be less extreme, but closer to the population mean, when observed a second time.[4] Thus, if significant original findings have a higher chance of being published, the literature will consist mainly of exaggerated

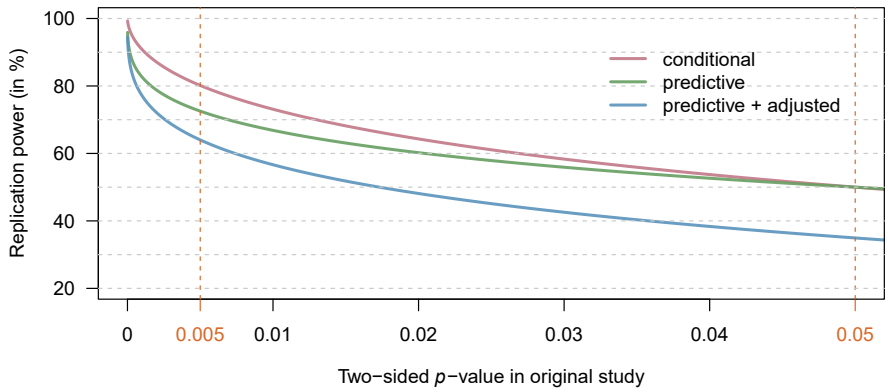

Figure 1: Replication power as a function of the *p*-value of the original study, if the replication sample size is equal to the original one. Shown are conditional power, predictive power, and predictive power adjusted for regression to the mean.

Samuel Pawel is a PhD candidate in biostatistics at UZH.

Simon Schwab is a postdoctoral fellow in the Center for Reproducible Science at UZH.

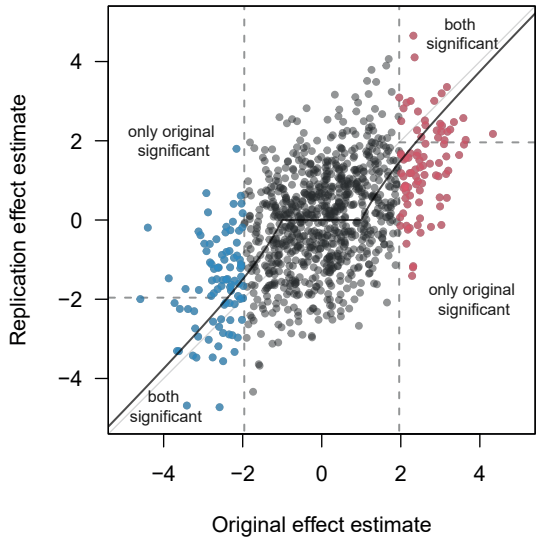

Figure 2: One thousand simulated pairs of original and replication effect estimates, all with standard error of 1. Significant original effect estimates are coloured red if positive and blue if negative. The black line shows the prediction of the replication effect estimate adjusted for regression to the mean.

effect estimates. Replication studies will thus, on average, lead to less impressive effect sizes.

To illustrate this phenomenon, Figure 2 shows a sample of 1,000 simulated pairs of original and replication effect estimates whose underlying true effect sizes are the same. The corresponding standard errors are fixed at 1, so a total of 178 original studies turn out to be significant. However, the corresponding replication effect sizes are on average 43% smaller and only 34% (60/178) also achieve significance.

If we want to "predict" the replication estimate based on the original estimate, the naive approach – which takes the original estimate at face value – is doomed to be biased. More sophisticated methods that adjust for regression to the mean should be used.[5] These methods take advantage of the fact that regression to the mean becomes less severe the larger the original effect estimate is relative to its measurement error (the larger the "signal" compared to the "noise"). The black line in Figure 2 shows the corresponding prediction adjusted for regression to the mean. Note that the amount of shrinkage of the prediction depends on the signal-to-noise ratio, which can be summarised by the *p*-value of the original study. The adjustment barely shrinks large effect estimates from very convincing original studies but applies more shrinkage to less convincing ones.[6] For our simulated data, the adjustment reduces the mean squared prediction error from 1.87 to 1.56 (17%).

But can we also adjust for regression to the mean in the calculation of replication power? Researchers often use an *ad hoc* approach and simply shrink the original effect estimate by a fixed percentage. For example, in the Camerer *et al.* project,[2] the original effect estimates have been reduced by 25% to calculate replication sample sizes. If we use the regression to the mean adjustment, an original *p*-value of 0.005 would result in 13% shrinkage, while an original *p*-value of 0.05 would result in 26% shrinkage. This adjustment reduces replication power further and is particularly pronounced for "borderline significant" original studies (see the blue curve in Figure 1).

So, how credible is a result with a *p*-value at or around 0.05? Chances are only 50:50 that a replication study that simply follows the original study design (including sample size) will be statistically significant. Adjusted for regression to the mean, the replication probability is even lower (35%). Even if we increase the sample size by a factor of 11, as in the Ackerman *et al.* replication study,[1] adjusted replication power is only 83%. These considerations show the need for more stringent *p*-value thresholds to trust (original) "out-of-the-blue" findings.[7]

R. A. Fisher interpreted "significance" at the traditional 0.05 threshold merely as an indication that an experiment was worth repeating.[8] This cautious view is more important than ever and is reflected in the recent trend of organisations, including the US National Academies of Sciences, Engineering, and Medicine (bit.ly/3p1AUkT) and the Royal Netherlands Academy of Arts and Sciences (bit.ly/2TpNDiR), to specifically support and conduct replication studies. ■

p-values and significance tests defined

The *p*-value "is defined as the probability, under the assumption of no effect or no difference (the null hypothesis), of obtaining a result equal to or more extreme than what was actually observed".[9] If the *p*-value is smaller than some pre-defined *significance level* (usually 0.05), the result is said to be *statistically significant*. The probability to obtain a statistically significant result is called the *power* of the test, which also depends on the true effect size and the sample size.

Disclosure statement

Held receives funding from the Swiss National Science Foundation (no. 189295). Pawel declares no competing interests. Schwab receives funding from the Foundation for Scientific Research at the University of Zurich (no. STWF-19-007).

References

1. Ackerman, J. M., Nocera, C. C. and Bargh, J. A. (2010) Incidental haptic sensations influence social judgments and decisions. *Science*, **328**(5986), 1712–1715.

2. Camerer, C. F., Dreber, A., Holzmeister, F. *et al.* (2018) Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, **2**(9), 637–644.

3. Goodman, S. N. (1992) A comment on replication, *p*-values and evidence. *Statistics in Medicine*, **11**(7), 875–879.

4. Senn, S. (2011) Francis Galton and regression to the mean. *Significance*, **8**(3), 124–126.

5. Copas, J. B. (1997) Using regression models for prediction: Shrinkage and regression to the mean. *Statistical Methods in Medical Research*, **6**, 167–183.

6. Pawel, S. and Held, L. (2020) Probabilistic forecasting of replication studies. *PLoS One*, **15**(4), e0231416.

7. Benjamin, D. J., Berger, J. O., Johannesson, M. B. *et al.* (2018) Redefine statistical significance. *Nature Human Behaviour*, **2**, 6–10.

8. Goodman, S. N. (2016) Aligning statistical and scientific reasoning. *Science*, **352**(6290), 1180–1181.

9. Goodman, S. N (1999) Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine*, **130**(12), 995–1004.