Is the *p*-curve always right-skewed in observational research?

We are thankful to Simonsohn, Nelson and Simmons (SNS) for the opportunity to directly reply to their blog post that discusses our article about *p*-hacking and the *p*-curve in observational research (Bruns and Ioannidis, 2016). It was a great pleasure to read their blog post that makes an important clarification. In this reply, we argue that if the *p*-curve is intended to identify any type of effect (causal or spurious) as proposed by SNS, the *p*-curve becomes a valid but irrelevant concept in observational research.

True (causal) effects and spurious relationships

Consider A and B having a common cause denoted by C (Figure 1). If we do not condition on C, we will find a correlation between A and B in observational data that expresses the spurious relationship between these two variables. This is the example used by SNS in their blog post with female partners being A, shotguns being B, and gender being C.

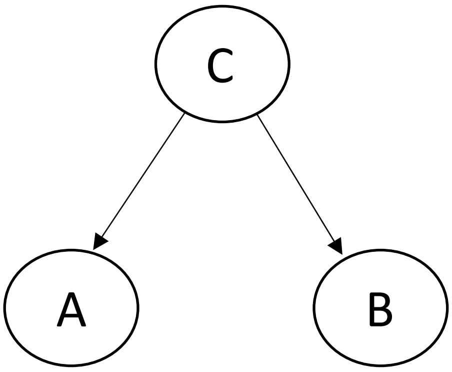

Figure 1: Causal graph of A, B, and C. A and B have a spurious relationship if we do not condition on C.

The original paper by SNS is written with a focus on experimental research. If A is the treatment and B the outcome, we can expect that we do not find an association between A and B if randomization worked properly. In their original paper SNS write:

“When a studied effect does exist (i.e., the null is false), the expected distribution of *p* values of independent tests is right-skewed, that is, more *p*s < .025 than .025 < *p*s < .05” (Simonsohn et al., 2014, p. 536).

In their blog post, SNS clarify that “a studied effect” can be any type of effect. This includes true (causal) effects, but also spurious relationships (e.g. confounded effects).[1] We are very happy that SNS agree with this point of our article: right-skewed *p*-curves cannot distinguish between true (causal) effects and spurious relationships.

SNS’s core point is that a right-skewed *p*-curve identifies only that there is some sort of effect and then an “expert” needs to be called who identifies whether the *p*-curve is right-skewed due to a causal or due to a spurious relationship. However, it is more than clear after many decades of observational research that the ability of “experts” to identify confounding is close to non-existent. Young and Karr (2011) have shown that of 52 high-profile claims in observational studies, 0 were

[1] One should mention that there are other types of spurious relationships than the classical example considered in Figure 1, such as spurious relationships due to, for example, unit roots (Granger and Newbold, 1974) or the use of ratios (Pearson, 1896). These spurious relationships will also result in right-skewed *p*-curves.

validated in experimental studies. All these observational effects had been deemed to be causal by some of the very best epidemiologists in the world.

Moreover, the nature of *p*-hacking in observational research is different from experimental research. *p*-hacking in many types of observational research relies predominantly on omitted-variable biases to produce statistically significant estimates. These omitted-variable biases introduce spurious relationships such as in our illustration in Figure 1 between A and B. What information does a right-skewed *p*-curve really provide if both this highly common type of *p*-hacking and true (causal) effects result in similarly right-skewed *p*-curves?

p-hacking in observational research

The *p*-curve needs to be right-skewed in some cases and uniformly distributed (or left-skewed) in other cases to be helpful in distinguishing between different cases, e.g. between *p*-hacking and true effects. However, *p*-hacking in observational research systematically produces right-skewed *p*-curves as shown in our article. *p*-hacking in many, probably most, fields of observational research is mainly based on specification searching, i.e. authors estimate different regression models by varying the set of control variables to search for statistically significant estimates. A famous econometrician described this approach as:

"Econometricians have found their Philosophers' Stone; it is called regression analysis and is used for transforming data into "significant" results!" (Hendry, 1980, p. 390)

As described in our article, this type of *p*-hacking is based on omitted-variable biases. Hence, the *t*-value of the coefficient of interest is drawn from a non-central *t*-distribution implying a right-skewed *p*-curve. In other words, this approach relies on utilizing many different spurious association (e.g. between A and B in Figure 1) to produce significant estimates. Of course, the *p*-hacked results are replicable in the sense that they may also result in right-skewed *p*-curves in a different sample, if the same type of confounding exists. This is simply how *p*-hacking works in observational research, it is just so easy to find a way to make the results fit expectations.

Right-skewed, right-skewed, right-skewed, …

We can even go a step further and make the point that the *p*-curve is always right-skewed in observational research irrespective of *p*-hacking or true (causal) effects. SNS write themselves in their blog post: "With observational data it's hard to identify exactly zero effects because there is always the risk of omitted variables, selection bias, long and difficult-to-understand causal chains, etc."

A striking example is Schuemie et al. (2014a, 2014b) that we discuss in our article. They show for the biomedical literature that the use of best-practice research designs results in right-skewed *p*-curves even if theory strongly suggests a null effect. Most likely, these right-skewed *p*-curves occur as the correct regression specification is simply unknown and some (small) omitted-variable biases may always remain in the analysis of observational data. Many other biases that abound in observational research can add to this generation of significant, but spurious results.

Conclusions

In our article we show that *p*-curves cannot distinguish between true (causal) effects and spurious relationships. SNS argue in their blog post that the *p*-curve is only meant to identify any type of effect irrespective of whether it is a true (causal) effect or a spurious relationship. If one follows this argumentation the *p*-curve may become irrelevant as the typical type of *p*-hacking in observational research (specification searching) also results in right-skewed *p*-curves. Moreover, the *p*-curve may even be always right-skewed in observational research, simply because some omitted-variable biases may always remain. What can we learn from the *p*-curve if the *p*-curve is always right-skewed?

It is important to note that the points made here do not generalize to experimental research if randomization works properly. The *p*-curve may be a useful tool to analyze experimental research.

References

Bruns, S. B., & Ioannidis, J. P. (2016). p-Curve and p-Hacking in Observational Research. *PLoS ONE*, 11(2), e0149144.

Granger, C. W., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, 2(2), 111-120.

Hendry, D. F. (1980). Econometrics-alchemy or science?. *Economica*, 47(188), 387-406.

Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution - On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs. *Proceedings of the Royal Society of London*, 60(359-367), 489-498.

Serghiou, S., Patel, C. J., Tan, Y. Y., Koay, P., & Ioannidis, J. P. (2015). Field-wide meta-analyses of observational associations can map selective availability of risk factors and the impact of model specifications. *Journal of Clinical Epidemiology*, 71, 58-67.

Schuemie, M. J., Ryan, P. B., Dumouchel, W., Suchard, M. A., & Madigan, D. (2014a). Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in Medicine*, 33, 209–218.

Schuemie, M. J., Ryan, P. B., Suchard, M. A., Shahn, Z., & Madigan, D. (2014b) Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15, 36–39.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2), 534-547.

Young, S. S., & Karr, A. (2011). Deming, data and observational studies. *Significance*, 8, 116–120.