This version: September 4th, 2014

SUPPLEMENTARY MATERIALS

Paper available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2377290

P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results

Uri Simonsohn
University of Pennsylvania

Leif D. Nelson
University of California, Berkeley

Joseph P. Simmons
University of Pennsylvania

Outline

Supplement 1. Robustness to non-normality (p.2-5)

Supplement 2. Robustness to heterogeneity of effect size (p.6-7)

Supplement 3. Trim-and-Fill performance when some $p>.05$ are observed (p.8-9)

Supplement 4. Alternative loss functions (p.10-11)

Supplement 5. R Code for all results in the paper http://www.p-curve.com/Supplement/Rcode

Supplement 1. Robustness to non-normality

The simulations behind all figures in the paper assume the underlying data are normal (or more precisely, that the difference of means test-statistic is distributed student, normality of the raw data is a sufficient but not necessary condition for this assumption).

On the one hand, there are good a-priori reasons to expect *p*-curve to be robust to deviations from normality. First, for a long time it has been known that comparing means assuming they are distributed student (e.g., t-test, ANOVA and regression) are quite robust to severe deviations from normality (Boneau, 1960; Pearson, 1931; Sawilowsky et al., 1992). Second, in earlier work we have showed that under the null (when the true effect is 0) *p*-curve is uniform even with distributions that look nothing like normal (see Supplement 2 in Simonsohn et al., 2014).Third, the two examples from the many-labs replication project had non-normal data (binary and likert scales), and yet *p*-curve obtained correct estimates (see Figure 5B).

On the other hand, there is work documenting that *extreme* skew/outliers can disrupt the validity of the t-test (see e.g., Keselman et al., 2004 and references within). There certainly are extreme situations where basic statistical tools break down, and hence where *p*-curve breaks down. Moreover, just because *under the null p*-curve has the shape one expects it to have assuming normality, it does not mean than when the null is false it will also. For us to recover effect size from *p*-curve, we need to know how right-skewed *p*-curve is expected to be under the alternative (when the data are not normal).

With this in mind we repeated the simulations from Figure 2, where we assess *p*-curve's ability to estimate effect size correcting for selective reporting of studies that are $p<.05$ results, but instead of simulating test-statistics drawn from the student distribution, we used actual data published in psychology papers to compute difference of means t-test and submit these results to *p*-curve.

We used a diverse set of 10 dependent variables that were reported in papers published in the data-posting journal *Judgment and Decision Making*. These were selected seeking to maximize variety of distributions. They include, for instance, count data from the Eurovision contest, perceived breaking speed of a car, and dollar donations in a dictator game. [1]

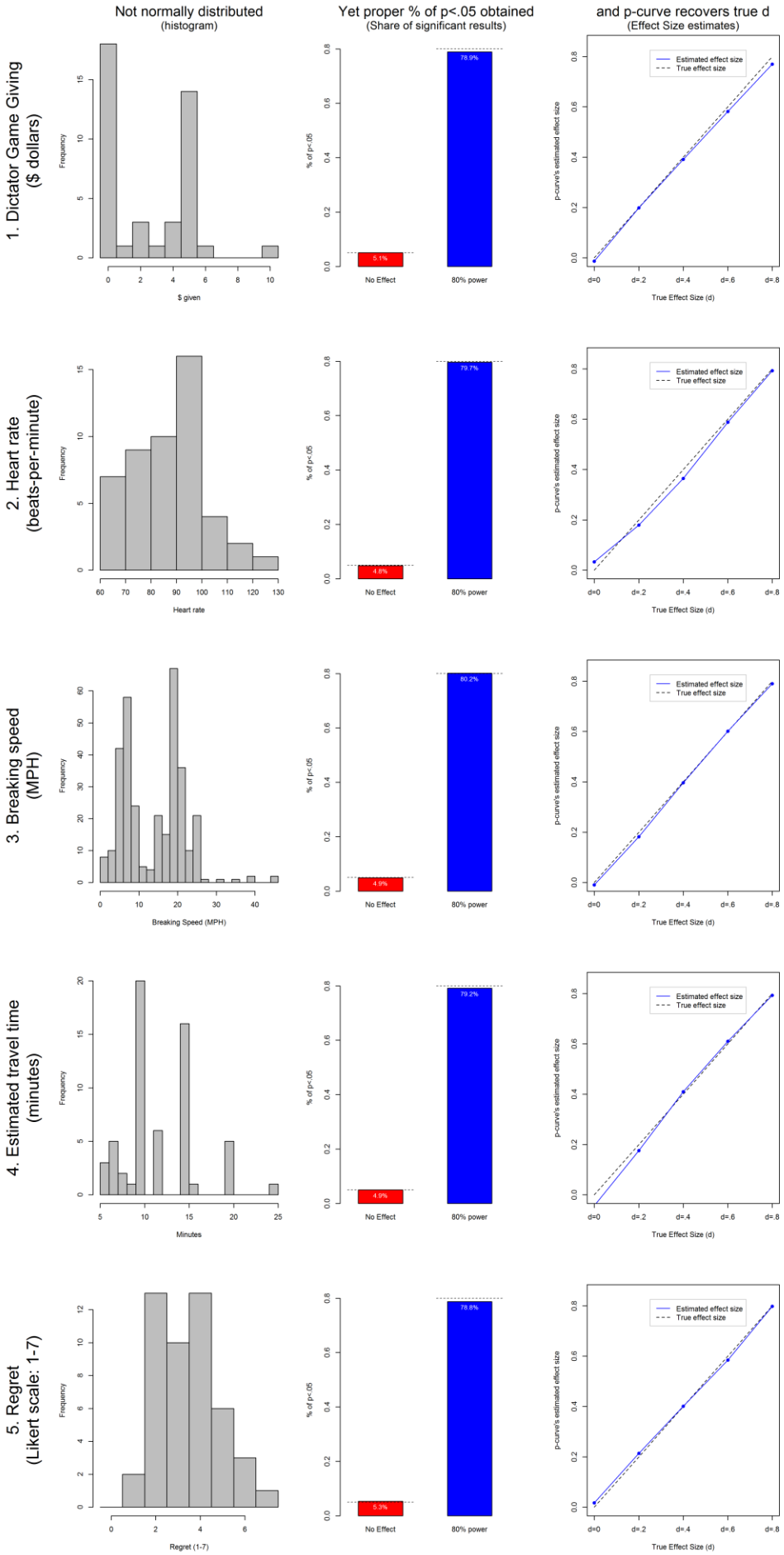
The approach was the following. For each variable we pooled the observations across conditions and drew –with replacement- two random samples of size n=20 each, conducted a t-test, and noted the resulting t-value and *p*-value. We repeated that several thousand times. That involves drawing under the null, both samples are coming from the same population, so we would expect 5% of results to be significant.

Then we modified this procedure by adding .91SDs to one of the samples before conducting the t-test. If data were normally distributed, an effect size of d=.91 should lead to 80% power. We hence paid attention to what percentage of the simulations were $p<05$. 80%?

Then we proceeded analogously but instead of adding d=.91, we added d=.2, kept all the significant t-values, and analyzed them with *p*-curve. If *p*-curve works properly, we should estimate d=.2. If non-normality affects *p*-curve then we should not. We then did that for d=.4, .6 and .8.

We summarize the results of those simulations in figure S1 below. For each variable we report three charts. The first consist of a histogram, useful for intuitively assessing the departure from normality of the variable in question. The second plots the percentage of simulated t-test that were significant when the null was true (d=0) and when the test was powered to 80% assuming normality. The third shows the estimated effect size when the true d was .2,.4,.6 and .8. The figures suggest *p*-curve performs just as well under normality assumptions, as it does under every-day departures from it.

[1] One of us used those data in a different project comparing frequentist with Bayesian confidence intervals (Simonsohn, 2014); details on data sources and variable selection are presented there.


Not normally distributed
(histogram)
Yet proper % of p<.05 obtained
(Share of significant results)
and p-curve recovers true d
(Effect Size estimates)
1. Dictator Game Giving ($ dollars)
2. Heart rate (beats-per-minute)
3. Breaking speed (MPH)
4. Estimated travel time (minutes)
5. Regret (Likert scale: 1-7)
Frequency
$ given
Heart rate
Breaking Speed (MPH)
Minutes
Regret (1-7)
% of p<.05
5.1%
78.9%
4.8%
79.7%
4.9%
80.2%
4.9%
79.2%
5.3%
78.8%
No Effect
80% power
p-curve's estimated effect size
Estimated effect size
True effect size
True Effect Size (d)
d=0
d=.2
d=.4
d=.6
d=.8

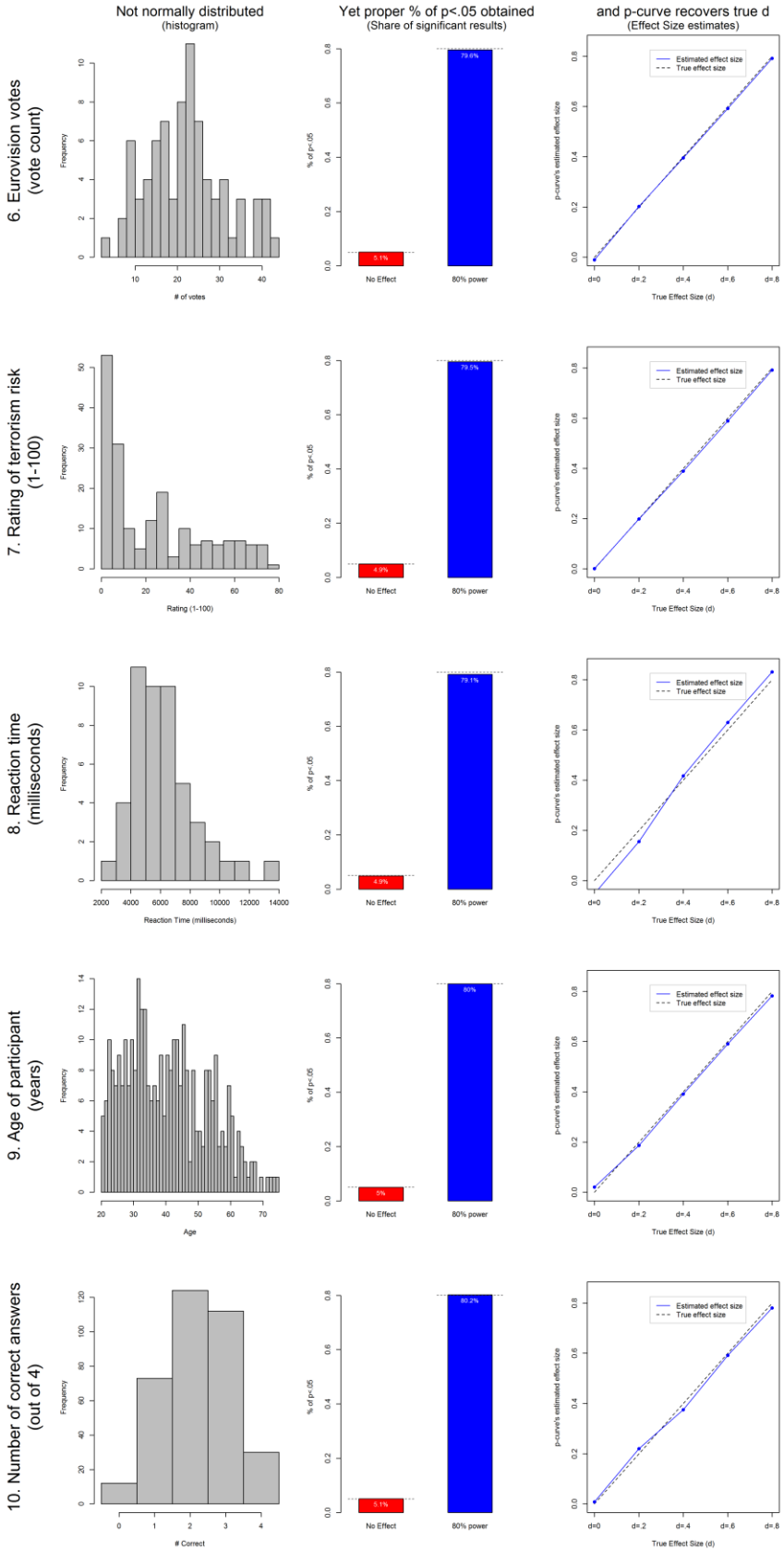


Figure S1. Non-impact of non-Normal Distributions on results.

Column 1 reports histogram of dependent variable selected, column B the percent of simulated t-test that are p<.05 when the null is true and when means differ by .91 SD, and column C the estimated effect size using p-curve.

Supplement 2. Robustness to heterogeneity of effect size

Figure 2C in the main paper reports results showing that *p*-curve recovers the average effect size when it varies across studies. Here we discuss a binary example that may facilitate building an intuition for how this works.

We create a stylized situation where half the studies have one underlying effect size, d_1, and the other half of studies have underlying effect size d_2. The average true underlying effect depends on how we weight effects d_1 and d_2. As we state in the paper, *p*-curve estimates the average true effect size for the set of statistically significant findings. Therefore, if studies have the same sample size, and there are just as many significant studies with underlying true effect d_1 and d_2, then *p*-curve will estimate as the effect size the simple average $(d_1+d_2)/2$. If there are more statistically significant studies from d_1 or d_2 studies, then the average deviates from this simple mean, and the same occurs if the sample sizes are different for studies with d_1 and d_2 (because *p*-curve weights proportionally to precision and hence is affected by n).

In Figure S2 we report results for several scenarios under which effects d_1 and d_2 may be studied. Panel A displays results when half the significant studies have and underlying true effect of d_1 and half of d_2, all studies have the same sample size (n=20). The true average effect size of published studies is d=.4 and *p*-curve correctly recovers that value. The naïve average published effect is dramatically off in all cases so we do not discuss it further.

Panel B considers a situation where the same number of studies are *attempted* for d_1 and d_2. When $d_1>d_2$, the subset of significant results submitted to *p*-curve will have more d_1 studies and hence the average true effect size of significant studies is greater than $(d_1+d_2)/2$. *P*-curve correctly recovers this average too.

For example, if all studies have n=20, d_1=.6 studies have 45% power, and d_2=.2 have 9% power. The set of significant studies, therefore, have about five d_1=.6 studies for every one d_2=.2 study, and hence the true average effect of the observed set is about 5/6*.6+1/6*.2~.54, the right most black bar in Panel B. *P*-curve correctly recovers that value. Panel C further emphasizes this point by considering a situation where sample size is greater for studies with bigger effects.

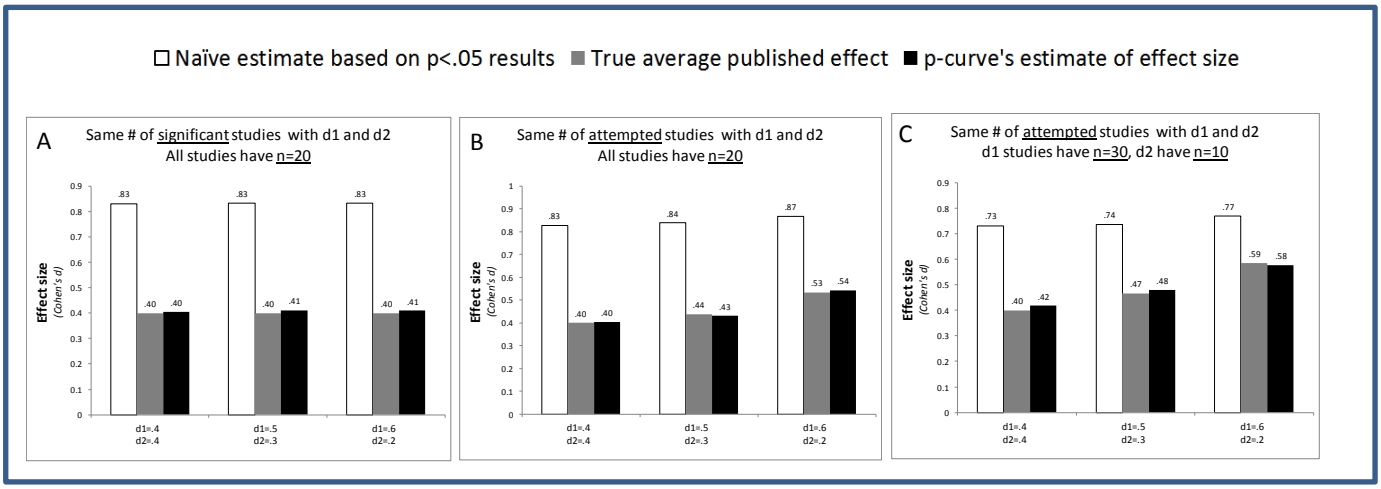


Figure S2. Estimated effect size when studies are a mixed of two underlying effect sizes d_1 and d_2.

Supplement 3. Trim-and-Fill performance when some *p*>.05 are observed

Our simulations in the paper and other supplementary materials have applied the Trim-and-Fill procedure only to the subset of statistically significant findings. We conducted simulations this way seeking to emulate reality, where the vast majority of failed studies are presumably never written up, let alone shared publically. Nevertheless, some failed studies are sometimes observable to meta-analysts: How much better does Trim-and-Fill perform in that case?

We modified the simulations from Supplement 2 to shed some light on this question. We considered a situation where the true effect size is δ=0, a meta-analysis is performed including 500 statistically significant results (as always: all of the same sign). Half of them had n=10 per cell, half had n=30. We then add in increments of 50, additional *n.s.* studies, chosen at random from the pool of all studies conducted. So we start with 0 *n.s.*, compute the naïve and Trim-and-Fill corrected effect size estimate, add 50 *n.s.* studies, recompute those effect sizes, and so on.

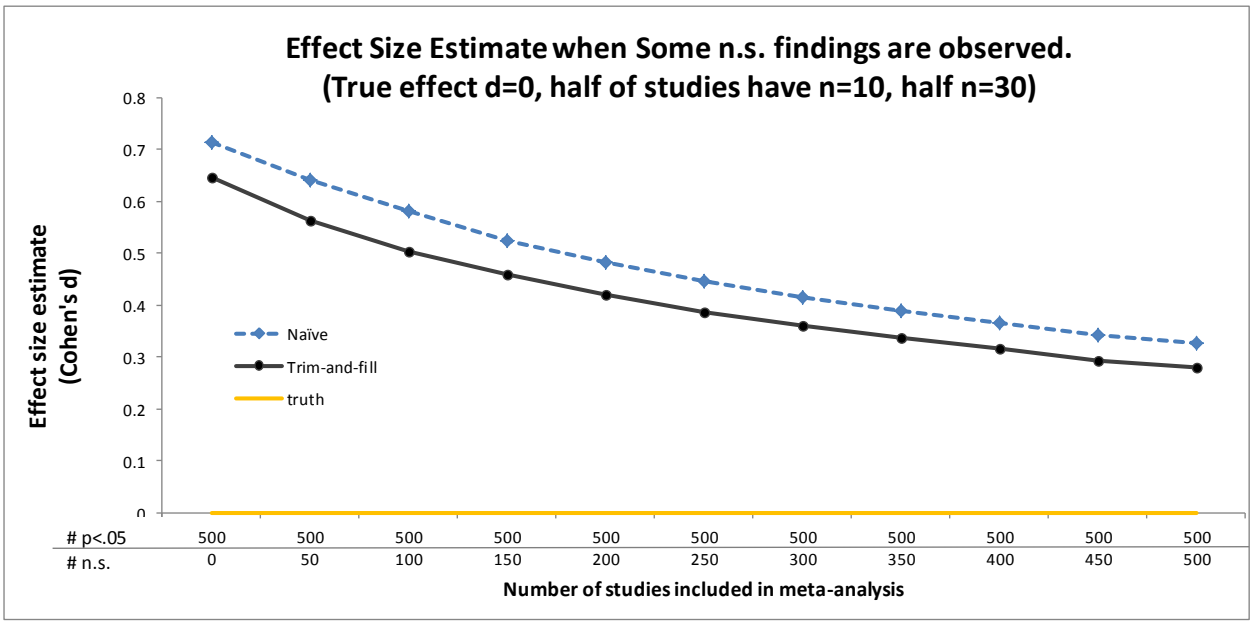


Figure S3 displays the results. For example, the right-most markers display the effect-size estimates when the meta-analysis is performed on a total of 1000 studies, 500 with $p<.05$, and 500 with $p>.05$. The figure shows that as we add more *n.s.* findings to the analysis, the naïve estimate of effect size drops: if we reduce the size of the file-drawer, we reduce the size of the bias induced by the file-drawer. [2]

The Trim-and-Fill estimates decrease also, but at a very similar rate. Adding *n.s.* studies helps the Trim-and-Fill not because it can better correct for bias with them, but because there is less bias to be corrected in the first place. Note that in all cases *p*-curve would correctly estimate the effect to be zero.

[2] Going from zero *n.s.* studies to just as many of them as there are $p<.05$ studies reduces the estimated effect sizes in about half, this makes intuitive sense, as half the studies are now on average $\hat{d}=0$, the average of x and 0 is x/2.

Supplement 4. Alternative loss functions.

Alternative 1. Anderson-Darling test

Instead of comparing the observed distribution of *pp*-values to the uniform distribution U[0,1] via the Kolmogorov-Smirnov test, one could use the Anderson-Darling test. In our simulations we obtained nearly identical results with this approach. We noted the loss function tended to be flatter around the minimum leading to somewhat less precise estimates (a-la Figure 4 in the paper).

Alternative 2. Overall χ^2 for evidential value based on Fisher's method

Let *k* be the number of significant results included in *p*-curve. One may combine the resulting *k pp*-values via Fisher's method for aggregating *p*-values. In particular, we have that if pp~U[0,1], then $\sum$-2ln(pp_i) ~ $\chi^2(2k)$.

If a given candidate effect *d* fits the data perfectly, then the *p*-value associated with the $\chi^2(2k)$ test would be *p*=.5. We can define as the loss function the absolute difference between the *p*-value of that $\chi^2(2k)$ test and .5, and use as our effect size estimator the one that makes that aggregate test as close as possible to .5.

One advantage of this approach is that it is more closely related to the approach we took to define evidential value when using *p*-curve for hypothesis testing (Simonsohn, et al., 2014). A second advantage is that it lends itself to a very natural approach to computing a confidence interval for the estimated effect size without using bootstrapping. One finds the effect size that makes the overall χ^2 test have *p*=.975 and *p*=.025.

Alternative 3. Overall Z test for evidential value based on Stouffer's method

A variation of Alternative 2 consists of Stouffer's methods to the *pp*-values, where each of *k pp*-value is converted into a Z-score, and then aggregated using $Z=\frac{\sum Z_i}{\sqrt{k}}$.

An advantage of this approach is that Stouffer's method allows weighting the different Zs differently. This would allow using *p*-curve for conducting random effect estimation. We leave it to future research to explore this promising possibility.

References

Boneau, C. A. (1960). "The Effects of Violations of Assumptions Underlying the T Test." *Psychological bulletin, 57*(1), 49.

Keselman, H., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). "The New and Improved Two-Sample T Test." *Psychological Science, 15*(1), 47-51.

Pearson, E. S. (1931). "The Analysis of Variance in Cases of Non-Normal Variation." *Biometrika, 23*(1-2), 114-133.

Sawilowsky, S. S., & Blair, R. C. (1992). "A More Realistic Look at the Robustness and Type Ii Error Properties of the< Em> T</Em> Test to Departures from Population Normality." *Psychological bulletin, 111*(2), 352.

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.