Estimating Population Mean Power Under Conditions of Heterogeneity and Selection for Significance

Jerry Brunner and Ulrich Schimmack
University of Toronto Mississauga

In scientific fields that use significance tests, statistical power is important for successful replications of significant results because it is the long-run success rate in a series of exact replication studies. For any population of published results, there is a population of power values of the statistical tests on which conclusions are based. We give exact theoretical results showing how selection for significance affects the distribution of statistical power in a heterogeneous population of significance tests. In a set of large-scale simulation studies, we compare four methods for estimating population mean power of a set of studies selected for significance (a maximum likelihood model, extensions of p-curve and p-uniform, & z-curve). The p-uniform and p-curve methods performed well with a fixed effects size and varying sample sizes. However, when there was substantial variability in effect sizes as well as sample sizes, both methods systematically overestimate mean power. When the assumptions of the maximum likelihood were satisfied, it produced the most accurate estimates for heterogeneity in effect sizes, but z-curve produced more accurate estimates when the assumptions of the maximum likelihood model were not met. We recommend the use of z-curve to estimate the typical power of significant results, which has implications for the replicability of significant results in psychology journals.

Keywords: Power estimation, Post-hoc power analysis, Publication bias, Maximum likelihood, Z-curve, P-curve, P-uniform, Effect size, Replicability, Meta-analysis

The purpose of this paper is to develop and evaluate methods for predicting the success rate if a set studies with significant results were replicated exactly. We call this statistical property, the average power of a set of studies. Average power can range from the criterion for a type-I error, if all significant results are false positives, to 100%, if the statistical power of original studies approaches 1. We agree with Simonsohn, Nelson, and Simmons (2014) that average power can be used to quantify the degree of evidential value in a set of studies. While studies with average power of 80% or more provide strong support for most hypotheses, studies with 20% power provide weak evidence and may contain a large number of false positive results or true positive results with negligible effect sizes.

Estimating average power of original studies is interesting because it is tightly connected with the outcome of replication studies (Greenwald, Gonzalez, Harris, & Guthrie, 1996; Yuan & Maxwell, 2005). To claim that a finding has been replicated, a replication study should reproduce a produce a significant result, and the probability of a successful replication is a function of statistical power. Thus, if reproducibility is a requirement of good science (Bunge, 1998; Popper, 1959), it follows that high statistical power is a necessary condition for quality science. Even if a different statistical approach is used, only studies with high true signal-to-noise ratios (i.e., non-centrality parameters) can produce consistent evidence of a predicted effect.

Information about the average power of studies is also useful because selection for significance increases the type-I error rate and inflates effect sizes (Ioannidis, 2008). However, these biases are relatively small if the original studies had high power. Knowledge about the average power of studies is also useful to adjust power analyses for the planning of future studies. If average power is high, replication studies can use the same sample sizes as original studies, but if average power is low, sample sizes need to be increased to avoid false negative results.

Given the practical importance of power for good science, it is not surprising that psychologists have started to examine the evidential value of results published in psychology journals. At present, two statistical methods have been used to make claims about the average power of psychological research; namely p-curve (www.p-curve.com) and z-curve (https://replicationindex.wordpress.com/), but so far neither method has been peer-reviewed.

Most of the ideas in this paper were developed jointly. An exception is the z-curve method, which is solely due to Schimmack. Brunner did all the programming and derived the proofs of the Principles. We would like to thank Dr. Jeffrey Graham for providing remote access to the computers in the Psychology Laboratory at the University of Toronto Mississauga. Thanks to Josef Duchesne for technical advice.

Statistical Power Before and After A Study Has Been Conducted

Before we proceed, we would like to clarify that statistical power of a statistical test is defined as the probability of *correctly* rejecting the null hypothesis (Neyman & Pearson, 1933). This probability depends on the sampling error of a study and the population effect size. The traditional definition of power does not consider effect sizes of zero (false positives) because the goal of a priori power planning is to ensure that a non-zero effect can be demonstrated.

However, our goal is not to plan future studies, but to analyze results of existing studies. For post-hoc power analysis, it is impossible to distinguish between true positives and false positives and to estimate the average power conditional on the unknown status of hypotheses (i.e., the null-hypothesis is true or false). Thus, we use the term average power as the probability of *correctly or incorrectly* rejecting the null-hypothesis (Sterling, Rosenbaum, & Weinkam, 1995). As a result, post-hoc average power includes an unknown percentage of false positives that have a probability equal to alpha (typically 5%) to reproduce a significant result in a replication attempt. At the same time, we believe that the strict null-hypothesis is rarely true in psychological research (Cohen, 1994).

It would be ideal if it were possible to estimate the power of a single statistical test that supports a particular finding. Unfortunately, well-documented problems with the "observed power" method suggest that the goal of estimating the power of an individual test may be out of reach (Boos & Stefanski, 2012; Hoenig & Heisey, 2001). The main problem is that estimates for a single result are too variable to be practically useful (Yuan & Maxwell, 2005; but see Anderson, Kelley, & Maxwell, 2017). Thus, we focus on estimating mean power of a set of studies. The number of studies has to be reasonably large to obtain useful estimates. We used a minimum of 15 studies for our simulations.

It is important to distinguish our undertaking from that of Cohen (1962) and follow-up studies by L.J. Chase and R. B. Chase (1976) and Sedlmeier and Gigerenzer (1989). In Cohen's classic survey of power in the *Journal of Abnormal and Social Psychology*, the results of the studies were not selected in any way. Power was never estimated. It was calculated exactly for a priori effect sizes deemed "small," "medium" and "large." If a "medium"effect size referred to the population mean (which Cohen never claimed), power at the mean effect size is still not the same as mean power.

Two Populations of Studies

We distinguish two populations of tests. One population contains all studies that have been conducted. This population contains significant and non-significant results. The other population contains the subset of studies that produced a significant result. We focus on the population of studies selected for significance for two reasons.

First, often non-significant results are not available because journals selectively publish significant results (Rosenthal, 1979; Sterling, 1959; Sterling et al., 1995). Second, only significant results are used as evidence for a theoretical prediction. It is irrelevant how many false positives were tested and not reported because they fortunately produced non-significant results (true negatives). Psychological theories rests on studies that produced significant results. Thus, only the evidential value of significant results is relevant for evaluations of the robustness of psychology as a science. In short, we are interested in statistical methods that can estimate the average power of a set of studies with significant results only.

The Study Selection Model

We developed a number of theorems that specify how selection for significance influences the distribution of power. These theorems are very general. They do not depend on the particular population distribution of power, the significance tests involved, or the Type I error probabilities of those tests. They do not even depend on the appropriateness of the tests or the assumptions of the tests being satisfied. The only requirement is that for every study with a specific population effect size, sample size, and statistical test, the probability of a result being selected is the true power of a study. We discuss the two most important theorems in detail. All six theorems are provided in the appendix, along with an illustration of the theorems by simulation.

First Theorem: *Population mean true power equals the overall probability of a significant result.*

Theorem 1 establishes the central importance of population mean power after selection for significance for predicting replication outcomes. Think of a coin-tossing experiment in which a large population of coins is manufactured, each with a different probability of heads; that is, these coins are not fair coins with equal probabilities for both sides. Also consider heads to be successes or wins. Repeatedly tossing the set of coins and counting the number of heads produces an expected value of the number of successes. For example, the experiment may yield 60% heads and 40% tails. While the exact probability of showing heads of individual coins are unknown, the observable success rate is equivalent to the mean power of all coins. Theorem 1 states that success rate and mean power are equivalent even if the set of coins is a subset of all coins. For example, assume all coins were tossed once and only coins showing heads were retained. Repeating the coin toss experiment, we would still find that the success rate for the set of selected coins matches the mean probabilities of the selected coins.

Second Theorem: *The effect of selection for significance is to multiply the probability of each power value by a quantity equal to the power value itself, divided by population mean power before selection. If the distribution of power is continuous, this statement applies to the probability density function.*

Figure 1 illustrates Theorem 2 for a simple, artificial example in which power before selection is uniformly distributed on the interval from 0.05 to 1.0. The corresponding distribution after selection for significance is triangular – a substantial change. In Figure 2, power before selection is less heterogeneous and higher on average. Consequently, the distributions of power before selection and after selection are much more similar. In both cases, though, mean true power after selection for significance is higher than mean true power before selection for significance.

Theorem 2 may seem overly simplistic and unrealistic. Few researchers conduct a study and give up after a first attempt produces a non-significant result. Instead they may try several times with slight variations of the study and not report studies that failed to produce significant results. For example, Morewedge, Gilbert, and Wilson (2014) explained that they did not report "some preliminary studies that used different stimuli and different procedures and that showed no interesting effects. (e.g., Morewedge, Gilbert, & Wilson, 2014). From a theoretical perspective, it is important that all studies tested the same hypothesis and that more than one non-significant finding was not reported. However, for the estimation of mean power of the studies that were selected, it is irrelevant that all studies tested the same hypotheses. Each study that was conducted by Morewedge et al. has an unknown true power to produce a significant result. Theorem 2 implies that the mean power of the studies that produced a significant result is greater than the mean power of the studies that were not selected as well as the mean power of the total set of studies; with the exception when all studies are false positives. Thus, the estimation methods that we tested can be used for realistic scenarios like the one described by Morewedge et al.

Figure 1. Uniform distribution of power before selection

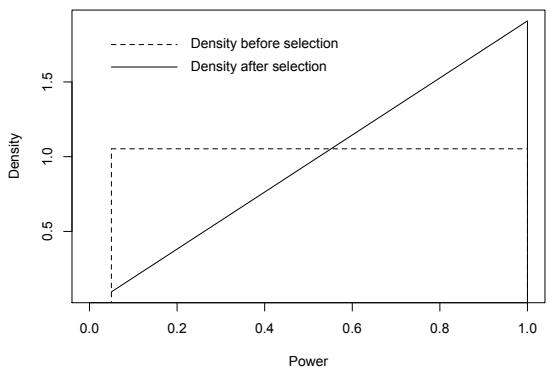
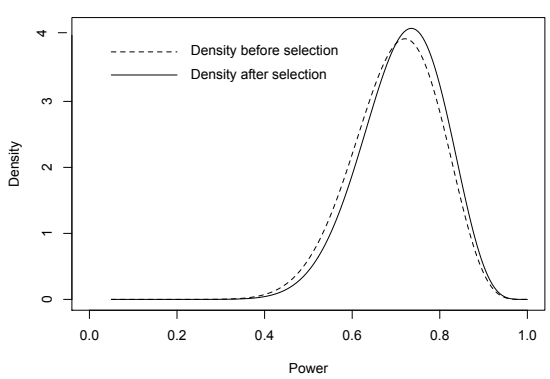


Figure 2. Chi-squared distribution of power before selection



Estimation Methods

In this section, we describe four methods for estimating population mean power after selection for significance under conditions of heterogeneity in sample size and effect size.

Notation and statistical background

To present our methods formally, it is necessary to introduce some statistical notation. Rather than using traditional notation from statistics that might make it difficult for non-statisticians to understand our method, we follow Simonsohn et al. (2014a), who employed a modified version of the S syntax (Becker, Chambers, & Wilks, 1988) to represent probability distributions. The S language is familiar to psychologists who use the R statistical software (R Core Team, 2012). The notation also makes it easier to implement our methods in R, particularly in the simulation studies.

The outcome of an empirical study is partially determined by random sampling error, which implies that statistical results will vary across studies. This variation is expected to follow a random sampling distribution. Each statistical test has its own sampling distribution. We will use the symbol T to denote a general test statistic; it could be a t-statistic, F, chi-squared, Z, or something more obscure. Assume an upper-tailed test, so that the null hypothesis will be

rejected at a specific significance level (usually .05), when the continuous test statistic T exceeds a critical value c.

Typically there is a sample of test statistic values, $T_1 ... T_k$, but when only one is being considered the subscript will be omitted. The notation p(t) refers to the probability under the null hypothesis that T is less than or equal to the fixed constant t. The symbol p would represent pnorm if the test statistic were standard normal, pf if the test statistic had an F-distribution, and so on. While p(t) is the area under the curve, d(t) is the value on the y axis for a particular t, as in dnorm. Following the conventions of the S language, the inverse of p is q, so that p(q(t)) = q(p(t)) = t.

Sampling distributions when the null-hypothesis are true are well-known to psychologists because they provide the foundation of null-hypothesis significance testing. Most psychologists are less familiar with non-central sampling distributions (see Johnson et al. 1995, for a detailed and authoritative treatment). When the null hypothesis is false, the area under the curve of the test statistic's sampling distribution is p(t,ncp), representing particular cases like pf(t,df1,df2,ncp). The initials ncp stand for "noncentrality parameter." This notation applies directly when T has one of the common non-central distributions like the noncentral t, F or chi-squared under the alternative hypothesis, but it can be extended to the distribution of any test statistic under any specific alternative, even when the distribution in question is technically not a non-central distribution. The non-centrality parameter is positive when the null hypothesis is false, and statistical power is a monotonically increasing function of the non-centrality parameter. This function is given explicitly by Power = 1 - p(c,ncp).

For the most important non-central distributions (Z, t, chisquared and F), the non-centrality parameter can be factored into the product of two terms. The first term is an increasing function of sample size, (n) and the second term is a function of the unknown parameters that reflects the standardized effect sizes (es). In symbols,

$$\texttt{ncp} = f_1(n) \cdot f_2(\texttt{es}). \qquad (1)$$

Thus, specification of effect sizes is determined by convenience and interpretability. This usage is consistent with that of Cohen (1988), who freely uses "effect size" to describe various functions of the model parameters, even for the same statistical test (Grissom & Kim, 2012). As an example of Equation (1), consider for example a standard F-test for difference between the means of two nor-mal populations with a common variance. After some simplification, the non-centrality parameter of the non-central F may be written as

$$\texttt{ncp} = n\,\rho\,(1-\rho)\,d^2,$$

where $n = n_1 + n_2$ is the total sample size, $\rho = \frac{n_1}{}$ the proportion of cases allocated to the first treatment, and $d = \frac{|\mu_1 - \mu_2|}{\sigma}$ is Cohen's (1988) *effect size* for the two-sample problem. This expression for the non-centrality parameter can be factored in various ways to match Equation 1; for example, $f_1(n) = n\,\rho\,(1-\rho)$ and $f_2(\texttt{es}) = \texttt{es}^2$. Note that this is just an example; Equation 1 applies to the non-centrality parameters of the non-central *Z*, *t*, chi-squared and *F* distributions in general. Thus for a given sample size and a given effect size, the power of a statistical test is

$$\text{Power} = 1 - \texttt{p}(c, f_1(n) \cdot f_2(\texttt{es})). \qquad (2)$$

The function $f_2(\texttt{es})$ can also be applied to sets of studies with different traditional effect sizes. For example, es could be Cohen's *d*, and the alternative effect size es′ could be the point-biserial correlation *r* (Cohen, 1988, p. 24). Symbolically, $\texttt{es}' = g(\texttt{es})$. Since the function $g(\texttt{es})$ is monotone increasing, a corresponding inverse function exists, so that $\texttt{es} = g^{-1}(\texttt{es}')$. Then Equation (2) becomes

$$\begin{aligned} \text{Power} &= 1 - \texttt{p}(c, f_1(n) \cdot f_2(\texttt{es})) \\ &= 1 - \texttt{p}(c, f_1(n) \cdot f_2\left(g^{-1}(\texttt{es}')\right)) \\ &= 1 - \texttt{p}(c, f_1(n) \cdot f_2'(\texttt{es}')), \end{aligned}$$

where f_2' just means another function f_2. That is, if the definition of effect size is changed (in a monotone way), the change is absorbed by the function f_2, and Equation (2) still applies.

After this introduction of notation and a basic introduction of power, non-centrality parameters, sample sizes, and effect sizes, we are ready to introduce four estimation methods for the estimation of mean power based on a set of studies that vary in power with known sample sizes and unknown population effect sizes. The four methods are pcurve, p-uniform, a Maximum Likelihood model, and zcurve.

Estimation Methods

The first two estimation methods are based on methods that were developed for the estimation of effect sizes. Our use of these methods for the estimation of mean power is an extension of these methods. Our simulation studies should not be considered tests of these methods for the estimation of effect sizes. We developed these methods simply because power is a function of effect size and sample size and sample sizes are known. Thus, only estimation of unknown effect sizes is needed to estimate power. In fact, power estimation is a simple additional step to compute power for each study as a function of the effect size estimate and the sample size of each study. These methods should have no problem in simulations with a fixed effect size. More interesting is their performance with heterogeneity in effect sizes.

P-curve 2.1 and p-uniform

A p-curve method for estimation of mean power is available online (www.p-curve.com). It is important to point out that this method differs from the p-curve method that we developed. The online p-curve method is called p-curve4.06. We built our p-curve method on the effect size p-curve method with the version code p-curve2.0 (Simonsohn et al., 2014). Hence, we refer to our p-curve method as p-curve2.1. P-uniform is very similar to p-curve (vanAsssen, vanAert, & Wicherts, 2014). Both methods aim to find an effect size that produces a uniform distribution of p-values between .05 and .00. Since we developed our p-uniform method for power estimation, the p-uniform has developed a better est-imation method. We conducted our studies before this new method was available. Thus, the performance of p-uniform in our simulation studies should not be interpreted as evidence for or against the new p-uniform method.

To find the best fitting effect size for a set of observed test statistics, p-curve 2.1 and p-uniform compute p-values for various effect sizes and chose the effect size that yields the best approximation of a uniform distribution. If the modified null hypothesis that effect size = es is true, the cumulative distribution function of the test statistic is the conditional probability

$$\begin{aligned} F_0(t) &= Pr\{T \le t | T > c\} \\ &= \frac{\mathtt{p}(t,\mathtt{ncp}) - \mathtt{p}(c,\mathtt{ncp})}{1 - \mathtt{p}(c,\mathtt{ncp})} \\ &= \frac{\mathtt{p}(t, f_1(n)\cdot f_2(\mathtt{es})) - \mathtt{p}(c, f_1(n)\cdot f_2(\mathtt{es}))}{1 - \mathtt{p}(c, f_1(n_i)\cdot f_2(\mathtt{es}))}, \end{aligned}$$

using $\mathtt{ncp} = f_1(n)\cdot f_2(\mathtt{es})$ as given in Equation 1. The corresponding modified p-value is

$$1 - F_0(T) = \frac{1 - \mathtt{p}(T, f_1(n)\cdot f_2(\mathtt{es}))}{1 - \mathtt{p}(c, f_1(n)\cdot f_2(\mathtt{es}))}.$$

Note that since the sample sizes of the tests may differ, the symbols p, n and c as well as T may have different refer-ents for $j = 1, \ldots, k$ test statistics. The subscript j has been omitted to reduce notational clutter.

If the modified null hypothesis were true, the modified p-values would have a uniform distribution. Both p-curve 2.1 and p-uniform choose as estimated effect size the value of es that makes the modified p-values most nearly uniform. They differ only in the criterion for deciding when uniformity has been reached.

P-curve 2.1 is based on a Kolmogorov-Smirnov test for departure from a uniform distribution, choosing the es value yielding the smallest value of the test statistic. P-uniform is based on a different criterion. Denoting by Pj the modified p-value associated with test j, calculate $Y = -\sum_{j=1}^{k} \mathrm{n}(P_j)$, where ln is the natural logarithm. If the P_j values were uni-formly distributed, Y would have a gamma distribution with expected value k, the number of tests. The P-uniform esti-mate is the modified null hypothesis effect size es that makes Y equal to k, its expected value under uniformity.

These technologies are designed for heterogeneity in sample size only, and assume a common effect size for all the tests. Given an estimate of the common effect size, estimated power for each test varies only as a function of sample size. which can be determined by Expression 2 because sample sizes are known. Population mean power can then be estimated by averaging the k power estimates.

Maximum likelihood model

Our maximum likelihood (ML) model also first estimates effect sizes and then combines effect size estimates with known sample sizes to estimate mean power. Unlike p-curve2.1 and p-uniform, the ML model allows for heterogeneity in effect sizes. In this way, the model is similar to Heges and Vevea's (1996) model for effect size estimation before selection for significance. To take selection for significance into account, the likelihood function of the ML model is a product of k conditional densities; each term is the conditional density of the test statistic T_j, given $N_j = n_j$ and $T_j > c_j$, the critical value.

Likelihood function. The model assumes that sample sizes and effect sizes are independent before the selection for significance. Suppose that the distribution of effect sizes before selection is continuous with probability density $g_\theta(\mathrm{es})$. This notation indicates that the distribution of effect sizes depends on an unknown parameter or parameter vector θ. In the appendix, it is shown that the likelihood function (a function of θ) is a product of k terms of the form

$$\frac{\int_0^\infty \mathtt{d}(t_j, f_1(n_j)\cdot f_2(\mathtt{es}))\, g_\theta(\mathtt{es})\, d\mathtt{es}}{\int_0^\infty \left[1-\mathtt{p}(c_j, f_1(n_j)\cdot f_2(\mathtt{es}))\right] g_\theta(\mathtt{es})\, d\mathtt{es}}, \qquad (3)$$

where the integrals denote areas under curves that can be computed with R's integrate function. The maximum likelihood estimate is the parameter value yielding the the highest product. To be applicable to actual data, the ML model has to make as-sumptions about the distribution of effect sizes. The ML model that was used in the simulation studies assumed a gamma distribution of effect sizes. A gamma distribution is defined by two parameters that need to be estimated based on the data. The effect sizes based on the most likely distribution are then combined with information about sample sizes to obtain power estimates for each study. An estimate of population mean power is then produced by averaging estimated power for the k significance tests. As shown in the appendix, the terms to be averaged are

$$\frac{\int_0^\infty \left[1-\mathtt{p}(c_j, f_1(n_j)\cdot f_2(\mathtt{es}))\right]^2 g_{\widehat{\theta}}(\mathtt{es})\, d\mathtt{es}}{\int_0^\infty \left[1-\mathtt{p}(c_j, f_1(n_j)\cdot f_2(\mathtt{es}))\right] g_{\widehat{\theta}}(\mathtt{es})\, d\mathtt{es}}. \qquad (4)$$

Z-curve

Z-curve follows a traditional meta-analyses that converts p-values into Z-scores as a common metric to integrate results from different original studies (Rosenthal, 1979; Stouffer, Suchman, DeVinney, Star, & Williams, 1949). The use of *Z*-scores as a common metric makes it possible to fit a sin-gle function to *p*-values arising from different statistical methods and tests. The method is based on the simplicity and tractability of power analysis for the one-tailed *Z*-test, in which the distribution of the test statistic under the alternative hypothesis is just a standard normal shifted by a fixed quantity that plays the role of a non-centrality parameter, and will be denoted by *m*. Input to the *Z*-curve is a sample of *p*-values, all less than $\alpha = 0.05$. These *p*-values are processed in several steps to produce an estimate.

1. *Convert p-values to Z-scores*. The first step is to imagine, for simplicity, that all the *p*-values arose from two-tailed *Z*-tests in which results were in the predicted direction. This is equivalent to an upper-tailed *Z*-test with significance level $\alpha/2$. In our simulations, alpha was set to .05, which results in a selection criterion of z = 1.96. The conversion to z-scores (Stouffer et al., 1949) consists of finding the test statistic *Z* that would have produced that *p*-value. The formula is

$$Z = \mathtt{qnorm}(1 - p/2). \qquad (5)$$

2. *Set aside* $Z > 6$. For convenience, we set aside extreme z-scores. This avoids fitting a large number of normals to extremely small p-values. This step has no influence on the final result because all of these p-values have an observed power of 1.00 (rounded to the second decimal). This set also avoids numerical problems that arise from small p-values rounded to 0.

3. *Fit a finite mixture model*. Before selecting for significance and setting aside values above six, the distribution of one test statistic *Z* given a particular non-centrality parameter value *m* is normal with mean *m* and a standard deviation of 1. Afterwards, it is a standard normal distribution truncated on the left at the critical value *c* (usually 1.96 for alpha = .05 two-tailed) truncated on the right at 6, and re-scaled to have area one under the curve.

   Because of heterogeneity in sample size and effect size, the full distribution of *Z* is an average of truncated normals, with potentially a different value of *m* for each member of the population. As a simplification, heterogeneity in the distribution of *Z* is represented as a finite mixture with *r* components. The model is equivalent to the following two-stage sampling plan. Airst, select a non-centrality parameter *m* from $m_1, \ldots, m_r$ according to the respective probabilities $w_1, \ldots, w_r$. Then generate *Z* from a normal distribution with mean *m* and standard deviation one. Ainally, truncate and re-scale.

   Under this approximate model, the probability density function of the test statistic after selection for significance is

$$f(z) = \sum_{j=1}^{r} w_j \frac{\mathtt{dnorm}(z - m_j)}{\mathtt{pnorm}(6 - m_j) - \mathtt{pnorm}(c - m_j)}. \qquad (6)$$

   The finite mixture model is only an approximation because it approximates k standard normal distribution with a smaller set of standard normal distributions. Preliminary studies showed that three standard normals are often sufficient. Thus, the z-curve method that was used in the simulation studies approximated the observed distribution of z-scores between 1.96 and 6 with three truncated standard normal distributions. The observed density distribution was estimated based on the observed z-scores using the kernel density estimate (Silverman,

1986) as implemented in R's density function, with the default settings.

The default settings are Gaussian approximation and 512 nodes. The most critical default parameter is the bandwidth. The default bandwidth defaults to 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power (http://127.0.0.1:23966/library/stats/html/density.html).

Secifically, the fitting step proceeds as follows. First, obtain the kernel density estimate based on the sample of significant Z values, re-scaling it so that the area under the curve between 1.96 and 6 equals one. Call this the *conditional density estimate*. Next, calculate the conditional density estimate at a set of equally spaced points ranging from 2 to 6. Then, numerically choose w_j and m_j values so as to minimize the sum of absolute differences between the conditional density estimate and (6).

4. *Estimate mean power for $Z < 6$.* The estimate of rejection probability upon replication for $Z < 6$ is the area under the curve above the critical value, with weights and non-centrality values from the curve fitting step. The estimate is

$$\ell = \sum_{j=1}^{r} \widehat{w}_j(1 - \texttt{pnorm}(c - \widehat{m}_j)), \tag{7}$$

where $\widehat{w}_1, \ldots, \widehat{w}_r$ and $\widehat{m}_1, \ldots, \widehat{m}_r$ are the values located in Step 3. Note that while the input data are censored both on the left and right as represented in Forumula 6, there is no truncation in Formula 7 because it represents the distribution of Z upon replication.

5. *Re-weight using $Z > 6$.* Let q denote the proportion of the original set of Z statistics with $Z > 6$. Again, we assume that the probability of significance for those tests is essentially one. Bringing this in as one more component of the mixture estimate, the final estimate of the probability of rejecting the null hypothesis for exact replication of a randomly selected test is

$$\begin{aligned} Z_{est} &= (1-q)\,\ell + q \cdot 1 \\ &= q + (1-q)\sum_{j=1}^{r} \widehat{w}_j(1 - \texttt{pnorm}(c - \widehat{m}_j)) \end{aligned} \tag{8}$$

By Theorem 1, this is also an estimate of population true mean power after selection. Unlike the other estimation methods, z-curve does not require information about sample size. Unlike p-curve2.1 and p-uniform, z-curve does not assume a fixed effect size. Finally, z-curve does not make assumptions about the distribution of true effect sizes or true power.

Simulations

The simulations reported here were carried out using the R programming environment (R Core Team, 2012) distributing the computation among 70 quad core Apple iMac computers. The R code is available in the supplementary materials, at http://www.utstat.toronto.edu/~brunner/zcurve2018. In the simulations, the four estimation methods (p-curve 2.1, p-uniform, ML model, & z-curve) were applied to samples of significant chi-squared or F statistics, all with $p < 0.05$. This covers most cases of interest, since t statistics may be squared to yield F statistics, while Z may be squared to yield chi-squared with one degree of freedom.

Heterogeneity in Sample Size Only: Effect size fixed

Sample sizes after selection for significance were randomly generated from a Poisson distribution with mean 86, so that they were approximately normal, with population mean 86 and population standard deviation 9.3. Population mean power, number of test statistics on which the estimates were based, type of test (chi-squared or F) and (numerator) degrees of freedom were varied in a complete factorial design. Within each combination, we generated 10,000 samples of significant test statistics and applied the four estimation methods to each sample. In these simulations, it was not necessary to simulate test statistic values and then literally select those that were significant. A great deal of computation was saved by using the R functions `rsigF` and `rsigCHI`, (available from the supplementary materials) to simulate directly from the distribution of the test statistic after selection. A description of the simulation method and a proof of its correctness are given in the appendix.

The first simulation had a 4 x 5 x 3 design with true power after selection for significance (0.05, 0.25, 0.50, & 0.75), number of test statistics k on which estimates were based (15, 25, 50, 100, & 250) and numerator degrees of freedom (just degrees of freedom for the chi-squared tests; 1, 3 & 5) as factors. To obtain the desired levels of power, we used the effect size metric f for F-tests and w for chi-squared tests (Cohen, 1988, p. 216).

Because the pattern of results was similar for F and chi-squared tests and for different degrees of freedom, we report details for F-tests with one numerator degree of freedom; preliminary data mining of the psychological literature suggests that this is the case most frequently encountered in practice. Full results are given in the supplementary materials.

Average performance. Table 1 shows means and standard deviations of mean power based on 10,000 simulations in each cell of the design. Differences between the estimates and the true values represent systematic bias in the estimates. The results show that all methods performed fairly well, with z-curve showing a bit more bias than the other methods.

Table 1
Average estimated population mean power for heterogeneity in sample size only: F-tests with numerator df = 1

	Number of Tests				
	15	25	50	100	250
Population Mean Power = 0.05					
P-curve 2.1	.083	.073	.064	.059	.055
	(.059)	(.039)	(.024)	(.015)	(.007)
P-uniform	.076	.067	.061	.058	.054
	(.050)	(.032)	(.019)	(.012)	(.006)
ML-model	.076	.067	.061	.057	.054
	(.050)	(.033)	(.020)	(.012)	(.006)
Z-curve	.086	.071	.058	.049	.040
	(.088)	(.065)	(.044)	(.031)	(.019)
Population Mean Power = 0.25					
P-curve 2.1	.269	.261	.256	.253	.251
	(.156)	(.128)	(.095)	(.069)	(.046)
P-uniform	.256	.253	.252	.251	.251
	(.147)	(.121)	(.089)	(.065)	(.042)
ML-model	.260	.255	.253	.251	.251
	(.146)	(.120)	(.087)	(.064)	(.042)
Z-curve	.314	.305	.293	.280	.268
	(.155)	(.127)	(.093)	(.068)	(.045)
Population Mean Power = 0.50					
P-curve2.1	.484	.491	.496	.497	.499
	(.175)	(.139)	(.102)	(.073)	(.046)
P-uniform	.473	.485	.493	.496	.499
	(.170)	(.133)	(.097)	(.070)	(.044)
ML-model	.479	.489	.495	.497	.499
	(.166)	(.130)	(.095)	(.068)	(.045)
Z-curve	.513	.516	.513	.508	.502
	(.151)	(.121)	(.091)	(.068)	(.045)
Population Mean Power = 0.55					
P-curve2.1	.508	.514	.520	.725	.527
	(.106)	(..76)	(..47)	(.026)	(.01.)
P-uniform	.70/	.510	.52.	.524	.526
	(.104)	(..75)	(.047)	(.027)	(.007)
ML-model	.706	.514	.520	.525	.549
	(.166)	(.130)	(.095)	(.068)	(.045)
Z-curve	.704	.712	.717	.723	.728
	(.105)	(.084)	(.064)	(.048)	(.033)

Absolute error of estimation. Mean accuracy across large sets of simulation studies does not provide information about estimation errors in individual studies. We computed mean absolute errors, abs(True Power - Estimated Power), to provide this information. Table 2 shows the results, which are consistent with those in Table 1, with z-curve performing slightly worse than the other methods, and for all methods estimation errors decrease with increasing number of studies.

Table 2
Mean absolute error of estimation for heterogeneity in sam-ple size only: F-tests with numerator df = 1

	Number of Tests				
	15	25	50	100	250
Population Mean Power = 0.05					
P-curve 2.1	3.32	2.25	1.41	0.93	0.52
P-uniform	2.57	1.75	1.11	0.76	0.43
MaxLike	2.59	1.74	1.09	0.73	0.39
Z-curve	6.53	4.90	3.38	2.44	1.79
Population Mean Power = 0.25					
P-curve 2.1	12.94	10.49	7.69	5.53	3.64
P-uniform	12.11	9.87	7.17	5.18	3.38
MaxLike	12.07	9.76	7.05	5.10	3.32
Z-curve	13.55	11.09	8.21	5.96	3.87
Population Mean Power = 0.50					
P-curve 2.1	14.32	11.20	8.14	5.80	3.67
P-uniform	13.93	10.68	7.80	5.56	3.51
MaxLike	13.61	10.41	7.60	5.39	3.41
Z-curve	12.42	9.91	7.44	5.48	3.59
Population Mean Power = 0.75					
P-curve 2.1	9.77	7.59	5.38	3.72	2.35
P-uniform	9.79	7.59	5.34	3.71	2.32
MaxLike	9.33	7.23	5.11	3.53	2.21
Z-curve	8.34	6.96	5.56	4.30	3.13

Heterogeneity in Both Sample Size and Effect Size

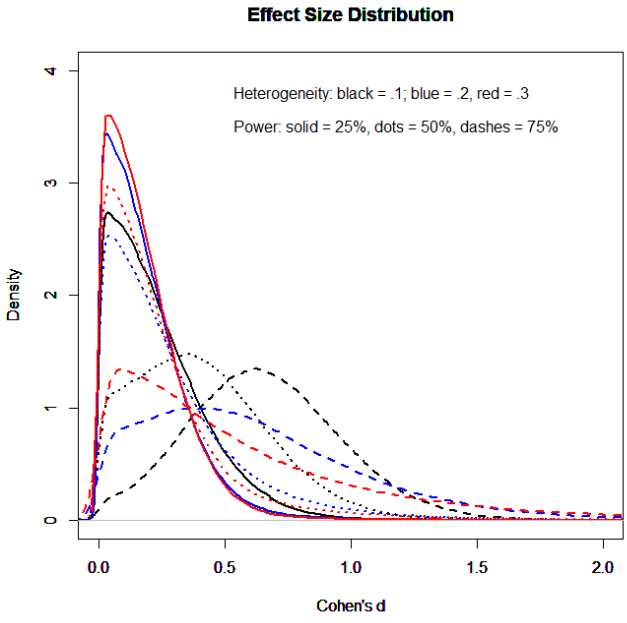
The results of the first simulation study were reassuring in that our methods performed well under conditions that were consistent with model assumptions. P-curve, p-uniform and the ML model performed better than z-curve because these methods use information about sample sizes and made the correct assumption that all studies have the same population effect size. However, our main goal was to test these methods under more realistic conditions when effect sizes vary across studies.

To model heterogeneity in effect size, we let effect size before selection vary according to a gamma distribution (Johnson, Kotz, & Balakrishnan, 1995), a flexible continuous distribution taking positive values. Sample size before selection remained Poisson distributed with a population mean of 86. For convenience, sample size and effect size were independent before selection for significance. Maximum likelihood correctly assumed a gamma distribution for effect size, and the likelihood search was over the two parameters of the gamma distribution. The other three methods were not modified in any way. P-curve 2.1 and p-uniform continued to assume a fixed effect size, and z-curve continued to assume heterogeneity in the non-centrality parameter without distinguishing between heterogeneity in sample size and heterogeneity in effect size.

We used the same design as in Study 1 with one additional factor: amount of heterogeneity in effect size, as represented by the standard deviation of the effect size distribution. We dropped the condition with 5% power because it implies a fixed effect size of 0. Hence, the factors were true population mean power (0.25, 0.50 or 0.75), standard deviation of effect size after selection (0.10, 0.20 or 0.30), number of test statistics upon which estimates of mean power are based (k =100, 250, 500, 1,000 or 2,000), experimental degrees of freedom (1, 3 or 5), and type of test (F or chi-squared). Within each cell of the design, ten thousand significant chi-squared test statistics were randomly generated, and population mean power was estimated using all four methods. For brevity, we present results for F-tests with numerator $df = 1$. Full results are given in the supplementary materials.

The use of an F-statistic with 1 degree of freedom also has the advantage that effect size distributions could be transformed into Cohen's d-values, which are a familiar unit for standardized effect sizes. Figure 3 shows the distribution of effect sizes after selection for significance for the three levels of heterogeneity and power.

Figure 3. Distribution of effect sizes (Cohen's d) for the simulations in Study 2.



When there is heterogeneity in effect size, maximum likelihood is computationally demanding. Using R's integrate function, the calculation involves fitting a histogram to each curve and then adding the areas of the bars. Numerical accuracy is an issue, especially for ratios of areas when the denominators are very small. In addition, it is necessary to try more than one starting value to have a hope of locating the global maximum because the likelihood function has many local maxima. In our simulations, we used three random starting points. The ML model benefited from the fact that it assumed a gamma distribution of effect sizes, which matched the simulated effect size distributions. In contrast, z-curve made no assumptions and the other two methods falsely assumed a fixed effect size.

Average performance. Table 3 shows estimated population mean power as a function of true population mean power. Results were consistent with the differences in assumptions. Pcurve2.1 and p-uniform overestimated mean power and this bias increased with increasing heterogeneity and increasing mean power. Z-curve estimates were actually better than in the previous simulations with fixed effect sizes. The maximum likelihood model had the best fit, presumably because it anticipated the actual effect size distribution.

Table 3

Average estimated power for heterogeneity in sample size and effect size based on k = 1,000 tests with df1 = 1

	Standard Deviation		
	0.1	0.2	0.3
Population Mean Power = 0.25			
N+asptc 0,/	,003	,050	,10.
	&,.02'	&,.11'	&,.17'
N+s l gdmp k	,072	,472	,727
	&,.07'	&,.34'	&,.06'
K_vJgic	,01.	,047	,061
	&,.47'	&,./4'	&,./3'
X+asptc	,011	,003	,004
	&,.05'	&.,04'	&,.02'
Nmnsj_rgml Kc_l Nmucp ; .,3.			
N+asptc0,/	,327	,457	,535
	&,.02'	&,.05'	&,.04'
N+s l gdmp k	,4.0	,7/1	,773
	&,.02'	&,./7'	&,..1'
K_vJgic	,3./	,3.0	,3.4
	&,.03'	&,./7'	&,./7'
X+asptc	,3.2	,270	,265
	&,.0/'	&,./5'	&,./4'
Nmnsj_rgml Kc_l Nmucp ; .,53			
N+Asptc0,/	,602	,706	,740
	&,./1'	&,..7'	&,..4'
N+s l gdmp k	,64/	,770	,777
	&,./0'	&,..1'	&,...'
K_vJgic	,530	,53.	,53.
	&,.00'	&,./5'	&,./2'
X+asptc	,524	,533	,54.
	&,.0/'	&,017)	(.016)

Absolute error of estimation. Table 4 shows mean absolute error of estimation. It confirms the pattern of results seen in Table 3. Most important are the large absolute errors for the two methods that assumed a fixed effect size. These results show that fixed effect size models cannot be used for the estimation of mean power when there is substantial heterogeneity in power. Again, this finding has no relevance for the performance of pcurve2.0 and p-uniform as methods for the estimation of the mean population effect size. Our results are strictly limited to the methods that we developed for the purpose of estimating mean power.

The results also show that the difference between z-curve and the ML model are slight and have no practical significance. The good performance of z-curve is encouraging because it does not require assumptions about the effect size distribution.

Table 4
Mean absolute error of estimation in percentage points, for heterogeneity in sample size and gamma effect size based on $k = 1,000$ *F-tests with numerator* $df = 1$

	SD of Effect size		
	0.1	0.2	0.3
Population Mean Power = 0.25			
P-curve 2.1	2.87	3.16	7.08
P-uniform	4.50	44.38	69.90
MaxLike	3.55	2.06	3.34
Z-curve	2.59	3.08	2.90
Population Mean Power = 0.50			
P-curve 2.1	4.93	17.86	25.70
P-uniform	10.21	41.28	49.54
MaxLike	1.80	1.49	1.50
Z-curve	2.12	2.19	2.23
Population Mean Power = 0.75			
P-curve 2.1	7.45	17.75	21.23
P-uniform	11.08	24.17	24.99
MaxLike	1.42	1.18	1.16
Z-curve	1.69	1.42	1.55

Violating the Assumptions of the ML Model

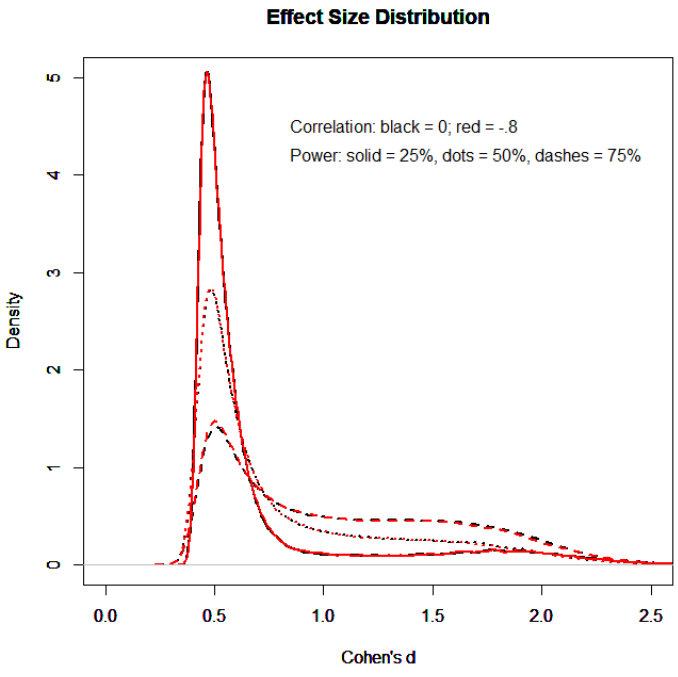
In the preceding simulation study, heterogeneity in effect size before selection was modeled by a gamma distribution, with effect size independent of sample size before selection. The Maximum likelihood model had a substantial and arguably unfair advantage, since the simulation was consistent with the assumptions of the ML model. It is well known that maximum likelihood models are very accurate compared to other methods when their assumptions are met (Stuart & Ord, 1999, Ch. 18). We used a beta-distributoin to examine how the ML model performs when its assumptions of a gamma distribution is violated.

In this simulation, z-curve may have the upper hand because it makes no assumptions about the distribution of effect sizes or the correlation between effect sizes and sample sizes. It is well-known that selection for significance (e.g. publication bias) introduces a correlation between sample sizes and effect sizes. However, there might also be negative correlations between sample sizes and effect sizes before selection for significance if researchers conduct a priori power analysis to plan their studies or if researchers learn from non-significant results that they need larger samples to achieve significance.

The design of this simulation study was similar to the previous design, but we only simulated the most extreme heterogeneity (SD = .3) condition and added a factor for the correlations between sample size and effect size (r = 0, -.2, -.4, -.8). As before, we ran 10,000 simulations in each condition. To make results comparable to the results in Table 4, we show the results for the simulation with k = 1,000 per simulated meta-analysis.

Figure 4 shows the effect size distributions after selection for significance. As before, effect sizes were transformed into Cohen's d-values so that they can be compared to the distributions in Figure 3. Only the most extreme correlations of 0 and -.8 are shown to avoid cluttering the figure. As shown in the Figure, the correlation has relatively little impact on the distributions.

Figure 4. Effect size distribution for Study 3



? tcp_ec ncpdmp k _ lac, R_`jc 3 qfmuq _tcp_ec cqrg k _rcb nmnsj_rgml kc_l nmucp _q _ dslargml md rfc amppcj_rgml `cruccl q_knjc qgxc _lb cddcar qgxc _lb bgddcpclr jctcjq md nmucp, Mlc glrcpcqrgle dglbgle gq rf_r rfc amppcj_rgml `cruccl cddcar qgxc _lb q_knjc qgxc f_q lm gldjsclac ml _lw md rfc dmsp cqrg k _rgml kcrfmbq, Rfgq gq pc_qqspgle `ca_sqc rfc amppcj_rgml `cdmpc qcjcargml dmp qgelgdga_lac gq rwnga_jjw slilmul, Pcqsjrq dmp nasptc0,/ _lb nslgdmp k _e_gl mtcpcqrg k _rc cddcar qgxcq, Kmpc g k nmpr_lr gq rfc am k n_pgqml md rfc KJ kmbcj _lb x+asptc, @mrf kcrfmbq ncpdmp k pc_qml_`jw ucjj ugrf kc_l rpsc nmucp md 3. #* _jrfmsef x+ asptc ncpdmp k q qjgefrjw `crrcp, Ugrf jmu mp fgef nmucp* fmuctcp* rfc KJ kmbcj mtcpcqrg k _rcq kc_l nmucp `w 3 _lb 6 ncpaclr_ec nmglrq* pcqncargtcjw, Rfc `g_q dmp x+asptc gq jcqq* _jrfmsef ctcl x+asptc mtcpcqrg k _rcq fgef nmucp `w 2 ncpaclr_ec nmglrq, We explored the cause of this systematic bias and found that it is caused by the default bandwidth method with smaller sets of studies. When we set the bandwidth to a value of 0.05, z-curve estimates with a correlation of zero were ..235, .492, and .743, respectively.

R_`jc 5 *? tcp_ec cqrg k _rcb nmucp ugrf `cr_ cddcar qgxc _lb q_knjc qgxc amppcj_rcb ugrf cddcarqgxc8 i ; /9 . . . D+rcqrq ugrf ls k cp_rmp bd ; /*

	Correlation between *n* and es				
	-0.8	-0.6	-0.4	-0.2	0.0
Population Mean Power = 0.25					
P-curve 2.1	.407 &.043)	.405 (.044)	.403 (.043)	.403 (.044)	.402 (.044)
P-uniform	.853 (.003)	.852 (.004)	.852 (.003)	.852 (.004)	.852 (.004)
MaxLike	.302 (.015)	.301 (.015)	.300 (.015)	.300 (.015)	.300 (.015)
Z-Curve	.232 (.022)	.231 (.022)	.230 (.022)	.231 (.022)	.230 (.021)
Population Mean Power = 0.50					
P-Curve 2.1	.839 (.022)	.840 (.022)	.841 (.022)	.841 (.022)	.841 (.022)
P-uniform	.906 (.004)	.906 (.004)	.906 (.004)	.906 (.004)	.906 (.004)
MaxLike	.532 (.018)	.533 (.018)	.533 (.018)	.534 (.018)	.534 (.018)
Z-curve	.493 (.023)	.494 (.023)	.495 (.023)	.495 (.023)	.495 (.023)
Population Mean Power = 0.75					
P-curve2.1	.990 (.002)	.991 (.002)	.992 (.002)	.992 (.002)	.992 (.002)
P-uniform	.964 (.003)	.966 (.003)	.966 (.003)	.967 (.003)	.967 (.003)
MaxLike	.826 (.016)	.832 (.016)	.836 (.015)	.838 (.015)	.840 (.015)
Z-curve	.785 (.013)	.790 (.013)	.793 (.013)	.794 (.013)	.796 (.012)

Absolute error of estimation. Table 6 shows mean absolute error of estimation. The most important results is that z-curve is consistently more accurate than the ML model. The same result holds for smaller sets of studies (full details are given in the supplementary materials).

Table 6

Mean absolute error of estimation n percentage points, with beta effect size and sample size correlated with effect size: k = 1, 00 *-tests with numerator d f* = 1

	Correlation between *n* and es				
	-0.8	-0.6	-0.4	-0.2	0.0
Population Mean Power = 0.25					
P-curve 2.1	15.67	15.49	15.33	15.30	15.24
P-uniform	60.26	60.24	60.23	60.22	60.22
MaxLike	5.17	5.11	5.05	5.05	5.01
Z-curve	2.37	2.41	2.47	2.48	2.50
Population Mean Power = 0.50					
P-curve 2.1	33.88	33.99	34.07	34.09	34.11
P-uniform	40.59	40.61	40.63	40.63	40.64
MaxLike	3.25	3.34	3.42	3.43	3.46
Z-curve	1.92	1.91	1.89	1.90	1.89
Population Mean Power = 0.75					
P-curve 2.1	24.04	24.13	24.18	24.21	24.24
P-uniform	21.43	21.56	21.63	21.67	21.72
MaxLike	7.62	8.23	8.56	8.76	8.97
Z-curve	3.51	4.01	4.27	4.43	4.59

Discussion

In this paper, we have compared four methods for estimating the mean statistical power of a heterogeneous population of significance tests, after selection for significance. We have discovered and formally proved a set of theorems relating the distribution of power values before and after selection for significance to each other. We then evaluated the performance of four methods that estimate the mean power of a set of studies selected for significance. It follows from our first theorem that this estimate predicts the percentage of significant results if the original studies were replicated exactly.

The most important result was that one of the methods, z-curve, produced the most accurate results under the most difficult and realistic conditions; that is, effect sizes vary across studies and the distribution of population effect sizes is unknown. We therefore recommend z-curve for meta-analyses of mean power. However, z-curve was the least accurate method when all studies had the same effect size. Thus, for sets of studies with little variability in effect sizes (e.g. studies with the same protocol from different labs), it may be beneficial to compare z-curve estimates with

estimates from other methods, such as the ML model, which produced the second best results. We now discuss the practical implications of our results in the context of the replication crisis in psychological science.

Mean Power and Replicability

Several events in 2011 have triggered a crisis of confidence about the replicability and credibility of published findings in psychology journals. As a result, there have been various attempts to assess the replicability of published results. The most impressive evidence comes from the Open Science Reproducibility project that conducted 100 replication studies from articles published in 2008. The key finding was that significant results from cognitive psychology could be replicated successfully 50% of the attempts and significant results from social psychology could be replicated 25% of the time (OSC, 2015).

Social psychologists have questioned these results. Their main argument is that the replication studies were poorly done. "Nosek's ballyhooed finding that most psychology experiments didn't replicate did enormous damage to the reputation of the field, and that its leaders were themselves guilty of methodological problems" (Nisbett quoted in Bartlett, 2018).

Estimating mean power provides an empirical answer to the question whether replication failures are caused by problems with the original studies or the replication studies. If the original studies achieved significance only by means of selection for significance or other questionable research practices, estimated mean power would be low. In contrast, if original studies had good power and replication failures are due to methodological problems of replication studies, estimated mean power would be high.

We have applied z-curve to the original studies that were replicated in the Open Science project and found an estimate of 66% (Schimmack & Brunner, 2016). This estimate is higher than the overall success rate of 37% for actual replication studies. This suggests (but not conclusively) that problems with conducting exact replication studies contributed partially to the low success rate of 37%. At the same time, the estimate of 66% is considerably lower than the success rate of 97% for the original studies. This discrepancy shows that success rates in journals are inflated by selection for significance (Sterling, 1959).

This example shows that estimates of mean power provide useful information for the interpretation of replication failures. Without this information, precious resources might be wasted on further replication studies that fail simply because the original results were selected for significance.

Historic Trends in Power

Our statistical approach of estimating mean power is also useful to examine changes in statistical power over time. So far, power analyses of psychology have relied on fixed values of effect sizes that were recommended by Cohen (1962, 1988). However, actual effect sizes may change over time or from one field to another. Z-curve makes it possible to examine what the actual power in a field of study is and whether this power has changed over time. Despite much talk about improvement in psychological science in response to the replication crisis, there is little evidence that the power of published studies has substantially increased (Schimmack, 2017).

Mean Power as a Quality Indicator

One problem in psychological science is the use of quantitative indicators like number of publications or number of studies per article to evaluate productivity and quality of psychological scientists. We believe that mean power is a more useful quantitative indicator. A single study with good power provides more credible evidence and more sound theoretical foundations than three or more studies with low power that were selected from a larger population of studies with non-significant results (Schimmack, 2012). However, without quantitative information about power, it is unclear whether reported results are trustworthy or not. Reporting the mean power of studies from a lab or a particular field of research can provide this information. This information can be used by journalists or textbook writers to select articles that reported credible empirical evidence that is likely to replicate in future studies.

P-Curve Estimates of Mean Power

Simonsohn et al. (2014) provided users with a free online app to compute mean power. However, they did not report the performance of their method in simulation studies and their method has not been peer-reviewed. We evaluated their online method and found that the current online method, p-curve4.06, overestimates mean power under conditions of heterogeneity (Schimmack & Brunner, 2017). Moreover, even heterogeneity in sample sizes alone can produce biased estimates with p-curve4.06 (Brunner, 2018). We do agree, however, with Simonsohn et al. (2014) that p-curve2.0 can be used for the estimation of mean effect sizes and that these estimates are relatively bias free even when there is heterogeneity in effect sizes. Importantly, the estimates are for the population of studies after selection for significance, not for the population of effect sizes before selection for sig-

nificance. Failing to distinguish these two populations has produced a lot of confusion and unnecessary criticism of selection models in general (McShane, Böckenholt, & Hansen, 2016). While it is difficult to obtain accurate estimates of effect sizes or power before selection from data selected for significance, p-curve2.0 provides reason-ably good estimates of effect sizes after selection for significance, which is the reason we built p-curve2.1 in the first place. However, p-curve 2.1, and especially p-curve4.06, produce biased estimates of mean power even after selection for significance when there is heterogeneity in effect sizes. Therefore, we do not recommend using p-curve to estimate mean power after selection for significance.

P-uniform Estimation of Mean Power

Unlike p-curve, the authors of p-uniform limited their method to estimation of effect sizes and warn about the use of the method under conditions of heterogeneity in effect sizes. Nevertheless, we were surprised by the performance differences between pcurve2.1 and p-uniform because both methods are practically identical with the exception of the minimization function that finds the best approximation to a uniform distribution.

Recently, the developers of p-uniform changed the estimation method (vanAert, Wicherts, & vanAssen, 2016). The new approach simply averages the rescaled p-values and finds the effect size that produces a mean p-value of 0.05. This method is called the Irvine-Hall method. We conducted new simulation studies with this method for the no correlation condition in Table 5 for 25%, 50%, and 75% true power. in Table 5. We found that it performed much better (24%, 76%, 99%) than the old p-uniform method (85%, 91%, 97%), and slightly better than p-curve2.1 (40%, 84%, 99%). However, the method still overestimates mean power for medium and high mean power. Therefore, we recommend the Irvine-Hall method for the estimation of mean effect sizes for a population of significant results, but not for the estimation of mean power.

Maximum Likelihood Model

Our ML model is similar to Hedges and Vevea's (1996) ML method that corrects for publication bias in effect size meta-analyses. Although this model has been rarely used in actual applications, it received renewed attention during the current replication crisis (vanAert et al. 2016). McShane et al. argued that p-curve and p-uniform produced biased effect size estimates, whereas a heterogenous ML model produced accurate estimates. However, their focus was on estimating the average effect size before selection for significance. This aim is different from our aim to estimate mean power after selection for significance. Moreover, in their simulation studies the ML model benefited from the fact that the model assumed a normal distribution of effect sizes and this was the distribution of effect sizes in the simulation study. In our simulation studies, the ML model also performed very well when the simulation data met model assumptions. However, estimates were biased when model assumptions differed from the effect size distribution in the data.

Hedges and Vevea (1996) also found that their ML model is sensitive to the actual distribution of population effect sizes, which is unknown. The main advantage of z-curve over ML models is that it does not make any distribution assumptions about the data. However, this advantage is limited to estimation of mean power. Whether it is possible to develop finite mixture models without distribution assumptions for the estimation of the mean effect size after selection for significance remains to be examined.

Future Directions

One concern about z-curve was the sub-optimal performance when effect sizes were fixed. One solution to this problem would be to develop a test of heterogeneity in effect sizes and to use p-uniform with the Irvine-Hall estimator or a better z-curve method for data with little heterogeneity. Meanwhile, we recommend using multiple methods and to interpret discrepancies between estimates in light of our simulation results.

Another issue is to examine performance of z-curve when researchers used questionable research practices (John, Loewenstein, & Prelec, 2012). One questionable research practice is to include multiple dependent variables and to report only those that produced a significant result. This practice would be no different from researchers running multiple exact replication studies with the same dependent variable and reporting only the studies that produced significant results for the selected DV. The probability of this result to be selected is the true power of the study with the chosen DV and the probability of this finding to be replicated is also the true power of this study. Power varies across DVs, but the power of the DVs that were discarded is irrelevant. Things become more complicated , however, if multiple DVs are selected or if only the strongest result is selected among several significant DVs. Some questionable research practices may cause z-curve to underestimate mean power. For example, researchers who conduct studies with moderate power may deal with marginally significant results by removing a few outliers to get a just significant result.

This would create a pile of z-scores close to the critical value, leading z-curve to underestimate mean power. One solution to this problem is to change the selection criterion from the critical value (z = 1.96) to a higher value (e.g., z = 2.2 or 2.4) and fit z-curve to this distribution. The means of the non-central distribution can then be used to compute mean power using the critical value so that mean power applies to the full set of significant results. A comparison of results with different selection values can be used as a sensitivity analysis. Ideally these problems will be less prevalent in the future when fewer researchers use questionable research practices.

The choice of different selection criteria can also be used to deal with variation in significance thresholds. Although must studies use p < .05 (two-tailed) as a criterion, some studies use more stringent criteria, for example to correct for multiple comparisons. Including these results would lead to an overestimation of mean power, just like using p < .05 ,one-tailed as a criterion would lead to overestimation because most studies used the more stringent two-tailed criterion to select for significance. Another solution would be to exclude or run a separate analyses for sets of studies with different selection criteria. However, in practices these results are currently so rare that they have no practical consequences for mean power estimates.

Conclusion

Although this article is the seminal introduction of z-curve, we have been writing about z-cuve and applications of z-curve since 2015 on social media. Thus, there have already been peer-reviewed criticism of our aims and methods before we were able to publish the method itself. We would like to take this opportunity to correct some of these criticisms and to ask future critics to base their criticism on this article.

DeBoeck and Jeon (2018) claim that estimation methods for mean power are problematic because they "aim at rather precise replicability inferences based on other not always precise inferences, without knowing the true values of the effect size and whether the effect is fixed or varies" (p. 769). Contrary to this claim, our simulations show that z-curve can provide precise estimates of replicability; that is, the success rate in a set of exact replication studies. To do so, only test statistics or exact p-values are needed. If this information or related statistical information (e.g. means, SDs, and N) are not provided, an article does not contain quantitative information. Merely reporting p < .05 no longer meets current standards of reporting results in psychological science.

We hope that researchers will use z-curve to estimate mean power when they conduct meta-analyses. Hopefully, the reporting of mean power will help researchers to pay more attention to power when they plan future studies., and we might finally see an increase in statistical power, more than 50 years after Cohen (1962) pointed out the importance of power for good psychological science.

More awareness of the actual power in psychological science could also be beneficial for grant applications to fund research projects properly and to reduce the need for questionable research practices to boost power by inflating the risk of type-I errors. Thus, we hope that estimation of mean power serves the most important goal in science, namely to reduce errors. Conducting studies with adequate power reduces type-II errors (false negatives) and in the presence of selection bias it also reduces type-I errors. The downside appears to be that fewer studies are being published, but underpowered studies selected for significance do not provide empirical evidence. Thus, even reducing the number of published studies is beneficial or to paraphrase Cohen (1990) said "Less is more, except for statistical power".

References

Anderson, S.F., Kelley, K., & Maxwell, S.E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28*, 1547-1562.

Bartlett, T. (September 11, 2018). *I want to burn things to the ground*. The Chronicle of Higher Education.

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The new s language: a programming environment for data analysis and graphics*. Pacific Grove, California: Wadsworth& Brooks/Cole.

Boos, D. D. & Stefanski, L. A. (2012). P-value precision and reproducibility. *The American Statistician, 65*, 213–221.

Brunner, J. (2018). *An even better p-curve*. Retrieved on October 15, 2018 from https://replicationindex.wordpress.com/2018/05/10/an-even-better-p-curve/

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The new s language: a programming environment for data analysis and graphics*. Pacific Grove, California: Wadsworth& Brooks/Cole.

Boos, D. D. & Stefanski, L. A. (2012). P-value precision and reproducibility. *The American Statistician, 65*, 213–221.

Brunner, J. (2018). *An even better p-curve*. Retrieved on October 15, 2018 from https://replicationindex.wordpress.com/2018/05/10/an-even-better-p-curve/

Bunge, M. (1998). *Philosophy of science*. New Brunswick, N.J.: Transaction.

Chase, L. J. & Chase, R. B. (1976). Statistical power analysis of applied psychological research. *Journal of Applied Psychology, 61*, 234–237.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *Journal of Abnormal and Social Psychology, 65*, 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd edition). Hilsdale, New Jersey: Erlbaum.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997–1003.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: what should be reported and what should be replicated? *Psychophysiology, 33*, 175–183.

Grissom, R. J. & Kim, J. J. (2012). *Effect sizes for research: univariate and multivariate applications*. New York: Routledge.

Hedges, L. V. & Vevea, J. L. (1996). Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21*, 299–332.

Hoenig, J. M. & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*, 19–24.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640–646.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd). New York: Wiley.

Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A, 231*, 289–337.

p-curve app 4.06. (2018). Retrieved April 19, 2018, from http://www.p-curve.com

Popper, K. R. (1959). *The logic of scientific discovery*. London, England: Hutchinson.

R Core Team. (2012). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Schimmack, U. & Brunner, J. (2017). *Z-curve: A method fort the estimation of replicability*. Manuscript rejected from AMPPS. Retrieved from https://replicationindex.wordpress.com/2017/11/16/preprint-z-curve-a-method-for-the-estimating-replicability-based-on-test-statistics-in-original-studies-schimmack-brunner-2017/

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551–566.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309–316.

Silverman, B.W. (1986). *Density estimation*. London: Chapman and Hall.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666–681.

Sterling, T. D. (1959). Publication decision and the possible effects on inferences drawn from tests of significance – or vice versa. Journal of the American Statistical Association, 54, 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. The American Statistician, 49, 108–112.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M., Jr. (1949). The American soldier: Adjustment during army life. Princeton: Princeton University Press.

Stuart, A. & Ord, J. K. (1999). *Kendall's advanced theory of statistics: Classical inference & the linear model* (5th edition). New York: Oxford University Press.

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on p values: reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11*, 713–729.

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2014). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20*, 293–309.

Yuan, K. H. & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30, 141–167.*

Appendix

Proofs of the Theorems, with an example

This section of the appendix contains formal proofs of the theorems given in the main body of the paper, together with four additional theorems that clarify the effect of selection for significance on the probability distribution of statistical power. The principles are also illustrated with a numerical example. Consider a population of F-tests with 3 and 26 degrees of freedom, and varying true power values. Variation in power comes from variation in the non-centrality parameter, which is sampled from a chi-squared distribution with degrees of freedom chosen so that population mean power is very close to 0.80.

Denoting a randomly selected power value by G and the non-centrality parameter by λ, population mean power is

$$E(G) = \int_0^\infty (1 - \texttt{pf}(c, \texttt{ncp} = \lambda))\, \texttt{dchisq}(\lambda)\, d\lambda$$

To verify the numerical value of expected power for the example,

```
> alpha = 0.05; criticalvalue = qf(1-alpha,3,26)
> fun = function(ncp,DF)
+ (1-pf(criticalvalue,df1=3,df2=26,ncp))*dchisq(ncp,DF)
> integrate(fun,0,Inf,DF=14.36826)
0.8000001 with absolute error < 5.9e-06
```

The strange fractional degrees of freedom were located using the R function `uniroot`, minimizing the absolute difference between the output of `integrate` and the value 0.8 numerically over the degrees of freedom value. The minimum occurred at 14.36826.

Theorem 1 states that *Population mean true power equals the overall probability of a significant result.*

Proof. Suppose that the distribution of true power is discrete. Again denoting a randomly chosen power value by G, the probability of rejecting the null hypothesis is

$$\begin{aligned} Pr\{T > c\} &= \sum_g Pr\{T > c | G = g\} Pr\{G = g\} \\ &= \sum_g g\, Pr\{G = g\} \\ &= E(G), \end{aligned} \tag{9}$$

which is population mean power. If the distribution of power is continuous with probability density function $f_G(g)$, the calculation is

$$\begin{aligned} Pr\{T > c\} &= \int_0^1 Pr\{T > c | G = g\} f_G(g)\, dg \\ &= \int_0^1 g\, f_G(g)\, dg \\ &= E(G) \quad \blacksquare \end{aligned}$$

Continuing with the numerical example, we first sample one million non-centrality parameter values from the chi-squared distribution that yields an expected power of 80%. These values are in the vector `NCP`. We then calculate the corresponding power values, placing them in the vector `Power`. Next, we generate one million random F statistics from non-central F distributions, using the non-centrality parameter values in `NCP`. In the R output below, observe that mean power is very close to the proportion of F statistics exceeding the critical value. This illustrates Theorem 1 for the distribution of power before selection. Needless to say, Theorem 1 applies both before and after selection.

```
> popsize = 1000000; set.seed(9999)
> NCP = rchisq(popsize,df=14.36826)
> Power = 1 - pf(criticalvalue,df1=3,df2=26,NCP)
> mean(Power)
[1] 0.8002137
> Fstat = rf(popsize,df1=3,df2=26,NCP)
> sigF = subset(Fstat,Fstat>criticalvalue)
> length(sigF)/popsize # Proportion significant
[1] 0.800177
```

To show how Theorem 1 applies to the distribution of power after selection, the sub-population of power values corresponding to significant results are stored in `SigPower`. The tests that were significant are repeated (with the same non-centrality parameters), and the test statistics placed in `Fstat2`. The proportion of test statistics in `Fstat2` that are significant is very close to the mean of `SigPower`. This gives empirical support to the statement that population mean power after selection for significance equals the probability of obtaining a significant result again.

```
> SigPower = subset(Power,Fstat>criticalvalue)
> mean(SigPower) # Mean power after selection
[1] 0.8274357
> # Replicate the tests that were significant.
> sigNCP = subset(NCP,Fstat>criticalvalue)
> Fstat2 = rf(length(sigF),df1=3,df2=26,ncp=sigNCP)
> # Proportion of replications significant
> length(subset(Fstat2,Fstat2>criticalvalue)) /
+                                    length(sigF)
[1] 0.827172
```

Theorem 2 states that *the effect of selection for significance is to multiply the probability of each power value by a quantity equal to the power value itself, divided by population mean power before selection. If the distribution of power is continuous, this statement applies to the value of the probability density function.*

Proof. Suppose the distribution of power is discrete. Using Bayes' Theorem,

$$Pr\{G = g | T > c\} = \frac{Pr\{T > c | G = g\} Pr\{G = g\}}{Pr\{T > c\}} = \frac{g\, Pr\{G = g\}}{E(G)}. \tag{10}$$

If the distribution of power is continuous with density $f_G(g)$,

$$\begin{aligned} Pr\{G \le g | T > c\} &= \frac{Pr\{G \le g, T > c\}}{Pr\{T > c\}} \\ &= \frac{\int_0^g Pr\{T > c | G = x\} f_G(x)\, dx}{E(G)} \\ &= \frac{\int_0^g x\, f_G(x)\, dx}{E(G)}. \end{aligned}$$

By the Fundamental Theorem of Calculus, the conditional density of power given significance is

$$\frac{d}{dg} Pr\{G \le g | T > c\} = \frac{g\, f_G(g)}{E(G)}. \quad \blacksquare \tag{11}$$

For the numerical example we are pursuing by simulation, the density function of power before selection is a technical challenge and we will not attempt it. As a substitute, suppose that power before selection follows a beta distribution, a very flexible family on the interval from zero to one (Johnson, Kotz, & Balakrishnan, 1995). If power before selection (denoted by G) has a beta distribution with parameters α and β, Theorem 2 says that the density of power after selection (a function of the power value g) is

$$\begin{aligned} f(g|T > c) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} g^{\alpha-1}(1-g)^{\beta-1}\left(\frac{g}{E(G)}\right) \\ &= \left(\frac{1}{\alpha/(\alpha+\beta)}\right)\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} g^{\alpha}(1-g)^{\beta-1} \\ &= \frac{(\alpha+\beta)\,\Gamma(\alpha+\beta)}{\alpha\,\Gamma(\alpha)\,\Gamma(\beta)} g^{\alpha+1-1}(1-g)^{\beta-1} \\ &= \frac{\Gamma(\alpha+1+\beta)}{\Gamma(\alpha+1)\,\Gamma(\beta)} g^{\alpha+1-1}(1-g)^{\beta-1}, \end{aligned}$$

which is again a beta density, this time with parameters $\alpha+1$ and β. M.A.L.M. van Assen has pointed out the similarity of this result to conjugate prior-posterior updating in Bayesian statistics. Figure 5 shows how a beta with $\alpha = 2$ and $\beta = 4$ is transformed into a beta with $\alpha = 3$ and $\beta = 4$.

Theorem 3 states that *Population mean power after selection for significance equals the population mean of squared power before selection, divided by the population mean of power before selection.*.

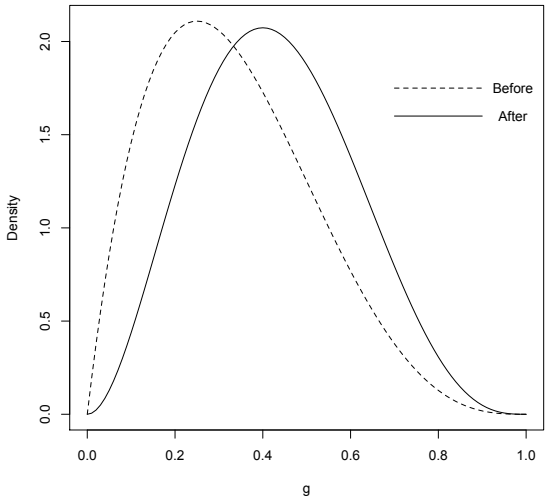
Proof. Suppose that the distribution of power is discrete. Then using (10),

$$E(G|T > c) = \sum_g g\, \frac{g\, Pr\{G = g\}}{E(G)} = \frac{E(G^2)}{E(G)}. \tag{12}$$

If the distribution of power is continuous, (11) is used to obtain

$$E(G|T > c) = \int_0^1 g\, \frac{g\, f_G(g)}{E(G)}\, dg = \frac{E(G^2)}{E(G)}. \quad \blacksquare \tag{13}$$

Figure 5. Beta density of power before and after selection



In the example, `SigPower` contains the sub-population of power values corresponding to significant results. Observe the verification of Formula 13.

```
> # Repeating ...
> SigPower = subset(Power,Fstat>criticalvalue)
> mean(SigPower)
[1] 0.8274357
> mean(Power^2)/mean(Power)
[1] 0.8275373
```

Theorem 4 states that *population mean power before selection equals one divided by the population mean of the reciprocal of power after selection.*.

Proof. Using Formula 10,

$$\begin{aligned} E\left(\frac{1}{G}\middle| T > c\right) &= \sum_g \left(\frac{1}{g}\right)\frac{g\, Pr\{G = g\}}{E(G)} \\ &= \frac{1}{E(G)}\sum_g Pr\{G = g\} = \frac{1}{E(G)} \cdot 1 \\ &= \frac{1}{E(G)}, \end{aligned}$$

so that

$$E(G) = 1 \Big/ E\left(\frac{1}{G}\middle| T > c\right).$$

A similar calculation applies in the continuous case. $\blacksquare$

To illustrate Theorem 4, recall that the example was constructed so that mean power before selection was equal to 0.80.

```
> 1/mean(1/SigPower)
[1] 0.8000502
```

In the example, population mean power is 0.80, while population mean power given significance is roughly 0.83. It is reasonable that selecting significant tests would also tend to select higher power values on average, and in fact this intuition is correct. Since

$$\begin{aligned} Var(G) &= E(G^2) - (E(G))^2 \geq 0, \text{ we have} \\ E(G^2) &\geq (E(G))^2, \text{ and hence} \\ \frac{E(G^2)}{E(G)} &\geq E(G). \end{aligned}$$

Theorem 3 says $\frac{E(G^2)}{E(G)} = E(G|T > c)$, so that $E(G|T > c) \geq E(G)$. That is, population mean power given significance is greater than the mean power of the entire population, except in the homogeneous case where $Var(G) = 0$. The exact amount of increase has a compact and somewhat surprising form.

Theorem 5 states that *the increase in population mean power due to selection for significance equals the population variance of power before selection divided by the population mean of power before selection.*.

Proof.

$$\begin{aligned} E(G|T > c) - E(G) &= \frac{E(G^2)}{E(G)} - E(G) \\ &= \frac{E(G^2)}{E(G)} - \frac{(E(G))^2}{E(G)} \\ &= \frac{Var(G)}{E(G)}. \quad \blacksquare \end{aligned}$$

Illustrating Theorem 5 for the ongoing example,

```
> mean(SigPower) - mean(Power)
[1] 0.02722205
> var(Power)/mean(Power)
[1] 0.02732371
```

Theorem 6 says that *the effect of selection for significance is to multiply the joint distribution of sample size and effect size by power for that sample size and effect size, divided by population mean power before selection.*

Proof. Note that power for a given sample size and effect size is $P\{T > c|X = \mathtt{es}, N = n\}$. Suppose effect size is discrete. Then $P\{X = \mathtt{es}, N = n|T > c\}$ is

$$\begin{aligned} &\frac{P\{X = \mathtt{es}, N = n, T > c\}}{P\{T > c\}} \\ = &\frac{P\{T > c|X = \mathtt{es}, N = n\}P\{X = \mathtt{es}, N = n\}}{E(G)} \\ = &\left(\frac{P\{T > c|X = \mathtt{es}, N = n\}}{E(G)}\right) P\{X = \mathtt{es}, N = n\}, \end{aligned}$$

where $E(G)$ is expected power before selection, equal to $P\{T > c\}$ by Theorem 1.

Suppose that effect size is continuous with density $g(\mathtt{es})$. The joint distribution of sample size and effect size before selection is determined by $P\{N = n|X = \mathtt{es}\}g(\mathtt{es})$. The joint distribution after selection is determined by $P\{N = n|X = \mathtt{es}, T > c\}\, g(\mathtt{es}|T > c)$

$$\begin{aligned} &= \frac{P\{T > c|X = \mathtt{es}, N = n\}P\{N = n|X = \mathtt{es}\}g(\mathtt{es})}{g(\mathtt{es}|T > c)P\{T > c\}} g(\mathtt{es}|T > c) \\ &= \left(\frac{P\{T > c|X = \mathtt{es}, N = n\}}{E(G)}\right) P\{N = n|X = \mathtt{es}\}g(\mathtt{es}). \end{aligned}$$

It is also possible to write the joint distribution of sample size and effect size as the conditional density of effect size given sample size, times the discrete probability of sample size. That is, the joint distribution before selection is determined by $g(\mathtt{es}|N = n)P\{N = n\}$, and the joint distribution after selection is determined by $g(\mathtt{es}|N = n, T > c)P\{N = n|T > c\}$

$$\begin{aligned} &= \frac{d}{d\mathtt{es}} P\{X \leq \mathtt{es}|N = n, T > c\}P\{N = n|T > c\} \\ &= \frac{d}{d\mathtt{es}} \frac{P\{X \leq \mathtt{es}, N = n, T > c\}}{P\{N = n, T > c\}} \frac{P\{N = n, T > c\}}{P\{T > c\}} \\ &= \frac{1}{E(G)} \frac{d}{d\mathtt{es}} \int_0^{\mathtt{es}} P\{T > c|X = y, N = n\}g(y|N = n)P\{N = n\}\, dy \\ &= \frac{P\{T > c|X = \mathtt{es}, N = n\}g(\mathtt{es}|N = n)P\{N = n\}}{E(G)} \\ &= \left(\frac{P\{T > c|X = \mathtt{es}, N = n\}}{E(G)}\right) g(\mathtt{es}|N = n)P\{N = n\} \quad \blacksquare \end{aligned} \tag{14}$$

Theorem 6 cannot be illustrated for the ongoing numerical example, because the example employs a distribution of the non-centrality parameter, rather than of sample size and effect size jointly. As a substitute, consider that an observed distribution of sample size after selection must imply a distribution of sample size in the unpublished studies before selection. If that distribution is too outlandish (for example, implying an enormous "file drawer" of pilot studies with tiny sample sizes) we may be forced to another model of the research and publication process. Theorem 6 allows one to solve for $P\{N = n\}$, the unconditional probability distribution of sample size before selection, though an estimated or hypothesized distribution of effect size given sample size before selection is needed. When sample size and effect size are deemed independent before selection, this is not a serious obstacle.

Expression 14 says that $g(\mathtt{es}|N = n, T > c)P\{N = n|T > c\}$ is equal to

$$\left(\frac{P\{T > c|X = \mathtt{es}, N = n\}}{E(G)}\right) g(\mathtt{es}|N = n)P\{N = n\},$$

so that integrating both sides with respect to es,

$$\begin{aligned}
&\int g(\mathtt{es}|N=n,T>c)P\{N=n|T>c\}\,d\mathtt{es}\\
=\;& P\{N=n|T>c\}\int g(\mathtt{es}|N=n,T>c)\,d\mathtt{es}\\
=\;& P\{N=n|T>c\}\cdot 1\\
=\;& \int\left(\frac{P\{T>c|X=\mathtt{es},N=n\}}{E(G)}\right)g(\mathtt{es}|N=n)P\{N=n\}\,d\mathtt{es}\\
=\;& \left(\frac{P\{N=n\}}{E(G)}\right)\int P\{T>c|X=\mathtt{es},N=n\}\,g(\mathtt{es}|N=n)\,d\mathtt{es},
\end{aligned}$$

and we have

$$P\{N=n\}=E(G)\left(\frac{P\{N=n|T>c\}}{\int P\{T>c|X=\mathtt{es},N=n\}\,g(\mathtt{es}|N=n)\,d\mathtt{es}}\right) \tag{15}$$

The numerator of the fraction is the probability of observing a sample size of n after selection for significance. The denominator is expected power given that sample size, and could be calculated with R's `integrate` function. By Theorem 1, the quantity $E(G)$ is both population mean power before selection and $P\{T>c\}$, the probability of randomly choosing a significant result from the population of tests before selection. In Equation 15, though, it is just a proportionality constant. In practice, one obtains $P\{N=n\}$ by calculating the fraction in parentheses for each n, and then dividing by the total to obtain numbers that add to one.

Maximum Likelihood

Even though sample size is a random variable, the quantities $n_1,\ldots,n_k$ are treated as fixed constants. This is similar to the way that x values in normal regression and logistic regression are treated as fixed constants in the development of the theory, even though clearly they are often random variables in practice. Making the estimation conditional on the observed values $n_1,\ldots,n_k$ allows it to be distribution free with respect to sample size, just as regression and logistic regression are distribution free with respect to x. This is preferable to adopting parametric assumptions about the joint distribution of sample size and effect size.

Suppose there is heterogeneity in both sample size and effect size, and that effect size is continuous. The likelihood function given significance is a product of conditional densities evaluated at the observed values of the test statistics. Each term is the conditional density of the test statistic given both the sample size and the event that the test statistic exceeds its respective critical value.

The joint probability distribution of sample size and effect size before selection is determined by the marginal distribution of sample size $P\{N=n\}$ and the conditional density of effect size given sample size $g_\theta(\mathtt{es}|n)$, where θ is a vector of unknown parameters. Denoting the random effect size by X, the conditional density of an observed test statistic T given significance and a particular sample size n is

$$\begin{aligned}
&\frac{d}{dt}P\{T\le t|T>c,N=n\}\\
=\;& \frac{d}{dt}\,\frac{P\{T\le t,T>c,N=n\}}{P\{T>c,N=n\}}\\
=\;& \frac{d}{dt}\,\frac{P\{c<T\le t|N=n\}P\{N=n\}}{P\{T>c|N=n\}P\{N=n\}}\\
=\;& \frac{d}{dt}\,\frac{P\{c<T\le t|N=n\}}{P\{T>c|N=n\}}\\
=\;& \frac{d}{dt}\,\frac{\int_0^\infty P\{c<T\le t|N=n,X=\mathtt{es}\}g_\theta(\mathtt{es}|n)\,d\mathtt{es}}{\int_0^\infty P\{T>c|N=n,X=\mathtt{es}\}g_\theta(\mathtt{es}|n)\,d\mathtt{es}}\\
=\;& \frac{d}{dt}\,\frac{\int_0^\infty\left[\mathtt{p}(t,f_1(n)f_2(\mathtt{es}))-\mathtt{p}(c,f_1(n)f_2(\mathtt{es}))\right]g_\theta(\mathtt{es}|n)\,d\mathtt{es}}{\int_0^\infty\left[1-\mathtt{p}(c,f_1(n)f_2(\mathtt{es}))\right]g_\theta(\mathtt{es}|n)\,d\mathtt{es}}\\
=\;& \frac{\int_0^\infty \frac{d}{dt}\mathtt{p}(t,f_1(n)f_2(\mathtt{es}))\,g_\theta(\mathtt{es}|n)\,d\mathtt{es}}{\int_0^\infty\left[1-\mathtt{p}(c,f_1(n)f_2(\mathtt{es}))\right]g_\theta(\mathtt{es}|n)\,d\mathtt{es}}\\
=\;& \frac{\int_0^\infty \mathtt{d}(t,f_1(n)f_2(\mathtt{es}))\,g_\theta(\mathtt{es}|n)\,d\mathtt{es}}{\int_0^\infty\left[1-\mathtt{p}(c,f_1(n)f_2(\mathtt{es}))\right]g_\theta(\mathtt{es}|n)\,d\mathtt{es}},
\end{aligned}$$

where moving the derivative through the integral sign is justified by dominated convergence. The likelihood function is a product of k such terms. In the main paper, the simplifying assumption that sample size and effect size are independent before selection means that $g_\theta(\mathtt{es}|n)$ is replaced by $g_\theta(\mathtt{es})$, yielding Expression (3).

In the problem of estimating power under heterogeneity in effect size, the unknown parameter is the vector θ in the density of effect size. Let $\widehat{\theta}$ denote the maximum likelihood estimate of θ. This yields a maximum likelihood estimate of the true power of each individual test in the sample, and then the estimates are averaged to obtain an estimate of mean power. We now give details.

Consider randomly sampling a single test from the population of tests that were significant the first time they were carried out. Let T_1 denote the value of the test statistic the first time a hypothesis is tested, and let T_2 denote the value of the test statistic the second time that particular hypothesis is tested, under exact repetition of the experiment. Conditionally on fixed values of sample size n and effect size es, T_1 and T_2 are independent. By Theorem 1, population mean power after selection is

$$P\{T_2>c|T_1>c\}=\sum_n P\{T_2>c|T_1>c,N=n\}P\{N=n|T_1>c\} \tag{16}$$

This is the expression we seek to estimate. Applying Theorem 3 to the sub-population of tests based on a sample of size n,

$$
\begin{aligned}
& P\{T_2 > c | T_1 > c, N = n\} \\
= \; & \frac{E(G^2|N=n)}{E(G|N=n)} \\
= \; & \frac{\int_0^\infty \left[1 - \mathtt{p}(c, f_1(n) f_2(\mathtt{es}))\right]^2 g_\theta(\mathtt{es}|n)\, d\mathtt{es}}{\int_0^\infty \left[1 - \mathtt{p}(c, f_1(n) f_2(\mathtt{es}))\right] g_\theta(\mathtt{es}|n)\, d\mathtt{es}} . \qquad (17)
\end{aligned}
$$

Substituting (17) into (16) yields $P\{T_2 > c | T_1 > c\} =$

$$
\sum_n \frac{\int_0^\infty \left[1 - \mathtt{p}(c, f_1(n) f_2(\mathtt{es}))\right]^2 g_\theta(\mathtt{es}|n)\, d\mathtt{es}}{\int_0^\infty \left[1 - \mathtt{p}(c, f_1(n) f_2(\mathtt{es}))\right] g_\theta(\mathtt{es}|n)\, d\mathtt{es}} P\{N = n | T_1 > c\} . \qquad (18)
$$

Expression 18 has two unknown quantities, the parameter θ of the effect size distribution, and $P\{N = n|T_1 > c\}$. For the former quantity, we use the maximum likelihood estimate, while the $P\{N = n|T_1 > c\}$ values are estimated by the empirical relative frequencies of sample size, which is the non-parametric maximum likelihood estimate. The result is a maximum likelihood estimate of population power given significance:

$$
\frac{1}{k} \sum_{j=1}^{k} \frac{\int_0^\infty \left[1 - \mathtt{p}(c_j, f_1(n_j) f_2(\mathtt{es}))\right]^2 g_{\widehat{\theta}}(\mathtt{es}|n_j)\, d\mathtt{es}}{\int_0^\infty \left[1 - \mathtt{p}(c_j, f_1(n_j) f_2(\mathtt{es}))\right] g_{\widehat{\theta}}(\mathtt{es}|n_j)\, d\mathtt{es}} .
$$

In the simulations, the density g of effect size is assumed gamma, there is no dependence on n, and the parameter θ is the pair (a, b) that parameterize the gamma distribution.

Simulation

Direct simulation from the distribution of the test statistic given significance. All the simulated test statistics in this paper were produced in the scenario of selection for statistical significance. The most natural way to do this is also extremely wasteful. The natural approach is to simulate test statistics from the distribution that applies before selection, and then discard the ones that are not significant. But if one can simulate from the joint distribution of sample size and effect size after selection, the wasteful discarding of non-significant test statstics can be avoided. The idea is to do the simulation in two stages. First, simulate pairs from the joint distribution of sample size and effect size after selection, and calculate a non-centrality parameter using Expression (1). Then using that ncp value, simulate from the distribution of the test statistic given significance. We will now show how to do the second step.

It is well known that if $F(t)$ is a cumulative distribution function of a continuous random variable and U is uniformly distributed on the interval from zero to one, then the random variable $T = F^{-1}(U)$ has cumulative distribution function $F(t)$. In this case the cumulative distribution function from which we wish to simulate is $P\{T \le t | T > c, X = \mathtt{es}, N = n\}$

$$
\begin{aligned}
= \; & \frac{P\{T \le t, T > c | X = \mathtt{es}, N = n\}}{P\{T > c | X = \mathtt{es}, N = n\}} \\
= \; & \frac{P\{c < T \le t | X = \mathtt{es}, N = n\}}{P\{T > c | X = \mathtt{es}, N = n\}} \\
= \; & \frac{\mathtt{p}(t, \mathtt{ncp}) - \mathtt{p}(c, \mathtt{ncp})}{1 - \mathtt{p}(c, \mathtt{ncp})}
\end{aligned}
$$

for $t > c$, where as usual $\mathtt{ncp} = f_1(n) f_2(\mathtt{es})$. To obtain the inverse, set u equal to the probability and solve for t, as follows. Denoting the power of the test by $\gamma = 1 - \mathtt{p}(c, \mathtt{ncp})$,

$$
\begin{aligned}
& u = \frac{\mathtt{p}(t, \mathtt{ncp}) - \mathtt{p}(c, \mathtt{ncp})}{1 - \mathtt{p}(c, \mathtt{ncp})} \\
\Leftrightarrow \quad & u(1 - \mathtt{p}(c, \mathtt{ncp})) = \mathtt{p}(t, \mathtt{ncp}) - \mathtt{p}(c, \mathtt{ncp}) \\
\Leftrightarrow \quad & \mathtt{p}(t, \mathtt{ncp}) = u(1 - \mathtt{p}(c, \mathtt{ncp})) + \mathtt{p}(c, \mathtt{ncp}) \\
\Leftrightarrow \quad & \mathtt{p}(t, \mathtt{ncp}) = \gamma u + 1 - \gamma \\
\Leftrightarrow \quad & t = \mathtt{q}(\gamma u + 1 - \gamma, \mathtt{ncp}) .
\end{aligned}
$$

Accordingly, let U be a Uniform (0,1) random variable. The significant test statistic is

$$
\begin{aligned}
T &= \mathtt{q}(\gamma U + 1 - \gamma, \mathtt{ncp}) \\
&= \mathtt{q}(1 + \gamma(U - 1), \mathtt{ncp}) \\
&= \mathtt{q}(1 - \gamma(1 - U), \mathtt{ncp}) .
\end{aligned}
$$

Since $1 - U$ also has a Uniform (0,1) distribution, one may proceed as follows. For a given sample size and effect size, first calculate the non-centrality parameter $\mathtt{ncp} = f_1(n) f_2(\mathtt{es})$, and use that to compute the power value $\gamma = 1 - \mathtt{p}(c, \mathtt{ncp})$. Then calculate the significant test statistic

$$
T = \mathtt{q}(1 - \gamma U, \mathtt{ncp}) , \qquad (19)
$$

where U is a pseudo-random variate from a Uniform (0,1) distribution. In R, the process can be applied to a vector of ncp values and a vector of independent U values of the same length.

Again, this is the second step. The first step is to simulate a collection of ncp values using the desired joint distribution of sample size and effect size after selection for significance. Naturally, simulation is is easiest if sample size and effect size come from well-known distributions with built-in random number generation, and if sample size and effect size are specified to be independent after selection. In one of our simulations, sample size and effect size after selection were correlated. The next section describes how this was done.

Correlated sample size and effect size. The following is a description of how correlated sample sizes and effect sizes were produced in the simulations described in "Violating the assumptions of the ML Model." Let effect size X have density $g_\theta(\mathtt{es})$, where θ represents a vector of parameters for the distribution of effect size. Conditionally on $X = \mathtt{es}$, let

the distribution of sample size be Poisson distributed with expected value $\exp(\beta_0 + \beta_1 \texttt{es})$. This is standard Poisson regression. Simulation from the joint distribution is easy. One simply simulates an effect size es according to the density g, computes the Poisson parameter $\lambda = \exp(\beta_0 + \beta_1 \texttt{es})$, and then samples a value n from a Poisson distribution with parameter λ. The challenge is to choose the parameters θ, β_0 and β_1 so that after selection, (a) the population mean power has a desired value, and at the same time (b) the population correlation between sample size and effect size has a desired value. Population mean power is $\gamma =$

$$\int_0^\infty \sum_n [1 - \texttt{p}(c, f_1(n) f_2(\texttt{es}))] \, P\{N = n | X = \texttt{es}\} g_\theta(\texttt{es}) d\texttt{es} .$$

Given values of θ, β_0 and β_1, this expression can be calculated by numerical integration; recall that $P\{N = n|X = \texttt{es}\}$ is a Poisson probability.

The population correlation between sample size and effect size is

$$\rho = \frac{E(XN) - E(X)E(N)}{SD(X)\, SD(N)} ,$$

where $SD(\cdot)$ refers to the population standard deviation of something. The quantities $E(X)$ and $SD(X)$ are direct functions of θ. The standard deviation of sample size $SD(N) = \sqrt{E(N^2) - [E(N)]^2}$, where

$$\begin{aligned} E(N) &= E(E[N|X]) \\ &= \int_0^\infty E[N|X = \texttt{es}] \, g_\theta(\texttt{es}) d\texttt{es} \\ &= \int_0^\infty e^{\beta_0 + \beta_1 \texttt{es}} g_\theta(\texttt{es}) d\texttt{es} \end{aligned}$$

and

$$\begin{aligned} E(N^2) &= E(E[N^2|X]) \\ &= E(Var(N) + E(N)^2 | X) \\ &= \int_0^\infty \left(e^{\beta_0 + \beta_1 \texttt{es}} + e^{2\beta_0 + 2\beta_1 \texttt{es}} \right) g_\theta(\texttt{es}) d\texttt{es} . \end{aligned}$$

Finally,

$$\begin{aligned} E(XN) &= \int_0^\infty \sum_n \texttt{es}\, n \, P\{N = n|X = \texttt{es}\} g_\theta(\texttt{es}) d\texttt{es} \\ &= \int_0^\infty \texttt{es}\, E(N|X = \texttt{es}) g_\theta(\texttt{es}) d\texttt{es} \\ &= \int_0^\infty \texttt{es}\, e^{\beta_0 + \beta_1 \texttt{es}} g_\theta(\texttt{es}) d\texttt{es} . \end{aligned}$$

All these expected values can be calculated by numerical integration using R's `integrate` function, so that the correlation ρ can be evaluated for any set of θ, β_0 and β_1 values.

In our simulation of correlated sample size and effect size, $g_\theta(\texttt{es})$ was a beta density, re-parameterized so that $\theta = (\mu, \sigma^2)$ consisted of the mean μ and variance σ^2. Conditionally on effect size, sample size was Poisson distributed with expected value $\exp(\beta_0 + \beta_1 \texttt{es})$. We set the variance of effect size σ^2 to a fixed value of 0.09, so that the standard deviation of effect size after selection was 0.30, a high value. Given any mean effect size μ and slope β_1, the parameter β_0 (the intercept of the Poisson regression) was adjusted so that expected sample size at the mean value was equal to 86: $\beta_0 = \ln(86) - \beta_1 \mu$.

With these constraints, the population mean power γ and correlation ρ were a function of the two free parameters μ and β_1. Let γ_0 be a desired value of mean power; for example, $\gamma_0 = 0.5$. Let ρ_0 be a desired value of the correlation between sample size and effect size; for example, $\rho_0 = -0.8$. Values of μ and β_1 were locating by numerically minimizing the function $f(\mu, \beta_1) = |\gamma - \gamma_0| + |\rho - \rho_0|$. We used R's optim function.