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

Jerry Brunner and Ulrich Schimmack

September 1, 2016

Abstract

Statistical power is a necessary condition for replicability. In any population of published results, there is a population of power values for the statistical tests on which conclusions are based. We show how this distribution is affected by publication bias, the tendency to suppress non-significant findings. In a set of large-scale simulation studies, we compare methods for estimating the mean of a heterogeneous population of power values, based only on significant results. Maximum likelihood is most accurate when assumptions about the distribution of effect size are correct. Without such assumptions, the most successful method is a new one: z-curve. We describe and validate a conservative bootstrap confidence interval for z-curve that allows the method to be applied to small samples as well as large samples.

1 Introduction

Science is built on a mixture of trust and healthy skepticism. On one hand, scientists who read and cite published work trust the authors, reviewers, and editors to ensure that most reported results provide credible evidence based on objective empirical studies. On the other hand, scientists also insist that results be reproducible; they or other researchers should be able to repeat an empirical study and obtain the same results. This way, the facts on which scientific knowledge is based cannot be in doubt for long. In particular, false positives will be promptly discovered when replication studies fail to reproduce the original results. Reproducibility is acknowledged to be a requirement of good science (Popper 1934, Bunge 1998).

In recent years, numerous replication failures have shown that published results are much less reproducible than psychologists would like to believe. This issue has been called the "replicability crisis," and extends across a variety of fields in the social and biomedical sciences (Hirschhorn, Lohmueller, Byrne and Hirschhorn 2002, Ioannidis 2008, Simmons, Nelson and Simonsohn 2011, Begley and Ellis 2012, John, Lowenstein and Prelec 2012, Begley 2013, Chang and Li 2015, Baker 2016). In Psychology, the Open Science Collaboration (OSR) project attempted to replicate the primary findings of 100 studies, mostly in the areas of cognitive and social psychology (OSR, 2016). For three studies, the “effect” was a failure to find significant results. The remaining 97 studies reported support for a hypothesis with a significant or at least marginally significant result. However, the corresponding replication studies produced only 37% significant results. This low success rate has created heated debates, especially in social psychology where the success rate was only 25%.

One possible source of resistance to the OSC results is that in spite of indications of persistently low statistical power in psychological research (Cohen 1962, Sedlmeier and Gigerenzer 1989), there may be lack of awareness about the true probability of obtaining a statistically significant result in a psychological study. The reason may be that journals have a tendency publish mostly significant results, a condition that has been called “publication bias” (Sterling, 1959; Sterling, Rosenbaum and Weinkam, 1995). As a result, psychological journals create a false expectation that significant results are to be expected. However, the success rate in journals is inflated because non-significant results often remain unpublished and end up in Rosenthal’s (1969) proverbial file drawer. As the size of researchers’ file drawers is unknown, it is unclear how likely it is that an empirical study produces a significant result and how likely it is that a published significant result will reproduce a significant result in a replication study.

Studies like the OSC replication project produce valuable information, but they are very costly in terms of time and resources. The objective of this paper is to introduce methods for assessing the typical replicability of a body published findings without literally repeating any of the studies. Instead, estimates are based on statistical results reported in the articles. Our approach applies to scientific disciplines like Psychology, where most findings are established by rejecting null hypotheses based on tests of statistical significance. This convention applies to 97 out of 100 studies in the OSC project. Replicating a finding means collecting another set of data from the same population with exactly the same procedure and sample size, and obtaining significant results in the same direction again. We define the replicability of a finding as the probability of replication. Because of the caveat “in the same direction,” replicability is technically less than or equal to the probability of significance. Except when the null hypothesis is true or almost true, the difference is small.

Consider a population of empirical findings. Each finding in the population has been validated by a test of significance, and every test has its own probability of being significant; that is, there is a population of power values. Now suppose that one finding is randomly selected from the population. The study is repeated exactly, and the same statistical test that originally validated the finding is carried out on the new data. We show (Principle 1 of Section 2) that the probability of obtaining significant results from this two-stage procedure is exactly equal to the population mean power value. This means that population mean power sets an upper bound on replicability, and an estimate of mean power is a generous estimate of average replicability. It is important to note that this assumes *exact* replication, including the subject population and possible unreported failures of experimental control. Thus, while mathematically the probability of replication is often barely less than the power, in practice we expect good estimates of mean power to be noticeably higher than actual replicability. When estimated power is low, actual replicability is likely to be even worse.

In this article, we introduce four methods for estimating population mean power based on published results. It is important to distinguish this undertaking from that of Cohen (1962) and Sedlmeier and Gigerenzer's (1989) follow-up. In Cohen's classic survey of power in the Journal of Abnormal and Social Psychology, power was calculated exactly for effect sizes deemed "small," "medium" and "large," using the observed sample sizes and designs of the studies. 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. In fact, by Jensen's inequality (Billingsley 1986, p. 283) power at the mean effect size is greater than mean power.

Two of the methods we consider – Simonsohn, Nelson and Simmons' (2014b) p-curve and van Assen, van Aert, and Wicherts' (2014) p-uniform – were developed to correct for publication bias in meta-analyses of effect sizes. Both methods assume a fixed population effect size. Simmonsohn et al. have extended their method to estimate power in the restricted setting of a single fixed power value for the entire population, implying homogeneity in sample size as well as effect size (`www.p-curve.com`). We introduce two additional methods that are explicitly designed to estimate power for heterogeneous data under publication bias. We use extensive simulation studies to compare all four methods for a wide variety of scenarios. Finally, we apply all four methods to the studies from the OSC reproducibility project.

1.1 Statistical Power

The power of a statistical test (Neyman and Pearson, 1933; Lehman, 1959; Cohen, 1988) is the probability of correctly rejecting the null hypothesis. Power can be calculated exactly for any chosen set of parameter values, without using sample data in any way. This may be done before data are collected in order to choose sample size (Cohen 1988, Desu and Raghavarao 1990), or with published studies to assess average power assuming effects of a designated magnitude (Cohen 1962, Sedlmeier and Gigerenzer 1989). But since true effect size is never known exactly, the power of a reported statistical test is an unknown quantity. Our objective is to estimate the population mean of such quantities.

The formal definition of power implies that power is not defined for a population effect size of zero. In this case the null hypothesis is false, and will be rejected incorrectly with probability equal to the significance criterion, usually 0.05. As the probability of the null hypothesis being true is unknown, this definition of power would make power estimation impossible in principle. One way around the problem is to assume that the null hypothesis is never exactly true (Sterling et al., 1995). Another solution is to extend the definition of power so that it is defined even when the null hypothesis is true. In the end, there are no practical implications for power estimation because power approaches the significance criterion in the limit as effect sizes approach zero. In the limit, a highly underpowered study with a true effect is as unlikely to replicate a significant result as a Type I error. Assuming the usual 0.05 significance level, power equals 0.05 if the null hypothesis is true.

1.1.1 Selection for significance

There is a tendency for non-significant results not to appear in the published literature, a condition that has been called "publication bias" (Sterling, 1959; Sterling, Rosenbaum and Weinkam, 1995). By severing the relationship between average power and apparent success rate, publication bias can create a false expectation that significant results are to be expected, leading to inflated estimates of of power (Francis, 2012; Schimmack, 2012). For example in the OSC data, significant results were initially reported in 93 out of 100 studies; four were "marginally significant," and three were presented as null results. This would suggest an average power of 93%, very different from the 37% success rate on actual replication. This shows the importance of allowing for publication bias. Though some non-significant results are reported, we suspect that they may often be selected to make a point and so to be quite unrepresentative of the population from which they are taken. It is safest to discard them. Thus, estimates will be based upon a sub-population of tests that are statistically significant.

1.1.2 Observed power

The difficulty of estimating true power based on a single study is well documented (Boos and Stefanski 2012, Gerard, Smith and Weerakkody 1998, Gillett 1994, Hoenig and Heisey 2001, Thomas 1997, Yuan and Maxwell 2005). One problem is that the observed power method relies on the observed effect size as an estimate of the population effect size to compute power, and observed effect size is severely inflated by publication bias. Even if the bias in effect size could be corrected on average, the resulting estimates of power are too variable to be practically useful for a single study. However, low precision does not mean that estimates from individual studies are useless. In fact, psychologists routinely report standardized effect sizes in small samples, often with very large confidence intervals. The point estimate is not very informative, but it can be used for meta-analyses, and meta-analyses of effect sizes can produce precise estimates of population parameters because sampling error decreases as the number of studies in a meta-analysis increase. The same principle applies to statistical power. Although sampling error creates too much noise for the interpretation of observed power in a single study, meta-analyses can provide valuable information about power in the presence of publication bias (Francis, 2012; Schimmack, 2012).

1.1.3 Heterogeneity

Since power is a function of effect size and sample size, estimates of effect size lead immediately to estimates of power. Furthermore, some methods of estimating effect size explicitly take publication bias into account. In our view, the most promising of these are the p-curve method of Simonsohn, Nelson and Simmons (2014b) and the p-uniform method of van Assen, van Aert, and Wicherts (2014). Once an estimate of the population effect size has been found, it is straightforward to use this parameter to compute an estimated power for each study. Averaging these quantities produces an estimate of population mean power.

Estimates of effect size generally assume that a single quantity is being estimated. In contrast, our interest is in a setting where not only the sample sizes, but the effect sizes, the topics being investigated and the statistical tests employed are all subject to sampling variation. That is we wish to estimate population mean power not just assuming selection for significance, but also under *heterogeneity* — that is, assuming that each test in the population has its own true power (a fixed, unknown number), and they all might be different.

It has been suggested (Ioannidis and Trikalinos 2007, Kepes, Banks, McDaniel and Whetzel 2012, Simonsohn et al. 2014b, van Assen et al. 2014) that methods developed for homogeneity may be applied to the heterogeneous situation by subsetting the data into tests with approximately the same effect size, or even the same true power. Our objection to this idea is that it involves too much unverified guesswork. It is impossible to know a priori which tests belong in the same cluster. This would be especially true if true effect size varies continuously, a very plausible state of affairs. Moreover, this approach is not feasible when the set of studies is large as, for example, in an analysis of an entire journal over a multi-year period.

We propose two novel methods for estimating mean power in the challenging scenario where effect sizes and sample sizes are heterogeneous and publication bias is present. The methods are maximum likelihood, (which we view as the default method of estimation in Statistics) and a new method we call z-curve. We test these in a wide range of simulation studies. Furthermore, we compare our methods to p-curve and p-uniform to examine the robustness of these methods for homogeneous data when heterogeneity is present. As previous simulations have focused on effect size estimation, our simulations provide the first test of these methods for the estimation of power and replicability.

1.2 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 use and extend an innovative notation of Simonsohn, Nelson and Simmons (2014a), who employ a modified version of the S syntax (Becker, Chambers and Wilks, 1988) to represent probability distributions. The S language is familiar to psychologists who conduct data analysis using the R statistical software (R core team, 2012). It 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 significance level α (usually $\alpha = 0.05$), when the continuous test statistic T exceeds a critical value c. Typically there is a sample of test statistic values $T_1, \ldots, T_k$, but when only one is being considered the subscript will be omitted. The notation $\texttt{p}(t)$ 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 $\texttt{p}(t)$ is the area under the curve, $\texttt{d}(t)$ is height of the curve above the x-axis, as in `dnorm`. Following the conventions of the S language, the inverse of `p` is `q`, so that $\texttt{p}(\texttt{q}(t)) = \texttt{q}(\texttt{p}(t)) = t$.

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, Kotz and Balakrishnan, 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 $\texttt{p}(t\texttt{,ncp})$, representing particular cases like $\texttt{pf}(t\texttt{,df1,df2,ncp})$. The initials `ncp` stand for "non-centrality parameter." This notation applies directly when T has one of the common non-central distributions like the non-central t, F or chi-squared under the alternative hypothesis, but it extends 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 $= 1 - \texttt{p}(c\texttt{,ncp})$.

For the most important non-central distributions (Z, t, chi-squared 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, and the second term is a function of the unknown parameters that reflects how wrong the null hypothesis is. In symbols,

$$\texttt{ncp} = f_1(n) \cdot f_2(\texttt{es}). \tag{1.1}$$

In this equation, n is the sample size and `es` is *effect size.* While sample size is observable, effect size is a function of unknown parameters and can never be known exactly. The quantities that are computed from sample data and commonly called "effect size" are properly *estimates* of `es`.

As we use the term, effect size refers to any function of the model parameters that equals zero when the null hypothesis is true, and assumes larger and larger positive values as the null hypothesis becomes more false. From this perspective, all reasonable definitions of effect size for a particular statistical model are deterministic monotone functions of one another and so the choice of which one to use is determined by convenience and interpretability. This usage is consistent in spirit with that of Cohen (1988), who freely uses "effect size" to describe various functions of the model parameters, even for the same statistical test. Also see Grissom and Kim (2012).

As an example of Equation (1.1), consider for example a standard F-test for difference between the means of two normal populations with a common variance. After some simplification, the non-centrality parameter of the non-central F may be written

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

where $n = n_1 + n_2$ is the total sample size, $\rho = \frac{n_1}{n}$ is 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 1.1; 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 1.1 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

$$\text{Power} = 1 - \texttt{p(}c, f_1(n) \cdot f_2(\texttt{es})\texttt{)}. \tag{1.2}$$

The function $f_2(\texttt{es})$ is particularly convenient because it will accommodate any reasonable definition of effect size. Details are given in the technical supplement.

2 Two Populations of Power

Consider a population of independent statistical tests. Each test has its own power value, a true probability of rejecting the null hypothesis determined by the sample size, procedure and true parameter values. The tests are conducted. Significant results are published and become available as data. Non-significant results go into the mythical "file drawer" of Rosenthal (1979). This means that there are are two populations of power values: the original population, and the sub-population corresponding to the tests that happened to be statistically significant.

Selection for significance (publication bias) does not change the power values of individual studies. However, the population of studies in the set of studies selected for significance differs from the original population of studies without selection for significance. The reason is that selection for significance tends to select studies with higher power. For example, a study with 80% power is more likely to end up in the sample of studies selected for significance than a study with 20% power.

Probability models may often be clarified by thinking of them as games of chance. Designing a study and selecting a hypothesis to test corresponds to manufacturing a roulette wheel that may not be perfectly balanced. The numbers on the wheel are p-values, and $p < 0.05$ is a win. Running the study and collecting data corresponds to spinning the wheel. The unique balance and other physical properties of the wheel determine the probability of a win; this corresponds to the power of the test. Performing the statistical analysis corresponds to examining the number that comes up on the wheel and noting whether $p < 0.05$. A large number of wheels are manufactured and spun once. This is the population before selection. The wheels that yield wins are put on display; this is the population after selection. Naturally, there is a tendency for wheels with a higher chance of winning to be put on display. The wheels that yield losing numbers are sent to warehouses (the file drawer), or more likely to landfill. All records are destroyed. Even the number of losing wheels is suppressed.

Spinning all the wheels on display a second time would take a great deal of effort, but if we did so we could record the proportion of wins. This would not be the true probability of significance, but if the number of wheels on display is large it would be close. Spinning all the wheels a third time would yield another proportion of wins, presumably close to the first. Repeating this impossibly tedious exercise a large number of times and averaging the proportions would give the true probability of a win for the wheels on display. The objective of this paper is to estimate this important unknown quantity using only the numbers that appeared on first spin.

We now give a set of fundamental principles connecting the probability distribution power before selection to its distribution after selection. These principles 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 each power value in the population is the probability that the corresponding test will be significant. The supplementary materials contain proofs and a numerical example.

Principle 1 *Population mean power equals the overall probability of a significant result.*

Principle 1 applies equally to the population of studies before and after selection. Because it applies after selection, this principle establishes the link between replicability and population mean power. If a single published result is randomly selected and the study is repeated exactly, the probability of obtaining another significant result equals population mean power after selection. In terms of the roulette wheel analogy, this is a two-stage game. The first stage is to select a wheel at random from those on display, and the second stage is to spin the wheel. Principle 1 says that the probability of winning the game is exactly the mean probability of a win for the wheels on display.

Principle 2 *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.*

For example, suppose that before selection 80% of studies have power equal to 0.10 and 20% have power equal to 0.60. Table 1 shows the distribution of power before and after selection. Expected (population mean) power before selection is $0.10 * 0.8 + 0.60 * 0.2 = 0.20$. After selection there are still the same two power values, but their probabilities change. To obtain the probability that power equals 0.10 after selection, multiply 0.8 by the power value 0.1, and divide by the expected power before selection of 0.20. The resulting probability after selection is $0.8 * 0.1/0.2 = 0.40$. In the technical supplement, Principle 2 is used to derive the remaining principles.

Table 1: Illustration of Principle 2

	Probability	
Power value	Before selection	After selection
0.10	0.80	0.40
0.60	0.20	0.60

Principle 3 shows how population mean power after selection is related to population mean power before selection. In simulation studies, Principle 3 allows the distribution of power before selection to be chosen so that expected power after selection is exactly equal to some desired value.

Principle 3 *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.*

It is also possible to go backwards from power after selection to mean power before selection, again without knowing the full distributions. In Principle 4, the reciprocal of power refers to one divided by the power value. Naturally this quantity has a population mean.

Principle 4 *Population mean power before selection equals one divided by the population mean of the reciprocal of power after selection.*

Although we do not pursue the topic in this paper, Principle 4 opens the door to estimating mean power before selection using only significant results.

Selection for significance is often called "publication bias" (Sterling 1959, Sterling et al. 1995), and it has indisputable drawbacks. However, it does increase average power because tests with higher power are more likely to be selected. Principle 5 quantifies the increase.

Principle 5 *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.*

Because variances cannot be negative, population mean power after selection for significance is always greater than or equal to population mean power before selection, with equality occurring only in the homogeneous case where the population variance of power before selection is equal to zero. The greatest increases in mean power will occur when the distribution of power before selection is heterogeneous, and average power is low.

3 Estimation Methods

In this section, we describe four methods for estimating population mean power under conditions of heterogeneity, after selection for statistical significance.

3.1 P-curve and p-uniform estimation of mean power

The p-curve (Simonsohn et al. 2014b) and p-uniform (van Assen et al. 2014) methods are designed for estimating effect sizes in meta-analyses where there is a single fixed effect size, but possibly varying sample sizes. We adapted them slightly to produce estimates of mean power, again for the setting of heterogeneity in sample size but not effect size.

Both p-uniform and p-curve are based on the idea that p-values are uniformly distributed when the null hypothesis is true. Originally, the test statistics were used to test the null hypothesis that the effect size is zero, and they all rejected that null hypothesis. Now the set of significant test statistics is used to test a *modified* null hypothesis that the effect size equals some specified non-zero value. If the modified null hypotheses were true, the resulting p-values would again have a uniform distribution. To find the best fitting effect size for a set of observed test statistics, p-curve and p-uniform compute p-values for various effect sizes and chose the effect size that yields the best approximation of a uniform distribution. The main difference between p-curve and p-uniform is the criterion used to pick the best fitting effect size.

If the modified null hypothesis that effect size = $\texttt{es}$ is true, the cumulative distribution function of the test statistic is the conditional probability

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

using $\texttt{ncp} = f_1(n) \cdot f_2(\texttt{es})$ as given in Equation 1.1. The corresponding modified p-value (which Simonsohn et al. would call the pp-value) is

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

Note that since the sample sizes of the tests may differ, the symbols $\texttt{p}$, n and c as well as T may have different referents 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 and p-uniform choose as estimated effect size the value of $\texttt{es}$ that makes the modified p-values most nearly uniform. They differ only in the criterion for deciding when uniformity has been reached.

P-curve is based on a Kolmogorov-Smirnov test for departure from a uniform distribution, choosing the $\texttt{es}$ value yielding the smallest value of the test statistic. P-uniform is based on a different criterion. Denoting by P_j the modified p-value associated with test j, calculate $Y = -\sum_{j=1}^{k} \ln(P_j)$, where ln is the natural logarithm. If the P_j values were uniformly distributed, Y would have a Gamma distribution with expected value k, the number of tests. The P-uniform estimate is the modified null hypothesis effect size $\texttt{es}$ 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 $\widehat{\texttt{es}}$ of the common effect size, estimated power for each test is solely determined by sample size. Using Expression 1.2, the estimated power of test j is $1 - \texttt{p(}c_j\texttt{,}f_1(n_j) \cdot f_2(\widehat{\texttt{es}})\texttt{)}$. Population mean power can then be estimated by averaging the k power estimates. This natural way of estimating mean power is merely implicit in the papers by van Assen et al. (2014) and Simonsohn et al. (2014b).

As previous simulation studies have shown that these methods work well when their assumptions are met, we expect both p-uniform and p-curve to produce accurate estimates of mean power in simulations with fixed effect sizes. However, simulation studies with heterogeneity in effect sizes show that both methods are likely to overestimate the average population effect size when effect sizes are heterogeneous (van Aert et al., in press). Thus, we expect that these methods will provide inflated estimates of replicability in simulations with heterogeneous effect sizes.

3.2 Maximum likelihood estimation of mean power

The method of Maximum Likelihood (Fisher, 1922; also see the historical account by Aldrich, 1997) is a general method for the estimation of an unknown parameter by finding the parameters value that makes the observed data most probable. For any set of observed data, the statistical assumptions allow calculation of the probability of obtaining the observed the data (or for continuous distributions, the probability of obtaining data in a tiny region surrounding the observed data). The *likelihood function* expresses this probability as a function of the unknown parameter. Geometrically, the likelihood function is a curve, and estimation proceeds by finding the highest point on the curve. The maximum likelihood estimate is the parameter value yielding that maximum. The case of multi-parameter estimation is analogous, with the curve being replaced by a convoluted surface in higher dimension. When data are consistent with the model assumptions, maximum likelihood generally yields more precise parameter estimates than other methods, especially for large samples (Lehmann and Casella, 1998). It is fair to say that maximum likelihood is the default method of estimation in Statistics for parametric models.

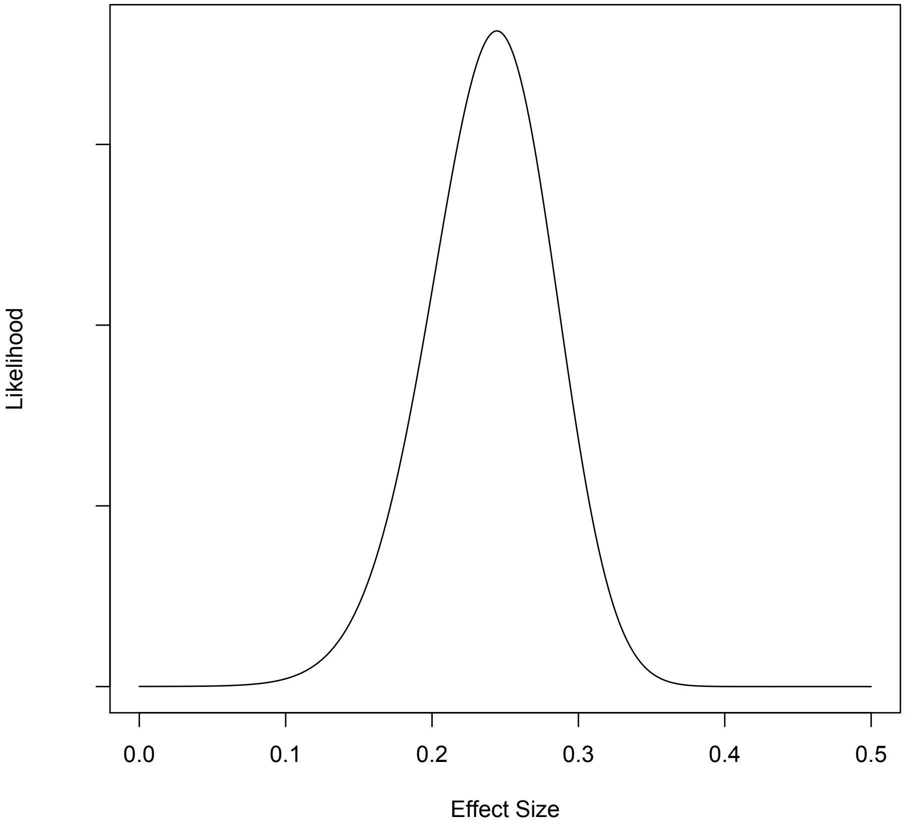
For simplicity, first consider the case of heterogeneity in sample size but not effect size. In this case the single unknown parameter is effect size (`es`), and the likelihood function is based on the conditional probability of observing the data given selection for significance. Denoting the observed test statistic values by $t_1, \ldots, t_k$, the likelihood function is a product of k terms of the form

$$\frac{\mathtt{d}(t_j\,\mathtt{,}\,f_1(n_j)\cdot f_2(\mathtt{es})\mathtt{)}}{1-\mathtt{p}(c_j\,\mathtt{,}\,f_1(n_j)\cdot f_2(\mathtt{es})\mathtt{)}}, \tag{3.1}$$

where because of selection for significance, all the t_j values are greater than their respective critical values c_j. Expression 3.1 becomes the likelihood of Hedges (1984) for the case of a two-sample t-test.

As an example, consider a one-way ANOVA with four treatment groups, equal sample sizes, and a “medium” value of 0.25 for Cohen’s (1988, p. 275) effect size `f`. As shown in the technical supplement, $\mathtt{ncp} = f_1(n)\cdot f_2(\mathtt{es}) = n\cdot \mathtt{es}^2$ for this problem, where n is the total sample size. Figure 1 shows the likelihood function for a simulated set of $k = 25$ F statistics. In this example, the sample sizes before selection varied about a mean of twenty per treatment. The likelihood function reaches its maximum when effect size equals 0.244; this is the maximum likelihood estimate. It is quite close to the true value of 0.25.

Figure 1: Likelihood Function for 25 F-tests With True Effect Size = 0.25



In general, the maximum likelihood estimate of $\texttt{es}$ is the effect size value that makes the likelihood function greatest. Denote it by $\widehat{\texttt{es}}$. The estimated probability of significance for each study is obtained by

$$\text{Estimated Power} = 1 - \texttt{p(}c_j\texttt{,}f_1(n_j) \cdot f_2(\widehat{\texttt{es}})\texttt{)},$$

and then as for p-curve and p-uniform, the estimated power values are averaged to produce a single estimate of mean power.

Now include heterogeneity in effect size as well as sample size. If sample size and effect size before selection are independent, selection for significance induces a mild relationship between sample size and effect size, since tests that are low in both sample size and effect size are under-selected, while tests high in both are over-selected. Suppose that the distribution of effect size before selection is continuous with probability density $g_\theta(\texttt{es})$. This notation indicates that the distribution of effect size depends on an unknown parameter or parameter vector θ. In the technical supplement, it is shown that the likelihood function (a function of θ) is a product of k terms of the form

$$\frac{\int_0^\infty \texttt{d(}t_j\texttt{,}f_1(n_j) \cdot f_2(\texttt{es})\texttt{)}\, g_\theta(\texttt{es})\, d\texttt{es}}{\int_0^\infty \left[1 - \texttt{p(}c_j\texttt{,}f_1(n_j) \cdot f_2(\texttt{es})\texttt{)}\right] g_\theta(\texttt{es})\, d\texttt{es}}, \tag{3.2}$$

where the integrals denote areas under curves that can be computed with R's `integrate` function. Again, the maximum likelihood estimate is the value of θ for which the value of the product is highest. Denote the maximum likelihood estimate by $\widehat{\theta}$. Typically $\widehat{\theta}$ is a single number or a pair of numbers.

As before, an estimate of population mean power is produced by averaging estimated power for the k significance tests. It is shown in the technical supplement that the terms to be averaged are

$$\frac{\int_0^\infty \left[1 - \texttt{p(}c_j\texttt{,}f_1(n_j) \cdot f_2(\texttt{es})\texttt{)}\right]^2 g_{\widehat{\theta}}(\texttt{es})\, d\texttt{es}}{\int_0^\infty \left[1 - \texttt{p(}c_j\texttt{,}f_1(n_j) \cdot f_2(\texttt{es})\texttt{)}\right] g_{\widehat{\theta}}(\texttt{es})\, d\texttt{es}}, \tag{3.3}$$

an expression that also follows from an informed application of Principle 3.

3.3 Z-curve

In this section we describe a new estimation method called Z-curve. It follows a traditional meta-analyses that converts p-values into Z-scores as a common metric to integrate results from different original studies (Stouffer, Suchman, DeVinney, Star and Williams, 1949; Rosenthal, 1979). The use of Z-scores as a common metric makes it possible to fit a single function to p-values arising from widely 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 we will denote by m m (Heisey and Hoenig, 2001). As described the technical supplement, m is the non-centrality parameter for the one-tailed Z-test. Input to the Z-curve is a sample of p-values from two-sided or other non-directional tests, 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 $\alpha/2 = 0.025$. 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

$$Z = \texttt{qnorm(}1 - p/2\texttt{)}. \tag{3.4}$$

2. *Set aside* $Z > 6$. We assume that p-values in this range come from tests with power essentially equal to one. To avoid numerical problems arising from p-values that are approximately zero, we set them aside for now and bring them back in the final step.

3. *Fit a finite mixture model.* Before selecting for significance and setting aside values above six, the distribution of the test statistic Z given a particular non-centrality parameter value m is normal[1] with mean m. Afterwards, it is a normal distribution truncated on the left at the critical value c (usually 1.96) 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. First, 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. Finally, re-scale so that the area under the curve equals one.

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

$$f(z) = \sum_{j=1}^{r} w_j \, \frac{\texttt{dnorm(}z - m_j\texttt{)}}{\texttt{pnorm(}6 - m_j\texttt{)} - \texttt{pnorm(}c - m_j\texttt{)}}, \tag{3.5}$$

   for $c < z < 6$.

   The finite mixture model is only an approximation. If the true probability density function of Z given significance were known, the approximation could be optimized by choosing $w_1, \ldots, w_r$ and $m_1, \ldots, m_r$ to bring (3.5) as close to the true density as possible. Since the true density is unknown, we use a kernel density estimate (Silverman, 1986) as implemented in R's `density` function, with the default settings.

[1] This statement would be exactly true if the p-values really came from one-sided Z-tests as suggested in Step 1. In practice it is an approximation.

Specifically, 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 $c = 1.96$ 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 (3.5).

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

$$\ell = \sum_{j=1}^{r} \widehat{w}_j (1 - \texttt{pnorm(}c - \widehat{m}_j\texttt{)}), \tag{3.6}$$

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 3.5, there is no truncation in Formula 3.6 because it represets 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

$$\begin{aligned} Z_{est} &= (1-q)\,\ell + q \cdot 1 \qquad (3.7) \\ &= q + (1-q) \sum_{j=1}^{r} \widehat{w}_j (1 - \texttt{pnorm(}c - \widehat{m}_j\texttt{)}) \end{aligned}$$

By Principle 1, this is the estimate of population mean power after selection for significance.

4 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[2]. The R code is available in the supplementary materials. In the simulations, the four estimation methods (p-curve, p-uniform, maximum likelihood and 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.

[2] We would like to thank Dr. Jeffrey Graham for providing remote access to the machines in the Psychology Laboratory at the University of Toronto Mississauga. Thanks to Josef Duchesne for technical advice.

4.1 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 (Johnson, Kemp and Kotz, 2005). 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 simulating directly from the distribution of the test statistic after selection; details are given in the technical supplement.

Effect sizes were selected to yield population mean power values after selection of 0.05, 0.25, 0.50 or 0.75. For F-tests, we used Cohen's (1988, p. 275) effect size metric $\mathbf{f}$. For chi-squared tests, we used $\mathbf{w}$ (Cohen, 1988, p. 216). The number of test statistics k on which estimates were based was 15, 25, 50, 100 or 250. Numerator degrees of freedom (just degrees of freedom for the chi-squared tests) were one, three or five. Because the pattern of results was similar for F and chi-squared tests and for different degrees of freedom, we give 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 2 shows means and standard deviations of estimated population mean power after selection. Differences between the mean estimates and the true values represent bias in estimation. We conclude that all methods performed fairly well, with z-curve showing a bit more bias than the other methods.

Absolute error of estimation It is desirable for average estimates to be close to the true values, but still positive and negative errors may cancel. More interesting is how close the estimate is on average to the true value being estimated. Table 3 shows mean absolute error of estimation for F-tests with one numerator degree of freedom; full results are givn in the supplementary materials. As expected, all the methods become more accurate with larger numbers of tests. Though the differences are fairly small, Z-curve is least accurate when mean power is low, and most accurate when mean power is high. Maximum likelihood has a slight edge over the other methods under most circumstances, except that z-curve sometimes does better when population mean power is moderate to high and the estimates are based on a small number of tests.

Testing differences in accuracy Because results like the ones in in Table 2 are based on random number generation, some of the apparent differences could be due to chance. Thus we find ourselves applying statistical tests to an investigation of statistical tests. Within each of the 20 combinations of power and number of tests, there are six potential pairwise comparisons of mean absolute error. These comparisons were carried out using large-sample

Table 2: Means and standard deviation of estimated population mean power for heterogeneity in sample size only: F-tests with numerator $df = 1$

			Mean						*Standard Deviation*		
					Population Mean Power = 0.05						
		Number of Tests						Number of Tests			
	15	25	50	100	250		15	25	50	100	250
P-curve	0.083	0.073	0.064	0.059	0.055	P-curve	0.059	0.039	0.024	0.015	0.007
P-uniform	0.076	0.067	0.061	0.058	0.054	P-uniform	0.050	0.032	0.019	0.012	0.006
MaxLike	0.076	0.067	0.061	0.057	0.054	MaxLike	0.050	0.033	0.020	0.012	0.006
Z-curve	0.086	0.071	0.058	0.049	0.040	Z-curve	0.088	0.065	0.044	0.031	0.019
					Population Mean Power = 0.25						
		Number of Tests						Number of Tests			
	15	25	50	100	250		15	25	50	100	250
P-curve	0.269	0.261	0.256	0.253	0.251	P-curve	0.156	0.128	0.095	0.069	0.046
P-uniform	0.256	0.253	0.252	0.251	0.251	P-uniform	0.147	0.121	0.089	0.065	0.042
MaxLike	0.260	0.255	0.253	0.251	0.251	MaxLike	0.146	0.120	0.087	0.064	0.042
Z-curve	0.314	0.305	0.293	0.280	0.268	Z-curve	0.155	0.127	0.093	0.068	0.045
					Population Mean Power = 0.50						
		Number of Tests						Number of Tests			
	15	25	50	100	250		15	25	50	100	250
P-curve	0.484	0.491	0.496	0.497	0.499	P-curve	0.175	0.139	0.102	0.073	0.046
P-uniform	0.473	0.485	0.493	0.496	0.499	P-uniform	0.170	0.132	0.097	0.070	0.044
MaxLike	0.479	0.489	0.495	0.497	0.499	MaxLike	0.166	0.130	0.095	0.068	0.043
Z-curve	0.513	0.516	0.513	0.508	0.502	Z-curve	0.151	0.121	0.091	0.068	0.045
					Population Mean Power = 0.75						
		Number of Tests						Number of Tests			
	15	25	50	100	250		15	25	50	100	250
Pcurve	0.728	0.736	0.742	0.747	0.749	Pcurve	0.128	0.098	0.069	0.048	0.030
Puniform	0.721	0.732	0.740	0.746	0.748	Puniform	0.126	0.097	0.067	0.047	0.029
MaxLike	0.728	0.736	0.742	0.747	0.749	MaxLike	0.121	0.093	0.065	0.045	0.028
Zcurve	0.704	0.712	0.717	0.723	0.728	Zcurve	0.105	0.084	0.064	0.048	0.033

Table 3: Mean absolute error of estimation 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	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	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	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	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

Table 4: Number of times row method is significantly more accurate than column method: Heterogeneity in sample size only

		Chi-squared tests							***F*-tests**				
						$df = 1$							
		PC	PU	ML	ZC	Total		PC	PU	ML	ZC	Total	
P-curve	(PC)	0	0	0	14	14	P-curve	(PC)	0	0	0	13	13
P-uniform	(PU)	15	0	0	14	29	P-uniform	(PU)	15	0	0	13	28
MaxLike	(ML)	20	16	0	16	52	MaxLike	(ML)	20	17	0	14	51
Z-curve	(ZC)	5	4	3	0	12	Z-curve	(ZC)	7	5	4	0	16
						$df = 3$							
P-curve	(PC)	0	0	0	15	15	P-curve	(PC)	0	0	0	13	13
P-uniform	(PU)	16	0	0	15	31	P-uniform	(PU)	15	0	0	14	29
MaxLike	(ML)	20	15	0	16	51	MaxLike	(ML)	20	16	0	15	51
Z-curve	(ZC)	5	2	2	0	9	Z-curve	(ZC)	6	4	3	0	13
						$df = 5$							
P-curve	(PC)	0	0	0	15	15	P-curve	(PC)	0	0	0	13	13
P-uniform	(PU)	15	0	1	16	32	P-uniform	(PU)	14	0	0	14	28
MaxLike	(ML)	20	15	0	17	52	MaxLike	(ML)	20	16	0	15	51
Z-curve	(ZC)	3	2	2	0	7	Z-curve	(ZC)	6	4	3	0	13

two-sided matched Z-tests with a Bonferroni correction, yielding a joint 0.001 significance level for the 120 tests.

Table 4 shows the number of times that the row method was significantly more accurate than the column method by this stringent criterion. There are 6 sub-tables, one for each combination of type of test (chi-squared or F) and degrees of freedom. For F-tests, *df* refers to the numerator (experimental) degrees of freedom. Note that the Bonferroni correction was applied separately to each sub-table. In all, Table 4 summarizes the results of 720 tests. Full details are given in the supplementary materials.

In each sub-table of Table 4, the most accurate method overall is maximum likelihood, followed by p-uniform. When maximum likelihood lost a comparison it was usually to z-curve. As one would expect from the general theory od maximum likelihood estimation (Lehmann and Casella 1998, Ch. 6), maximum likelihood performed particularly well when estimates were based on a large number of tests. It is important to recognize, however, the the differences in average estimation error are fairly small. We conclude that although maximum likelihood performs best, all the methods yield reasonable estimates when effect sizes are homogeneous.

4.2 Heterogeneity in Both Sample Size and Effect Size

To model heterogeneity in effect size, we let effect size before selection vary according to a gamma distribution (Johnson, Kotz and 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. Maximum likelihood correctly assumed that effect size is gamma distributed, and the likelihood search was over the two parameters of the gamma distribution. The other 3 methods were not modified in any way. P-curve 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 carried out a simulation experiment like the one in Section 4.1, with one additional factor: amount of heterogeneity in effect size, as represented by the standard deviation of the effect size distribution. 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), type of test (F or chi-squared), and experimental degrees of freedom (1, 3 or 5). 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.

When there is heterogeneity in effect size, Maximum Likelihood is computationally demanding. The areas under many curves must be calculated numerically; see Expression 3.2. Using R's `integrate` function, the calculation involves fitting a histogram to each curve and then adding the areas of the bars. It is slow, and some of the curves are very skewed and razor thin. Numerical accuracy is an issue, especially for ratios of areas when the denominators are very small. In addition, the likelihood function has many local maxima, and it is necessary to try more than one starting value to have a hope of locating the global maximum. In our simulations, we used three random starting points. More would have been better, but the computational burden was too great for a simulation study. As a result, we consider the performance of maximum likelihood to be under-stated.

Average performance Table 5 shows means and standard deviations of estimated population mean power as a function of true population mean power and the standard deviation of effect size size. In this table the estimates were based on $k = 1,000$ test statistics, and good accuracy may be anticipated. P-uniform broke down completely for higher heterogeneity in effect size, with most estimates close to one regardless of the true value. Notice how a mean p-uniform estimate at the maximum value of one produces a standard deviation of zero. For moderate to high mean power, the p-curve also produces an over-estimate on average, with the problem becoming most severe when mean power and heterogeneity in effect size are both high. Maximum likelihood and z-curve perform much better.

Absolute error of estimation Table 6 shows mean absolute error of estimation. It confirms the inaccuracy of p-uniform, and suggests that p-curve may be competitive with

Table 5: Means and standard deviations of estimated power for heterogeneity in sample size and effect size based on 1,000 F-tests with numerator *df* = 1

	Mean				*Standard Deviation*		
Population Mean Power = 0.25							
	SD of Effect Size				SD of Effect Size		
	0.1	0.2	0.3		0.1	0.2	0.3
P-curve	0.225	0.272	0.320	P-curve	0.024	0.033	0.039
P-uniform	0.294	0.694	0.949	P-uniform	0.029	0.056	0.028
MaxLike	0.230	0.269	0.283	MaxLike	0.069	0.016	0.015
Z-curve	0.233	0.225	0.226	Z-curve	0.027	0.026	0.024
Population Mean Power = 0.50							
	SD of Effect Size				SD of Effect Size		
	0.1	0.2	0.3		0.1	0.2	0.3
P-curve	0.549	0.679	0.757	P-curve	0.024	0.027	0.026
P-uniform	0.602	0.913	0.995	P-uniform	0.024	0.019	0.003
MaxLike	0.501	0.502	0.506	MaxLike	0.025	0.019	0.019
Z-curve	0.504	0.492	0.487	Z-curve	0.026	0.026	0.025
Population Mean Power = 0.75							
	SD of Effect Size				SD of Effect Size		
	0.1	0.2	0.3		0.1	0.2	0.3
P-curve	0.824	0.928	0.962	P-curve	0.013	0.009	0.006
P-uniform	0.861	0.992	1.000	P-uniform	0.012	0.003	0.000
MaxLike	0.752	0.750	0.750	MaxLike	0.022	0.017	0.014
Z-curve	0.746	0.755	0.760	Z-curve	0.021	0.017	0.016

Maximum Likelihood and z-curve when heterogeneity and true mean power are both low, but not otherwise.

Testing differences in accuracy Table 6 is a sub-table, giving results when estimates are based on $k = 1,000$ tests. The full table for F-tests with numerator *df* = 1 has 3 levels of power, 3 levels of the standard deviation of effect size, and 5 levels of number of tests. Within each of these 45 combinations, there are 6 pairwise comparisons of the 4 estimation methods. The resulting 270 matched Z-tests were protected with a Bonferroni correction at the joint 0.001 significance level.

Table 7 counts the wins for all three *df* values, and for chi-squared tests as well as F-tests. The clear winner is Maximum Likelihood, followed by z-curve, p-curve and p-uniform in that order. When other methods beat maximum likelihood, it was almost always when heterogeneity in effect size and true population power were both low. This is consistent with Table 5, in which Maximum Likelihood performs better when mean power is moderate to high.

Table 6: Mean Absolute Error of estimation for heterogeneity in sample size and 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	2.87	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	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	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

4.3 Violating the Assumptions

In Section 4.2, heterogeneity in effect size before selection was modeled as a gamma distribution, with effect size independent of sample size before selection. Maximum likelihood had a substantial and arguably unfair advantage, since it assumed exactly the correct distribution for effect size. Also, sample size and effect size before selection were independent in both the simulations and in the assumptions of maximum likelihood. It is well known that when its assumptions are correct, maximum likelihood is very accurate compared to other methods (Lehmann and Casella 1998, Ch. 6). When assumptions are incorrect however, there are no genera theoretical results and the performance of maximum likelihood must be assessed on a case-by-case basis.

To test the robustness of maximum likelihood to assumptions, we conducted a smaller-scale simulation limited to F-tests with numerator degrees of freedom equal to one. Effect size after selection had a beta distribution rather than a gamma before selection. Though the beta distribution covers the interval zero to one and thus lacks the long right tail of the gamma, still the maximum value of one is more than more than twice Cohen's (1988, p. 287) large effect size of $\mathbf{f} = 0.4$. We made sample size and effect size non-independent, connecting them by a Poisson regression. This induced varying population correlations between sample size and effect size. Negative correlations would be expected, because of some researchers doing power analyses to select sample size, or otherwise having a sense of the sample sizes required for significance in their fields of study.

Table 7: Number of times row method is significantly more accurate than column method: Heterogeneity in sample size and effect size

		Chi-squared tests							***F*-tests**				
						df = 1							
		PC	PU	ML	ZC	Total			PC	PU	ML	ZC	Total
P-curve	(PC)	0	45	0	0	45	P-curve	(PC)	0	45	4	0	49
P-uniform	(PU)	0	0	0	0	0	P-uniform	(PU)	0	0	0	0	0
MaxLike	(ML)	41	45	0	33	119	MaxLike	(ML)	40	45	0	31	116
Z-curve	(ZC)	45	45	8	0	98	Z-curve	(ZC)	42	45	10	0	97
						df = 3							
P-curve	(PC)	0	45	4	1	50	P-curve	(PC)	0	45	5	4	54
P-uniform	(PU)	0	0	0	0	0	P-uniform	(PU)	0	0	5	0	5
MaxLike	(ML)	40	44	0	34	118	MaxLike	(ML)	40	40	0	34	114
Z-curve	(ZC)	40	45	7	0	92	Z-curve	(ZC)	39	45	7	0	91
						df = 5							
P-curve	(PC)	0	45	5	4	54	P-curve	(PC)	0	45	5	6	56
P-uniform	(PU)	0	0	0	0	0	P-uniform	(PU)	0	0	5	1	6
MaxLike	(ML)	40	45	0	36	121	MaxLike	(ML)	40	40	0	34	114
Z-curve	(ZC)	38	45	5	0	88	Z-curve	(ZC)	38	42	8	0	88

In our simulations, the variance of effect size after selection was fixed at 0.30, the high heterogeneity value in Section 4.2. Sample size after selection was Poisson distributed with expected value $\exp(\beta_0 + \beta_1 \texttt{es})$. Mean effect size after selection and the parameters β_0 and β_1 were selected to achieve (a) Desired population mean power after selection, (b) Desired population correlation between effect size and sample size after selection, and (c) Population mean sample size of 86 after selection at the mean effect size. Details are given in the Technical Supplement.

Three values of population mean power (0.25, 0.50 and 0.75), five values of the number of test statistics k (100, 250, 500, 1000 and 2000) and five values of the correlation between sample size and effect size (0.0, -0.2, -0.4, -0., -0.8) were varied in a factorial design, with ten thousand simulated data sets in each combination of values. All four estimation methods were applied to each simulated data set, with three random starting values for maximum likelihood.

Table 8 shows means and standard deviations of estimated population mean power as a function of true population mean power and the standard deviation of effect size. In this table, the estimates were based on $k = 1,000$ test statistics. Maximum likelihood tends to overestimate power when true power is high or low but not as much when true power equals 0.5. Correlation between sample size and effect size does not appear to matter much. P-curve and p-uniform produce estimates that are much too high on average.

Table 9 shows mean absolute error of estimation when estimates are based on $k = 1,000$

Table 8: Means and standard deviations of estimated power with beta effect size and correlated sample size and effect size: $k = 1,000$ F-tests with numerator $df = 1$

	Mean						*Standard Deviation*				
Population Mean Power = 0.25											
	Correlation						Correlation				
	-0.8	-0.6	-0.4	-0.2	0.0		-0.8	-0.6	-0.4	-0.2	0.0
P-curve	0.407	0.405	0.403	0.403	0.402	P-curve	0.043	0.044	0.043	0.044	0.044
P-uniform	0.853	0.852	0.852	0.852	0.852	P-uniform	0.003	0.004	0.003	0.004	0.004
MaxLike	0.302	0.301	0.300	0.300	0.300	MaxLike	0.015	0.015	0.015	0.015	0.015
Z-curve	0.232	0.231	0.230	0.231	0.230	Z-curve	0.015	0.015	0.015	0.015	0.015
Population Mean Power = 0.50											
	Correlation						Correlation				
	-0.8	-0.6	-0.4	-0.2	0.0		-0.8	-0.6	-0.4	-0.2	0.0
P-curve	0.839	0.840	0.841	0.841	0.841	P-curve	0.022	0.022	0.022	0.022	0.022
P-uniform	0.906	0.906	0.906	0.906	0.906	P-uniform	0.004	0.004	0.004	0.004	0.004
MaxLike	0.532	0.533	0.533	0.534	0.534	MaxLike	0.018	0.018	0.019	0.019	0.019
Z-curve	0.493	0.494	0.495	0.495	0.495	Z-curve	0.023	0.023	0.023	0.023	0.023
Population Mean Power = 0.75											
	Correlation						Correlation				
	-0.8	-0.6	-0.4	-0.2	0.0		-0.8	-0.6	-0.4	-0.2	0.0
Pcurve	0.990	0.991	0.992	0.992	0.992	Pcurve	0.002	0.002	0.002	0.002	0.002
Puniform	0.964	0.966	0.966	0.967	0.967	Puniform	0.003	0.003	0.003	0.003	0.003
MaxLike	0.826	0.832	0.836	0.838	0.840	MaxLike	0.016	0.016	0.015	0.015	0.015
Zcurve	0.785	0.790	0.793	0.794	0.796	Zcurve	0.013	0.013	0.013	0.012	0.012

test statisitics. It shows maximum likelihood to be consistently less accurate than z-curve, though not as bad as p-curve and p-uniform. Correlation between sample size and effect size appears to have little effect. Table 9 shows the results only for $k = 1,000$ test statistics, but results are very similar for the other values of k. Full details are given in the supplementary materials.

Within each of the $5 \times 3 \times 5 = 75$ combinations of correlation between sample size and effect size, population mean power and number of tests on which the estimates are based, there are six pairwise comparisons of mean absolute error for the four estimation methods. The resulting 450 matched Z-tests were protected with a Bonferroni correction at the joint 0.001 significance level. The full set of Z statistics may be found in the supplementary materiaals.

Table 10 counts the wins. While Table 9 shows results just for estimates based on $k = 1,000$ tests, Table 10 pools the results for all five values of k, because they were extremely similar. These results show that when the distributional assumptions of maximum likelihood are violated, it can be less accurate than z-curve. Maximum likelihood still beat p-curve and

Table 9: Mean Absolute Error of estimation with beta effect size and correlated sample size and effect size: $k = 1,000$ F-tests with numerator $df = 1$

	Correlation				
	-0.8	-0.6	-0.4	-0.2	0.0
Population Mean Power = 0.25					
P-curve	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	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	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

p-uniform in every comparison, as did z-curve.

Table 10: Number of times row method is significantly more accurate than column method with beta effect size and correlated sample size and effect size: F-tests with numerator $df = 1$

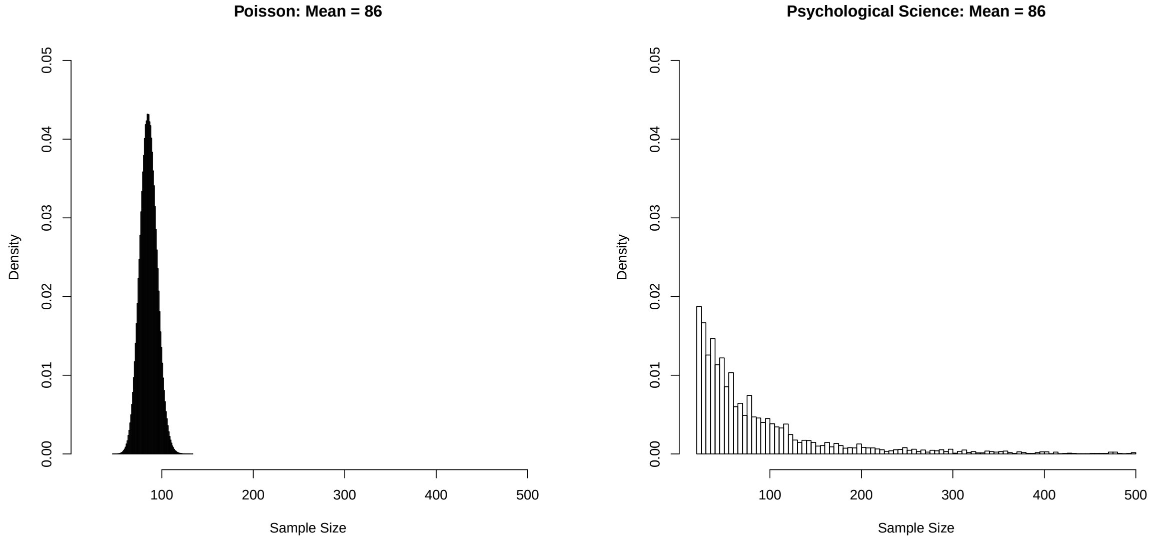
	P-curve	P-uniform	MaxLike	Z-curve	Total
P-curve	0	50	0	0	50
P-uniform	25	0	0	0	25
MaxLike	75	75	0	5	155
Z-curve	75	75	69	0	219

4.4 Full Heterogeneity

When population mean power in a field of study is being estimated, there will typically be heterogeneity not just in sample size and effect size, but also in the tests on which estimates are based. The distribution of sample size is unlikely to be Poisson, the distribution of effect size will not be gamma and the null hypothesis will be true with non-zero probability. Our full heterogeneity simulation examines the performance of the four methods in this situation. Given the performance of p-curve and p-uniform in the previous scenario, we do not expect these methods to perform well. A more important question is how z-curve and maximum likelihood perform when they are faced with full heterogeneity.

Sample size and degrees of freedom In the simulations so far, sample sizes have been Poisson distributed. While the Poisson distribution is a widely accepted model for count data (Johnson, Kemp and Koch, 2005), sample size may be more dispersed and skewed than the Poisson in practice when a variety of research designs are employed. Figure 2 compares the Poisson distribution with mean 86 to a histogram of 7,000 approximate sample sizes based on denominator degrees of freedom in the journal *Psychological Science* (give years). These are preliminary data and not a random sample, but we believe they are closer to reality than the Poisson when a full range of topics is being investigated.

Figure 2: Poisson versus *Psychological Science* Sample Sizes



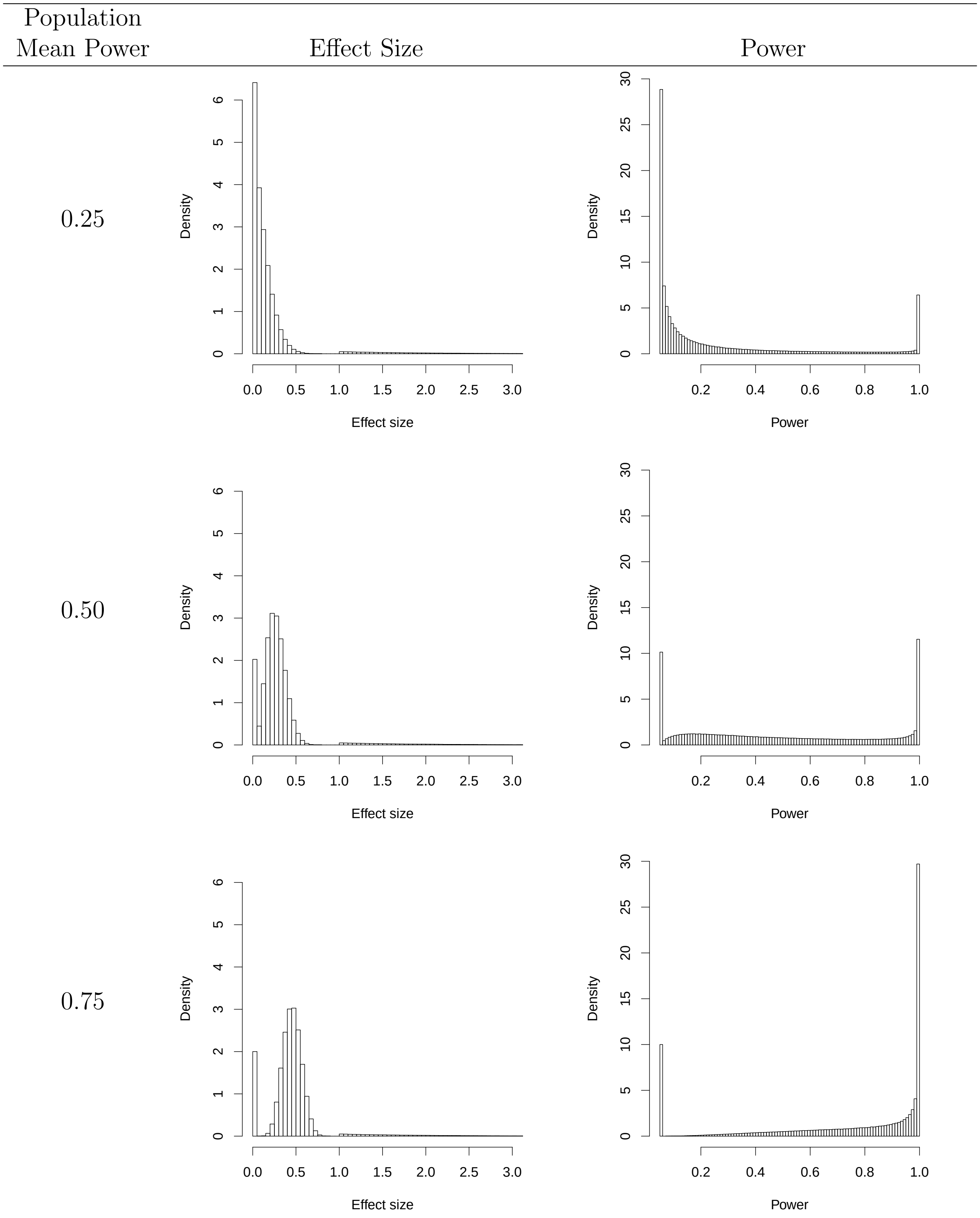
The *Psychological Science* data consist of 7,000 pairs of numerator and denominator degrees of freedom. Actual sample sizes were not collected in this preliminary attempt at data mining, so sample size was approximated by $n = df_1 + df_2 + 1$. Numerator degrees of freedom were limited to ten or fewer, and the data were edited so that sample size ranged from 20 to 500, with a mean of 86.

In this simulation, eighty percent of the tests were F-tests, and twenty percent were chi-squared. For the F-tests, (df_1, df_2) pairs were randomly sampled with replacement from the *Psychological Science* data. The degrees of freedom for the chi-squared tests were randomly sampled with replacement from the df_1 values. Sample size was selected with replacement, independently of degrees of freedom.

Effect size In this set of simulations, effect size has a mixed continuous-discrete distribution. With probability 0.10, effect size equals zero, so that the null hypothesis is exactly true. With probability 0.05, effect size has a standard exponential distribution shifted by one; in this case the minimum effect size is over twice Cohen's (1988) "high" value, representing manipulation checks and other "findings" that are too good to be true. The other 0.85 probability is devoted to a beta distribution, with parameters chosen to make population mean power after selection either 0.25, 0.50 or 0.75. No special attempt was made to hold the standard deviation of effect size constant, but all values were above Section 4.2's high value of 0.30. Sample size and effect size are independent after selection, so that before selection they are non-independent.

Figure 3 shows the distribution of effect size after selection and and the resulting distribution of power after selection. It is evident that the effect of heterogeneity in sample size and effect size is increased heterogeneity in power. Since power is bounded by 0.05 and one, its distribution is forced to the extremes.

Figure 3: Distributions of effect size and power after selection under full heterogeneity



Average performance The p-curve, p-uniform, maximum likelihood and z-curve methods were used to estimate the means of the power distributions depicted in Figure 3. Maximum likelihood continued to assume a gamma distribution for effect size, and three sets of random starting values for the gamma parameters were employed. Table 11 shows means and standard deviations of the estimates. The p-uniform method yields estimates that are much too high. P-curve also over-estimates mean power, though to a much lesser degree than p-uniform. Over-estimation by p-curve is more pronounced when true population mean power is high. Maximum likelihood and z-curve also yield mildly biased estimates, though not in a consistent direction across conditions.

Table 11: Means and standard deviations of estimated population mean power under full heterogeneity

			Mean						*Standard Deviation*		
					Population Mean Power = 0.25						
		Number of Tests							Number of Tests		
	100	250	500	1000	2000		100	250	500	1000	2000
P-curve	0.280	0.280	0.283	0.288	0.292	P-curve	0.072	0.051	0.037	0.027	0.020
P-uniform	0.691	0.776	0.823	0.856	0.877	P-uniform	0.155	0.107	0.077	0.054	0.039
MaxLike	0.267	0.267	0.268	0.269	0.269	MaxLike	0.046	0.029	0.020	0.015	0.012
Z-curve	0.251	0.240	0.234	0.232	0.230	Z-curve	0.064	0.042	0.032	0.025	0.020
					Population Mean Power = 0.50						
		Number of Tests							Number of Tests		
	100	250	500	1000	2000		100	250	500	1000	2000
P-curve	0.561	0.571	0.577	0.581	0.585	P-curve	0.063	0.040	0.029	0.020	0.015
P-uniform	0.807	0.861	0.891	0.911	0.923	P-uniform	0.090	0.060	0.042	0.030	0.022
MaxLike	0.473	0.468	0.465	0.463	0.462	MaxLike	0.054	0.035	0.025	0.019	0.015
Z-curve	0.517	0.505	0.497	0.491	0.487	Z-curve	0.071	0.047	0.035	0.026	0.020
					Population Mean Power = 0.75						
		Number of Tests							Number of Tests		
	100	250	500	1000	2000		100	250	500	1000	2000
Pcurve	0.828	0.836	0.840	0.842	0.844	Pcurve	0.034	0.020	0.014	0.010	0.007
Puniform	0.921	0.945	0.956	0.964	0.968	Puniform	0.035	0.022	0.015	0.011	0.008
MaxLike	0.740	0.736	0.734	0.731	0.730	MaxLike	0.045	0.030	0.022	0.016	0.012
Zcurve	0.764	0.756	0.750	0.745	0.740	Zcurve	0.042	0.030	0.023	0.018 0	.014

Absolute error of estimation Table 12 shows mean absolute differences between the estimates and mean power, multiplied by 100. The p-uniform estimates are unacceptable, and p-curve is clearly less accurate than maximum likelihood or z-curve.

Table 12: Mean Absolute Error of Estimation under Full Heterogeneity

	Number of Tests				
	100	250	500	1000	2000
Population Mean Power = 0.25					
P-curve	6.27	4.68	4.05	4.00	4.25
P-uniform	44.14	52.57	57.35	60.56	62.67
MaxLike	3.87	2.66	2.23	2.03	1.99
Z-curve	5.13	3.53	2.95	2.60	2.43
Population Mean Power = 0.50					
P-curve	7.39	7.21	7.67	8.10	8.50
P-uniform	30.67	36.14	39.13	41.06	42.30
MaxLike	4.81	3.84	3.67	3.74	3.79
Z-curve	5.93	3.78	2.81	2.23	1.98
Population Mean Power = 0.75					
P-curve	7.88	8.62	8.99	9.24	9.41
P-uniform	17.11	19.48	20.61	21.36	21.84
MaxLike	3.67	2.61	2.16	2.03	2.07
Z-curve	3.64	2.45	1.81	1.48	1.38

Within each of the 15 combinations of power and number of tests, there are six potential pairwise comparisons of mean accuracy. These comparisons were carried out using large-sample two-sided matched Z-tests with a Bonferroni correction at the joint 0.001 level. As would be anticipated from Table 12, p-uniform was significantly less accurate than the other methods in all comparisons, and p-curve was significantly less accurate than maximum likelihood and z-curve in all comparisons.

Table 13 counts significant wins and losses; z-curve prevails over maximum likelihood by a score of seven to six. Five of maximum likelihood's six wins occur when the true population mean power is 0.25. In this setting, the z-curve estimate appears to settle down to 0.23 rather than 0.25 as the number of tests k on which the estimate is based increases. This is not a serious error in practice. Note that while the distributional assumptions of maximum likelihood are violated in this simulation, it still performs approximately as well as z-curve.

4.5 A conservative bootstrap confidence interval for z-curve

Estimates should always be accompanied by confidence intervals, to give an idea of their precision. For z-curve, the most natural choice is a bootstrap confidence interval. The bootstrap (Efron 1981, Efron and Tibshirani 1993) is based on re-sampling from the observed

Table 13: Number of times row method is significantly more accurate than column method under full heterogeneity

	P-curve	P-uniform	MaxLike	Z-curve	Total
P-curve	0	15	0	0	15
P-uniform	0	0	0	0	0
MaxLike	15	15	0	6	36
Z-curve	15	15	7	0	37

data with replacement, calculating a statistic on each re-sampled data set, and using the histogram of the resulting values as an approximation to the sampling distribution of the statistic. In this case the statistic is the z-curve estimate. Our choice is the percentile confidence interval method, which assumes that the sampling distribution of the estimate is symmetric, and centered on the quantity being estimated. Here, we re-sampled test statistics and computed z-curve estimates $B = 500$ times. The 95 percent bootstrap confidence interval ranges from the 2.5 percentile to the 97.5 percentile of the estimates.

Especially when samples are small, it is important to verify that a proposed 95% confidence interval contains the true value 95% of the time. This is called the *coverage* of the confidence interval. In a pilot study, we found that the coverage of the 95% bootstrap confidence interval was sometimes less than 95%. For example, notice in Table 11 that the mean estimate for power = 0.25 and $k = 2,000$ is 0.23 rather than 0.25. The sampling distribution of the z-curve estimate is nicely symmetric as required by the bootstrap method, but it is centered on 0.23 and not 0.25. The resulting coverage of the confidence interval is roughly 84% when it should be 95. With increasing volume of data, the width of the confidence interval would shrink and the coverage would decrease to zero.

Reviewing the average z-curve estimates from all the simulations, we determined that the the bias of the z-curve estimate is seldom more than two percentage points, and never more than two percentage points for larger samples. Thus an easy fix of the confidence interval is to decrease the lower limit by 0.02 and increase the upper limit by 0.02. This yields our *conservative bootstrap confidence interval*.

We tested the conservative bootstrap confidence interval in the setting of full heterogeneity, with 10,000 simulated datasets in each combination of three values of true population mean power (again, the distributions in Figure 3), and seven values of the number of test statistics, ranging from $k = 25$ to $k = 2,000$.

Table 14 gives the coverage values. Even for $k = 25$ its performance is respectable. The table shows that the conservative bootstrap confidence interval is indeed conservative under most circumstances. When the estimates are based on larger numbers of test statistics, it behaves more like a 99 percent confidence interval. For estimates based on fewer than 25 test statistics, it might be helpful to increase the correction factor from 0.02 to 0.025.

Table 14: Coverage of the 95% conservative bootstrap confidence interval

Population	Number of Tests						
Mean Power	25	50	100	250	500	1000	2000
0.25	95.78	97.13	98.02	98.69	98.76	98.35	97.95
0.50	94.58	95.51	96.79	98.27	99.11	99.28	99.15
0.75	93.21	94.81	96.83	98.85	99.37	99.73	99.58

Table 15 shows mean upper and lower confidence limits. The upper limit is the top number in each cell, and the lower limit is the bottom number. For example, when the true population mean power is 0.75 and the z-curve estimate is based on $k = 100$ test statistics, the average confidence interval will range from 0.65 to 0.85. This may be sufficient precision for some purposes, but it is desirable to base estimates on a larger number of test statistics if possible.

Table 15: Average Upper and Lower Confidence limits

Population	Number of Tests						
Mean Power	25	50	100	250	500	1000	2000
0.25	0.54	0.46	0.40	0.35	0.32	0.30	0.29
	0.06	0.09	0.11	0.14	0.16	0.17	0.17
0.50	0.76	0.71	0.67	0.62	0.58	0.56	0.55
	0.26	0.32	0.36	0.39	0.41	0.42	0.43
0.75	0.89	0.87	0.85	0.83	0.81	0.80	0.79
	0.55	0.61	0.65	0.67	0.68	0.69	0.69

5 Application to the Replication Project data

Of the 100 original studies in the OSC (2015) Replication Project, three were null results (failures to reject the null hypothesis), and in an additional four studies the original result was only "marginally" significant, with p-values ranging from 0.051 to 0.073. These were set aside, because the methods discussed in this paper require $p < 0.05$. Of the remaining 93 studies, five were eliminated for other reasons, leaving 88. Of these, thirty-four tests or 39% were significant on replication.

Most of the test statistics for the originally reported tests were F or chi-squared. The remainder were converted by squaring t statistics to obtain Fs, and squaring Z statistics t obtain chi-squared with one degree of freedom. Input to z-curve was simply the set of p-values. For the other three methods, test statistics were divided into subsets according to the type of test (F or chi-squared) and the (numerator) degrees of freedom. Estimates were calculated for each subset, and then combined as a weighted sum, using the observed proportions of the subsets as weights.

The estimates of population mean power were 0.68 for p-curve, 0.76 for p-uniform, 0.59 for maximum likelihood and 0.66 for z-curve. The 95% confidence interval for z-curve was from 0.49 to 0.79. While $k = 88$ test statistic is not a large number, it is still notable that all the estimates of population mean power were substantially greater than the observed replication rate. In retrospect this should not be surprising. The estimation methods and simulations assume that hypotheses are tested once, published if and only if the results are significant, and then the studies leading to significant results are replicated exactly in every detail. When these conditions are not satisfied, the probability of significance upon replication will typically be diminished. More detail is given in the Discussion section.

6 Discussion

In this paper, we have compared four methods for estimating population mean power after selection for significance: p-curve, p-uniform, maximum likelihood and z-curve. P-curve (Simonsohn et al., 2014b) and p-uniform (van Assen et al., 2014) are slight adaptations of methods for estimating a fixed effect size. Maximum likelihood is a generic approach to estimation for any parametric model, and z-curve is new. Based on a set of large-scale simulation studies, we conclude that z-curve is the most accurate method when there is substantial heterogeneity in effect size and the distribution of effect size is unknown. It is also the most convenient, requiring only a set of p-values as input. Estimates should be accompanied by confidence intervals. We describe a conservative bootstrap confidence interval for z-curve and verify by simulation that it has good coverage even for small samples.

In a meta-analysis of studies testing exactly the same hypothesis with very similar subject populations, it is reasonable to assume that effect size is a single fixed constant, while sample size of course may vary. This is the setting for which p-curve and p-uniform were designed. Here, all the methods performed reasonably well in our simulations. The most accurate method was maximum likelihood, followed by p-uniform. Then we introduced heterogeneity in effect size. In this situation, maximum likelihood estimates are based on a parametric model for the distribution of effect size, and also for the relationship between sample size and effect size. We carried out another large-scale simulation experiment in which effect size was gamma distributed and independent of sample size before selection. Maximum likelihood made full use of these features. When heterogeneity in effect size was moderate to high, maximum likelihood was by far the most accurate method in spite of numerical difficulties. The next most accurate was z-curve, which performed acceptably but not as well as maximum likelihood. The effect of high heterogeneity on p-uniform was particularly severe, leading to very high estimated mean power almost regardless of the true value.

In practice, the probability distribution of effect size will never be known, and effect size may well be related to sample size. To test the robustness of maximum likelihood, we conducted a study in which effect size was beta distributed (limited to values between zero and one, in contrast to the assumed right-skewed gamma distribution), and the population correlation between sample size ranged from zero to -0.8. Maximum likelihood continued to assume a gamma distribution for effect size and zero correlation between sample size and effect size. Here, z-curve was clearly more accurate than maximum likelihood, which in turn still out-performed p-curve and p-uniform. There is clear evidence that maximum likelihood estimation of power is sensitive to violation of distributional assumptions; correlation between sample size and effect size had little effect. In another simulation where effect size was right skewed but not gamma distributed, z-curve and maximum likelihood performed about equally well. We conclude that since the distribution of effect size is always unknown and moderate heterogeneity in effect size cannot be ruled out, the preferred method of estimating population mean power from published results is z-curve.

Some important statistical features of z-curve require further investigation. One is the question of independence. In all he simulations, the input p-values were independent. While z-curve does not formally assume independent inputs, the bootstrap confidence interval definitely does. Further simulations could provide reassurance (or raise a warning flag) about the performance of the method when clusters of p-values come from tests conducted on the same raw data set. Another unresolved issue is how well the method performs for tests that do not have one of the common non-central distributions under the alternative hypothesis. The most important case is in classical repeated measures ANOVA, where many test statistics have central F distributions when the null hypothesis is true, but multiples of a central F when the null hypothesis is false and power is greater than 0.05. Preliminary results are encouraging, but a full simulation study is needed.

When we applied z-curve and the other methods to the OSC replication data, we obtained estimates that were not extremely different from one another, suggesting moderate heterogeneity in effect size. The estimate for z-curve was 0.66 compared to a much lower empirical replication rate of 0.39. Because the estimate was based on a fairly modest number of studies, the confidence interval was wide, ranging from 0.49 to 0.79. Though the 39% replication rate is also subject to sampling error, it is clear that mean power was quite a bit greater than actual replicability.

While the gap between mean power and replicability is greater than we anticipated, still it is be expected. Both the simulations and the theory behind the estimation methods are based on a simplified and idealized model of the research process, one that makes strong assumptions about how published test statistics are generated. First, it is assumed that hypotheses are formulated in advance, and results are published if and only if $p < 0.05$. But as we know too well, not all submitted papers are published. Any tendency on the part of reviewers to select "interesting" (that is, unexpected) results for publication may at the same time select phenomena with small or zero effect sizes.

Second, it is assumed that studies are replicated exactly in every detail, so that when a result is significant and is reported, the power of the test on replication remains identical to what it was in the original study. The replications in the OSC project were certainly not exact in most cases. The subject populations were usually different, and even sample sizes were sometimes changed — for reasons that no doubt seemed good at the time. There is evidence that in general, the effect sizes in replication studies tend to be lower than in the original studies (need reference). This probably held true in the OSC project too. For a multitude of possible reasons including weaker expectancy effects (Rosenthal, 1966) in the replication studies, the mean power of the tests on replication in the OSC project may well have been less than they were in the original studies. We suspect that this is true of replications in general, and is not limited to the OSC replication project.

Third, the estimation methods and the simulations make no allowance for the kind of exploratory statistical analysis that capitalizes on chance, and makes statistical significance a near certainty. The terms "vibration effects" (Ioannidis, 2008), "False-Positive Psychology" (Simmons, Nelson and Simonsohn, 2011) "p-hacking" (Simonsohn et al., 2014a) and "Questionable Research Practices" (Schimmack, 2012) have been used. Here, we will call it p-hacking. We have no doubt that the practice of p-hacking is widespread and that it reduces the average true power of published studies. The question is how it influences *estimates* of power.

Here are two competing hypotheses. One possibility is that p-hacking is just another form of selection for significance, essentially searching at random through a set of possibilities much like the original unselected population, and stopping once it finds a test that is significant. In this case, estimation methods that allow for selection (publication bias) also automatically allow for p-hacking. Simonsohn et al. (2014b) report some simulations that suggest optimism about the effect of p-hacking upon p-curve estimates of effect size. Similar but larger-scale simulations are needed for mean power.

Another perspective on p-hacking is less comforting. In selection for significance, it may be that if one test is not significant, the next will happen upon a large effect, or employ a large sample size. The resulting power will be higher. But in p-hacking, the data analyst is testing roughly the same hypothesis over and over in different ways (perhaps using different covariates or discarding different "outliers") until something works. True (as opposed to estimated) effect size will not change much, and neither will sample size. In the end, average power will not be affected much at all by a selection process that essentially selects everything. The positive effects of selection for significance upon average power will be lost, and the power of published findings will be distributed much like power before selection. By Principle 5, the effect of p-hacking upon mean power will be largest when power before selection is widely dispersed and low on average. Unfortunately, this is the setting where p-hacking would be most helpful for professional advancement. It is quite possible that the discrepancy between estimated mean power and replication rate in the OSC data arose from p-hacking.

Whatever the source of the discrepancy, it is clear that that estimates of mean power provide conservative estimates of replicability. When they are low, one may be assured that replicability is still lower. By mining the published literature for genuine probability samples of p-values and then applying z-curve, it will be possible to determine where the replicability problem is most acute in the scientific literature. It would be prohibitively expensive to do this by literally replicating random samples of studies. Estimating mean power is the answer, and z-curve is the current state of the art.

References

[1] Aldrich, J. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science*, 12, 162-176.

[2] Baker, M. (2016) Is there a reproducibility crisis? *Nature* 533, 452454.

[3] Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The new S Language: A programming environment for data analysis and graphics.* Pacific Grove, California: Wadsworth& Brooks/Cole.

[4] Begley, C.G., (2013) Reproducibility: Six red flags for suspect work. *Nature* 497, 433434.

[5] Begley, C, G. and Ellis, L. M. (2012) Drug development: Raise standards for preclinical cancer research. *Nature* 483, 531-533.

[6] Billingsley, P. (1986). *Probability and measure.* New York: Wiley.

[7] Bishop, Y. M. M., Feinberg, S. E. and Holland, M. M. (1975). *Discrete multivariate analysis.* Cambridge, Mass.: MIT Press.

[8] Bollen, K. A. (1989), *Structural equations with latent variables,* New York: Wiley.

[9] Boos, D. D. and Stefnski, L. A. (2012). P-value precision and reproducibility. *The American Statistician* 65, 213-221.

[10] Chang, A. C. and Li, P. (2015) Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not", *Finance and Economics Discussion Series 2015- 083.* Washington, D.C.: Board of Governors of the Federal Reserve System.

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

[12] Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (2nd Edition), Hilsdale, New Jersey: Erlbaum.

[13] Desu, M. M. and Raghavarao, D. (1990). *Sample size methodology.* New York: Academic Press.

[14] Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* 68, 589599

[15] Efron, R. and Tibshirani, R. (1993) *An introduction to the bootstrap.* New York: Chapman and Hall.

[16] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A* 222 309-368.

[17] Fisher, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society of London. Series A* 121, 654-673.

[18] Gerard, P. D., Smith, D. R. & Weerakkody, G. (1998). Limits of retrospective power analysis. Journal of Wildlife Management, 62, 801 - 807.

[19] Greenwald, A. G., Gonzalez, R., Harris, R. J., and Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology*, 33, 175183.

[20] Gillett, R. (1994). Post hoc power analysis. *Journal of Applied Psychology* 79, 783785.

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

[22] Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, 9, 6185.

[23] Hirschhorn, J. N., Lohmueller, K., Byrne, E., Hirschhorn K. (2002) A comprehensive review of genetic association studies. *Genetics in Medicine* 4, 4561.

[24] Hoenig, J. M. and Heisey, D.M (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician* 55, 19-24.

[25] Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, 19(5), 640-646.

[26] Ioannidis, J. P., and Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4, 245253

[27] John, L. K., Lowenstein, G. and Prelec, D. (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23 517-523

[28] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd. Edition). New York: Wiley.

[29] Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). *Univariate discrete distributions.* (3d Edition). Hoboken, N.J.: Wiley.

[30] Kepes, S., Banks, G. C., McDaniel, M., and Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods*, 15, 624662

[31] Lehman, E. L. (1959). Testing statistical hypotheses. New York: Wiley.

[32] Lehman, E. L. and Casella, G. (1998) *Theory of point estimation* (2nd. Edition). New York: Springer.

[33] Lehman, E. L. and Romano, J. P. (2010) *Testing statistical hypotheses.* (3d Edition). New York: Wiley.

[34] Lindsay, B. G. and Roeder, K. (2008). Uniqueness of estimation and identifiability in mixture models. *Canadian Journal of Statistics* 21, 139-147.

[35] McCullagh, P. and Nelder, J. A. (1989) *Generalized linear models.* (2nd Edition). New York: Chapman and Hall.

[36] Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A* 231, 289337.

[37] Patinak, P. B. (1949). The non-central χ^2 and F distributions and their applications. *Biometrika*, 36, 202-232.

[38] Pinsky, M. A. and Karlin S. (2011). *An introduction to stochastic modeling.* San Diego: Academic Press.

[39] Popper, K. R. (1959). *The logic of scientific discovery.* English translation by Popper of *Logik der Forschung* (1934). London: Hutchinson.

[40] Posavac, E. J. (2002). Using p values to estimate the probability of a statistically significant replication. *Understanding Statistics*, 1, 101112.

[41] R Core Team (2012). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

[42] Rosenthal, R. (1966) *Experimenter effects in behavioral research.* New York: Appleton-Century-Crofts.

[43] Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin*, 86(3), 638.

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

[45] Silverman, B. W. (1986) *Density Estimation.* London: Chapman and Hall.

[46] Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22 1359-1366.

[47] Simonsohn, U., Nelson, L. D. and Simmons, J. P. (2014a). *P*-curve: A key to the file drawer. *Journal of experimental psychology: General*, 143, 534-547.

[48] Simonsohn, U., Nelson, L. D. and Simmons, J. P. (2014b). *p*-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspectives on Psychological Science*, 9, 666-681.

[49] 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.

[50] Sterling, T. D., Rosenbaum, W.L. and 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, Vol.1: Adjustment during Army Life.* Princeton University Press, Princeton.

[51] Stuart, A. and Ord, J. K. (1999). *Kendall's Advanced Theory of Statistics, Vol. 2: Classical Inference & the Linear Model* (5th ed.). New York: Oxford University Press.

[52] Thomas, L. (1997) Retrospective Power Analysis. *Conservation Biology* 11, 276-280.

[53] van Assen, M. A. L. M., van Aert, R. C. M. and Wicherts, J. M. (2014) Meta-analysis using effect size distributions of only statistically significant studies. *Psychological methods* 1-18.

[54] Yuan, K. H. and Maxwell, S. (2005) On the post hoc power in testing mean differences. *Journal of educational and behavioral statistics* 30, 141-167.