Shrinkage, False Discovery Rate, and False Sign Rate Estimation when precision varies across units

Matthew Stephens[1*],
1 Department of Statistics and Department of Human Genetics, University of Chicago, Chicago, IL, USA
∗ **E-mail: Corresponding mstephens@uchicago.edu**

Abstract

Introduction

Suppose that we measure, with error, a series of "effects", $\beta_1, \ldots, \beta_J$. To take just one concrete example, β_j could be the difference in the mean (log) expression levels of gene j ($j = 1, \ldots, J$) between 2 conditions. In this case, a measurement of the effect might be the difference in sample means obtained in the two conditions. We will let $\hat{\beta}_j$ denote the measured value of β_j, and assume that each measurement comes with an associated standard error, s_j. A key aim here will be to take proper account of the fact that some measurements may be more precise than others: that is, to take proper account of variation in s_j across j.

A common goal, particularly in genomic studies, is to identify which β_j differ from zero. This is commonly tackled by first computing an effect size estimate ($\hat{\beta}_j$) and its standard error (s_j), converting this to a Z score ($Z_j = \hat{\beta}_j/s_j$) and a corresponding p value (p_j), testing $H_j : \beta_j = 0$. Then standard methods (e.g. the qvalue package) can be used to estimate False Discovery Rates at any given threshold.

There are two issues with this approach that I would like to address here. The first is that it really does not take proper account of the measurement errors. To see this, consider an example where half the measurements are quite precise, and the other half are really, really, poor. Intuitively, the poor measurements tell us nothing, and any sane analysis should effectively ignore them. However, in a standard FDR-type analysis, these poor measurements add "noise" and affect estimated FDRs. This is because the p values from the poor measurements will be effectively uniformly distributed, and some will be significant at any given threshold.

The second issue is that directly modeling the p values, say via non-parametric methods, without taking account of their precision, can lead to unrealistic distributions being fitted. Put another way, because z scores are the result of adding noise to some distribution, the range of distributions they can take is limited Using entirely non-parametric methods loses this information. The solution is to model beta hat as a convolution of some distribution g and an error component.

The initial goal of the ASH (Adaptive SHrinkage) project is to provide simple, generic, and flexible methods to derive "shrinkage-based" estimates and credible intervals for unknown quantities $\beta = (\beta_1, \ldots, \beta_J)$, given only estimates of those quantities ($\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_J)$) and their corresponding estimated standard errors ($s = (s_1, \ldots, s_J)$).

Although shrinkage-based estimation can be motivated in various ways, our key goal here is to combine information across the multiple measurements $j = 1, \ldots, J$ to improve inference for each individual β_j. By improved inference, we mean both improved average accuracy of point estimates, which is the traditional focus of shrinkage-based methods, *and* improved assessments of uncertainty.

By "adaptive" shrinkage we have two key properties in mind. First, the appropriate amount of shrinkage is determined from the data, rather than being pre-specified. Second, the amount of shrinkage undergone by each $\hat{\beta}_j$ will depend on the standard error s_j: measurements with high standard error will undergo more shrinkage than measurements with low standard error.

Given that shrinkage estimation is widely recognized as a powerful tool, there are suprisingly few software packages for performing the simplest type of shrinkage estimation considered here. (There are

more packages for the more complex setting of covariance estimation, where shrinkage is perhaps still more important.) The only package we have found that provides anything similar to the functionality provided here is mixfdr (Muralidharan). Compared with mixfdr, the key features of ashr are that it i) focuses on allowing for variation in the standard deviation of each observation; ii) constrains the underlying density to be unimodal (and possibly symettric). NOTE: should emphasise these differences in the examples.

As an important special case, these methods address the "multiple comparisons" setting, where interest usually focuses on which β_j can be confidently inferred to be non-zero. Such problems are usually tackled by computing a p value for each j, often by applying a t test to $\hat{\beta}_j/s_j$, and then applying a generic procedure, such as that of Benjamini and Hochberg (1995?) or Storey (2001?), designed to control or estimate the false discovery rate (FDR) or the positive FDR (Storey, 2001?). In essence we aim to provide analagous generic methods that work directly with two numbers for each measurement $(\hat{\beta}_j, s_j)$, rather than a single number (e.g. the p value, or t statistic). Working with these two numbers has two important benefits: first, it permits estimation and not only testing; second, the uncertainty in each measurement $\hat{\beta}_j$ can be more fully accounted for, reducing the impact of "high-noise" measurements (large s_j) that can reduce the effectiveness of a standard FDR analysis.

The potential for shrinkage-based estimation to address the multiple comparisons setting has been highlighted previously, including Greenland and Robins (1991), Efron (2008) and Gelman et al (2012). [Note, check also Louis, JASA, 1984]

It is common in statistics that you measure many "similar" things imperfectly, and wish to estimate their values. The situation arises commonly in the kinds of genomics applications I am often involved in, but also in other areas of statistics. In genomics, for example, a very common goal is to compare the mean expression (activity) level of many genes in two conditions. Let μ_j^0 and μ_j^1 denote the mean expression of gene j ($j = 1, \ldots, J$) in the two conditions, and define $\beta_j := \mu_j^0 - \mu_j^1$ to be the difference. Typically expression measurements are made on only a small number of samples in each condition - sometimes as few as one sample in each condition. Thus the error in estimates of μ_j^0 and μ_j^1 is appreciable, and the error in estimates of β_j still greater.

A fundamental idea is that the measurements of β_j for each gene can be used to improve inference for the values of β for other genes.

Methods

The two main assumptions we will make here are that * these effects are exchangeable: that is (by de Finetti's theorem) they can be thought of as being independent and identically distributed from some unknown distribution $g(\beta)$. * these effects are symmetric and unimodal about 0.

Specifically we will assume a parametric form for g as a mixture of 0-centered normal distributions

A key issue I want to address here is that if these measurements are made with different precisions, then we want to take this into account in our analysis. We can do this by assuming that the likelihood for β is a normal, with mean $\hat{\beta}_p$ and standard deviation s_p. (Note that this is equivalent to the likelihood we would get if we "observed" data $\hat{\beta}_p \sim N(\beta_p, s_p)$.)

The methods are based on treating the vectors $\hat{\beta}$ and s as "observed data", and then performing inference for β from these observed data, using a standard hierarchical modelling framework to combine information across $j = 1, \ldots, J$.

Specifically, we assume that the true β_j values are independent and identically distributed from some distribution $g(\cdot; \pi)$, where π is a hyper-parameter to be estimated. Then, given β, we assume that $(\hat{\beta}_j, s_j)$ are independent across j, and depend only on β_j. Putting these together, the joint model for the

unobserved β and the observed $\hat{\beta}, s$ is:

$$p(\hat{\beta}, s, \beta|\pi) = \prod_j g(\beta_j;\pi)p(\hat{\beta}_j, s_j|\beta_j) \quad (1)$$

$$= \prod_j g(\beta_j;\pi)L(\beta_j;\hat{\beta}_j, s_j). \quad (2)$$

The specific choices of g and L are described below.

We fit this hierarchical model using the following "Empirical Bayes" approach. First we estimate the hyper-parameters π by maximizing the likelihood

$$L(\pi;\hat{\beta}, s) := p(\hat{\beta}, s|\pi) = \int p(\hat{\beta}, s, \beta|\pi)d\beta.$$

Then, given this estimate $\hat{\pi}$, we compute the conditional distributions

$$p(\beta_j|\hat{\pi}, \hat{\beta}, s) \propto g(\beta_j;\pi)L(\beta_j;\hat{\beta}_j, s_j).$$

In principle we would prefer to take a full Bayes approach that accounts for uncertainty in π, but, at least for now, we compromise this principle for the simplicity of the EB approach. [Note: a Variational Bayes version of this is also implemented, and may become our preferred approach after testing]

[put picture of hierarchical model here]

The conditional distributions $p(\beta_j|\hat{\pi}, \hat{\beta}, s)$ encapsulate uncertainty in the values for β_j, combining information across $j = 1, \ldots, J$. The combining of the information occurs through estimation of π, which involves all of the data, and it is

These conditional distributions can be conveniently summarized in various ways, including point estimates (e.g. the posterior means or medians), and credible intervals/regions.

Efron (2008) states "most of the z-values near zero come from null genes" His main aim in making this assumption is to estimate an empirical null though (not assume N(0,1) for the null) rather than to impose identifiability.

By modeling the z scores directly under the alternative, rather than the z scores as being the truth plus noise, maybe that is what is most problematic? Because the resulting distribution of z scores is just not possible.

Note that [**?**] models z scores as something plus noise under both H_0 and H_1, which avoids this problem. (DOes the same maybe apply to modeling beta, rather than z scores, when the errors vary?)

Rice and Spiegelhalter - BRCA data?

The key components of this hierarchical model are the distribution g and the likelihood $L(\beta_j;\hat{\beta}_j, s_j)$. We make the following choices for these.

1. The likelihood for β_j is normal, centered on $\hat{\beta}_j$, with standard deviation s_j. That is,

$$L(\beta_j;\hat{\beta}_j, s_j) \propto \exp[-0.5(\beta_j - \hat{\beta}_j)^2/s_j^2]. \quad (**)$$

2. The distribution $g(\cdot;\pi)$ is a mixture of zero-centered normal distributions,

$$g(\cdot;\pi) = \sum_{k=1}^{K} \pi_k N(\cdot; 0, \sigma_k^2).$$

In practice, we currently fix the number of components K to be large, and take the variances $\sigma_1 < \sigma_2 < \cdots < \sigma_K$ to be fixed, and vary from very small (possibly 0), to very large – sufficiently large that typically $\hat{\pi}_K = 0$.

The choice of normal likelihood seems natural, and indeed it can be motivated in multiple ways. For example, we can write $p(\hat{\beta}_j, s_j|\beta_j) = p(\hat{\beta}_j|s_j, \beta_j)p(s_j|\beta_j)$. Now, if we are willing to assume that s_j alone contains no information about β_j, or equivalently that $p(s_j|\beta_j)$ does not depend on β_j, then $L(\beta_j) \propto p(\hat{\beta}_j|s_j, \beta_j)$, and if $\hat{\beta}_j|s_j, \beta_j \sim N(\beta_j, s_j^2)$, as is often asymptotically the case, then we obtain the likelihood (**) above.

An alternative motivation is to think of this as a normal approximation to the likelihood from the raw data D_j that were used to compute $\hat{\beta}_j$ and s_j. Then if we observed these data the likelihood for β would be $p(D_j|\beta_j)$, and a Taylor series expansion of the log likelihood around the maximum likelihood estimate $\hat{\beta}_j$ yields

$$l(\beta_j) \approx l(\hat{\beta}_j) + 0.5 * (\beta_j - \hat{\beta}_j)^2 l''(\hat{\beta}_j).$$

[Fill in details?]

Using a mixture of normal distributions for g also seems very natural: mixtures of normals provide a flexible family of distributions able to provide a good approximation to any true underlying g; and when combined with the normal likelihood they give an analytic form for the conditional distribution $p(\beta_j|\hat{\pi}, \hat{\beta}_j, s_j)$ (also a mixture of normals). The constraint that these normals be centered at zero may seem initially less natural. Certainly this constraint could be relaxed in principle. However, we view it as a convenient way to impose an assumption that g is unimodal with its mode at 0, which we view as a plausible assumption in many settings, and one which may be helpful to avoid "overfitting" of g. (Using normal distributions centered at 0 also imposes an assumption that g is symmetric about zero, which we view as less plausible, but represents a compromise between simplicity and flexibility. In cases where this assumption seems wildly inappropriate one could perhaps improve results by applying the model separately to positive and negative values of $\hat{\beta}$.)

Finally, using a large number of normal components with a wide range of variances, rather than, say, a smaller number of components with the variances to be estimated, is simply for computational convenience. With fixed variances there exists a very simple EM algorithm to maximise the likelihood in π. Obtaining maximum likelihood estimates for the variances could certainly be implemented with a little more work, but it is unclear whether this would result in practically-important gains in many situations of interest.

0.1 The local False Sign Rate

The local False Discovery Rate (lfdr) for observation j is

$$\text{lfdr}_j = p(\beta_j = 0|\hat{\beta}, s). \tag{3}$$

The lfdr terminology comes from using "discovery" to refer to rejecting the null ($H_j : \beta_j = 0$). Specifically lfdr_j is the probability that, if we reject H_j, it is a "false discovery".

As pointed out by [**?**], there are settings where $\beta_j = 0$ is implausible, in which case the lfdr is not a useful concept: if every β_j is non-zero then there is no such thing as a false discovery and the lfdr will be identically 0 for all methods. Gelman et al suggest that in such settings we might replace the concept of a false discovery with the concept of an "error in sign". The idea is that, in settings where $\beta_j = 0$ is implausible, the most fundamental inference objective is to ask which β_j are positive and which are negative. Indeed, even in settings where some β_j are exactly zero, it could be argued that identifying which are positive and which negative is fundamentally more interesting and useful than identifying which are non-zero.

Following these ideas, we define the local False Sign Rate (lfsr) for observation j as

$$\text{lfsr}_j := \min[p(\beta_j \geq 0|\hat{\beta}, s), p(\beta_j \leq 0|\hat{\beta}, s)]. \tag{4}$$

Thus lfsr_j gives the probability that we would get the sign of β_j wrong if we were to make our best guess. (Note that we count it as an error to state that β is positive or negative when it is truly zero.) To

illustrate, suppose for concreteness that the minimum is achieved by the first term, $p(\beta_j \geq 0|\hat{\beta}, s) = 0.05$ say. Then our best guess would be that β is negative, and the probability that we have made an error in sign would be 0.05. We note that the idea of focussing on tail areas, rather than point null hypotheses, has a long history (e.g. Altham? others?).

Note that $\text{lfsr}_j \geq \text{lfdr}_j$ because both the events $\beta_j \geq 0$ and $\beta_j \leq 0$ include the event $\beta_j = 0$. Thus, lfsr gives an upper bound for lfdr.

0.2 Computation Outline

As outlined above, we fit the model using the following Empirical Bayes procedure: 1. Estimate π by maximizing the likelihood $L(\pi)$. 2. Compute the conditional distributions $p(\beta_j|\hat{\beta}_j, s_j, \hat{\pi})$.

Using a normal likelihood $L(\beta_j)$, and assuming g to be a mixture of normal distributions with fixed variances, yields a simple EM algorithm for estimating π in Step 1, and simple analytic forms for the conditional distributions in Step 2.

Results

0.3 Improved conservative estimation of π_0

To illustrate, we provide simulation results for two scenarios, the first being a "difficult" case where many non-zero β are too close to zero to be reliably detected, and the second being an "easier" case where most non-zero β are sufficiently different from zero that they can be reliably detected.

Specifically we use the following distributions for the non-zero β:

Simulation 1:

$$f_1(\cdot) = 0.4N(0, 0.25^2) + 0.2N(0, 0.5^2) + 0.2N(0, 1^2), 0.2N(0, 2^2) \tag{5}$$

Simulation 2:

$$f_1(\cdot) = N(0, 4^2) \tag{6}$$

Emphasize that although pi0 is unidentifiable, the underlying distribution g is estimable quite accurately.

0.4 Local False Sign Rate

There are two reasons to use the lfsr instead of the lfdr: it is more generally meaningful (e.g. it applies whether or not zero effects exist), and estimation of lfsr is more robust to modeling assumptions.

- Illustration that estimated lfsr is more robust.
- Illustration that adjusted estimate of lfsr is still more robust..
- Compare estimated lfsr with true lfsr for difficult simulation case, to show nearly achieves Bayes Risk?

0.5 Accounting for measurement precision improves estimates of FDR/fsr

- Illustration of effects of contamination on FDR by low-quality data comparison with ash.
- Comparison between exchangeable effects vs Exchangeable standardized effects; Hedenberg data.
- Possibly difference between CIs obtained by ash and usual CIs (e.g. "of intervals where local fsr is ¡0.05, what proportion are the sign correct"?).
- Asymmetric case (Half uniforms)

Things to emphasise for paper: - the number of components is not critical; the unimodal constraint is enough. You can increase the number of components arbitrarily and the likelihood remains bounded.

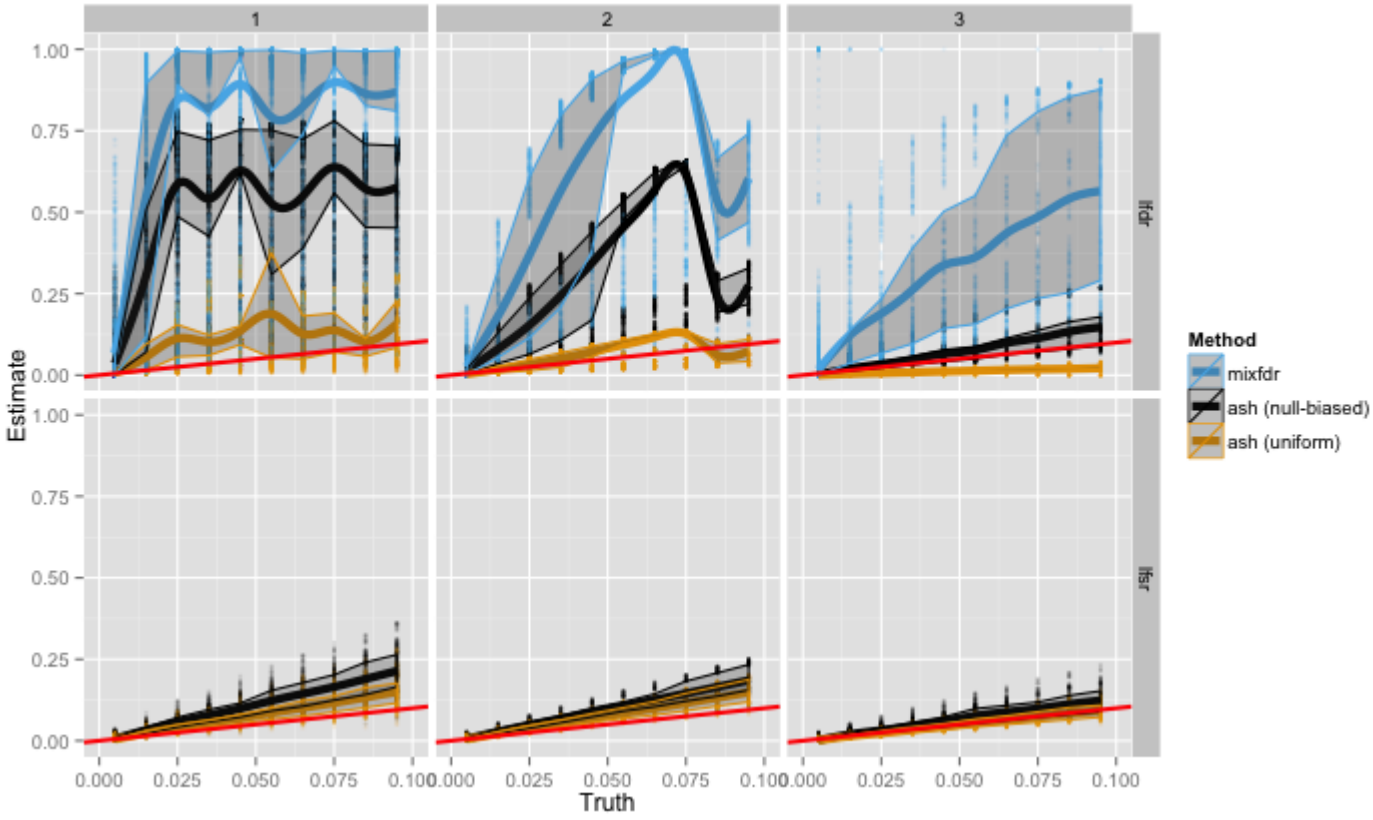

Figure 1. Figure showing lfsr is more robust than lfdr

- the ability to incorporate item-specific measurement error - the ability to compare a model in which effects are proportional to error with model in which effects

- unimodal distributions could be argued to be a primary motivation for shrinkage

- bayesian coverage intervals (attempt to) give much stronger guarantess than standard confidence intervals. Eg among the 0.95 intervals excluding 0, less than 5

Some advantages of the unimodal constraint: -more statsitically efficient (if true); show with small sample sizes. - less sensitive to number of components? Can allow number to tend to infinity? - may be less sensitive to local optima? could demonstrate this by looking at convergence more carefully, and comparing results with random starting points.

- it allows us to easily just vary sigma on a grid, and fit pi, which makes allowing different noise levels really easy!

0.6 The number of components is not critical

OR - this section could be more generally about selecting between models (normal, uniforms etc) using log-likelihood.

Because of the uni-modal constraint, the number of mixture components is generally not critical, at least provided it is "sufficiently large". Indeed, as the number of components tends to infinity, the

likelihood is bounded above, and even for large numbers of components and small amounts of data the inferred underlying distributions tend not to be too crazy, showing few signs of "overfitting".

However, it is true that the normal distributions tend to yield smoother estimated underlying distributions. Similarly, using the posterior mean for π, rather than the maximum likelihood estimate, tends to lead to somewhat smoother fitted distributions. In addition, with small amounts of data the underlying distributions are inevitably not well determined by the data, and the fits may vary depending on the underlying assumptions made (e.g. number of components or distribution of the components); however, even then shrinkage-based estimates of the betas can be relatively robust.

0.7 Do we need a point mass at zero?

In some settings it is the convention to focus on testing whether $\beta_j = 0$. However some dislike this focus, objecting that it is unlikely to be the case that $\beta_j = 0$ exactly. For example, when comparing the average expression of a gene in human samples vs chimp samples, it might be considered unlikely that the expression is *exactly* the same in both. Whether or not $\beta_j = 0$ is considered unlikely may depend on the context. However, in most contexts, finite data cannot distinguish between $\beta_j = 0$ and β_j being very close to zero. Thus finite data cannot usually convince a skeptic that β_j is exactly zero, rather than just very small. In contrast it is easy to imagine data that would convince a doubter that β_j is truly non-zero. In this sense there is an assymetry between the inferences "β_j is zero" and "β_j is non-zero", an assymetry that is reflected in the admonition "failure to reject the null does not imply it to be true".

Thus any analysis that purports to distinguish between these cases must be making an assumption.

Consider two analyses of the same data, using two different "priors" g for β_j, that effectively differ only in their assumptions about whether or not β_j can be exactly zero. For concreteness, consider

$$g_1(\cdot) = \pi\delta_0(\cdot) + (1-\pi)N(\cdot; 0, \sigma^2)$$

and

$$g_2(\cdot) = \pi N(\cdot; 0, \epsilon^2) + (1-\pi)N(\cdot; 0, \sigma^2).$$

If ϵ^2 is sufficiently small, then these priors are "approximately the same", and will lead to "approximately the same" posteriors and inferences in many senses. To discuss these, let p_j denote the posterior under prior g_j. Then, for any given (small) δ, we will have $p_1(|\beta_j| < \delta) \approx p_2(|\beta_j| < \delta)$. However, we will not have $p_1(\beta_j = 0) \approx p_2(\beta_j = 0)$: the latter will always be zero, while the former could be appreciable.

What if, instead, we examine $p_1(\beta_j > 0)$ and $p_2(\beta_j > 0)$? Again, these will differ. If this probability is big in the first analysis, say $1-\alpha$ with α small, then it could be as big as $1-\alpha/2$ in the second analysis. This is because if $p_1(\beta_j > 0) = 1-\alpha$, then $p_1(\beta_j = 0)$ will often be close to α, so for small ϵ $p_2(\beta_j)$ will have mass α near 0, of which half will be positive and half will be negative. Thus if we do an analysis without a point mass, but allow for mass near 0, then we may predict what the results would have been if we had used a point mass.

Let's try: "'r beta.ash.pm = ash(ss$betahat$, ssbetasd, usePointMass=TRUE) print(beta.ash.pm) print(beta.ash.auto) plot(beta.ash.auto$localfsr$, $beta.ash.pm$localfsr,main="comparison of ash localfsr, with and without point mass",xlab="no point mass", ylab="with point mass",xlim=c(0,1),ylim=c(0,1)) abline(a=0,b=1) abline(a=0,b=2) "'

Our conclusion: if we simulate data with a point mass, and we analyze it without a point mass, we may underestimate the lfsr by a factor of 2. Therefore, to be conservative, we might prefer to analyze the data allowing for the point mass, or, if analyzed without a point mass, multiply estimated false sign rates by 2. In fact the latter might be preferable: even if we analyze the data with a point mass, there is going to be some unidentifiability that means estimating the pi value on the point mass will be somewhat unreliable, and we still might underestimate the false sign rate if we rely on that estimate. TO THINK ABOUT: does multiplying the smaller of Pr(¡0) and Pr(¿0) by 2, and adding to Pr(=0) solve the problem in either case?

0.8 Side notes on Multiple comparisons

Note on multiple comparisons: it isn't really a "problem" but an "opportunity". This viewpoint also espoused by Greenland and Robins. It isn't the number of tests that is relevant (the false discovery rate at a given threshold does not depend on the number of tests). It is the *results* of the tests that are relevant.

Performing multiple comparisons, or multiple tests, is often regarded as a "problem". However, here we regard it instead as an opportunity - an opportunity to combine (or "pool") information across tests or comparisons.

Imagine that you are preparing to perform 1 million tests, each based on a Z score that is assumed to be $N(0, 1)$ under the null. You first order these tests randomly, and begin by performing the first test, which returns a Z score of 4. At this point you are interrupted by a friend, who asks how the analysis is going. "It's early days, but looking promising" you reply. Well, who wouldn't? If the aim is to find lots of significant differences, a strong first result is surely a good outcome.

At this point you have reason to expect that many of the subsequent tests also output strong results.

Now consider two alternative scenarios for the remaining 999,999 tests. In the first scenario, the remaining tests produce Z values that fit very well with the null, closely following a standard normal distribution; in the second scenario a large proportion of the remaining tests, say 50 percent, show outcomes that lie outside of $[-4, 4]$.

If your friend enquired after your analysis again, your response would surely differ in the first scenario ("Oh, it didn't pan out so well after all") vs the second ("It went great"). Further, in the first scenario, if your friend pressed you further about the results of the first test, you would likely, I think, be inclined to put them down to chance. In contrast, in the second scenario, the first test turned out to be, as you hoped, a harbinger of good things to come, and in this scenario you would likely regard that test as likely corresponding to a true discovery.

The key point is that it is the *outcomes* of the tests, not the *number* of tests, that impacts interpretation of that first test.

(Some may by pondering whether the fact that you are about to perform another 999,999 tests should be considered pertinent in responding to your friend. Our view is that it is irrelevant. The standard frequentist framework would disagree, because it requires the analyst to consider hypothetical repetitions of the "experiment", and so the fact that the experiment consists of a million tests is pertinent. However, this argument is a distraction from the main point.)

Indeed, we believe that the practice of focussing on the *number* of tests performed is

Focussing on the number of tests performed can be seen as an approximation.

The standard argument is that, when performing multiple tests, some will be significant just by chance.

Acknowledgments

Statistical analyses were conducted in the R programming language [**?**], Figures produced using the ggplot2 package [**?**], and text prepared using LaTeX.

Figure Legends

Tables

Supporting Information Legends

Supplementary material can be found in **Supplementary Information S1.**