BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 55, Number 1, January 2018, Pages 31–55
http://dx.doi.org/10.1090/bull/1597
Article electronically published on October 4, 2017

STATISTICAL PROOF? THE PROBLEM OF IRREPRODUCIBILITY

SUSAN HOLMES

Abstract. Data currently generated in the fields of ecology, medicine, climatology, and neuroscience often contain tens of thousands of measured variables. If special care is not taken, the complexity associated with statistical analysis of such data can lead to publication of results that prove to be irreproducible.

The field of modern statistics has had to revisit the classical hypothesis testing paradigm to accommodate modern high-throughput settings. A first step is correction for multiplicity in the number of possible variables selected as significant using multiple hypotheses correction to ensure false discovery rate (FDR) control (Benjamini, Hochberg, 1995). FDR adjustments do not solve the problem of double dipping the data, and recent work develops a field known as post-selection inference that enables inference when the same data is used both to choose and to evaluate models.

It remains that the complexity of software and flexibility of choices in tuning parameters can bias the output toward inflation of significant results; neuroscientists recently revisited the problem and found that many fMRI studies have resulted in false positives.

Unfortunately, all formal correction methods are tailored for specific settings and do not take into account the flexibility available to today's statisticians. A constructive way forward is to be transparent about the analyses performed, separate the exploratory and confirmatory phases of the analyses, and provide open access code; this will result in both enhanced reproducibility and replicability.

1. Introduction

Statistics is the main inferential tool used in science and medicine. It provides a process for drawing conclusions and making decisions in spite of the uncertainty inherent in any data, which are necessarily collected under constraints of cost, time, and measurement precision.

Mathematics is based on logical deduction, using arguments of the type:

$$(A \Rightarrow B) \iff (\neg B \Rightarrow \neg A)$$

then the observation of $(\neg B)$ makes A impossible.

Statisticians are willing to pay "some chance of error to extract knowledge" (J. W. Tukey [87]) using induction as follows.

> If, given $(A \Rightarrow B)$, then the existence of a small ϵ such that $P(B) < \epsilon$ tells us that A is **probably** not true.

Received by the editors June 15, 2017.
2010 *Mathematics Subject Classification.* Primary 62F03, 62F15.
This work was supported by a Stanford Gabilan fellowship.

This translates into an inference which suggests that if we observe data X, which is very unlikely if A is true (written $P(X|A) < \epsilon$), then A is not plausible.[1]

This is usually rephrased in terms of a null hypothesis $A = H_0$ and a cutoff level for a probability to be considered small, which is called the *significance level* α. Assume that if H_0 is true, $P(X) < \alpha$. Under this assumption, if we observe X, we can **infer** that H_0 can be **rejected** as improbable.

This is the framework of classical statistical hypothesis testing that originated with Fisher, Neyman, and Pearson [55]. Their theory is founded on the premise that starting with one hypothesis H_0 (e.g., "A is true"), the scientist designs an appropriate experiment, and collects data X.

1.1. **A cutoff for significance.** One computes the probability that the data observed could occur under that hypothesis; this is called the p-value.

If the probability $P(X|H_0)$ is small (say smaller than $\epsilon = \alpha = 0.05$),[2] we reject the plausibility of H_0; we call this a **discovery**.

In what is considered one of statistic's breakthrough papers, entitled "The arrangement of field experiments" [31], Fisher presented the argument in favor 0.05 as the result of considering 20 years of data:

> It will illustrate the meaning of tests of significance if we consider for how many years the produce (i.e., results) should have been recorded in order to make the evidence convincing. First, if the experimenter could say that in twenty years experience with uniform treatment the difference in favour of the acre treated with manure had never before touched 10 per cent, the evidence would have reached a point which may be called the verge of significance. This level, which we may call the 5 per cent. point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials. ... If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent. point), or one in a hundred (the 1 per cent. point).

However, modern experiments produce high-throughput data of tens of thousands of measurements in less than a week. For example, modern neuroscience involves inspecting images containing hundreds of thousands of voxels, any of which could be associated with a phenomenon of interest. These lead to multiple hypotheses and variables on which statisticians have to make exponentially many preprocessing choices [19]. This allows the possibility for scientists to choose their outcomes after looking at the data. The use and abuse of statistics in these high-throughput settings has led to an outpouring of negative press about the scientific method, and p-values in particular [56].

I will argue that new statistical, mathematical, and computational tools, and scientific transparency facilitate valid inference in this modern paradigm, without overburdening researchers.

This review starts with a few examples that will help focus our attention on the problem of multiplicity and the current so-called "crisis in reproducibility". I will then explain what the essence of statistical testing is, defining p-values and null

[1] We do not say here that the probability of A is low; as we will see in a standard frequentist setting, either A is true or not and fixed events do not have probabilities. In the Bayesian setting we would be able to state a probability for A.

[2] Statisticians use a prespecified α as their significance level or false positive rate.

hypothesis through a simple example. I will explain how some of the methods for controlling false discovery rates work and will provide a few examples of applying these new ideas to the problem of multiplicity in the context of variable selection. These new approaches attempt to adjust for some of the flexibilities of modern statistical methods, and their justification requires some mathematical subtleties in asymptotic analyses and probability theory on which I will not dwell; I recommend the reader consult the original papers. The last section will give strategies that improve reproducibility, such as careful design and computational tools, and show that transparency can act as surrogate for proof.

2. Dangers of flexibility

2.1. **A simple case of flexibility.** Consider this example from immunology that involves looking at a specific site in a protein which is tested to see if it is an epitope (i.e., creates an allergic reaction) using an ELISA test.[3] The outcome will be 1 or 0, 1 indicating a reaction. Due to the nature of the test, sometimes a 1 occurs by chance (a false positive), and we are told that this happens with probability $\frac{1}{100}$ (the false positive rate). Let the null hypothesis H_0 be "the site does not create a reaction", i.e., is not an epitope. The data X are collected from 50 different people, four of whom have positive ELISA tests at that position. If the null hypothesis is true, the chance of observing four or more 1's is well approximated by a Poisson distribution with parameter $\frac{1}{2}$ (Law of small numbers), where the Poisson distribution with parameters λ is the probability measure on $\{1, 2, 3, \ldots\}$ which has $P_\lambda(j) = e^{-\lambda}\lambda^j/j!$. Thus in this data

$$P\{4 \text{ or more positive test results}|H_0\} = \sum_{j=4}^{\infty} e^{-1/2}(1/2)^j/j! = 0.00175.$$

This seems an unlikely event and standard practice will reject the null hypothesis H_0 with a p-value equal to 0.00175. The above describes the way things are supposed to be done. In reality, the immunologist who gave us the data might well have tried several potential positions and reported the site with the largest numbers of 1's.

Now, if we find out that the immunologists had actually started by exploring 60 potential positions whose cumulative scores for the 50 patients are represented in Figure 1, then the computations and conclusion change. Does this flexibility alter the evidence in favor of a possible epitope?

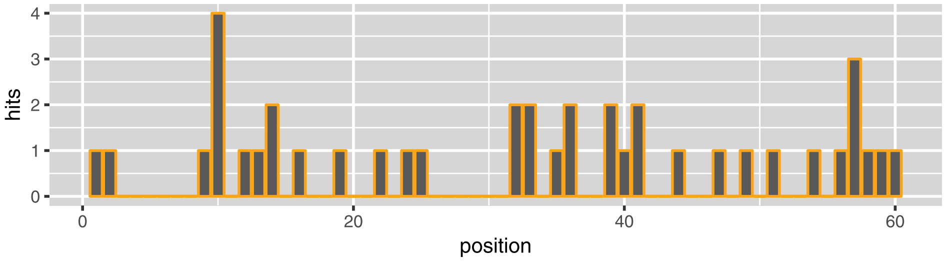

Figure 1. Output of an ELISA array for 50 patients in 60 positions

[3] Enzyme-Linked Immunosorbent asSAy, see `https://en.wikipedia.org/wiki/ELISA`

The computation of the p-value would then have to factor in that the position was chosen as the best out of 60. We order the data values $x_1, x_2, \ldots, x_{60}$, renaming them with indices that indicate their order, $x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(60)}$, so that $x_{(1)}$ denotes the smallest and $x_{(60)}$ the largest of the counts over the 60 positions.

The maximum value being as large as 4 is the complementary event of having all 60 independent counts be smaller or equal to 3 with the Poisson of rate $\lambda = 0.5$.

$$\begin{aligned} P(x_{(60)} \geq 4) &= 1 - P(x_{(60)} \leq 3) = 1 - \prod_{i=1}^{60} P(x_i \leq 3) \\ &= 1 - \left(\sum_{k=0}^{3} \frac{e^{-\lambda} \lambda^k}{k!} \right)^{60} = 1 - 0.900 = 0.10 \end{aligned}$$

The p-value is no longer small enough to justify rejecting the null hypothesis of the positions not being an epitope. We usually take the cutoff for significance or rejection of H_0 to be $\alpha = 0.05$ or 0.01. We see that flexibility in the choice of the position changed the conclusion of the test.

Following this simple computation comes an interesting mathematical aside. An elegant part of probability deals analytically with the maximum of independent random variables. If the original n measurements followed a normal distribution and $\max_n = X_{(n)}$ is their maximum, the theory says

$$P\{X_{(n)} \leq \sqrt{2 \log n - \log \log n + 2x}\} \sim e^{\frac{-e^{-x}}{\sqrt{2\pi}}}, \qquad \text{for any fixed } x, -\infty < x < \infty.$$

In practice, we would find x so the right-hand side is 0.99 and use that x on the left-hand side to fix the confidence bound. The right-hand side is called the Gumbel or extreme value distribution. It has a universal quality; many different underlying distributions for X_i have the same limit.

For the maximum of independent Poisson (λ) random variables, it would be nice if we could find a_n and b_n so that when n is large,

$$P\{\frac{X_{(n)} - a_n}{b_n} \leq x\} \sim e^{-e^{-x}}. \tag{2.1}$$

Surprisingly here, the discreteness of the Poisson outcomes really matters. It can be proved that there is no possible such a_n, b_n for Poisson or even rounded normal random quantities to achieve a relationship of the form in equation (2.1). Number theory issues have to be dealt with, however there are still useful approximations available; see [15], [21]. Of course, today it is straightforward to use Monte Carlo to approximate distributions, although tail probabilities and the calculation of probabilities of rare events still pose challenging problems.

Unfortunately, although much progress has been made using the type of extreme value theory shown above, it is not sufficient to deal with all the flexibilities that we encounter.

2.2. **Are most research findings false?** The last few years have seen the questioning of biomedical research as it has been practiced for over 40 years. Medical research is usually published if the p-values show an improvement or the significance of a result is smaller than $\alpha = 0.05$ or sometimes $\alpha = 0.01$. However, many authors have questioned the validity and reproducibility of results that ignore the file drawer effect [70]: a natural bias that occurs when experiments which do not show statistically significant differences cannot be published. John Ioannidis [44] went so far as to challenge that "most published research findings are false"—his work has had strong repercussions in the biomedical world where the pressure to publish has economic underpinnings. His paper focuses on how application of the hypothesis testing framework could result in a preponderance of false positive results in the medical literature. The key argument is analogous to population based disease screening (think TB test). A very sensitive and specific diagnostic may still have a low positive predictive value if the disease is not prevalent in the population. His premise is that significant hypotheses are rare, and he takes a prevalence of possible significant results at 1%, meaning that the null hypothesis is true for 99% of the hypotheses being considered, and 1% of the tested hypotheses are indeed false and should be rejected. Statistical power measures how likely a test is to reject a test when it should, i.e., the true positive rate.[4] If scientists test 1,000 hypotheses with a statistical power of 80%, then there will be $1,000 \times 1\% \times 80\% = 8$ true alternative hypotheses correctly detected. Even though the false positive rate is much lower, the prevalence of null hypotheses is much higher. With a false positive error rate of $\alpha = 0.05$, we expect $1000 \times 0.99 \times 0.05 = 49.5$ null hypotheses will be incorrectly rejected (we will call these false discoveries). So 49.5/(57.5)= 86% of rejected hypotheses will actually be null. Assuming selective reporting of positive results, a lower prevalence of true alternatives or other sources of bias in the publication process could make this estimate even worse.

These numbers were challenged by a subsequent analysis based on collecting all the p-values in more than 5,000 papers in medical journals [46]. Their model for p-values uses a mixture model and leads them to a less dramatic false discovery rate closer to 20%; however, this actually depends on other parameters, such as the power of the studies. Many studies today suffer from being underpowered, and there is a lack of rigor in experimental design; this is a real problem [18] that we will not delve into here. There are some things that can be done to adjust for some of the flexibility, and we will give details in the next section. But first a little reminder that statistics is a field with its own subtleties where even some of the most brilliant mathematicians can go astray.

2.3. Orion and the star constellations.

David Mumford [62] suggested in a paper entitled "Intelligent design found in the sky with $p < 0.001$" that the positioning of the stars in the Orion constellation were so particular that only intelligent design could explain them. However, in his calculations he starts by asking what the probability is that the stars of Orion are so aligned. This argument suffers from the blade-of-grass paradox, as put by Diaconis (*New York Times*, 1990):

> If you were to stand in a field and reach down to touch a blade of grass, there are millions of grass blades that you might touch. But you will, in fact, touch one of them. The a priori fact that the blade you touch will be any particular one has an extremely tiny probability, but such an occurrence must take place if you are going to touch a blade of grass.[5]

[4] The statistical power of a test is often written $1-\beta = 1-P$ (do not reject $|H_0$ is false), where β denotes the false negative rate.

[5] A similar argument by Einstein was related by Wigner in his autobiography: "Life is finite. Time is infinite. The probability that I am alive today is zero. In spite of this, I am now alive. Now how is that?" None of his students had an answer. After a pause, Einstein said, "Well, after the fact, one should not ask for probabilities."

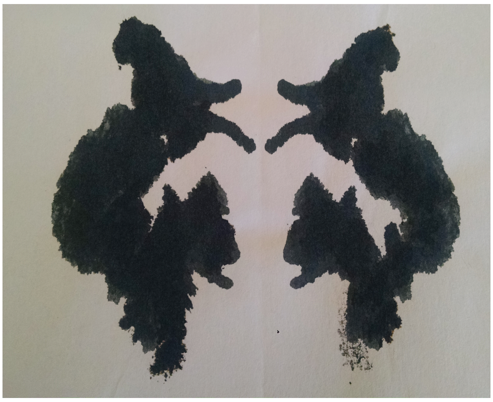

Figure 2. Modern nonparametric visualization and computational tools enable flexibility and many choices of possible endpoints with multiple opportunities for Pareidolia: these are at the core of the problems facing statistical inference today.

Why not ask what the chances of an equilateral triangle or other formation are? There is a lot of flexibility in that choice. Approaches where the endpoint is not made specific before looking at the data can lead to magical thinking [22] and are especially dangerous when data visualizations are involved, as Pareidolia[6] becomes a real problem (see Figure 2 for an illustration of Pareidolia).

2.4. **The Bible codes.** Witztum, Rips, and Rosenberg [92] used a statistical test to show surprising proximities between equidistant letter sequences (ELS) of names of famous rabbis and their known dates of birth and death. This work was endorsed by David Kazhdan, Joseph Bernstein, Hillel Furstenberg, and Ilya Piatestski-Shapiro, who wrote in the forward to the book:

> The present work represents serious research carried out by serious investigators.

McKay, Bar-Natan, Bar-Hillel, and Kalai [60] showed the existence of a large flexibility in the ELS experiment due to the different possible spellings of names and dates in Hebrew. The authors provide plenty of evidence for biased data-selection in this case. One thing to keep in mind is that this flexibility does not result in the addition of small effects but, as in the maximum problem in the previous section, multiplicative effects, and multiplying many numbers smaller than 1 can result in very small probabilities indeed.

3. Adjustments for multiple hypotheses

Formalizing the problem of multiple testing requires a bit more terminology. We have encountered the false positive rate α, sometimes called type I error and defined as the probability of rejecting H_0 even though it is true.

The type II error or false negative rate is the probability of not rejecting a hypothesis that is in fact false, and it is denoted β. Ideally, we want tests with α small (0.05 or less) and β large (larger than 0.8). High-throughput data give us the

[6] See `https://en.wikipedia.org/wiki/Pareidolia` for more details.

opportunity to consider m different potential hypotheses $M_1, M_2, \ldots, M_m$; this is called the simultaneous inference problem.

3.1. **Signal detection.** The question of whether there is a signal in m measurements or if at least one of the m hypotheses can be rejected was addressed in 1953 by John Tukey[7] in an unpublished report, now available in his collected works [88]. He viewed his procedure as a second level of significance testing and used an index now called *higher criticism* [24].

The method is aimed at the case when the M_i are independent and we want to test the joint null (all M_i's are true). Then the coefficient

$$HC_m = \sqrt{m}\frac{(\text{Fraction Significant at level } \alpha) - \alpha}{\sqrt{\alpha \times (1-\alpha))}}.$$

enables us to decide if there is a signal in the data. Tukey suggested 2 as a good threshold by analogy with a normal distribution's cutoff at two standard deviations. For instance, if we set α to be 0.05 and see 25 out $m = 400$ hypotheses significant, then the HC would be 2.29, enough to conclude that there was a signal in the data. The higher criticism article ([24]) generalizes Ingster's normal mixture model [43] which formalizes the intersection null hypothesis as

$$H_0: \qquad X_i \sim \phi_0, 1 \le i \le m,$$

to be tested against an alternative hypothesis formed as a mixture

$$H_A: \qquad X_i \sim (1-\epsilon)\phi_0 + \epsilon\phi_\mu,$$

where ϕ_0 is the density of the standard normal and ϕ_μ the density of the $N(\mu, 1)$. In fact, writing $\epsilon = m^{-\beta}$ for $\frac{1}{2} < \beta < 1$ allows the control of how small the nonstandard part of the mixture is. Detection will also depend on the size of μ. Suppose we take $\mu = \sqrt{2r\log(m)}$ for some $r, 0 < r < 1$. Ingster [43] shows asymptotically, when m is large, the detection boundary can be written

$$\begin{array}{ll} r > \rho^*(\beta) & H_0 \text{ and } H_A \text{ are distinguishable,} \\ & \qquad\qquad \text{where } \rho^*(\beta) = \left\{ \begin{array}{ll} \beta - \frac{1}{2}, & \frac{1}{2} < \beta \le \frac{3}{4}, \\ (1-\sqrt{1-\beta})^2, & \frac{3}{4} < \beta < 1. \end{array} \right. \\ r < \rho^*(\beta) & H_0 \text{ and } H_A \text{ are indistinguishable,} \end{array}$$

This detection theory was developed in [24] to generalize the mixture detection method and generalize to the case of rare and weak signals, which do not fall under the restricted class considered by [43]. Donoho and Jin [24] show that by using the criteria

$$HC_m^* = \max_{0<\alpha_0\le\alpha} \sqrt{m}\frac{(\text{Fraction Significant at level } \alpha) - \alpha}{\sqrt{\alpha \times (1-\alpha))}},$$

one can resolve the problem of testing whether there are a few (r out of m) variables with nonzero means or if they all have in fact mean zero.

Through their specific example, Donoho and Jin make clear that there are domains where it is possible to detect a signal but not identify which of the hypotheses to reject. In fact for certain cases, it is possible to map out a phase diagram [24] showing how detectability and localization depend on the strength of the signal (how big the nonzero means are), how many nonzero means there are, and how many tests are carried out in total.

[7] Tukey trained as a topologist: some mathematicians understand statistics very well.

The last 10 years have seen extensions of this approach to addressing modern problems where many more variables than observations (large p, small n) are studied. These extend the standard analysis of variance type framework [2] or time series where the variables/hypotheses are dependent [35, 36].

These articles all show that the more difficult and interesting question in practice is to find which of the M_i to reject; this is called the estimation or localization problem. Here we review some of the foundational work on that problem. Figures 3 and 4 represent the histograms of p-values for $20,000$ genes.

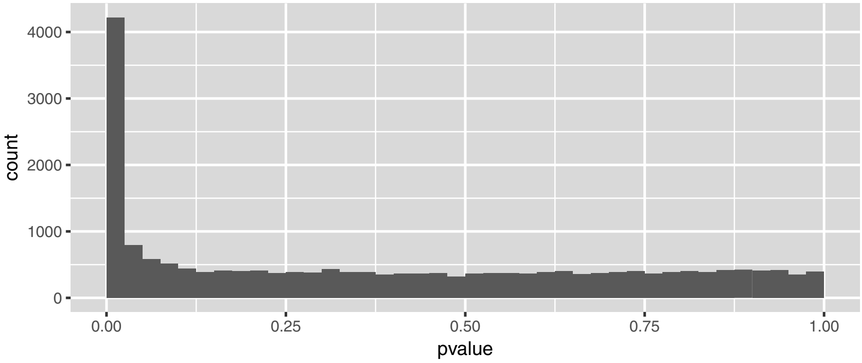

FIGURE 3. A histogram of p-values from a gene expression experiment involving more than 20,000 genes measured through counting RNA-sequencing reads [40].

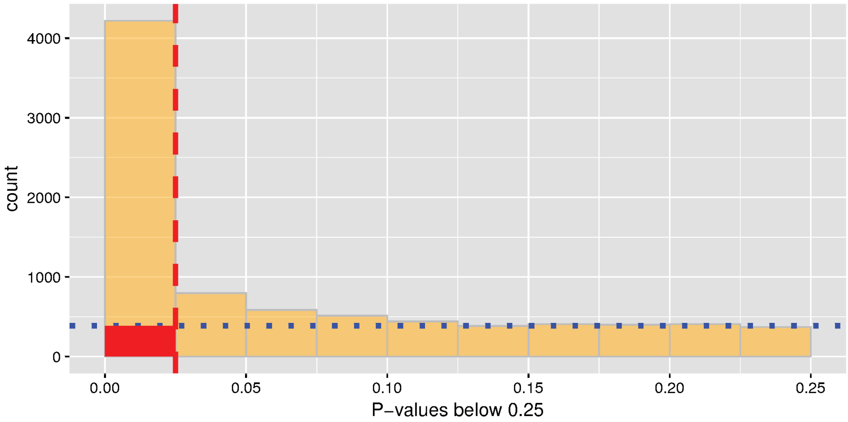

FIGURE 4. The red area in the lower left side represents a possible estimate of the false discovery rate from one instance of the p-value histogram (only the left side is shown to maximize resolution) with the false positive rate set at $\alpha = 0.025$. The dotted line shows the estimated p-value density if the null hypothesis is true. This is a close-up magnification of the data shown in Figure 3 (see [40] for a complete description).

3.2. **Correction of p-values through control of the FDR.** We start with a simple but important mathematical fact: under the null hypotheses the probability distribution of the p-values (called the null distribution) associated with a valid test statistic is uniformly distributed between 0 and 1 (represented by the horizontal dotted line in Figure 4) . Suppose that in fact m_0 of the hypotheses are truly null. For some of these, we may mistakenly reject null hypotheses. Suppose we actually reject the hypothesis R times and call these the discoveries. V among them are actually not true discoveries but correspond to random uniform p-values smaller than the cutoff α we specified (α, the false positive rate is often fixed at 0.05 or 0.01). Replacing the cutoff α by α/m provides what is known as the Bonferroni correction for multiple testing. This correction controls the family-wise error rate (FWER), defined to be the probability that V is nonzero. In contrast, define the false discovery rate (FDR) as the expected value of the false discovery proportion (FDP) V/R.[8] Benjamini and Hochberg [8] remarked that it is "desirable to control the expected proportion of errors among the rejected hypotheses", and they defined a procedure similar to Simes [73] that controls the FDR. It is important to notice that the FDR is an expected value and is not associated to just one hypotheses but to the ensemble of m hypotheses being tested in the simultaneous inference. Here is a step-by-step description of the Benjamini and Hochberg (BH) method for controlling the FDR:

- Fix a level γ under which we want to bound the FDR.
- Sort the observed p-values, $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ and compare $p_{(k)}$ to $k \cdot \gamma/m$.
- Let $r = \max\{p_{(k)} \leq k \cdot \gamma/m\}$ and reject the first r hypotheses $M_{(1)}, \ldots, M_{(r)}$.

This is illustrated in Figure 5 which shows the ranked p-values for the lowest 6,000 out of 20,000 genes of an RNA-seq experiment (see details in [40]) as they compare to the line $k \cdot \gamma/m$ with γ taken to be 0.10.

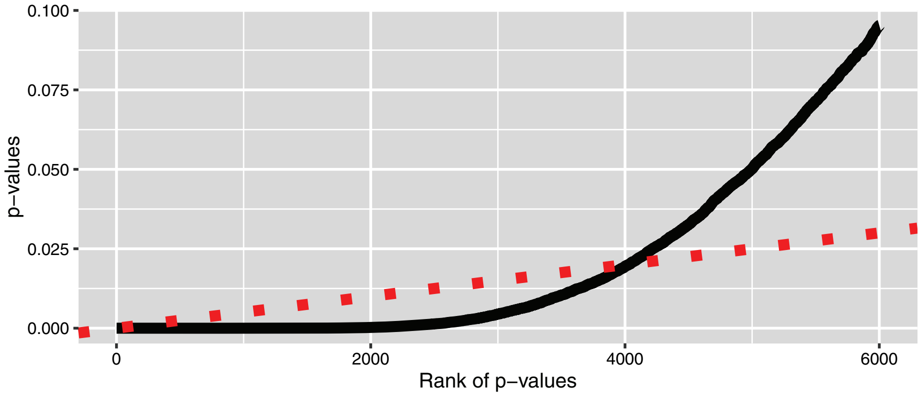

FIGURE 5. The BH method applied to the RNA-seq data comparing airway smooth muscle cell lines as described in [40] and whose p-values are plotted in Figure 3 suggests taking the first 4,100 p-values as significant.

[8] In all these definitions, we will make the convention that $\frac{0}{0} = 0$.

This procedure was designed to work for independent tests. Later theoretical work shows that it works even if the test statistics are dependent. Under independence of the hypotheses, this procedure leads to FDR $= m_0 \cdot \gamma/m$. Under positive dependence FDR $= m_0 \cdot \gamma/m$ (see [10]) and general dependencies FDR $\leq (1 + \frac{1}{2} + \ldots + \frac{1}{m}) \cdot m_0 \cdot \gamma/m$.

3.3. **Mixture model for testing.** The advent and importance of adjustments for multiple hypotheses testing is important for gene expression studies, such as microarrays or RNA-seq, where tens of thousands of genes are tested in two different conditions such as normal and cancer cells. Scientists were eager to conclude that they were able to find significant differences in gene expression. Unfortunately, many of these discoveries were later shown to be impossible to replicate [45].

One of the approaches statisticians [29, 63, 77, 78] developed to address the multiple testing correction problem in this context was to introduce a mixture model. Careful inspection of the histogram of p-values in Figure 3 suggests that a mixture model is appropriate. It seems that a proportion of the p-values are clustered at zero, whereas a remaining proportion (π_0) are uniformly spread throughout the $[0, 1]$ interval.

This can be described as a hierarchical model; there is a hidden layer above the observed data, and a latent variable is used to distinguish the truly differentially expressed genes from the others.

Write as a vector the statistics measured in the m experiments

$$T = (T_1, T_2, \ldots, T_m)$$

and define an indicator M_k for the kth hypothesis $H_0^{(k)}$ such that $M_k = 0$ if $H_0^{(k)}$, the null hypothesis is actually true. In the case when $H_0^{(k)}$ is true T_k will come from a density f_0, we write $T_k \sim f_0$; otherwise, $M_k = 1$ and $T_k \sim f_1$. We assume further the "hierarchical" parameter π_0, such that M_k are Bernoulli random variables with probability $1 - \pi_0$. We can write that the distribution of T_k's:

$$T_k \text{ has the density } f(t) = \pi_0 f_0(t) + (1 - \pi_0) f_1(t).$$

If we consider the T_k's to be the p-values, $f_0(p) = 1$, and we have

$$f(p) = \pi_0 + (1 - \pi_0) f_1(p) \text{ (see Storey [77] or Efron [26] for details).}$$

What we really want to know is the probability that a hypothesis is true:

$$P(M_k | T_k = t). \tag{3.1}$$

It is important to note that in the traditional frequentist context model, parameters and hypotheses are not random, and the expression in (3.1) would be meaningless. However a Bayesian can use the mixture model and the Bayes theorem to compute this probability:

$$P(M_k = 0 | T_k = t) = \frac{\pi_0 f_0(t)}{f(t)} = \frac{\pi_0 f_0(t)}{\pi_0 f_0(t) + (1 - \pi_0) f_1(t)}.$$

Now controlling the positive FDR (pFDR), defined as

$$E\left(\frac{V}{R} | R > 0\right) = \mathrm{P}(H_k = 0 | R_k = 1) = \frac{EV}{ER} \text{ (also denoted Fdr by Efron [29]),}$$

becomes a Bayesian classification problem. This is explained by Storey [77] showing that in fact we are trying to decide whether a hypothesis belongs to a the null group or not.

For $R > 0, V$ is distributed as a binomial on R trials with probability Fdr. If we define $\{\Gamma_\alpha\}_{0\leq\alpha\leq 1}$ to be a nested set of significant regions such that $\Gamma_{\alpha'} \subset \Gamma_\alpha$ for $\alpha' < \alpha$, then Storey [77, Theorem 1] shows that the positive FDR is

$$\mathrm{pFDR}(\Gamma_\alpha) = \Pr(M_k = 0|T_k \in \Gamma_\alpha) = \int_{\Gamma_\alpha} \Pr(M_k = 0|T_k = t) f(t|t \in \Gamma_\alpha) dt.$$

The significance regions $\Gamma_\alpha = \left\{ t : \frac{\pi_0 f_0(t)}{f(t)} \leq \alpha \right\}$ minimize the Bayes error of classifying $\{M_k\}$ using the standard $R_k(T_k, \delta(q))$ classification procedure for which $\mathrm{Fdr} = P(M_k = 0|R_k = 1) = q$. We see here that many of the difficulties in framing the multiple hypotheses testing problem vanish by using the Bayesian formulation, and we have access to the exact quantity of interest pFDR as a misclassification rate.

Of all rules $R_k(T_k, \delta)$ with $\mathrm{Fdr} = q$, the Bayes classifier has maximum power to make discoveries $P(R_k = 1|M_k)$ and minimum false negative probability: $P(M_k = 1|R_k = 0)$; see [26] for a book-long treatment.

Practical Note: The procedures for controlling the pFDR(Fdr) are now in widespread use, and several implementations in the open source statistical software program R make them accessible, among them:

- `locfdr` package estimates π_0, f_0, f_1 and computes the Fdr of the rejection rule $R_k = \mathbb{I}_{z\leq Z}$.
- `qvalue` package in R computes the pFDR of the rejection rule: $R_k = \mathbb{I}_{p_k \leq p}$.

$$P(H_k = 0|p_k \leq p) = \frac{H_k = 0, p_k \leq p}{P(p_k \leq p)} = \frac{P(p_k \leq p|H_k = 0)\pi_0}{P(p_k \leq p)},$$

$$\widehat{q\text{value}} = \frac{p \times \hat{\pi}_0}{\#\{k : P_k \leq p\}/m}, \qquad \hat{\pi}_0 = 2 \times \#\{k : 0.5 \leq P_k\}$$

Note that this is how the blue line in Figure 4 was drawn.

3.4. **There is more to the data than p-values.** In fact, it is the statisticians' hubris that is to blame: boiling down complex multidimensional problems to one number is definitely hubris. All the methods described in the above sections depended on the correction of the p-values taken on their own. Recent work has shown that improvements can be made by using **more** of the available data. Originally, one has the variance of the test statistics and even useful covariates that can inform the classification into true and false positives. Stephens [75] uses unimodality of the underlying effect and estimates of the standard errors as well as the effect sizes; using more of the information improves the power of his procedure. The estimation of the unimodal distribution only involves solving a convex optimization problem, and it enables more accurate inferences provided that the assumptions hold. His method is available in the R package `ashr`. This method uses enriched information and facilitates the estimation of actual effect sizes, providing credible regions for each effect in addition to measures of significance.

Ignatiadis et al. [42] propose a weighting scheme that also enriches the p-value information using covariates independent of the p-values under the null hypothesis. This method also increases the statistical power $(1-\beta)$ to discover associations and is implemented in the `IHW` Bioconductor package.

4. Model selection

There is also need to provide inference (p-values or confidence statements) in a slightly different framework than the simultaneous inference and multiple testing explored above. Modern statistical applications, such as predicting a response variable with nonparametric regression, require marginal inferences on multiple parameters selected after a preliminary exploration of the data.

Consider a toy example: data such as $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is collected. A preliminary plot shows the y_i is approximately a quadratic function of x_i. Now, a statistical test of the quadratic coefficient is carried out. How should the preliminary plotting be accounted for?

This is called post-selective inference, a subject that has seen healthy development in the last five years. Post hoc analyses were studied by Tukey and Scheffe in the 1950s to deal with inference in Analysis of Variance (ANOVA) where one significant contrast was chosen from the data; a complete review of methods before 1995 can be found in [72]. However, the era of high-dimensional data is upon us and *double dipping* [50] the same data set at the selection and evaluation steps invalidates standard analytics: the statistics are not independent of the selection criteria under the null hypothesis. Several groups have been investigating rigorous approaches to tackle this issue.

4.1. **Post-selective inference after model selection for standard linear models.** Classification and regression were traditionally set in a fixed model context. A **response** y is the variable statisticians attempt to predict from a given set of variables. The model can be written $y = f(x_1, x_2, \ldots, x_p) + \epsilon$, where ϵ is the noise, often supposed to have expectation 0 and be modeled as an independent identically distributed random variable. However, the growing size of p and in particular the nonparametric nature of the cases where $p > n$ have introduced new procedures (for a full review see for instance [38]) where a subset of the explanatory variables is chosen to ensure a sparser, regularized model. Thus for instance forward stepwise regression is a sequential procedure that steps through the possible predictors and aims to provide a parsimonious set of variables. LASSO regression imposes a penalty on nonzero coefficients in a regression and ensures that the model has a lower variance and less overfitting than a standard least squares model would. Adaptive regression models are stochastic and, as pointed out by Berk and coauthors [11], pose serious challenges to the classical inferential paradigm:

> Posed so starkly, the problems with statistical inference after variable selection may well seem insurmountable. At a minimum, one would expect technical solutions to be possible only when a formal selection algorithm is (1) well-specified (1a) in advance and (1b) covering all eventualities, (2) strictly adhered to in the course of data analysis, and (3) not improved on by informal and post-hoc elements. It may, however, be unrealistic to expect this level of rigor in most data analysis contexts, with the exception of well-conducted clinical trials. The real challenge is therefore to devise statistical inference that is valid following any type of variable selection, be it formal, informal, post hoc, or a combination thereof.

The authors propose a valid Post-Selective Inference (PoSI) with family-wise error guarantees that account for all types of variable selection. This involves ignoring the variables whose coefficients in the model are estimated at zero. Suppose we want to use the equation merely to describe association between predictors and response variables. Here the predictors are associated to each other and to the response. The model $\hat{M}$ is random and depends on the observed response y. A relevant parameter in this case is r, the "effective degrees of freedom", which in the simplest case of a full model with equality in variances, is $n-p$, but can be many more if the model is chosen among submodels. The consequences of their theory of PoSI in the linear model case is that by making the standard confidence bound (the ones often used follow a $t_{1-\alpha/2,r}$ distribution), using a larger constant $K(X, M, \alpha, r)$ than that provided by the standard t, one can provide simultaneity protection for up to $p \cdot 2^{p-1}$ parameters $\beta_j \cdot M$, and K depends strongly on the predictor matrix X as the asymptotic bound for $K(X_{n\times p}, M, \alpha, r)$ with $d = \text{rank}(X)$ ranges between the minimum of $\sqrt{2\log d}$ achieved for orthogonal designs on the one hand, and $\sqrt{d}$ on the other hand. This wide asymptotic range suggests that computation is critical for problems with large numbers of predictors. In the classical case $d = p$, their current computational methods are feasible only up to about $p \simeq 20$.

In fact, the key in the above discussion is to declare that M is random and to present an honest way to account for randomness in all hypotheses: such a model is available through the Bayesian paradigm described in the next section.

4.2. **Bayesian selective inference.** In practical situations, statisticians start by finding interesting parameters—a random process—then they want to provide inferences for these selected parameters. In fact, the attentive reader will realize that there is a real challenge posed in providing a frequentist interpretation of post-selective model selection. The randomness of the model inhibits a long term frequency interpretation of statements such as those of the following [11, equation 4.6]

$$P\left\{\beta_{j,\hat{M}} \in CI_{j,\hat{M}}(K) \;\; \forall \hat{M}\right\}.$$

That, in the long run, the random confidence intervals CI contain the true $\beta_{j\hat{M}}$ for the model-selected parameters. This can be simply overcome by doing what Storey did for multiple testing and posing the problem in a Bayesian perspective. One can compute the probabilities conditioning on the coefficients being nonzero.

Yeketuli [93] developed such a framework for accounting for truncation of the data in the selection process. Here I present a very elegant example that captures a subtlety due to the conditional nature of the inferences and truncation that occurs.

4.2.1. *Example from* [93]. Suppose the goal is to predict a student's true academic ability from their observed/tested ability but only for the students admitted to college. The following toy model is used as our thought experiment:

- The students have a true (unknown) academic ability $\theta_i \sim \mathcal{N}(0,1)$; notice the Bayesian can take a true parameter to have a distribution.
- Their academic ability observed in high school is $Y_i \sim \mathcal{N}(\theta_i, 1)$.
- Only students with $Y_i > 0$ are admitted to college.

$f_S(\theta_i, y_i)$ is needed to compute the prediction error

$$\int_{\theta_i=-\infty}^{\infty} \int_{y_i=0}^{\infty} f_S(\theta_i, y_i) \cdot (\theta_i - \delta(y_i))^2 dy_i d\theta_i.$$

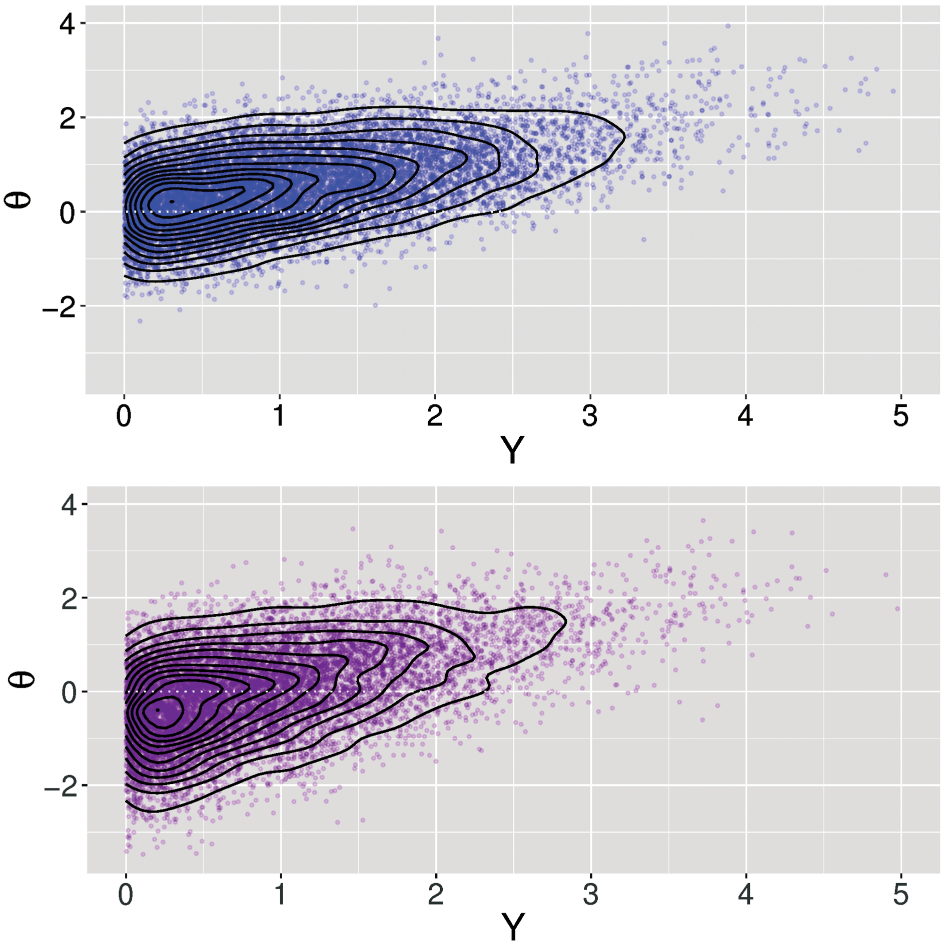

Figure 6. Simulation of high school students according to the two different sampling schemes: the college students on the top are drawn conditional on having $Y > 0$, and the high school students in lower plot have Y's that come from a truncated normal. I have used exactly the same parameters for the estimation of the density contours in each case. There are 12 bins and one can see that the lower plot has its center below the dotted white horizontal axis, whereas in the upper plot it is above.

For the college professor predicting θ for a student in their class, the joint distribution (θ, Y) of a random college student is found by by generating (θ, y) for a random high school student and selecting (θ, y) only if $0 < y$. Thus the joint distribution is

$$f_S(\theta, y) \propto \frac{\exp(-\frac{\theta^2}{2})\exp(-\frac{(\theta-y)^2}{2})}{P(Y > 0)} \propto \exp(-\frac{\frac{(\theta-y)^2}{2}}{2 \times \frac{1}{2}}).$$

Now we look at the high school counselor who is assigned to only counsel students whose (unknown) ability would allow them to attend college. The distribution seen comes from $\theta_i \sim \mathcal{N}(0,1)$, and the Y_i used to predict θ_i is drawn from $\mathcal{N}(\theta,1)$ and is truncated by $Y_i > 0$. The joint density becomes:

$$f_S(\theta, y) \propto \frac{\exp(-\frac{\theta^2}{2})\exp(-\frac{(\theta-y)^2}{2})}{P(Y>0|\theta)}.$$

See that the difference is subtle, but one can distinguish the two densities as shown in Figure 6; in the college population case one divides by $P(Y > 0)$ which will be half the density, whereas in the high school population the probability $P(Y > 0|\theta)$ cannot be computed in closed form but this decreases in θ.

This example can be generalized. Suppose θ is the parameter, Y is the data, and Ω is the data sample space. $\pi(\theta)$ is the prior distribution for the parameter of interest and $f(y|\theta)$ is the likelihood function. The multiple parameters, for which inference may or may not be provided, are actually multiple functions of $\theta : h_1(\theta), h_2(\theta)$. For each $h_i(\theta)$ there is a given subset $S^i_\Omega \subset \Omega$ such that inference is provided for $h_i(\theta)$ only if $y \in S^i_\Omega$ is observed.

Yeketuli's formulation enables a clear description of selection-adjusted Bayesian inference, with $\pi_S(\theta)$ a selection-adjusted prior distribution, and as the selection adjusted likelihood Yeketuli uses the truncated distribution of $Y|\theta$,

$$f_S(y|\theta) = \mathbb{I}_{S_\Omega} \cdot f(y|\theta)/P(Y \in S_\Omega|\theta).$$

Then the Bayes rules are based on the selection-adjusted posterior distribution

$$\pi_S(\theta|y) \propto \pi_S(\theta) \cdot f_S(y|\theta).$$

This distribution fully characterizes post-selection inference; again a clear Bayesian formulation enables a clarification of the correct procedures and their properties. The authors of [64] have recently also harnessed this approach in an application to variable selection. They rely on an approximation to the full truncated likelihood and can approximate the maximum-likelihood estimate as the Maximum A Posteriori (MAP) estimate corresponding to a constant prior.

4.3. **Post-selective inference using a geometric approach.** Jonathan Taylor, Robert Tibshirani, and collaborators [27, 32, 37, 51, 58, 81–85, 90] have approached the same problem of inference after variable selection through a geometric approach.

A very readable account appears in PNAS [81], where the authors take the example of deciding which mutations occurring in an HIV-strain DNA sequence are significant predictors of drug resistance [69].

Using a standard linear regression model, with response $Y \in \mathbb{R}^n$ and predictors X_j, $j = 1, \ldots, p$,

$$\hat{Y} = \mathbf{X}\beta, \qquad Y \sim N_n(\theta,\ \Sigma).$$

The standard model estimates the parameters by computing

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ ||Y - X\beta||_2^2. \tag{4.1}$$

To obtain a more interpretable and robust model, one wants to take a subset $M \subset \{1, \ldots, p\}$ of the predictors. Each subset M leads to a different model corresponding to the assumption $\hat{Y} = \mathbf{X}_M\beta^M$, where $\mathbf{X}_M$ denotes the matrix consisting of columns

X_j for $j \in M$. Then, it is customary to report tests of $H_{0,j}^M : \beta_j^M = 0$ for each coefficient in the model.

Assume that the selection has the effect of partitioning the sample space into polyhedral (or convex) sets. Jonathan Taylor and coauthors [51] prove the following important result.

Lemma 4.1. *Suppose that we observe $y \sim \mathcal{N}(\theta, 1)$, for any vector η, the condition that y is in the polyhedron $\{Ay \le b\}$ can be rewritten*

$$\{Ay \le b\} = \{\mathcal{V}^\ell \le \eta^T y \le \mathcal{V}^u(z), \mathcal{V}^0(z) \ge 0\}, \quad \text{where}$$

$$\begin{aligned} z &\doteq (\mathbb{I} - c\eta^T)y, \qquad c \doteq \Sigma\eta(\eta^T\Sigma\eta)^{-1}, \\ \mathcal{V}^\ell(z) &\doteq \max_{j:(Ac)_j<0} \frac{b_j - (Az)_j}{(Ac)_j}, \\ \mathcal{V}^u(z) &\doteq \min_{j:(Ac)_j>0} \frac{b_j - (Az)_j}{(Ac)_j}, \\ \mathcal{V}^0(z) &\doteq \min_{j:(Ac)_j=0} b_j - (Az)_j. \end{aligned}$$

This says that the distribution of $\eta^T y$ conditional on being in the polyhedron is

$$\mathcal{L}(\eta^T y|Ay \le b) = \mathcal{L}(\eta^T y|\{\mathcal{V}^\ell(z) \le \eta^T y \le \mathcal{V}^u(z), \mathcal{V}^0(z) \ge 0\}).$$

$\mathcal{V}^\ell(z), \mathcal{V}^u(z), \mathcal{V}^0(z)$ are independent of $\eta^T y$, thus the conditional distribution

$$\mathcal{L}(\eta^T y|\{A(m)y \le b(m), z = z_0\}) \text{ is a truncated normal.}$$

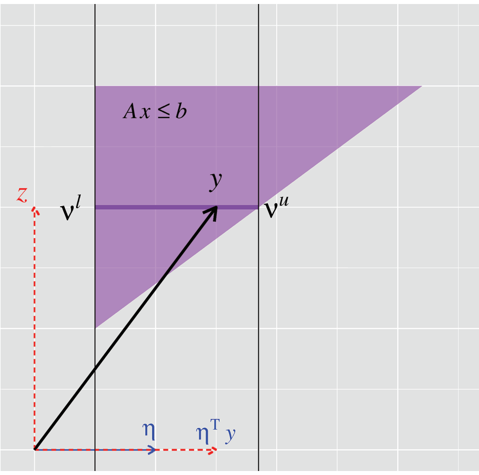

FIGURE 7. The polyhedral lemma, in the special case where $\Sigma = 1$. The bounds represented by the thin vertical lines define the interval of truncation $[V^\ell(z), V^u(z)]$ (in purple) to which the normal is constrained.

This means that by using the inverse cumulative distribution transform, one can find a statistic $F^z(\nu^T\theta)$ whose distribution is uniform which enables inferences on a linear contrast $\nu^T\theta$. A little more work is necessary—in fact several signs are possible, and the inferences conditional on the selected model (with the signs s fixed) being chosen as m with sign s have to be written $\{M(y) = m, s\}$. The overall inference is then made using a conditional distribution whose support is a series of disjoint intervals; for ample details see [51].

4.4. **Variable selection using simulated data: The knockoff method.** Another recent advance in variable selection is the idea of augmenting the data by adding to the existing explanatory variables. These new explanatory variables are constructed to have the same correlation between themselves as the original ones while being unrelated to the response variable Y. This method was proposed by Barber and Candès [6] and named the **knockoff** filter. For details and properties of the procedure, see their paper which contains two important ideas: data augmentation to assess a variable's influence on the response and the use of exchangeability under the null to estimate the false positive proportion (FPP).

The setting for this method is similar to the previous section where the goal is variable selection, for instance in the sparse regression LASSO [38] context described above. Given the observed response $\mathbf{y}$, we want to find a small number of explanatory variables from among the explanatory variables $\mathbf{X}_{n\times p}$, with $n > 2p$, and we want to find a vector β such that $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$. The LASSO provides a solution through

$$\hat{\beta}(\lambda) = \underset{\mathbf{b}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\}.$$

The new knock-off variables $\tilde{\mathbf{X}}$ are built using the cross product $\mathbf{\Sigma} = \mathbf{X}^t\mathbf{X}$ by constraining them to fulfill

$$\tilde{\mathbf{X}}^t\tilde{\mathbf{X}} = \mathbf{\Sigma}, \qquad \mathbf{X}^t\tilde{\mathbf{X}} = \mathbf{\Sigma} - \operatorname{diag}\{\mathbf{s}\},$$

where $\mathbf{s}$ is a p-dimensional nonnegative vector. In words, $\tilde{\mathbf{X}}$ exhibits the same covariance structure as the original $\mathbf{X}$. The correlations between distinct original and knock-off variables are constrained to be the same as those between the originals (because $\mathbf{\Sigma}$ and $\mathbf{\Sigma} - \operatorname{diag}\{\mathbf{s}\}$ are equal on off-diagonal entries),

$$\mathbf{X}_j^t\tilde{\mathbf{X}}_k = \mathbf{X}_j^t\mathbf{X}_k \qquad \text{for all } j \neq k.$$

However, comparing a feature $\mathbf{X}_j$ to its own knockoff $\tilde{\mathbf{X}}_j$, we have $\mathbf{X}_j^t\tilde{\mathbf{X}}_j = \Sigma_{jj} - s_j = 1 - s_j$, while $\mathbf{X}_j^t\mathbf{X}_j = \tilde{\mathbf{X}}_j^t\tilde{\mathbf{X}}_j = 1$. To maximize the true positive rate, and thus the power to detect signals, the authors choose the entries of $\mathbf{s}$ as large ensuring the variable $\mathbf{X}_j$ is not too similar to its knockoff $\tilde{\mathbf{X}}_j$. However, one also has to have $2\mathbf{\Sigma} - \operatorname{diag}\{\mathbf{s}\}$ be positive semidefinite ensuring that the matrix $\tilde{\mathbf{X}}$ can be constructed as

$$\tilde{\mathbf{X}} = \mathbf{X}\big(\mathbb{I} - \mathbf{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\}\big) + \tilde{\mathbf{U}}\mathbf{C}.$$

The next step is to calculate the point λ at which feature $\mathbf{X}_j$ first enters the model and to use that as the test statistic for variable j: $Z_j \doteq \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$, hoping it is large for signals and small for null variables. A clever strategy enables the authors to choose an appropriate threshold for variable selection, using the knockoff variables. They compute the statistics on the augmented $n \times 2p$ explanatory variable matrix $[\mathbf{X} \quad \tilde{\mathbf{X}}]$ (concatenating columnwise $\mathbf{X}$ and $\tilde{\mathbf{X}}$). This yields a $2p$-dimensional vector $(Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p)$. For $j \in \{1, \ldots, p\}$, they then define

$$W_j = \max(Z_j, \tilde{Z}_j) \cdot \begin{cases} +1, & Z_j > \tilde{Z}_j, \\ -1, & Z_j < \tilde{Z}_j. \end{cases}$$

A large positive value of W_j indicates that variable $\mathbf{X}_j$ enters the LASSO model early (at some large value of λ) and before its knock-off copy $\tilde{\mathbf{X}}_j$. This means that this variable belongs in the model. A threshold τ is then calculated using the knockoffs. Letting γ be the target FDR, define a data-dependent threshold τ as

$$\tau = \min\left\{ t \in \{\mathcal{W} : \frac{\#\{j : W_j \leq -t\}}{\max(\#\{\mathrm{j : W_j \geq t}\}, 1)} \leq \gamma \right\}$$

or $\tau = +\infty$ if this set is empty, where $\mathcal{W} = \{|W_j| : j = 1, \ldots, p\} \setminus \{0\}$ is the set of unique nonzero values attained by the $|W_j|$'s.

It is stated in [6, Lemma 1] that if we have $\varepsilon \in \{-1, +1\}^p$ a sign sequence independent of W, with $\varepsilon_j = +1$ for all nonnull j and ε_j identically and independently distributed with values in $\{-1, +1\}$ for null j, then $(W_1, \ldots, W_p)$ has the same joint distribution as $(W_1 \cdot \varepsilon_1, \ldots, W_p \cdot \varepsilon_p)$. This ensures that $\#\{j : \beta_j = 0, W_j \leq -t\}$ has the same distribution as $\#\{j : \beta_j = 0, W_j \geq t\}$ and validates their estimate of the true false discovery proportion FDP.

This is a promising tool which is generating a healthy development of follow-up methods that extend to more general nonparametric settings with $n < p$ [7, 48]. Arias-Castro and Chen [3] have recently shown that the FDP estimation proposed by [7] attains the selection boundary in the multiple testing phase diagram. The method for finding a threshold only depends on a symmetry argument and does not require the p-values to be truly uniformly distributed. They also show that this estimate of FDP does as well as the BH procedure in an asymptotic sense that they make precise and without the knowledge of the null distributions. This type of theoretical justification does involve some delicate mathematical work; e.g., generalizing the simple Gaussian mixture model to asymptotically generalized Gaussian tail cases. The authors [3] use an oracle procedure that exhibits the bounds any threshold method would have on minimizing the sum of the false discovery and nondiscovery proportions.

4.5. **Blinding in particle physics.**

Physicists have advocated a clever method to prevent unconscious bias in the choice of variables, tuning parameters, and data transformations called *blinding* [49]. As in the construction of double-blind clinical studies, in the presence of different groups of observations, one can hide the labels of the observations and even change the signs to certain variables. These changes are recorded and only revealed after the data analyses have been completed. The method is tuned to physics experiments, and an open research problem is to adapt it to other scenarios.

5. Best practices

5.1. **Reproducible data analyses; replicable scientific experimentation.**

Given the long list of defensive caveats, a quote and a message of hope from Fred Mosteller is in order:

> It is easy to lie with statistics, but a whole lot easier without them.

Statistics still has a huge role to play in validating scientific discoveries; however, a clear path needs to be traced between all the pitfalls.

Let's start by clarifying the vocabulary somewhat. In [54] the authors make a clear and useful difference between **reproducibility** and **replicability**; both are desirable in science but pertain to different steps in the scientific process. A piece of work involving statistical code and data is called reproducible if another scientist can take their code and data and reproduce the figures and statistics published in their article. A study is replicable if another scientist can redo the experiment in their laboratory according to the same experimental design, collecting different data, do a statistical analysis similar to the one done by the previous group, and come to the same conclusion.

For important medical decisions, it is crucial that data and the code be published and be openly accessible, as shown in the case discussed in [5] where the author uncovered confounding in the experimental design, confusion of the gene labels (off-by-one errors), and mixing of group labels (sensitive/resistant) in a series of studies of chemo-sensitivity of cancer patient cell lines to given chemotherapeutic agents.

Providing the complete computational workflow is a very important insurance policy against over-tuning the parameters and preprocessing so that a "clean" outcome and narrative can be published. As shown in [19], if one counts the number of possible analyses on the same data—allowing for the choice of up to nine outliers, different transformations of the data, choice from 40 different possible distances, and five different ordination methods—the result is more than 200 million possibilities. No multiple hypotheses correction can protect the user. The only feasible strategy is to allow the reader to run the code and make changes to check the robustness of the conclusions. Robustness is an important component to consider, and a large literature on robustness to distributional assumptions, the choice of statistical functional, and the existence of outliers exists. It seems that the lessons learned by these studies are not always understood by practitioners, and recent studies by neuroscientists [30] on the inflation of false positive rates concluded that the scientists were using parametric Gaussian assumptions without checking their validity. Similar dynamics drive Taleb's black swan argument [79].

It is certainly problematic that scientists often use black-box procedures they do not understand. This can be fixed by encouraging open access to all the code and data used in the analyses and providing educational resources. Once scientists have access to all the code and understand what their black boxes are doing, there will be a better understanding of the assumptions behind the procedures used and motivation to address sensitivity to unmet assumptions, for instance by using nonparametric robust permutation tests instead of parametric ones.

5.2. **A clear division between exploratory and confirmatory stages.** Tukey and Mallows [59] clearly lay out the modern paradigm recommending that separate data be analyzed for the confirmatory (testing) stages of data analysis. We all know we should not be double dipping the data. The ideal situation is to do the *detective work* and exploratory data analysis (EDA) on one set of data; once this data has told us which variables to use, which transformations to make, and which dimension reductions to implement, the critical and confirmatory analyses are done on a different set of observations. Setting up the confirmatory analyses with fixed algorithms and computations untouched by human hand (Donoho, personal communication) would be ideal and would guarantee the integrity of the conclusions.

However, applied statisticians, as opposed to mathematicians, do not live in an ideal world.

5.3. **Honest appraisal of current progress.** In the early 1980s, Diaconis [22] had carefully laid out suggestions for best practices for organizing exploratory analyses in order to avoid some of the pitfalls exposed above. It is certainly the case that transparency should be a priority. Authors should say clearly:

- What are the assumptions about the data used to find the null distribution?
- Was optimization over a certain set of possibilities used?
- Can an uncertainty measure be attached to an estimate?

Today, we know the biases created by the pressure to publish incites p-hacking [74]. We also know this can be avoided if authors are encouraged to publish all the variables and transformations used in their computations; thus disclosing more of the relevant information. The original experimental design should be registered, the exploratory analysis performed on one set of data, a clear algorithm decided upon, before the collection of a separate confirmatory set. The confirmatory analyses should be published or posted whether the results are positive (as hoped) or not.

Many authors have published papers "talking the talk": offering ideas on the careful prescriptions to follow, there are many papers published doing meta-analyses of all the publications and showing how much new approaches are needed. However, there are many more papers giving lists of prescriptions to heal the problem than there are papers available showing scientists who are actually "walking the walk". This is because the incentive structure has not been changed; publishing one's data and code is costly in terms of recognition, time, and money. From experience, I know that a complete data workflow paper with a confirmatory data set takes almost a year more to publish [19, 23]. This is not a problem where mathematics can help, but mathematicians have set an example and a publication record where transparency is key and theorems are proved: mathematicians *show their work* and so should statisticians.

Currently, we have no excuse. Open access is here, whole toolkits of open source statistical packages such as Bioconductor [41], Galaxy [34], IPython notebooks, [61] and other such systems are available, and publishing code, data, and the complete workflows should be standard practice.

The code to reproduce the analyses and figures in this article is available at `http://statweb.stanford.edu/~susan/papers/RR/`.

Multiplicity but not duplicity. Everything we have presented above addresses errors that are made in good faith. In fact if the data are dredged secretly, dropping points which do not support the narrative the scientists want to publish or biasing the sampling, none of the approaches presented above can work. Baggerly and Coombs [5] showed by doing their careful informatics and statistical forensics that certain studies in the world of cancer genetics were flawed with intent to dupe. No statistical procedure can counter data manipulations done in bad faith, all we can hope for is to achieve more transparency through careful accounting for, and recording of, all the choices made during an analysis.

These are cases where true scientific replication by independent groups and funded by science foundations and institutes should find their place in mainstream publication.

Acknowledgments

I am thankful to David Eisenbud for giving me the opportunity to review the recent statistical literature, and to Persi Diaconis, Kris Sankaran, Barry Mazur, David Donoho, and Wolfgang Huber for ongoing discussions on the many difficulties in improving the transparency of the scientific endeavor when statistics are involved.

About the author

Susan Holmes is professor of statistics at Stanford University. Trained in the French school of geometrical data analysis, she works on probabilistic and statistical methods for complex biological systems.

References

[1] Robert J. Adler and Jonathan E. Taylor, *Topological complexity of smooth random functions*, Lecture Notes in Mathematics, vol. 2019, Springer, Heidelberg, 2011. Lectures from the 39th Probability Summer School held in Saint-Flour, 2009; École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. MR2768175

[2] Ery Arias-Castro, Emmanuel J. Candès, and Yaniv Plan, *Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism*, Ann. Statist. **39** (2011), no. 5, 2533–2556, DOI 10.1214/11-AOS910. MR2906877

[3] Ery Arias-Castro and Shiyun Chen, *Distribution-free multiple testing*, Electron. J. Stat. **11** (2017), no. 1, 1983–2001, DOI 10.1214/17-EJS1277. MR3651021

[4] Ery Arias-Castro and Nicolas Verzelen, *Community detection in dense random networks*, Ann. Statist. **42** (2014), no. 3, 940–969, DOI 10.1214/14-AOS1208. MR3210992

[5] Keith A. Baggerly and Kevin R. Coombes, *Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology*, Ann. Appl. Stat. **3** (2009), no. 4, 1309–1334, DOI 10.1214/09-AOAS291. MR2752136

[6] Rina Foygel Barber and Emmanuel J. Candès, *Controlling the false discovery rate via knockoffs*, Ann. Statist. **43** (2015), no. 5, 2055–2085, DOI 10.1214/15-AOS1337. MR3375876

[7] Rina Foygel Barber and Emmanuel J Candès, *A knockoff filter for high-dimensional selective inference*, arXiv:1602.03574, 10 February 2016.

[8] Yoav Benjamini and Yosef Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, J. Roy. Statist. Soc. Ser. B **57** (1995), no. 1, 289–300. MR1325392

[9] Yoav Benjamini and Yosef Hochberg, *On the adaptive control of the false discovery rate in multiple testing with independent statistics*, J. Educ. Behav. Stat. **25** (2000), no. 1, 60–83.

[10] Yoav Benjamini and Daniel Yekutieli, *The control of the false discovery rate in multiple testing under dependency*, Ann. Statist. **29** (2001), no. 4, 1165–1188, DOI 10.1214/aos/1013699998. MR1869245

[11] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao, *Valid post-selection inference*, Ann. Statist. **41** (2013), no. 2, 802–837, DOI 10.1214/12-AOS1077. MR3099122

[12] Richard Berk, Lawrence Brown, and Linda Zhao, *Statistical inference after model selection*, J. Quant. Criminol. **26** (2009), no. 2, 217–236.

[13] Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès, *SLOPE—adaptive variable selection via convex optimization*, Ann. Appl. Stat. **9** (2015), no. 3, 1103–1140, DOI 10.1214/15-AOAS842. MR3418717

[14] Frank Bretz, Torsten Hothorn, and Peter Westfall, *Multiple comparisons using R*, CRC Press, 2016.

[15] Keith M Briggs, Linlin Song, and Thomas Prellberg, *A note on the distribution of the maximum of a set of Poisson random variables*, arXiv:0903.4373, 2009.

[16] Peter Bühlmann, *Statistical significance in high-dimensional linear models*, Bernoulli **19** (2013), no. 4, 1212–1242, DOI 10.3150/12-BEJSP11. MR3102549

[17] A. Buja and L. Brown, *Discussion: "A significance test for the lasso"*, Ann. Statist. **42** (2014), no. 2, 509–517, DOI 10.1214/14-AOS1175F. MR3210976

[18] Katherine SḂutton, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò, *Power failure: why small sample size undermines the reliability of neuroscience*, Nature Reviews Neuroscience **14** (2013), no. 5, 365–376.

[19] Benjamin Callahan, Diana Proctor, David Relman, Julia Fukuyama, and Susan Holmes, *Reproducible research workflow in R for the analysis of personalized human microbiome data*, Pacific Symposium on Biocomputing., vol. 21, NIH Public Access, 2016, p. 183.

[20] David R. Cox, *Discussion: Comment on a paper by Jager and Leek*, Biostatistics **15** (2014), no. 1, 16–8; Discussion 39–45.

[21] Anthony D'Aristotile, Persi Diaconis, and David Freedman, *On merging of probabilities*, Sankhyā Ser. A **50** (1988), no. 3, 363–380. MR1065549

[22] Persi Diaconis, *Magical thinking in the analysis of scientific data*, Annals of the New York Academy of Sciences **364** (1981), no. 1, 236–244.

[23] Daniel B. DiGiulio, Benjamin J. Callahan, Paul J. McMurdie, Elizabeth K. Costello, Deirdre J. Lyell, Anna Robaczewska, Christine L. Sun, Daniela S. A. Goltsman, Ronald J. Wong, Gary Shaw, David K. Stevenson, Susan P. Holmes, and David A. Relman, *Temporal and spatial variation of the human microbiota during pregnancy*, Proc. Natl. Acad. Sci. U. S. A. **112** (2015), no. 35, 11060–11065.

[24] David Donoho and Jiashun Jin, *Higher criticism for detecting sparse heterogeneous mixtures*, Ann. Statist. **32** (2004), no. 3, 962–994, DOI 10.1214/009053604000000265. MR2065195

[25] David Donoho and Jiashun Jin, *Higher criticism for large-scale inference, especially for rare and weak effects*, Statist. Sci. **30** (2015), no. 1, 1–25, DOI 10.1214/14-STS506. MR3317751

[26] Bradley Efron, *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction*, Institute of Mathematical Statistics (IMS) Monographs, vol. 1, Cambridge University Press, Cambridge, 2010. MR2724758

[27] Bradley Efron, *Estimation and accuracy after model selection*, J. Amer. Statist. Assoc. **109** (2014), no. 507, 991–1007, DOI 10.1080/01621459.2013.823775. MR3265671

[28] Bradley Efron and Robert Tibshirani, *Empirical Bayes methods and false discovery rates for microarrays*, Genetic Epidemiology **23** (2002), no. 1, 70–86.

[29] Bradley Efron, Robert Tibshirani, John D. Storey, and Virginia Tusher, *Empirical Bayes analysis of a microarray experiment*, J. Amer. Statist. Assoc. **96** (2001), no. 456, 1151–1160, DOI 10.1198/016214501753382129. MR1946571

[30] Anders Eklund, Thomas E. Nichols, and Hans Knutsson, *Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates*, Proc. Natl. Acad. Sci. U. S. A. **113** (2016), no. 28, 7900–7905.

[31] Ronald A. Fisher, *The arrangement of field experiments*, Breakthroughs in Statistics, Springer, 1992, Originally published 1926, pp. 82–91.

[32] William Fithian, Dennis Sun, and Jonathan Taylor, *Optimal inference after model selection*, 9 October 2014.

[33] Andrew Gelman and Keith O'Rourke, *Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values*, Biostatistics **15** (2014), no. 1, 18–23; Discussion 39–45.

[34] Jeremy Goecks, Anton Nekrutenko, and James Taylor, *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences*, Genome Biology **11** (2010), no. 8, 1.

[35] Peter Hall and Jiashun Jin, *Properties of higher criticism under strong dependence*, Ann. Statist. **36** (2008), no. 1, 381–402, DOI 10.1214/009053607000000767. MR2387976

[36] Peter Hall and Jiashun Jin, *Innovated higher criticism for detecting sparse signals in correlated noise*, Ann. Statist. **38** (2010), no. 3, 1686–1732, DOI 10.1214/09-AOS764. MR2662357

[37] Xiaoying Tian Harris, Snigdha Panigrahi, Jelena Markovic, Nan Bi, and Jonathan Taylor, *Selective sampling after solving a convex problem*, arXiv:1609.05609, 19 September 2016.

[38] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning*, 2nd ed., Springer Series in Statistics, Springer, New York, 2009. Data mining, inference, and prediction. MR2722294

[39] David C. Hoaglin, Frederick Mosteller, and John W. Tukey, *Exploring data tables, trends, and shapes*, John Wiley & Sons, 2011.

[40] Susan Holmes and Wolfgang Huber, *Modern statistics for modern biology*, Cambridge University Press, 2017, to appear.

[41] Wolfgang Huber, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D. Hansen, Rafael A. Irizarry, Michael Lawrence, Michael I. Love, James MacDonald, Valerie Obenchain, Andrzej K. Oleś, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K. Smyth, Dan Tenenbaum, Levi Waldron, and Martin Morgan, *Orchestrating high-throughput genomic analysis with Bioconductor*, Nat. Methods **12** (2015), no. 2, 115–121.

[42] Nikolaos Ignatiadis, Bernd Klaus, Judith B. Zaugg, and Wolfgang Huber, *Data-driven hypothesis weighting increases detection power in genome-scale multiple testing*, Nat. Methods **13** (2016), no. 7, 577–580.

[43] Yu. I. Ingster, *Some problems of hypothesis testing leading to infinitely divisible distributions*, Math. Methods Statist. **6** (1997), no. 1, 47–69. MR1456646

[44] John P. A. Ioannidis, *Why most published research findings are false*, Chance **18** (2005), no. 4, 40–47, DOI 10.1080/09332480.2005.10722754. MR2216666

[45] John P. A. Ioannidis, David B. Allison, Catherine A. Ball, Issa Coulibaly, Xiangqin Cui, Aedín C. Culhane, Mario Falchi, Cesare Furlanello, Laurence Game, Giuseppe Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort, *Repeatability of published microarray gene expression analyses*, Nature Genetics **41** (2009), no. 2, 149–155.

[46] Leah R. Jager and Jeffrey T. Leek, *An estimate of the science-wise false discovery rate and application to the top medical literature*, Biostatistics **15** (2014), no. 1, 1–12.

[47] Jana Janková and Sara van de Geer, *Honest confidence regions and optimality in high-dimensional precision matrix estimation*, arXiv:1507.02061, 8 July 2015.

[48] Lucas Janson, Rina Foygel Barber, and Emmanuel Candès, *EigenPrism: Inference for high-dimensional signal-to-noise ratios*, arXiv:1505.02097, 8 May 2015.

[49] Joshua R. Klein and Aaron Roodman, *Blind analysis in nuclear and particle physics*, Annu. Rev. Nucl. Part. Sci. **55** (2005), 141–163.

[50] Nikolaus Kriegeskorte, W. Kyle Simmons, Patrick S. F. Bellgowan, and Chris I. Baker, *Circular analysis in systems neuroscience: the dangers of double dipping*, Nat. Neurosci. **12** (2009), no. 5, 535–540.

[51] Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor, *Exact post-selection inference, with application to the lasso*, Ann. Statist. **44** (2016), no. 3, 907–927, DOI 10.1214/15-AOS1371. MR3485948

[52] Hannes Leeb, Benedikt M. Pötscher, and Karl Ewald, *On various confidence intervals post-model-selection*, Statist. Sci. **30** (2015), no. 2, 216–227, DOI 10.1214/14-STS507. MR3353104

[53] Jeffery T. Leek and Roger D. Peng, *What is the question?*, Science **347** (2015), no. 6228, 1314–1315.

[54] Jeffrey T. Leek and Roger D. Peng, *Opinion: Reproducible research can still be wrong: Adopting a prevention approach*, Proceedings of the National Academy of Sciences **112** (2015), no. 6, 1645–1646.

[55] E. L. Lehmann and Joseph P. Romano, *Testing statistical hypotheses*, 3rd ed., Springer Texts in Statistics, Springer, New York, 2005. MR2135927

[56] Jonah Lehrer, *The truth wears off*, The New Yorker **13** (2010), no. 52, 229.

[57] Ang Li and Rina Foygel Barber, *Accumulation tests for FDR control in ordered hypothesis testing*, J. Amer. Statist. Assoc. **112** (2017), no. 518, 837–849, DOI 10.1080/01621459.2016.1180989. MR3671774

[58] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani, *A significance test for the lasso*, Ann. Statist. **42** (2014), no. 2, 413–468, DOI 10.1214/13-AOS1175. MR3210970

[59] Colin L. Mallows and John W. Tukey, *An overview of techniques of data analysis, emphasizing its exploratory aspects*, Some Recent Advances in Statistics, Academic Press, London, 1982, pp. 111–172. MR773678

[60] Brendan McKay, Dror Bar-Natan, Maya Bar-Hillel, and Gil Kalai, *Solving the bible code puzzle*, Statistical Science (1999), 150–173.

[61] K. Jarrod Millman and Fernando Pérez, *Developing open-source scientific practice*, Implementing Reproducible Research. CRC Press, Boca Raton, FL (2014), 149–183.

[62] David Mumford, *Intelligent design found in the sky with* $p < 0.001$, Newsletter of the Swedish Mathematical Society (2009), `http://www.dam.brown.edu/people/mumford/beyond/papers/2009a--Orion-SMS.pdf`.

[63] Michael A. Newton, Christina M. Kendziorski, Craig S. Richmond, Frederick R. Blattner, and Kam-Wah Tsui, *On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data*, Journal of Computational Biology **8** (2001), no. 1, 37–52.

[64] Snigdha Panigrahi, Jonathan Taylor, and Asaf Weinstein, *Bayesian post-selection inference in the linear model*, arXiv:1605.08824, 28 May 2016.

[65] Prasad Patil, Roger D. Peng, and Jeffrey Leek, *A statistical definition for reproducibility and replicability*, bioRxiv:066803, 2016.

[66] Roger Peng, *The reproducibility crisis in science: A statistical counterattack*, Significance **12** (2015), no. 3, 30–32.

[67] Russell A. Poldrack and Krzysztof J. Gorgolewski, *Making big data open: data sharing in neuroimaging*, Nat. Neurosci. **17** (2014), no. 11, 1510–1517.

[68] Florian Prinz, Thomas Schlange, and Khusru Asadullah, *Believe it or not: how much can we rely on published data on potential drug targets?*, Nature Reviews Drug Discovery **10** (2011), no. 9, 712–712.

[69] Soo-Yon Rhee, Matthew J. Gonzales, Rami Kantor, Bradley J. Betts, Jaideep Ravela, and Robert W. Shafer, *Human immunodeficiency virus reverse transcriptase and protease sequence database*, Nucleic Acids Research **31** (2003), no. 1, 298–303.

[70] Robert Rosenthal, *The file drawer problem and tolerance for null results.*, Psychological Bulletin **86** (1979), no. 3, 638.

[71] Henry Scheffé, *A method for judging all contrasts in the analysis of variance*, Biometrika **40** (1953), 87–104, DOI 10.2307/2333100. MR0057504

[72] Juliet Popper Shaffer, *Multiple hypothesis testing*, Annu. Rev. Psychol. **46** (1995), 561.

[73] R. J. Simes, *An improved Bonferroni procedure for multiple tests of significance*, Biometrika **73** (1986), no. 3, 751–754, DOI 10.1093/biomet/73.3.751. MR897872

[74] Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons, *P-curve: a key to the file-drawer*, Journal of Experimental Psychology: General **143** (2014), no. 2, 534.

[75] Matthew Stephens, *False discovery rates: a new deal*, Biostatistics, 17 October 2016.

[76] John D. Storey, *A direct approach to false discovery rates*, J. R. Stat. Soc. Ser. B Stat. Methodol. **64** (2002), no. 3, 479–498, DOI 10.1111/1467-9868.00346. MR1924302

[77] John D. Storey, *The positive false discovery rate: a Bayesian interpretation and the q-value*, Ann. Statist. **31** (2003), no. 6, 2013–2035, DOI 10.1214/aos/1074290335. MR2036398

[78] John D. Storey and Robert Tibshirani, *Statistical significance for genomewide studies*, Proc. Natl. Acad. Sci. USA **100** (2003), no. 16, 9440–9445, DOI 10.1073/pnas.1530509100. MR1994856

[79] Nassim Nicholas Taleb, *The black swan: The impact of the highly improbable*, Random House, 2007.

[80] Jonathan Taylor and Robert Tibshirani, *Post-selection inference for L1-penalized likelihood models*, arXiv:1602.07358, 24 February 2016.

[81] Jonathan Taylor and Robert J. Tibshirani, *Statistical learning and selective inference*, Proc. Natl. Acad. Sci. USA **112** (2015), no. 25, 7629–7634, DOI 10.1073/pnas.1507583112. MR3371123

[82] Jonathan E. Taylor, Joshua R. Loftus, and Ryan J. Tibshirani, *Inference in adaptive regression via the Kac–Rice formula*, Ann. Statist. **44** (2016), no. 2, 743–770, DOI 10.1214/15-AOS1386. MR3476616

[83] Xiaoying Tian, Nan Bi, and Jonathan Taylor, *MAGIC: a general, powerful and tractable method for selective inference*, arXiv:1607.02630, 9 July 2016.

[84] Xiaoying Tian and Jonathan E. Taylor, *Selective inference with a randomized response*, arXiv:1507.06739, 24 July 2015.

[85] Ryan J. Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani, *Exact post-selection inference for sequential regression procedures*, J. Amer. Statist. Assoc. **111** (2016), no. 514, 600–620, DOI 10.1080/01621459.2015.1108848. MR3538689

[86] John W. Tukey, *The problem of multiple comparisons: Introduction and Parts A, B, and C*, Princeton University, 1953.

[87] John W. Tukey, *The philosophy of multiple comparisons*, Statistical Science (1991), 100–116.

[88] John W. Tukey, *The collected works of John W. Tukey. Vol. VIII*, Chapman & Hall, New York, 1994. Multiple comparisons: 1948–1983; With a preface by William S. Cleveland; With

a biography by Frederick Mosteller; Edited and with an introduction and comments by Henry I. Braun. MR1263027

[89] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure, *On asymptotically optimal confidence regions and tests for high-dimensional models*, Ann. Statist. **42** (2014), no. 3, 1166–1202, DOI 10.1214/14-AOS1221. MR3224285

[90] Stefan Wager, Wenfei Du, Jonathan Taylor, and Robert J. Tibshirani, *High-dimensional regression adjustments in randomized experiments*, Proc. Natl. Acad. Sci. USA **113** (2016), no. 45, 12673–12678, DOI 10.1073/pnas.1614732113. MR3576188

[91] Peter H. Westfall, A. Krishen, and Stanley S. Young, *Using prior information to allocate significance levels for multiple endpoints*, Stat. Med. **17** (1998), no. 18, 2107–2119.

[92] Doron Witztum, Eliyahu Rips, and Yoav Rosenberg, *Equidistant letter sequences in the book of genesis*, Statistical Science **9** (1994), no. 3, 429–438.

[93] Daniel Yekutieli, *Adjusted Bayesian inference for selected parameters*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **74** (2012), no. 3, 515–541, DOI 10.1111/j.1467-9868.2011.01016.x. MR2925372

[94] Hong Zhang and Zheyang Wu, *SetTest: Group testing procedures for signal detection and goodness-of-fit*, 2017, R package version 0.1.0.

Statistics Department, Sequoia Hall,, Stanford, California 94305
E-mail address: susan@stat.stanford.edu