Title
Topics in Conditional Inference

Author
Hung, Kenneth

Publication Date
2019

Peer reviewed|Thesis/dissertation

Topics in Conditional Inference

by

Ka Kin Kenneth Hung

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor William Fithian, Co-chair
Professor Emeritus David Aldous, Co-chair
Professor Steven N. Evans
Associate Professor Aditya Guntuboyina

Spring 2019

Topics in Conditional Inference

Abstract

Topics in Conditional Inference

by

Ka Kin Kenneth Hung

Doctor of Philosophy in Mathematics

University of California, Berkeley

Assistant Professor William Fithian, Co-chair

Professor Emeritus David Aldous, Co-chair

The modern data analysis process is rarely one-step, but instead paved with iterative exploratory data analyses and choices. Often data analysts are tempted to peek at the data before choosing the hypotheses to be tested. In other times, the vast amount of data is screened and not all information is accessible to the analysts. In either case, data analyses have to be carried post-selection, as a consequence of even the most innocuous exploratory data analyses. A particular method to conduct post-selection inference is conditional inference, with a few instances detailed in this work.

Chapter 2 — based on Hung and Fithian (2019a) — explores a scenario where the choice of null hypothesis is dependent on the very same data used in the test. Using conditional inference, we provide a test that adapts to the data, for whichever hypothesis is most sensible. As a consequence of the adaptivity, our test is also much more powerful than the classical approaches.

Chapter 3 — based on Hung and Fithian (2019b) — describes a meta-analysis where the data itself has been selected, but meaningful inference is nonetheless desired. Through conditional inference, we modified classical methods to provide post-selection inference.

Finally in Chapter 4, I present unpublished work investigating an optimal method of combining information from a post-selection original experiment and a replication experiment, a current common concern in experimental psychology.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank many people who have supported me over the past five years.

I would like to first thank my family. My fiancée Rebecca has been a constant source of emotional guidance and support, inspiring me to be brave and persevering as I went through the ups-and-downs and big decisions in my academic career. My parents have also been essential to my journey and instilled in me from a young age the curiosity that led me to this career. Both my parents and my sister Ellen encouraged me to pursue my interests abroad, despite the fact that meant we would be an ocean away.

I would also like to thank Will, who has been an excellent mentor. His patience facilitated my transition into the various fields of statistics as a pure mathematician. Will has always emphasized the importance of good writing as an academic and pushed me to empathize with the readers of our publications. My weekly meetings with Will were always the highlight of the week, with conversations about balance between mathematical elegance and practical considerations, often dotted with statistical philosophy! I will always be grateful for taking Stat 212 taught by Will, because without the serendipity of meeting him, I would not be where I am today.

I would also like to thank David Aldous, whose great teaching in Stat C205a provided me with a solid foundation in probability theory. He has also provided invaluable encouragement and advice leading to my decision to work on more statistical problems.

Finally, I would like to thank the many faculty, staff and fellow students at Berkeley. My completion of this dissertation would not be possible without the educational classes taught by the amazing faculty, all the administrative help from Victoria Lee, and the emotional and mathematical support from my officemates, Richard Zhang, Alexander Rusciano and Alexander Appleton.

Chapter 1

Introduction

With the advancement in data storage and collection, datasets are rapidly growing in both size and complexity. The traditional assumption of independence between the hypothesis under test and the data is far from valid: exploratory data analysis is no longer a simple prologue to the main data analysis, but a crucial part of the main analysis, well integrated as an iterative process exploring a "garden of forking paths" (Gelman and Loken, 2013). Such practice is not just common but also advisable — it is unwise to construct a statistical model or choose a hypothesis without first checking with data to ensure its appropriateness. However it violates the typical statistical assumption that the hypotheses are independent of the data, necessitating a new framework that allows repeated exploration of the data. Chapter 2 includes an instance of a data-dependent hypothesis, with a formal statistical test that remains valid despite the adaptivity.

Furthermore, with the plentiful datasets available, it is unfeasible to attend to every dataset. Analysts thus screen data and redirect their limited time and effort to the datasets deemed promising; likewise, consumers of these analyses also dedicate their limited attention to the stronger results. This nonetheless has a side effect — to screen is to look at the data, once again violating the independence assumption mentioned above. The datasets that passed the screening are affected by selection bias, and tend to exaggerate any signals therein. Chapters 3 and 4 explore a specific scenario where such screening and selection may arise in academic publications, and provide methods to negate the effect of selection bias in hypothesis testing.

1.1 Conditional Inference

While data-dependent hypotheses may at first sound unusual, they are commonplace in practice, for example when pilot studies are performed to generate hypotheses that are tested later on with fresh data. There is no inherent conceptual problem with testing these data-dependent hypotheses: intuitively, we understand that the test remains valid because the type I error rate is controlled for whatever hypothesis is selected, conditional on that hypothesis having been selected.

Conditional inference is well-established in the statistical literature as a means of con-

structing valid confidence intervals for parameters that were selected in a data-dependent way (e.g. Sampson and Sill, 2005; Weinstein, Fithian, and Benjamini, 2013; Yekutieli, 2012; Zöllner and Pritchard, 2007). Fithian, Sun, and Taylor (2014) generalized the intuition about pilot studies to argue that a test of a data-dependent hypothesis is valid, so long as the type I error rate is controlled conditioned on the portion of the data that generated the hypothesis. The idea is simple: roughly, if ϕ is a test (1 for rejection and 0 otherwise) that takes the data X, a data-dependent model $\mathcal{M}$ and a data-dependent hypothesis H, while controlling for the type I error rate conditional on the data-dependent part, i.e.

$$\mathbb{P}[\phi(X, \mathcal{M}, H) \mid \mathcal{M}, H] \leq \alpha, \qquad \text{where } \alpha \text{ is the significance level,}$$

then we have typical type I error rate control from the law of total expectation.

I demonstrate the applications of this method in Chapters 2 to 4, exemplifying its capabilities in Chapter 2 and simplicity in Chapters 3 and 4.

1.2 Adaptive Hypothesis

Many statistical experiments involve comparing multiple population groups. For example, a public opinion poll may ask which of several political candidates commands the most support; a social scientific survey may report the most common of several responses to a question; or, a clinical trial may compare binary patient outcomes under several treatment conditions to determine the most effective treatment. Having observed the "winner" (largest observed response) in a noisy experiment, it is natural to ask whether that candidate, survey response, or treatment is actually the "best" (stochastically largest response). Chapter 2 concerns the problem of *rank verification* — post hoc significance tests of whether the orderings discovered in the data reflect the population ranks. For exponential family models, we show under mild conditions that an unadjusted two-tailed pairwise test comparing the first two order statistics (i.e., comparing the "winner" to the "runner-up") is a valid test of whether the winner is truly the best. We extend our analysis to provide equally simple procedures to obtain lower confidence bounds on the gap between the winning population and the others, and to verify ranks beyond the first.

1.3 Selection Bias

Large-scale replication studies like the Reproducibility Project: Psychology (RP:P) provide invaluable systematic data on scientific replicability, but most analyses and interpretations of the data fail to agree on the definition of "replicability" and disentangle the inexorable consequences of known selection bias from competing explanations. We discuss three concrete definitions of replicability based on (1) whether published findings about the signs of effects are mostly correct, (2) how effective replication studies are in reproducing whatever true effect size was present in the original experiment, and (3) whether true effect sizes tend to diminish in replication. In Chapter 3, we apply techniques from multiple testing and post-selection inference to develop new methods that

answer these questions while explicitly accounting for selection bias. Re-analyzing the RP:P data, we estimate that 22 out of 68 (32%) original directional claims were false (upper confidence bound 47%); by comparison, we estimate that among claims significant at the stricter significance threshold 0.005, only 2.2 out of 33 (7%) were directionally false (upper confidence bound 18%). In addition, we compute selection-adjusted confidence intervals for the difference in effect size between original and replication studies and, after adjusting for multiplicity, identify five (11%) which exclude zero (exact replication). We estimate that the effect size declined by at least 20% in the replication study relative to the original study in 16 of the 46 (35%) study pairs (lower confidence bound 11%). Our methods make no distributional assumptions about the true effect sizes.

Chapter 4 bases on the same setup as Chapter 3, but takes a different direction. We develop a uniformly most powerful unbiased test (UMPU) and a uniformly minimum variance unbiased (UMVU) estimator for normally distributed measurements and sample correlation coefficients. We also demonstrate the performance gains in a simulation study.

Chapter 2

Rank Verification for Exponential Families

2.1 Introduction

Motivating Example: Iowa Republican Caucus Poll

Table 2.1 shows the result of a Quinnipiac University poll asking 890 Iowa Republicans their preferred candidate for the Republican presidential nomination (Quinnipiac University Poll Institute, 2016). Donald Trump led with 31% of the vote, Ted Cruz came second with 24%, Marco Rubio third with 17%, and ten other candidates including "Don't know" trailed behind.

Rank	Candidate	Result	Votes
1 *	Trump	31%	276
2 *	Cruz	24%	214
3 *	Rubio	17%	151
4 *	Carson	8%	71
5	Paul	4%	36
6	Bush	4%	36
7	Huckabee	3%	27
⋮	⋮	⋮	⋮

Table 2.1: Results from a February 1, 2016 Quinnipiac University poll of 890 Iowa Republicans. To compute the last column (Votes), we make the simplifying assumption that the reported percentages in the third column (Result) are raw vote shares among survey respondents. The asterisks indicate that the rank is verified at level 0.05 by a stepwise procedure.

Seeing that Trump leads this poll, several salient questions may occur to us: Is Trump

really winning, and if so by how much? Furthermore, is Cruz really in second, is Rubio really in third, and so on? Note that there is implicitly a problem of multiple comparisons here, because if Cruz had led the poll instead, we would be asking a different set of questions ("Is Cruz really winning," etc.). Indeed, the selection issue appears especially pernicious due to the so-called "winner's curse": given that Trump leads the poll, it more likely than not overestimates his support.

Nevertheless, if we blithely ignore the selection issue, we might carry out the following analyses to answer the questions we posed before at significance level $\alpha = 0.05$. We assume for simplicity that the poll represents a simple random sample of Iowa Republicans; i.e., that the data are a multinomial sample of size 890 and underlying probabilities $(\pi_{\text{Trump}}, \pi_{\text{Cruz}}, \ldots)$. (The reality is a bit more complicated: before releasing the data, Quinnipiac has post-processed it to make the reported result more representative of likely caucus-goers. The raw data is proprietary.)

1. *Is Trump really winning?* If Trump and Cruz were in fact tied, then Trump's share of their combined 490 votes would be distributed as Binomial $(490, 0.5)$. Because the (two-tailed) p-value for this pairwise test is $p = 0.006$, we reject the null and conclude that Trump is really winning.
2. *By how much?* Using an exact 95% interval for the same binomial model, we conclude Trump has at least 7.5% more support than Cruz (i.e., $\pi_{\text{Trump}} \geq 1.075\,\pi_{\text{Cruz}}$) and also leads the other candidates by at least as much.
3. *Is Cruz in second, Rubio in third, etc.?* We can next compare Cruz to Rubio just as we compared Trump to Cruz (again rejecting because 214 is significantly more than half of 365), then Rubio to Carson, and so on, continuing until we fail to reject. The first four comparisons are all significant at level 0.05, but Paul and Bush are tied so we stop.

Perhaps surprisingly, all of the three procedures described above are statistically valid despite their ostensibly ignoring the implicit multiple-comparisons issue. In other words, Procedures 1 and 2 control the Type I error rate at level α and Procedure 3 controls the familywise error rate (FWER) at level α. The remainder of this chapter is devoted to justifying these procedures for the multinomial family, and extending to analogous procedures in other exponential family settings. While methods analogous to Procedures 1 and 2 have been justified previously for balanced independent samples from log-concave location families (Gutmann and Maymin, 1987; Stefansson, Kim, and Hsu, 1988), they have not been justified in exponential families before now.

Generic Problem Setting and Main Result

Generically, we will consider data drawn from an exponential family model with density

$$X \sim \exp\left(\theta' x - \psi\left(\theta\right)\right) g\left(x\right), \tag{2.1}$$

with respect to either the Lebesgue measure on $\mathbb{R}^n$ or counting measure on $\mathbb{Z}^n$. We assume further that $g\left(x\right)$ is symmetric with respect to permutation, and Schur concave,

a mild technical condition defined in Section 2.2. In addition to the multinomial family, model (2.1) also encompasses settings such as comparing independent binomial treatment outcomes in a clinical trial, competing sports teams under a Bradley–Terry model, entries of a Dirichlet distribution, and many more; see Section 2.2 for these and other examples.

We will generically use the term *population* to refer to the treatment group, sports team, political candidate, etc. represented by a given random variable X_j. As we will see, $\theta_j \geq \theta_k$ if and only if X_j is stochastically larger than X_k; thus, there is a well-defined stochastic ordering of the populations that matches the ordering of the entries of θ. We will refer to the population with maximal θ_j as the *best*, the population with second largest θ_j as the *second best*, the one with maximal X_j as the *winner*, and the one with the second-largest X_j as the *runner-up*, where ties between observations are broken randomly to obtain a full ordering. Following the convention in the ranking and selection literature, we assume that if there are multiple largest θ_j, then one is arbitrarily marked as the best. Note that in cases where it is more interesting to ask which is the smallest population (for example, if X_j is the number of patients on treatment j who suffer a heart attack during a trial) we can change the variables to $-X$ and the parameters to $-\theta$; this does not affect the Schur concavity assumption.

Write the order statistics of X as

$$X_{[1]} \geq X_{[2]} \geq \cdots \geq X_{[n]},$$

where $[j]$ will denote the random index for the j-th order statistic. Thus, $\theta_{[j]}$ is the entry of θ corresponding to the j-th order statistic of X (so $\theta_{[1]}$ might *not* equal $\max_j \theta_j$, for example).

In each of the above examples, there is a natural exact test we could apply to test $\theta_j = \theta_k$ for any two *fixed* populations j and k. In the multinomial case, we would apply the conditional binomial test based on the combined total $X_j + X_k$ as discussed in the previous section. For the case of independent binomials we would apply Fisher's exact test, again conditioning on $X_j + X_k$. These are both examples of a generic UMPU pairwise test in which we condition on the other $n-2$ indices (notated $X_{\setminus\{j,k\}}$) and $X_j + X_k$, and reject the null if X_j is outside the $\alpha/2$ and $1-\alpha/2$ quantiles of the conditional law $\mathcal{L}_{\theta_j=\theta_k}(X_j \mid X_j + X_k, X_{\setminus\{j,k\}})$. Crucially, this null distribution does not depend on the value of θ provided that $\theta_j = \theta_k$. We call this test the (two-tailed) *unadjusted pairwise test* since it makes no explicit adjustment for selection. Similarly, inverting this test for other values of $\theta_j - \theta_k$ yields an *unadjusted pairwise confidence interval*. (To avoid trivialities in the discrete case, we assume these procedures are appropriately randomized at the rejection thresholds to give exact level-α control.)

Generalizing the procedures described in Section 2.1 we obtain the following:

1. *Is the winner really the best?* To test the hypothesis $H : \theta_{[1]} \leq \max_{j \neq [1]} \theta_j$: Carry out the unadjusted pairwise test comparing the winner to the runner-up. If the test rejects at level α, reject H and declare that the winner is really the best.
2. *By how much?* To construct a lower confidence bound for $\theta_{[1]} - \max_{j \neq [1]} \theta_j$: Construct the unadjusted pairwise confidence interval comparing the winner to the

runner-up, and report the lower confidence bound obtained for $\theta_{[1]} - \theta_{[2]}$ if it is nonnegative, report $-\infty$ otherwise.

3. *Is the runner-up really the second best, etc.?* Continue by comparing the runner-up to the second runner-up, again using the unadjusted pairwise test, and so on down the list comparing adjacent values. Stop the first time the test does not reject; if there are j rejections, declare that

$$\theta_{[1]} > \theta_{[2]} > \cdots > \theta_{[j]} > \max_{k>j} \theta_{[j]}$$

Procedures 2 and 3 are conservative stand-ins for exact, but slightly more involved, conditional inference procedures. In particular, as we will see, reporting $-\infty$ in Procedure 2 is typically much more conservative than is necessary.

We now state our main theorem: under a mild technical assumption, Procedures 1–3 described above are statistically valid, even accounting for the selection.

Theorem 1. *Assume the model* (2.1) *holds and* $g(x)$ *is a Schur-concave function. Then:*

1. *Procedure 1 has exact level* α *conditional on* H *being true (conditional on the best population not winning), and marginally has level* $\alpha \cdot \mathbb{P}(H \text{ is true}) \leq \alpha\left(1 - \frac{1}{n}\right)$.
2. *Procedure 2 gives a conservative* $1-\alpha$ *lower confidence bound for* $\theta_{[1]} - \max_{j \neq [1]} \theta_j$.
3. *Procedure 3 is a conservative stepwise procedure with FWER no larger than* α.

Note that Theorem 1 implies that we could actually replace α with $\frac{n}{n-1}\alpha$ to obtain a more powerful version of Procedure 1 when n is not too large.

We define Schur-concavity and discuss its properties in Section 2.2. Because any log-concave and symmetric function is Schur-concave, Theorem 1 applies to all of the cases discussed above. The proof combines the conditional selective-inference framework of Fithian, Sun, and Taylor (2014) with classical multiple-testing methods, as well as new technical tools involving majorization and Schur-concavity.

Note that these procedures make an implicit adjustment for selection because they use two-tailed, rather than one-tailed, unadjusted tests. If we instead based our tests on an independent realization $X^* = (X_1^*, \ldots, X_n^*)$ then, for example, Procedure 1 could use a right-tailed version of the unadjusted pairwise test. In the case $n = 2$, Procedure 1 amounts to a simple two-tailed test of the null hypothesis $\theta_1 = \theta_2$, and it is intuitively clear that a one-tailed test would be too liberal. More surprising is that, no matter how large n is, Procedures 1–3 require no further adjustment beyond what is required when $n = 2$.

Related work

Rank verification has been studied extensively in the ranking and selection literature. See Gupta and Panchapakesan (1971, 1985) for surveys of the subset selection literature. The two main formulations of ranking and selection are closely related to procedures for multiple comparisons with the best treatment (Edwards and Hsu, 1983; Hsu, 1984), but

more powerful methods are available in some cases for procedures involving only the first sample rank, the problem of comparisons with the sample best; see Hsu (1996) for an overview and discussion of the relationships between these problems.

Comparisons with the sample best have been especially well-studied and the validity of Procedures 1 and 2 have been established in a different setting: balanced independent samples from log-concave location families. Gutmann and Maymin (1987) prove the validity of Procedure 1 in this setting, and Bofinger (1991), Karnnan and Panchapakesan (2009), and Maymin and Gutmann (1992) give similar results for other models including scale and location-scale families. Stefansson, Kim, and Hsu (1988) provide an alternative proof for the validity of Procedure 1 in the same setting, leading to a lower confidence bound analogous to that of Procedure 2; interestingly, the proof involves a very early application of the partitioning principle, later developed into fundamental technique in multiple comparisons (Finner and Strassburger, 2002). These results use very different technical tools than the ones we use here, require independence between the different groups (ruling out, for example, the multinomial family), and do not address the exponential family case. Because most exponential families are not location-scale families (the Gaussian being a notable exception), and because our results involve more general dependence structures, both our proof techniques and our technical results are complementary to the techniques and results in the above works.

For the multinomial case, Gupta and Nagel (1967), discussed in Section 2.3, remain the state of the art in finite-sample tests; Gupta and Wong (1976) discuss related approaches for Poisson models. Berger (1980) mentions an alternative, simpler rule which performs a binomial test on each population, but its power does not necessarily increase as the size m of observations increases in cases like Multinomial$(m; 2/3, 1/3, 0, \ldots, 0)$. Nettleton (2009) proves validity for an asymptotic version of the winner-versus-runner-up test, and Gupta and Liang (1989) consider an empirical Bayes approach for selecting the best binomial population wherein a parametric prior distribution is assumed for the success probabilities for the different populations. Ng and Panchapakesan (2007) discuss an exact test for a modified problem in which the maximum count is fixed instead of the total count; that is, we sample until the leading candidate has at least m votes. As Section 2.3 shows, our test can be much more powerful than the one in Gupta and Nagel (1967), especially if there are many candidates, because of the way our critical rejection threshold for $X_{[1]} - X_{[2]}$ adapts to the data. Thus, our work closes a significant gap in the ranking and selection literature, extending the result of Gutmann and Maymin (1987) and others to new families like the multinomial, independent binomials, and many others.

Outline

Section 2.2 defines Schur concavity, and gives several examples satisfying this condition. Section 2.3 justifies Procedure 1 and compares its power to that of Gupta and Nagel (1967). Sections 2.4 and 2.5 justify Procedures 2 and 3 respectively, and Section 2.6 concludes.

2.2 Majorization and Schur concavity

Definitions and basic properties

We start by reviewing the notion of *majorization*, defined on both $\mathbb{R}^n$ and $\mathbb{Z}^n$.

Definition 1. *For two vectors a and b in $\mathbb{R}^n$ (or $\mathbb{Z}^n$), suppose sorting the two vectors in descending order gives $a_{(1)} \geq \cdots \geq a_{(n)}$ and $b_{(1)} \geq \cdots \geq b_{(n)}$. We say that $a \succeq b$ (a majorizes b) if for $1 \leq i < n$,*

$$a_{(1)} + \cdots + a_{(i)} \geq b_{(1)} + \cdots + b_{(i)}, \quad \text{and}$$
$$a_{(1)} + \cdots + a_{(n)} = b_{(1)} + \cdots + b_{(n)}.$$

This forms a partial order in $\mathbb{R}^n$ (or $\mathbb{Z}^n$).

Intuitively, majorization is a partial order that monitors the evenness of a vector: the more even a vector is, the "smaller" it is. There are two properties of majorization that we will use in the proofs.

Lemma 2.

1. *Suppose $(x_1, x_2, x_3, \ldots)$ and $(x_1, y_2, y_3, \ldots)$ are two vectors in $\mathbb{R}^n$. Then*

$$(x_1, x_2, x_3, \ldots) \succeq (x_1, y_2, y_3, \ldots) \text{ if and only if } (x_2, x_3, \ldots) \succeq (y_2, y_3, \ldots).$$

2. *(Principle of transfer) If $x_1 > x_2$ and $t \geq 0$, then*

$$(x_1 + t, x_2, x_3, \ldots) \succeq (x_1, x_2 + t, x_3, \ldots).$$

If $t \leq 0$, the majorization is reversed.

Proof.

1. The property follows from an equivalent formulation of majorization listed in Marshall, Olkin, and Arnold (2010), where $x \succeq y$ if and only if

$$\sum_{j=1}^{n} x_n = \sum_{j=1}^{n} y_n \quad \text{and} \quad \sum_{j=1}^{n} (x_j - a)_+ \geq \sum_{j=1}^{n} (y_j - a)_+ \text{ for all } a \in \mathbb{R}.$$

2. Proved in Marshall, Olkin, and Arnold (2010). □

Definition 2. *A function g is Schur-concave if $x \succeq y$ implies $g(x) \leq g(y)$.*

A Schur-concave function is symmetric by default since $a \succeq b$ and $b \succeq a$ if and only if b is a permutation of the coordinates of a. Conversely a symmetric and log-concave function is Schur-concave (Marshall, Olkin, and Arnold, 2010). Interestingly, Gupta, Huang, and Panchapakesan (1984) also show that, in the context of independent location families, Schur concavity of the probability density is equivalent to monotone likelihood ratio.

Examples

Many common exponential family models have Schur-concave carrier densities. Below we give a few examples:

Example 1 (Independent binomial treatment outcomes in a clinical trial)**.** If each of n different treatments are applied to m patients independently, the number of positive outcomes X_j for treatment j is Binomial (m, p_j). The best treatment would be the treatment with the highest success probability p_j. The joint distribution of X is given by

$$p(x) \propto \exp\left(\sum_j x_j \log \frac{p_j}{1-p_j}\right) \frac{1}{x_1!\,(m-x_1)! \cdots x_n!\,(m-x_n)!}$$

The carrier measure above is Schur-concave. The unadjusted pairwise test in this family is Fisher's exact test.

Example 2 (Competitive sports under the Bradley–Terry model)**.** Suppose n players compete in a round robin tournament, where player j has ability θ_j, and the probability of player j winning against player k is

$$\frac{e^{\theta_j - \theta_k}}{1 + e^{\theta_j - \theta_k}} = \frac{e^{(\theta_j - \theta_k)/2}}{e^{(\theta_j - \theta_k)/2} + e^{(\theta_k - \theta_j)/2}}.$$

Let Y_{jk} be an indicator for the match between player j and k, where we take $Y_{jk} = 1$ if j beats k and $Y_{jk} = 0$ if k beats j. For symmetry, we will also adopt the convention that $Y_{jk} + Y_{kj} = 1$. Thus the joint distribution of $Y = (Y_{jk})_{j \neq k}$ is

$$p(y) \propto \exp\left(\sum_j 2\theta_j \sum_{k \neq j} y_{jk}\right) = \exp\left(2\theta' x\right),$$

where $x_j = \sum_{k \neq j} y_{jk}$. In other words, if X_j is the number of wins by player j, then $X = (X_1, \ldots, X_n)$ is a sufficient statistic with distribution

$$p(x) = \exp\left(2\theta' x\right) g(x),$$

where $g(x)$ is a function that counts the number of possible tournament results giving the net win vector x. A bijection proof shows that x is indeed Schur-concave. Therefore, we can use Procedures 1–3 to compare player qualities.

After conditioning on $U(X) = (X_1 + X_2, X_3, \ldots, X_n)$, and under the assumption $\theta_1 = \theta_2$, every feasible configuration of Y is equally likely. If n is not too large (say, no more than 40 players), we can find the conditional distribution of X_1 by enumerating over the configurations; for larger n, computation might pose a more serious problem, requiring us for example to compute the p-value using Markov Chain Monte Carlo techniques (Besag and Clifford, 1989).

Example 3 (Comparing the variances of different normal populations)**.** Suppose there are n normal populations with laws $N(\mu_j, \sigma_j^2)$ and m independent observations from each of them. The sample variance for population j can be denoted as R_j. By Cochran's theorem, $(m-1)\,R_j \sim \sigma_j^2 \chi_{m-1}^2$, and thus the joint distribution of R is

$$\begin{aligned} r &\sim \prod_{j=1}^{n} \left(\frac{(m-1)\,r_j}{\sigma_j^2} \right)^{(m-3)/2} e^{-(m-1)r_j/2\sigma_j^2} 1_{\{r_j>0\}} \\ &\propto \exp\left(-\frac{m-1}{2\sigma_1^2} r_1 - \cdots - \frac{m-1}{2\sigma_n^2} r_n \right) \prod_{j=1}^{n} r_j^{(m-3)/2} 1_{\{r>0\}}. \end{aligned}$$

The carrier measure is $\prod_{j=1}^{n} r_j^{(m-3)/2} 1_{\{r>0\}}$, which is Schur-concave. Thus, we can use Procedures 1–3 to find populations with the smallest or largest variances. In this example, the distribution of $X_1/(X_1+X_2)$ conditional on $(X_1+X_2, X_3, \ldots, X_4)$ is distributed as $\mathrm{Beta}(m/2, m/2)$ under the null, or equivalently X_1/X_2 is conditionally distributed as $F_{m,m}$; hence a (two-tailed) F-test is valid for comparing the top two populations.

2.3 Verifying the Winner: Is the Winner Really the Best?

First, we justify the notion that the population with largest θ_j is also the largest population in stochastic order:

Theorem 3. *For a multivariate exponential family with a symmetric carrier distribution, $X_1 \geq X_2$ in stochastic order if and only if $\theta_1 \geq \theta_2$.*

Proof. It suffices to prove the "if" part, as the "only if" part can be follows from swapping the role of θ_1 and θ_2. For any fixed a, and $x_1 \geq a$ and $x_2 < a$, we have $x_1 > x_2$ and

$$\exp\left(\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n - \psi(\theta)\right) g(x) \geq \exp\left(\theta_1 x_2 + \theta_2 x_1 + \cdots + \theta_n x_n - \psi(\theta)\right) g(x).$$

Integrating both sides over the region $\{x :\ x_1 \geq a, x_2 < a\}$ gives

$$\mathbb{P}\left[X_1 \geq a, X_2 < a\right] \geq \mathbb{P}\left[X_1 < a, X_2 \geq a\right].$$

Now adding $\mathbb{P}\left[X_1 \geq a, X_2 \geq a\right]$ to both probabilities gives

$$\mathbb{P}\left[X_1 \geq a\right] \geq \mathbb{P}\left[X_2 \geq a\right],$$

meaning that X_1 is greater than X_2 in stochastic order. □

Before proving our main result for Procedure 1, we give the following lemmas, the first of which clarifies a key idea in the proof, and the second is needed for a sharper bound in (2.2).

Lemma 4 (Berger, 1982)**.** *If p_j are valid p-values for testing null hypothesis H_{0j}, then $p_* = \max_j p_j$ is a valid p-value for the union null (i.e. disjunction null) hypothesis $H_0 = \bigcup_j H_{0j}$.*

Proof. Under H_0, one of the H_{0j} is true; without loss of generality assume it is H_{01}. Then,

$$\mathbb{P}[p_* \leq \alpha] \leq \mathbb{P}[p_1 \leq \alpha] \leq \alpha.$$

Therefore p_* is a valid p-value for the union null hypothesis. □

Lemma 5. *If $\theta_1 \geq \max_{j \neq 1} \theta_j$, then $\mathbb{P}[1 \text{ wins}] \geq \frac{1}{n}$.*

Proof. We can prove so with a coupling argument: for any sequence $x_1, x_2, \ldots, x_n$, define $\tau(x) = \{\tau(x_j)\}_{j=1,\ldots,n}$, obtained by swapping x_1 with the largest value in the sequence x. Hence

$$\exp(\theta_1 \tau(x_1) + \cdots + \theta_n \tau(x_n) - \psi(\theta)) g(X) \geq \exp(\theta_1 x_1 + \cdots + \theta_n x_n - \psi(\theta)) g(X).$$

If we integrating both sides over $\mathbb{R}^n$ (or $\mathbb{Z}^n$ in the case of counting measure), the right hand side gives 1. Since τ is an n-to-1 mapping, the left hand side is n times the integral over $\{x_1 \geq \max_{j>1} x_j\}$. In other words,

$$n\mathbb{P}[1 \text{ wins}] \geq 1$$

as desired.

In the case of counting measure, the above argument follows if a subscript is attached to identical observations uniformly to ensure strict ordering. □

We are now ready to prove our result for Procedure 1, restated here for reference.

Part 1 of Theorem 1. *Assume the model (2.1) holds and $g(x)$ is a Schur-concave function. Procedure 1 (the unadjusted pairwise test) has level α conditional on the best population not winning.*

Proof. Let j^* denote the (fixed) index of the best population, so $\theta_{j^*} \geq \max_{j \neq j^*} \theta_j$. The type I error — the probability of incorrectly declaring any other j to be the best — is

$$\mathbb{P}\left[\bigcup_{j \neq j^*} \text{declare } j \text{ best}\right] \leq \sum_{j \neq j^*} \mathbb{P}[\text{declare } j \text{ best} \mid j \text{ wins}]\, \mathbb{P}[j \text{ wins}],$$

recalling that ties are broken randomly, so there is only one winner in any realization. Thus, it is enough to bound $\mathbb{P}_\theta[\text{declare } j \text{ best} \mid j \text{ wins}] \leq \alpha$, for each $j \neq j^*$, and for all θ with $j^* \in \arg\max_j \theta_j$. Then we will have

$$\mathbb{P}\left[\bigcup_{j \neq j^*} \text{declare } j \text{ best}\right] \leq \sum_{j \neq j^*} \alpha \cdot \mathbb{P}[j \text{ wins}] = \alpha \mathbb{P}[j^* \text{ does not win}] \leq \frac{n-1}{n}\alpha, \quad (2.2)$$

where the last inequality follows from Lemma 5.

We start by assuming that we are working with the Lebesgue measure rather than the counting measure (eliminating the possibility of ties). The necessary modification of the proof for the counting measure case is provided at the end of this proof.

To minimize notational clutter, we consider only the case where the winner is 1, i.e. $X_1 \geq \max_{j>1} X_j$. Furthermore, we will denote the runner-up with 2. This is not necessarily true, but we will use it as a shorthand to simplify our notation. For other cases, the following proof remains valid under relabeling and can thus be applied. In this case, we will test the null hypothesis $H_{01} : \theta_1 \leq \max_{j>1} \theta_j$, which is the union of the null hypotheses $H_{01j} : \theta_1 \leq \theta_j$ for $j \geq 2$. For each of these we can construct an exact p-value p_{1j}, which is valid under H_{01j} conditional on A_1, the event that X_1 is the winner. Hence by Lemma 4, a test that rejects when $p_{1*} = \max_j p_{1j} \leq \alpha$ is valid for H_{01} conditional on A_1. Procedure 1 performs an unadjusted pairwise test comparing X_1 to X_2. Hence it is sufficient to show that $p_{12} = p_{1*}$ and that rejecting when $p_{12} \leq \alpha$ coincides with the unadjusted pairwise test.

Our proof has three main parts: (1) deriving p_{1j} for each $j \geq 2$, (2) showing that $p_{12} \geq p_{1j}$ for each $j \geq 2$, and (3) showing that p_{12} is an unadjusted pairwise p-value.

Derivation of p_{1j} Following the framework in Fithian, Sun, and Taylor (2014), we first construct the p-values by conditioning on the selection event where the winner is 1:

$$A_1 = \left\{ X_1 \geq \max_{j>1} X_j \right\}.$$

For convenience, we let

$$D_{jk} = \frac{X_j - X_k}{2} \quad \text{and} \quad M_{jk} = \frac{X_j + X_k}{2}.$$

We then re-parametrize to replace X_1 and X_j with D_{1j} and M_{1j}. The distribution is now an exponential family with sufficient statistics $D_{1j}, M_{1j}, X_{\setminus\{1,j\}}$ and corresponding natural parameters $\theta_1 - \theta_j, \theta_1 + \theta_j, \theta_{\setminus\{1,j\}}$. We now consider

$$\mathcal{L}_{\theta_1 - \theta_j = 0} \left(D_{1j} \mid M_{1j}, X_{\setminus\{1,j\}}, A_1 \right). \tag{2.3}$$

We can rewrite the selection event in terms of our new parameterization as

$$\begin{aligned} A_1 &= \{X_1 \geq X_j\} \cap \left\{ X_1 \geq \max_{k \neq 1,j} X_k \right\} \\ &= \{D_{1j} \geq 0\} \cap \left\{ D_{1j} \geq \max_{k \neq 1,j} X_k - M_{1j} \right\}. \end{aligned}$$

The conditional law of D_{1j} in (2.3), in particular, is a truncated distribution.

$$\begin{aligned} p\left(d_{1j} \mid M_{1j}, X_{\setminus\{1,j\}}, A_1\right) &\propto \exp\left((\theta_1 - \theta_j)\, d_{1j} + \theta_2 X_2 + \cdots + (\theta_1 + \theta_j)\, M_{1j} + \cdots + \theta_n X_n \right) \\ &\quad g\left(M_{1j} + d_{ij}, X_2, \ldots, M_{ij} - d_{ij}, \ldots X_n\right) 1_{A_1} \\ &\overset{(a)}{\propto} g\left(M_{1j} + d_{1j}, X_2, \ldots, M_{1j} - d_{1j}, \ldots X_n\right) 1_{A_1}, \end{aligned}$$

where at step (a), conditioning on $X_{\setminus\{1,j\}}$ and M_{1j} removes dependence on $\theta_{\setminus\{1,j\}}$ and $\theta_1 + \theta_j$ respectively, while $\theta_1 - \theta_j$ is taken to be 0 under our null hypothesis. Note that we consider this as a one-dimensional distribution of D_{1j} on $\mathbb{R}$, where M_{1j} and $X_{\setminus\{1,j\}}$ are treated as fixed.

The p-value for H_{01j} is thus

$$p_{1j} = \frac{\int_{D_{1j}}^{\infty} g\left(M_{1j} + z, X_2, \ldots, M_{1j} - z, \ldots, X_n\right)\, dz}{\int_{\max\{X_2 - M_{1j}, 0\}}^{\infty} g\left(M_{1j} + z, X_2, \ldots, M_{1j} - z, \ldots, X_n\right)\, dz}. \tag{2.4}$$

Finally, by construction, p_{1j} satisfies

$$\mathbb{P}_{H_{01j}}\left[p_{1j} < \alpha \;\middle|\; M_{1j}, X_{\setminus\{1,j\}}, A_1\right] \leq \alpha \quad \text{a.s.},$$

Marginalizing over $M_{1j}, X_{\setminus\{1,j\}}$,

$$\mathbb{P}_{H_{01j}}\left[p_{1j} < \alpha \mid A_1\right] \leq \alpha.$$

Therefore these p_{1j} are indeed valid p-values.

Demonstration that $p_{1*} = p_{12}$ We now proceed to show that p_{12}, the p-value comparing the winner to the runner-up, is the largest of all p_{1j}. Without loss of generality, it is sufficient to show that $p_{12} \geq p_{13}$.

From the first part of this proof, both p-values are constructed by conditioning on $X_{\setminus\{1,2,3\}}$. Upon conditioning these, (X_1, X_2, X_3) follows an exponential family distribution, with carrier distribution

$$g_{X_4,\ldots,X_n}\left(X_1, X_2, X_3\right) = g\left(X_1, \ldots, X_n\right),$$

here $X_4, \ldots, X_n$ are used in the subscript as they are conditioned on and no longer considered as variables. The first point in Lemma 2 says that the function $g_{X_4,\ldots,X_n}$ is Schur-concave as well. We have reduced the problem to the case when $n = 3$: we can apply the result for $n = 3$ to $g_{X_4,\ldots,X_n}$ to yield $p_{12} \geq p_{13}$ for $n > 3$.

We have reduced to the case when $n = 3$. The p-values thus are

$$p_{12} = \frac{\int_{D_{12}}^{\infty} g\left(M_{12} + z, M_{12} - z, X_3\right)\, dz}{\int_{0}^{\infty} g\left(M_{12} + z, M_{12} - z, X_3\right)\, dz},$$
$$p_{13} = \frac{\int_{D_{13}}^{\infty} g\left(M_{13} + z, X_2, M_{13} - z\right)\, dz}{\int_{\max\{X_2 - M_{13}, 0\}}^{\infty} g\left(M_{13} + z, X_2, M_{13} - z\right)\, dz}$$

The maximum in the denominator of p_{13} prompts us to consider two separate cases. First, we suppose $X_2 < M_{13}$. Changing variables such that the lower limits of both

integrals in the numerator are 0, we can re-parametrize the integrals above to give

$$\begin{aligned} p_{12} &= \frac{\int_0^\infty g\left(X_1+z, X_2-z, X_3\right)\, dz}{\int_0^\infty g\left(M_{12}+z, M_{12}-z, X_3\right)\, dz} \\ &= \frac{\int_0^\infty g\left(X_1+z, X_2-z, X_3\right)\, dz}{\int_{-D_{12}}^\infty g\left(X_1+z, X_2-z, X_3\right)\, dz}, \\ p_{13} &= \frac{\int_0^\infty g\left(X_1+z, X_2, X_3-z\right)\, dz}{\int_0^\infty g\left(M_{13}+z, X_2, M_{13}-z\right)\, dz} \\ &= \frac{\int_0^\infty g\left(X_1+z, X_2, X_3-z\right)\, dz}{\int_{-D_{13}}^\infty g\left(X_1+z, X_2, X_3-z\right)\, dz}. \end{aligned}$$

To help see the re-parametrization, each of these integrals can be thought of in terms of integrals along segments and rays. For example p_{12} can be represented in terms of integrals A and B in Figure 2.1. Specifically,

$$p_{12} = \frac{B}{A+B}$$

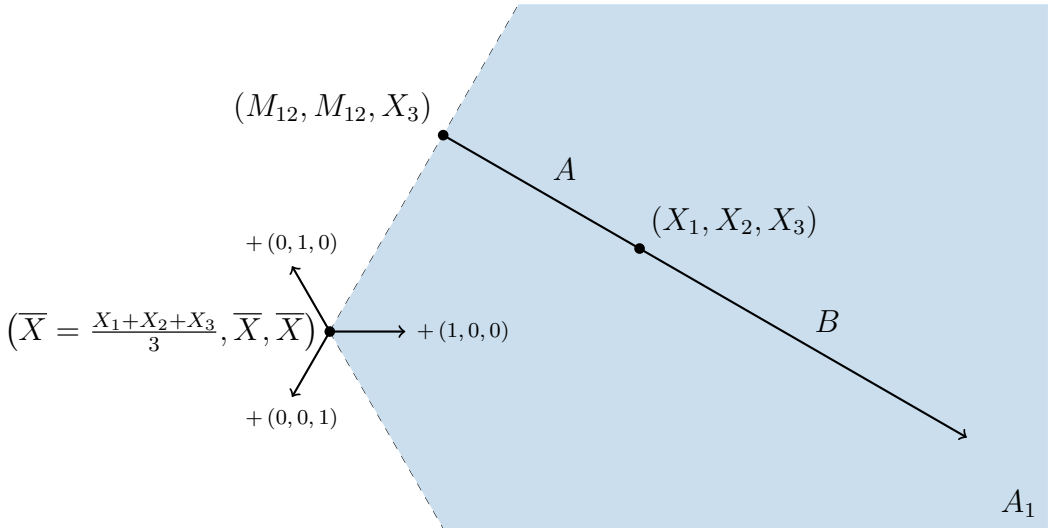


Figure 2.1: The p-value p_{12} can be written in terms of integral A along the segment and B along the ray. The diagram is drawn a level set of $x_1 + x_2 + x_3$. The green region represents the selection event A_1.

Figure 2.2 has both the p-values shown on the same diagram. Proving $p_{12} \geq p_{13}$ is the same as proving

$$\frac{B}{A+B} \geq \frac{D}{C+D} \quad \Longleftrightarrow \quad \frac{B}{A} \geq \frac{D}{C}.$$

We will prove so by extending A to include $\tilde{A}$ on the diagram. We denote the sum $A+\tilde{A}$ as A'. Formally,

$$A' = \int_{-D_{13}}^0 g\left(X_1+z, X_2-z, X_3\right)\, dz \geq \int_{-D_{12}}^0 g\left(X_1+z, X_2-z, X_3\right)\, dz = A. \qquad (2.5)$$

It is thus sufficient to show that $B \geq D$ and $C \geq A'$.

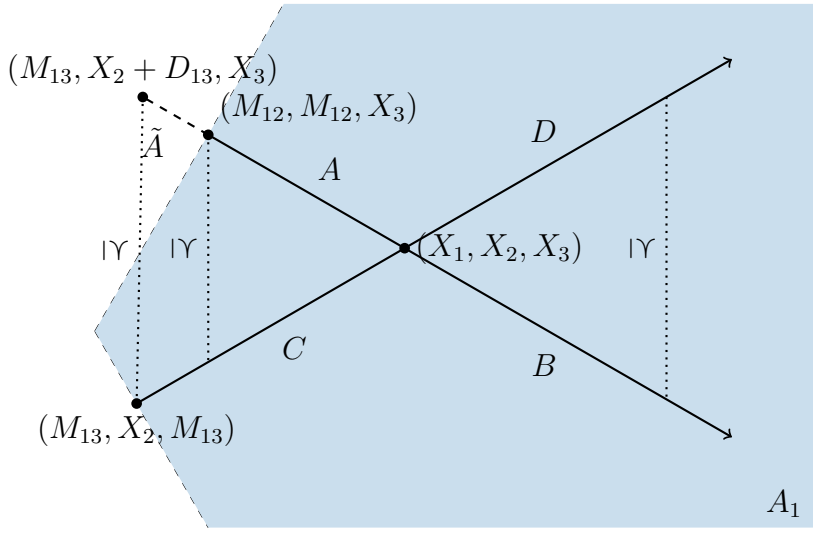


Figure 2.2: The p-value p_{12} can be written in terms of integral A along the segment and B along the ray; and p_{13} in terms of C and D. A' would refer to the sum of A with the dashed line portion labeled as $\tilde{A}$, formally explained in Equation (2.5). The majorization relation is indicated by the dotted line.

Indeed from the second point in Lemma 2 we have

$$(X_1 + z, X_2 - z, X_3) \succeq (X_1 + z, X_2, X_3 - z)$$

for $z \leq 0$ and the majorization reversed for $z \geq 0$. This majorization relation is indicated as the dotted line in Figure 2.2. So Schur-concavity shows that

$$g\left(X_1 + z, X_2 - z, X_3\right) \leq g\left(X_1 + z, X_2, X_3 - z\right)$$

for $z \leq 0$, and the inequality reversed for $z \geq 0$. Taking integrals on both sides yields the desired inequality.

For the second case where $X_2 \geq M_{13}$, the segment C will reach the line $x_1 = x_2$ first before it reaches $x_1 = x_3$, ending at $(X_2, X_2, X_1 - X_2 + X_3)$ instead. But we can still extend A by $\tilde{A}$ to (X_2, X_1, X_3). The rest of the proof follows. In either cases, $p_{12} \geq p_{13}$, or in generality, $p_{12} \geq p_{1j}$ for $j > 1$. In other words, $p_{12} = p_{1*}$.

p_{12} is an unadjusted pairwise p-value Before conditioning on A_1, the distribution in (2.3) is symmetric around 0 under $\theta_1 = \theta_j$. Since the denominator of p_{12} integrates over half of this symmetric distribution, it is always equal to $1/2$. Thus, the one-sided conditional test at level α is equivalent to the one-sided unadjusted test at level $\alpha/2$, or equivalently the two-sided unadjusted pairwise test at level α.

Modification for counting measure Now suppose the exponential family is defined on the counting measure instead. If ties are broken independently and randomly, the end points on the rays can be considered as "half an atom" if the coordinates are integers (or a smaller fraction of an atom in case of a multi-way tie). The number of atoms on each ray is the same (after the extension $\tilde{A}$) and the atoms on each ray can be paired up in exactly the same way as illustrated in Figure 2.2, with the inequalities above still holding for each pair of the atoms. Summing these inequalities yields our desired result. □

Power Comparison in the Multinomial Case

As the construction of this test follows Fithian, Sun, and Taylor (2014), it uses UMPU selective level-α tests for the pairwise p-values. This section compares the power of our procedure to the best previously known method for verifying multinomial ranks, by Gupta and Nagel (1967). They devise a rule to select a subset that includes the maximum π_j. In other words, if the selected subset is $J(X)$, it guarantees

$$\mathbb{P}\left[\arg\max_j \pi_j \in J(X)\right] \geq 1-\alpha. \tag{2.6}$$

This is achieved by finding an integer d, as a function on m, n and α, and selecting the subset

$$J(X) = \left\{j : X_j \geq \max_k X_k - d\right\}.$$

We take $d(m,n,\alpha)$ to be the smallest integer such that (2.6) holds for any π; Gupta and Nagel (1967) provide an algorithm for determining d.

Subset selection is closely related to testing whether the winner is the best. In particular, we can define a test that declares j the best whenever $J(X) = \{j\}$. If $J(X)$ satisfies (2.6), this test is valid at level α. We next compare the power of the resulting test against the power of our Procedure 1 in a multinomial example with $\pi \propto \left(e^\delta, 1, \ldots, 1\right)$, for several combinations of m and n.

Figure 2.3 gives the power curves for $\text{Multinomial}(m, \pi)$ and

$$\pi \propto \left(e^\delta, 1, \ldots, 1\right),$$

for various combinations of m and n. For their method, we use $\alpha = 0.05$; but in light of the extra factor of $\frac{n-1}{n}$ in (2.2), we will apply the selective procedure with $\frac{n}{n-1}\alpha$ such that the marginal type I error rate of both procedures are controlled at α. Their test coincides with our test at $n = 2$; however as n grows, the selective test shows significantly more power than Gupta and Nagel's test.

To interpret, e.g., the upper right panel of Figure 2.3, suppose that in a poll of $m = 50$ respondents, one candidate enjoys 30% support and the other $n - 1 = 9$ split the remainder ($\delta = \log\frac{0.3}{0.7/9} \approx 1.35$). Then our procedure has power approximately 0.3 to detect the best candidate, while Gupta and Nagel's procedure has power around 0.1.

To understand why our method is more powerful, note that both procedures operate by comparing $X_{[1]} - X_{[2]}$ to some threshold, but the two methods differ in how that

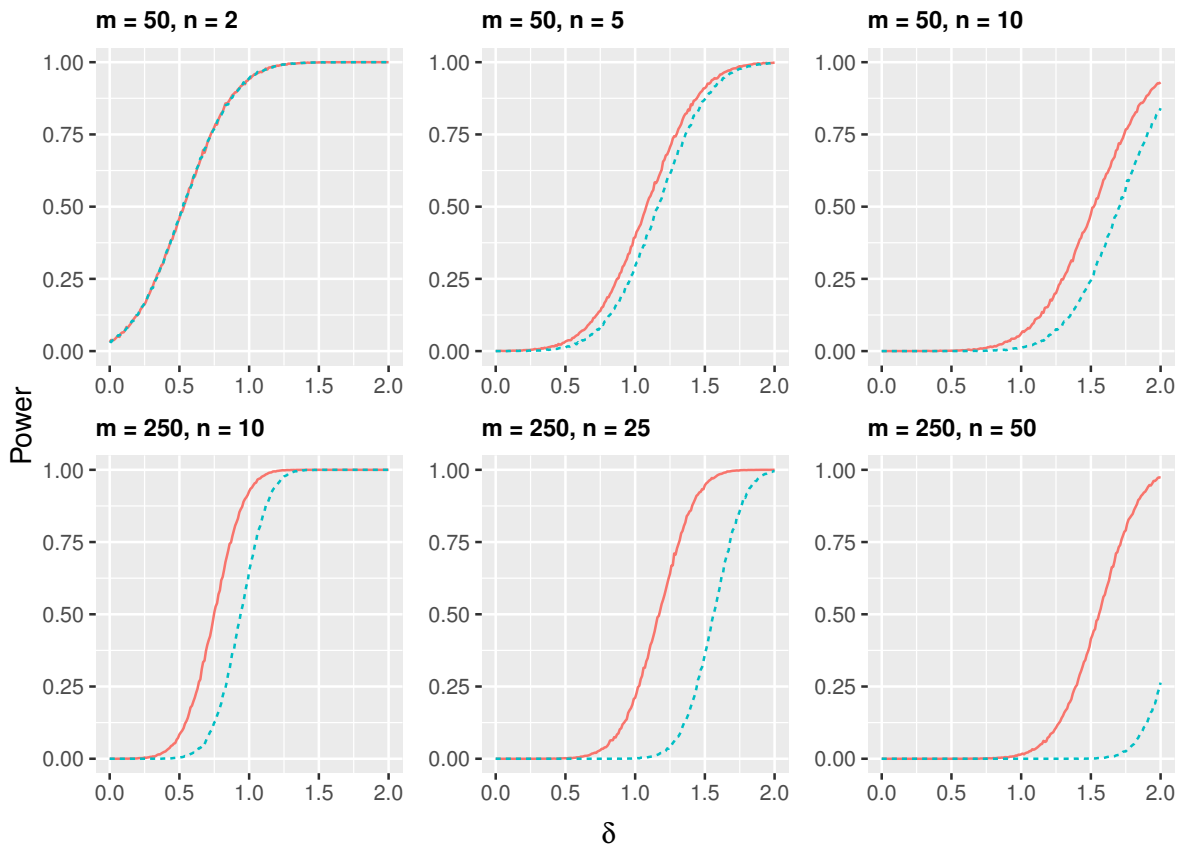


Figure 2.3: Power curves as a function of δ. The plots in the first row all have $m = 50$ and the second row $m = 250$. The solid line and the dashed line are the power for the selective test and Gupta and Nagel's test, respectively.

threshold is determined. The threshold from Gupta and Nagel (1967) is fixed and depends on m and n alone, whereas in our procedure the threshold depends on $X_{[1]} + X_{[2]}$, a data-adaptive choice.

The difference between the two methods is amplified when n is large and $\pi_{(1)} \ll 1/2$. In that case, d from Gupta and Nagel is usually computed based on the worst-case scenario $\pi = \left(\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0\right)$; i.e. d is the upper α quantile of

$$X_1 - X_2 \sim m - 2 \cdot \text{Binomial}\left(m, \frac{1}{2}\right) \approx \text{Normal}\left(0, m\right).$$

Thus $d \approx \sqrt{m} z_\alpha$, where z_α is the upper α quantile of a standard Gaussian. On the other hand, our method defines a threshold based on the upper $\frac{n}{n-1} \cdot \frac{\alpha}{2}$ quantile of

$$X_1 - X_2 \mid X_1 + X_2 \sim X_1 + X_2 - 2 \cdot \text{Binomial}\left(X_1 + X_2, \frac{1}{2}\right),$$

which is approximately $\sqrt{X_1 + X_2} z_{\alpha/2}$. If $\pi_{(1)} \ll 1/2$ then with high probability $X_1 + X_2 \ll m$, making our test much more liberal.

2.4 Confidence Bounds on Differences: By How Much?

By generalizing the above, we can construct a lower confidence bound for $\theta_{[1]} - \max_{j \neq [1]} \theta_j$. Here we provide a more powerful Procedure 2' first. We will proceed by inverting a statistical test of the hypothesis $H^{\delta}_{0[1]} : \theta_{[1]} - \max_{j \neq [1]} \theta_j \leq \delta$, which can be written as a union of null hypotheses:

$$H^{\delta}_{0[1]} = \bigcup_{j \neq [1]} H_{0[1]j} : \theta_{[1]} - \theta_j \leq \delta.$$

By Lemma 4, we can construct selective exact one-tailed p-values $p^{\delta}_{[1]j}$ for each of these by conditioning on $A_{[1]}$, $M_{[1]j}$ and $X_{\setminus\{[1],j\}}$, giving us an exact test for $H_{0[1]}$ by rejecting whenever $\max_{j \neq [1]} p^{\delta}_{[1]j} < \alpha$.

Theorem 6. *The p-values constructed above satisfy $p^{\delta}_{[1][2]} \geq p^{\delta}_{[1]j}$ for any $j \neq [1]$.*

Proof. Again we start with assuming $X_1 \geq X_2 \geq \max_{j>2} X_j$ for convenience. The p-values in question are derived from the conditional law

$$\mathcal{L}_{\theta_1 - \theta_j = \delta} \left(D_{1j} \mid M_{1j}, X_2, \ldots, X_n, A \right),$$

which is the truncated distribution

$$\begin{aligned} p\left(d_{1j}\right) &\propto \exp\left(\left(\theta_1 - \theta_j\right) d_{1j} + \theta_2 X_2 + \cdots + \left(\theta_1 + \theta_j\right) M_{1j} + \cdots + \theta_n X_n\right) \\ &\quad g\left(M_{1j} + d_{1j}, X_2, \ldots, M_{1j} - d_{1j}, \ldots X_n\right) 1_{A_1} \\ &\propto \exp\left(\delta d_{1j}\right) g\left(M_{1j} + d_{1j}, X_2, \ldots, M_{1j} - d_{1j}, \ldots X_n\right) 1_{A_1}. \end{aligned}$$

The p-values thus are

$$p^{\delta}_{1j} = \frac{\int_{D_{1j}}^{\infty} \exp\left(\delta z\right) g\left(M_{1j} + z, X_2, \ldots, M_{1j} - z, \ldots, X_n\right)\, dz}{\int_{\max\{X_2 - M_{1j}, 0\}}^{\infty} \exp\left(\delta z\right) g\left(M_{1j} + z, X_2, \ldots, M_{1j} - z, \ldots, X_n\right)\, dz}.$$

As before in Part 1 of Theorem 1, the conditioning reduces to the case where $n = 3$. Once again it is sufficient to show that $p_{12} \geq p_{13}$. We have the same two cases. If $X_2 < M_{13}$, then

$$\begin{aligned} p^{\delta}_{12} &= \frac{\int_0^{\infty} \exp\left(\delta\left(z + D_{12}\right)\right) g\left(X_1 + z, X_2 - z, X_3\right)\, dz}{\int_{-D_{12}}^{\infty} \exp\left(\delta\left(z + D_{12}\right)\right) g\left(X_1 + z, X_2 - z, X_3\right)\, dz} \\ &= \frac{\int_0^{\infty} \exp\left(\delta z\right) g\left(X_1 + z, X_2 - z, X_3\right)\, dz}{\int_{-D_{12}}^{\infty} \exp\left(\delta z\right) g\left(X_1 + z, X_2 - z, X_3\right)\, dz} \\ p^{\delta}_{13} &= \frac{\int_0^{\infty} \exp\left(\delta\left(z + D_{13}\right)\right) g\left(X_1 + z, X_2, X_3 - z\right)\, dz}{\int_{-D_{13}}^{\infty} \exp\left(\delta\left(z + D_{13}\right)\right) g\left(X_1 + z, X_2, X_3 - z\right)\, dz} \\ &= \frac{\int_0^{\infty} \exp\left(\delta z\right) g\left(X_1 + z, X_2, X_3 - z\right)\, dz}{\int_{-D_{13}}^{\infty} \exp\left(\delta z\right) g\left(X_1 + z, X_2, X_3 - z\right)\, dz}. \end{aligned}$$

The same argument in Figure 2.2 shows that $p^{\delta}_{12} \geq p^{\delta}_{13}$. This is again true for the case where $X_2 \geq M_{13}$ as well. □

In other words, Procedure 2' can be summarized as: Find the minimum δ such that $p^{\delta}_{[1][2]} \leq \alpha$. And by construction, Procedure 2' gives exact $1-\alpha$ confidence bound for $\theta_{[1]} - \max_{j \neq [1]} \theta_j$.

Part 2 of Theorem 1. *Assume the model* (2.1) *holds and* $g(x)$ *is a Schur-concave function. Procedure 2 (the lower bound of unadjusted pairwise confidence interval) gives a conservative* $1-\alpha$ *lower confidence bound for* $\theta_{[1]} - \max_{j \neq [1]} \theta_j$.

Proof. When Procedure 2 reports $-\infty$ as a confidence lower bound, it is definitely valid and conservative. It remains to show that when Procedure 2 reports a finite confidence lower bound, it is smaller than the confidence lower bound reported by Procedure 2'.

If Procedure 2 reports a finite confidence lower bound δ^*, then $\delta^* \geq 0$. Also

$$\frac{\alpha}{2} = \frac{\int_{D_{12}}^{\infty} \exp(\delta^* z)\, g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz}{\int_{-\infty}^{\infty} \exp(\delta^* z)\, g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz} \tag{2.7}$$

as Procedure 2 is constructed from an unadjusted two-tail pairwise confidence interval. However, as $\delta^* \geq 0$, we have

$$\frac{\int_{-\infty}^{0} \exp(\delta^* z)\, g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz}{\int_{0}^{\infty} \exp(\delta^* z)\, g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz} \leq 1$$
$$\frac{\int_{-\infty}^{\infty} \exp(\delta^* z)\, g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz}{\int_{0}^{\infty} \exp(\delta^* z)\, g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz} \leq 2.$$

Multiplying this to (2.7), we have

$$\alpha \geq \frac{\int_{D_{12}}^{\infty} \exp(\delta^* z)\, g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz}{\int_{0}^{\infty} \exp(\delta^* z)\, g(M_{12}+z, X_2, \ldots, M_{12}-z, \ldots, X_n)\, dz},$$

indicating that δ^* is smaller than the confidence bound that Procedure 2' would report. Hence δ^* is a valid and conservative. □

Note that Procedure 2 reporting $-\infty$ in case of $\delta^* \leq 0$ is rather extreme. In reality, we can always just adopt Procedure 2' in the case when Procedure 1 rejects. In fact, by Procedure 2', the multinomial example for polling in Section 2.1 can give a stronger lower confidence bound, that $\pi_{\text{Trump}} / \max_{j \neq \text{Trump}} \pi_j \geq 1.108$ (Trump leads the field by at least 10.8%).

2.5 Verifying Other Ranks: Is the Runner-Up Really the Second Best, etc.?

Often we will be interested in verifying ranks beyond the winner. More generally, we could imagine declaring that the first j populations are all in the correct order, that is

$$\theta_{[1]} > \cdots > \theta_{[j]} > \max_{k>j} \theta_{[k]}. \tag{2.8}$$

Let j_0 denote the largest j for which (2.8) is true. Note that j_0 is both random and unknown, because it depends on both the data and population ranks. Procedure 3 declares that $j_0 \geq j$ if the unadjusted pairwise tests between $X_{[k]}$ and $X_{[k+1]}$, reject at level α for *all* of $k = 1, \ldots, j$.

In terms of the Iowa polling example of Section 2.1, we would like to produce a statement of the form "Trump has the most support, Cruz has the second-most, and Rubio has the third-most." Procedure 3 performs unadjusted pairwise tests to ask if Cruz is really the runner-up upon verifying that Trump is the best, and if Rubio is really the second runner-up upon verifying that Cruz is the runner-up, etc., until we can no longer infer that a certain population really holds its rank.

While we aim to declare more populations to be in the correct order, declaring too many populations, i.e. out-of-place populations, to be in the right order is undesirable. It is possible to consider false discovery rate (the expected portion of out-of-place populations declared) here, but we restrict our derivation to FWER (the probability of having any out-of-place populations declared).

Formally, let $\hat{j}_0$ denote the number of ranks validated by a procedure (the number of rejections). Then the FWER of $\hat{j}_0$ is the probability that too many rejections are made; i.e. $\mathbb{P}\left[\hat{j}_0 > j_0\right]$. For example, suppose that the top three data ranks and population ranks coincide, but not the fourth ($j_0 = 3$). Then we will have made a Type I error if we declare that the top five ranks are correct ($\hat{j}_0 = 5$), but not if we declare that the top two are correct ($\hat{j}_0 = 2$). In other words, $\hat{j}_0$ is a lower confidence bound for j_0.

To show that Procedure 3 is valid, we will prove the validity of a more liberal Procedure 3', described in Algorithm 1. Procedure 3 is equivalent to Procedure 3' for the most part, except that Procedure 3 conditions on a larger event $\left\{X_{[j]} \geq \max_{k>j} X_{[k]}\right\}$ in Line 7.

Theorem 7. *Procedure 3' is a stepwise procedure that an estimate $\hat{j}_0$ of j_0 at the FWER controlled at α, where j_0 is given by*

$$j_0 = \max_j \left\{\theta_{[1]} > \cdots > \theta_{[j]} > \max_{k>j} \theta_{[k]}\right\}.$$

Proof. We will first show that Procedure 3' falls into the sequential goodness-of-fit testing framework proposed by Fithian, Taylor, and Tibshirani (2015). We thus analyze Procedure 3' as a special case of the BasicStop procedure on random hypothesis, described in the same paper. This enables us to construct valid selective p-values and derive Procedure 3'.

Application of the sequential goodness-of-fit testing framework Upon observing $X_{[1]} \geq \cdots \geq X_{[n]}$, we can set up a sequence of nested models

$$\mathcal{M}_1(X) \subseteq \cdots \subseteq \mathcal{M}_n(X), \qquad \text{where } \mathcal{M}_j(X)^c = \left\{\theta : \theta_{[1]} > \cdots > \theta_{[j]} > \max_{k>j} \theta_{[k]}\right\}.$$

Algorithm 1. Procedure 3', a more liberal version of Procedure 3

input : $X_1, \ldots, X_n$
output: $\hat{j}_0$, an estimate for j_0
`# Initialization`
$\tau_j \leftarrow [j]$;
`# Consider` τ_j `as part of the observation and the fixed realization of the random index` $[j]$
$X_{\tau_0} \leftarrow \infty$;
$j \leftarrow 0$;
rejected $\leftarrow$ **true**;
while rejected **do**
  $j \leftarrow j + 1$;
  $D_{\tau_j} \leftarrow X_{\tau_j} - X_{\tau_{j+1}}$;
  Set up the distribution of $D_{\tau_j \tau_{j+1}}$, conditioned on
    - the variables $X_{\tau_1}, \ldots, X_{\tau_{j-1}}, X_{\tau_{j+2}}, \ldots, X_{\tau_n]}$, and
    - the event $\left\{X_{\tau_{j-1}} \geq X_{\tau_j} \geq \max_{k>j} X_{\tau_k)}\right\}$;

  `# The distribution of` $D_{\tau_j \tau_{j+1}}$ `depends only on` $\theta_{\tau_j} - \theta_{\tau_{j+1}}$ `now`
  test $H_0 : \theta_{\tau_j} - \theta_{\tau_{j+1}} \leq 0$ against $H_1 : \theta_{\tau_j} - \theta_{\tau_{j+1}} > 0$ according to the distribution of $D_{\tau_j \tau_{j+1}}$;
  Set rejected as the output of the test;
end
$\hat{j}_0 \leftarrow j - 1$;

If we define the j-th null hypothesis as

$$\widetilde{H}_{0j} : \theta_{[j]} \leq \max_{k>j} \theta_{[k]},$$

then $\widetilde{H}_{01}, \ldots, \widetilde{H}_{0j}$ are all false if and only if $\theta \notin \mathcal{M}_j(X)$.

In other words, $\mathcal{M}_j(X)$ is a family of distributions that does not have all first j ranks correct. As we will see later, each step in Procedure 3' is similar to testing $\widetilde{H}_{0j}$, stating that without the first j ranks correct, it is hard to explain the observations. Thus, returning $\hat{j}_0 = j$ amounts to rejecting $\widetilde{H}_{01}, \ldots, \widetilde{H}_{0j}$, or equivalently determining that the models $\mathcal{M}_1(X), \ldots, \mathcal{M}_j(X)$ do *not* fit the data.

While the null hypotheses $\widetilde{H}_{0j}$ provided intuition in the setting up the nested models, they are rather cumbersome to work with. Inspired by Fithian, Taylor, and Tibshirani (2015), we will instead consider another sequence of random hypothesis that are more closely related to the nest models,

$$H_{0j} : \theta \in \mathcal{M}_j(X),$$

or equivalently, that $\theta_{[1]}, \ldots, \theta_{[j]}$ are not the best j parameters in order.

Adapting this notation, the FWER can be viewed as $\mathbb{P}\left[\text{reject } H_{0(j_0+1)}\right]$.

Special case of the BasicStop procedure While impractical, Procedure 3' can be thought of as performing all n tests first, producing a sequence of p-values p_j, and returning

$$\hat{j}_0 = \min\{j : p_j > \alpha\} - 1. \tag{2.9}$$

This is a special case of the BasicStop procedure. Instead of simply checking that Procedure 3' fits all the requirement for FWER control in BasicStop, we will give the construction of Procedure 3', assuming that we are to estimate j_0 with BasicStop.

In general, the FWER for BasicStop can be rewritten as $\mathbb{P}[p_{j_0+1} \leq \alpha]$. This is however difficult to analyze, as j_0 itself is random and dependent on X, thus we break the FWER down as follows:

$$\begin{aligned}
\mathbb{P}[p_{j_0+1} \leq \alpha] &= \sum_j \mathbb{P}[p_{j_0+1} \leq \alpha \mid j_0 = j]\,\mathbb{P}[j_0 = j] \\
&= \sum_j \mathbb{P}[p_{j+1} \leq \alpha \mid j_0 = j]\,\mathbb{P}[j_0 = j] \\
&= \sum_j \mathbb{P}[p_{j+1} \leq \alpha \mid \theta \in \mathcal{M}_{j+1}(X) \setminus \mathcal{M}_j(X)]\,\mathbb{P}[j_0 = j].
\end{aligned}$$

We emphasize here that θ is *not* random, but $\mathcal{M}_{j+1}$ is. Thus it suffices to construct the p-values such that

$$\mathbb{P}[p_j \leq \alpha \mid \theta \in \mathcal{M}_j(X) \setminus \mathcal{M}_{j-1}(X)] \leq \alpha \quad \text{for all } j. \tag{2.10}$$

Considerations for conditioning By smoothing, we are free to condition on additional variables in (2.10). A logical choice that simplified (2.10) is conditioning on the variables $\mathcal{M}_{j-1}(X)$ and $\mathcal{M}_j(X)$. Note that the choice of the model $\mathcal{M}_j(X)$, once again, based solely on the random indices $[1], \ldots, [j]$, so conditioning on both $\mathcal{M}_{j-1}(X)$ and $\mathcal{M}_j(X)$ is equivalent to conditioning on the random indices $[1], \ldots, [j]$, which in turns is equivalent to conditioning on the σ-field generated by the partition of the observation space X

$$\left\{ \left\{ X_{\tau_1} \geq \cdots \geq X_{\tau_j} \geq \max_{k>j} X_{\tau_k} \right\} : \tau \text{ is any permutation of } (1, \ldots, n) \right\},$$

or colloquially, the set of all possible choices of $[1], \ldots, [j]$. Within each set in this partition, the event $\{\theta \in \mathcal{M}_j(X) \setminus \mathcal{M}_{j-1}(X)\}$ is simply $\left\{\theta_{\tau_1} > \cdots > \theta_{\tau_j} \text{ and } \theta_{\tau_j} \leq \max_{k>j} \theta_{\tau_k}\right\}$, a trivial event.

As a brief summary, we want to construct p-values p_j such that

$$\mathbb{P}_{\substack{\theta_{\tau_1} > \cdots > \theta_{\tau_j} \\ \theta_{\tau_j} \leq \max_{k>j} \theta_{\tau_k}}} \left[p_j \leq \alpha \;\middle|\; X_{\tau_1} \geq \cdots \geq X_{\tau_j} \geq \max_{k>j} X_{\tau_k} \right].$$

Construction of the p-values To avoid the clutter in the subscripts, we will drop the τ in the subscript. Hence our goal is now

$$\mathbb{P}_{\substack{\theta_1>\cdots>\theta_j \\ \theta_j\le\max_{k>j}\theta_k}}\left[p_j \le \alpha \;\middle|\; X_1 \ge \cdots \ge X_j \ge \max_{k>j} X_k\right]$$

Construction of p_j for other permutations τ can be obtained similarly.

There are many valid options for p_j (such as constant α). We will follow the idea in the proof of Part 1 of Theorem 1 here. p_j is intended to test $H_{0j} : \theta \in \mathcal{M}_j(X)$, which is equivalent to the union of the null hypotheses:

1. $\theta_k \le \theta_{k+1}$ for $k = 1, \ldots, j-1$, and
2. $\theta_j \le \theta_k$ for $k = j+1, \ldots, n$. (The union of these null hypotheses is $\widetilde{H}_{0j}$.)

Since the joint distribution of X, restricted to $\{X_1 \ge \cdots \ge X_j \ge \max_{k>j} X_k\}$, remains in the exponential family, we can construct the p-values for each of the hypotheses above by conditioning on the variables corresponding to the nuisance parameters here, similar to the proof of Part 1 of Theorem 1. Then we can take p_j as the maximum of such p-values.

For the hypothesis $H_{0jk} : \theta_j \le \theta_k$, we can construct p_{jk}, by considering the survival function of the conditional law

$$\begin{aligned}
&\mathcal{L}_{\theta_j=\theta_k}\left(D_{jk} \;\middle|\; \left\{X_1 \ge \cdots \ge X_j \ge \max_{\ell>j} X_\ell\right\}, X_{\setminus\{j,k\}}, M_{jk}\right) \\
&= \mathcal{L}_{\theta_j=\theta_k}\left(D_{jk} \;\middle|\; \left\{X_{j-1} \ge X_j \ge \max_{\substack{\ell>j \\ \ell\ne k}} X_\ell \text{ and } X_j \ge M_{jk}\right\}, X_{\setminus\{j,k\}}, M_{jk}\right)
\end{aligned}$$

Once again, $X_{j+1} = \max_{\ell>j} X_\ell$ is simply shorthand for simplifying our notation. Now the p-values are similar to the ones in Equation (2.4), for $k > j$:

$$p_{jk} = \frac{\int_{D_{jk}}^{X_{j-1}} g\left(X_1, \ldots, M_{jk}+z, \ldots, M_{jk}-z, \ldots, X_n\right)\, dz}{\int_{\max\left\{X_{j+1}-M_{jk},0\right\}}^{X_{j-1}} g\left(X_1, \ldots, M_{jk}+z, \ldots, M_{jk}-z, \ldots, X_n\right)\, dz}.$$

We can graphically represent p_{jk} in Figure 2.4, a diagram analogous to Figure 2.2.

We have $p_{j(j+1)} \ge \max_{k>j} p_{jk}$ by Section 2.3: the upper truncation for X_j can be represented by cropping Figure 2.2 along a vertical line, shown in Figure 2.4. Considering $p_{j(j+1)}$ is sufficient in rejecting all the H_{0jk}. We will take $p_{j*} = p_{j(j+1)}$, noting that this is the p-value that Procedure 3' would produce. In fact, p_{j*} is also the p-value we would have constructed if we were to reject only $\widetilde{H}_{0j}$.

Upon constructing p_j, one should realize that the p-values for testing $\theta_k \le \theta_{k+1}$ would have been constructed in earlier iterations of BasicStop, as p_{k*}. In other words, $p_j = \max_{k\le j} p_{k*}$ is the sequence of p-values that works with BasicStop. However, from (2.9),

$$\hat{j}_0 = \min\left\{j : \max_{k\le j} p_{k*} > \alpha\right\} - 1 = \min\left\{j : p_{j*} > \alpha\right\} - 1,$$

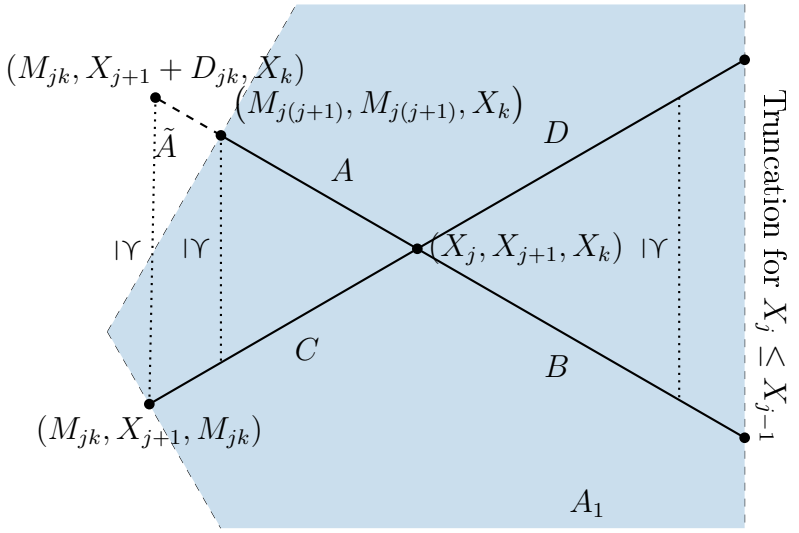


Figure 2.4: The two p-values constructed corresponds to taking integrals of g along these segments, that lie on a level set of $x_j + x_{j+1} + x_k$. The dashed line corresponds to extension in (2.5). The dotted line on the far right is the truncation that enforces $X_j < X_{j-1}$.

so it is safe to apply BasicStop to p_{j*} directly, yielding Procedure 3'. □

Part 3 of Theorem 1. *Assume the model* (2.1) *holds and* $g(x)$ *is a Schur-concave function. Procedure 3 is a conservative stepwise procedure with FWER no larger than* α.

Proof. The p-values $p_{j(j+1)}$ obtained in Procedure 3' are always smaller than their counterpart in Procedure 3, as the upper truncation at X_{j-1} is on the upper tail. Therefore Procedure 3 is conservative and definitely valid. □

2.6 Discussion

Combining ideas from conditional inference and multiple testing, we have proven the validity of several very simple and seemingly "naive" procedures for significance testing of sample ranks. In particular, we have shown that an unadjusted pairwise test comparing the winner with the runner-up is a valid significance test for the first rank. Our result complements and extends pre-exisiting analogous results for location and location-scale families with independence between observations. Our approach is considerably more powerful than previously known solutions. We provide similarly straightforward conservative methods for producing a lower confidence bound for the difference between the winner and runner up, and for verifying ranks beyond the first.

Claims reporting the "winner" are commonly made in the scientific literature, usually with no significance level reported or an incorrect method applied. For example, Uhls and Greenfield (2012) asked $n = 20$ elementary and middle school students which of seven personal values they most hoped to embody as adults, with "Fame" (8 responses) being

the most commonly selected, with "Benevolence" (5 responses) second. The authors' main finding — which appeared in the abstract, the first paragraph of the paper, and later a CNN.com headline (Alikhani, 2011) — was that "Fame" was the most likely response, accompanied by a significance level of 0.006, which the authors computed by testing whether the probability of selecting "Fame" was larger than 1/7. The obvious error in the authors' reasoning could have been avoided if they had performed an equally straightforward two-tailed binomial test of "Fame" vs. "Benevolence," which would have produced a p-value of 0.58.

Reproducibility

A git repository containing with the code generating the image in this chapter is available at `https://github.com/kenhungkk/verifying-winner`.

Chapter 3

Replicability Assessment

3.1 Introduction

Replicability crisis

Growing concerns about selection bias, p-hacking, and other questionable research practices (QRPs) have raised urgent questions about the reliability of scientific findings. While concerns about replicability cut across scientific disciplines, psychologists have led large-scale efforts to assess the replicability of their own field. The largest and most systematic of these efforts has been the Reproducibility Project: Psychology (RP:P),[1] a major collaboration by several hundred psychologists to replicate a representative sample of 100 studies published in 2008 in three top psychology journals, *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*.[2]

While the RP:P dataset is an invaluable resource, scientists disagree on how to quantify or measure replicability (Amrhein, Korner-Nievergelt, and Roth, 2017; Goodman, Fanelli, and Ioannidis, 2016). Open Science Collaboration (OSC; 2015) reported three main metrics: it found that 64% (= 1 − 36%) of the replication studies did not find statistically significant results in the same direction as the original studies, that 53% (= 1 − 47%) of 95% confidence intervals for the replication studies do not contain the point estimates for their corresponding original studies, and that 83% of the effect size estimates declined from original studies to replications. All three summary statistics were widely reported as indicating a dire crisis for the credibility of experimental psychology research. For example, the *Washington Post* reported that RP:P "affirms that the skepticism [of published results] was warranted" (Achenbach, 2015); the *Economist* noted that OSC "managed to replicate satisfactorily the results of only 39% of the studies investigated" (2016); and the *New York Times* reported that "more than half of the

[1] In some parts of the literature, "reproducibility" has taken on a computational connotation, meaning only that other scientists can repeat the analysis using the original study's data; we will lean toward the more unambiguous term "replicability."

[2] The test statistics, effect sizes and most pertinent information are all publicly available on at the Open Science Foundation website at `https://osf.io/ezcuj/`.

findings did not hold up when retested" (Carey, 2015).

This negative gloss was challenged in a comment by Gilbert et al. (2016b), who criticized both the fidelity of some of the replications' experimental designs and the aptness of the metrics reported by Open Science Collaboration (2015). In particular, Gilbert et al. pointed out that, because there is sampling error in the replication point estimates, we should not expect 95% of the estimates to fall into the replication confidence intervals even under ideal conditions. Moreover, any small or large variations in the true effect sizes between the original and replication studies could further deflate the expected fraction of "successful replications," as measured in this way. Gilbert et al. concluded that "OSC seriously underestimated the reproducibility of psychological science," sparking further debate between defenders of OSC's conclusions (Anderson et al., 2016; Nosek and Gilbert, 2016; Srivastava, 2016) and the critics (Gilbert et al., 2016a,c).[3]

To determine whether OSC truly underestimated replicability, we must first pin down the rather slippery question of what "replicability" actually is. Although the three metrics used by OSC are simply descriptive statistics that do not purport to estimate any explicitly defined underlying quantity, we can loosely characterize the 64%, 53% and 83% numbers respectively as qualitative answers to three questions:

False directional claims. *What fraction of the original studies were erroneous* in claiming that the true effect was nonzero, in the claimed direction (positive or negative)? Gelman and Tuerlinckx (2000) called such mistakes *type S* errors.

Effect shift. *How much do the effect sizes shift* from the original study to the replication study? We call the discrepancy between the original and replication effect *effect shift*.

Effect decline. *What fraction of the effect sizes decline?* More precisely, what fraction of the true effect sizes shift in a direction opposite to the original claims when the studies were replicated, and by how much?

The first question concerns a type of *false discovery rate* (FDR) of the statistical hypotheses, viewing the field of social psychology as a collective enterprise in large-scale multiple testing: it quantifies the fraction of findings that would be confirmed if the exact same studies could be carried out again with much larger samples from the same populations. The second question concerns a basic form of repeatability: whether scientists are typically successful in closely replicating each others' experimental conditions, so that the true effect being measured is stable across different experiments. The third question builds upon the second question: whether true effect sizes tend systematically to attenuate in replications. An overall trend of declining true effects could suggest various interpretations, including systematic biases in the original experiments or failures by the replication teams to reproduce key experimental conditions that produced the original effects.

[3]While much of the ensuing discussion focused on the question of whether the confidence interval metric 53% is too pessimistic, analogous criticisms apply to the "significant replications" metric of 64% as well: the replication studies could be underpowered even when a true effect is present.

As we will see, however, none of the three reported metrics can be taken at face value as *estimates* of the answers to the corresponding questions, due to the confounding factor of pervasive selection bias. By using techniques from multiple testing and post-selection inference, we will develop methods to rigorously address these questions without assuming a model for the prior distribution of effect sizes. For the RP:P data we estimate the rate of false directional claims at roughly 32% among studies with $p < 0.05$, which would be considered unacceptably high in most multiple testing applications. By contrast, among studies with $p < 0.005$, a lower threshold proposed by Benjamin et al. (2018), our estimate drops to 7%, with an upper confidence bound of 18%. We also compute confidence intervals for the effect shift in each individual study pair and find that, after adjusting for multiplicity, about 11% of the intervals exclude zero, an idealized null hypothesis of perfect replication. For effect decline, we find in aggregate that 35% of the true effects declined, and 35% declined by at least 20%.

In addressing each question, we define our estimands in terms of the true effects present in the statistical populations actually sampled in each study. Because some studies may be biased or lack external validity — for example, because of flaws in the study design, or because survey participants are unrepresentative of the broader population of scientific interest — these effect sizes may not reflect the latent scientific quantities the experiments purport to measure. Uncovering such discrepancies is beyond the reach of data analysis alone, but we should keep them in mind as we interpret the results.

The role of selection bias

The RP:P data shows unmistakable signs of selection for statistically significant findings in the original experiments: 91 of the 100 results replicated by OSC were statistically significant at the 0.05 level in the original study and four of the others had "marginally significant" p-values between 0.05 and 0.06. This is due partly to publication bias (that the studies might not have been published, or the results discussed, if the p-values had not been significant), but also partly to OSC's method for choosing which results to replicate. Each OSC replication team selected a "key result" from the last experiment presented in the original paper, and evidently most teams chose a significant finding as the key result (justifiably so, since positive results usually draw the most attention from journal readers and the outside world). Figure 3.1 shows the empirical distribution of p-values from the original and replication studies.

The resulting selection bias in the original studies leads to many well-known and predictable pathologies, such as systematically inflated effect size estimates, undercoverage of (unadjusted) confidence intervals, and misleading answers from unadjusted meta-analyses. Indeed, most of the phenomena reported by OSC, including the three metrics discussed above, could easily be produced by selection bias alone. This would be true *even if there are few false directional claims, all replications are exact, and true effects do not decline*, as illustrated in the following simulation study.

Example 4. Consider a stylized setting where all experiments (both original and replication) have an identical effect size θ, producing an unbiased Gaussian estimate with

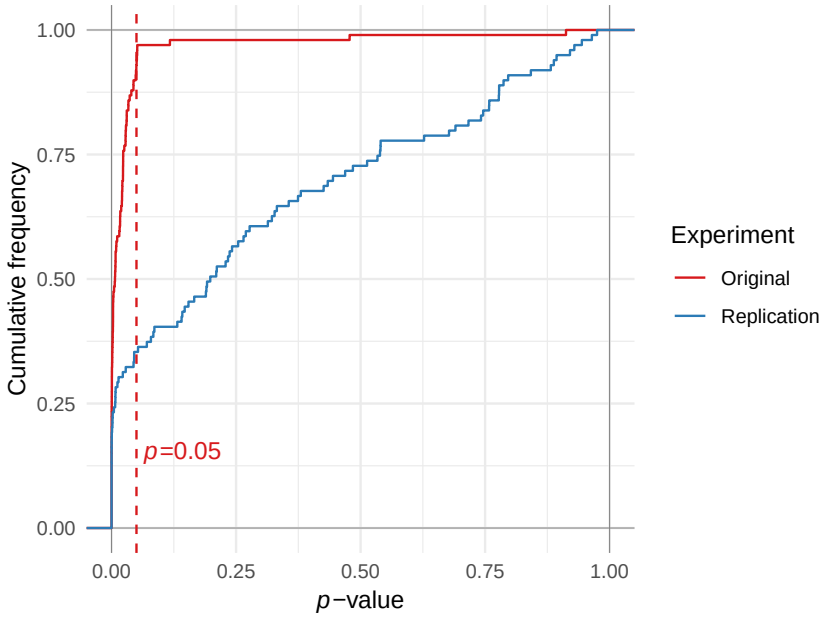


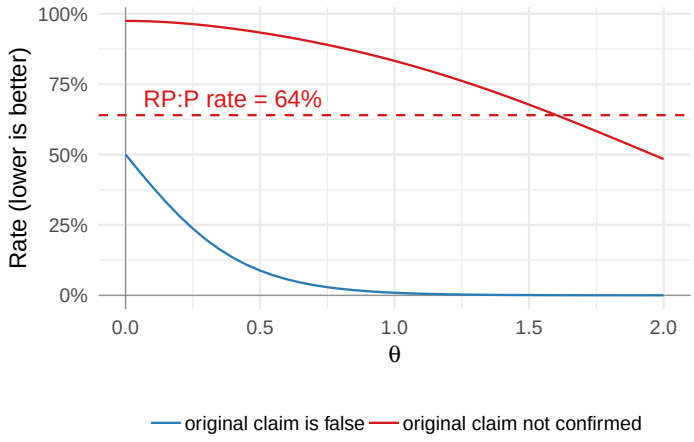
Figure 3.1: The empirical distribution of the original and replication p-values. Nearly all of the original p-values (in red) are smaller than 0.05.

standard error 1. Assume, however, that we observe only study pairs for which the original study is significant at level 0.05.

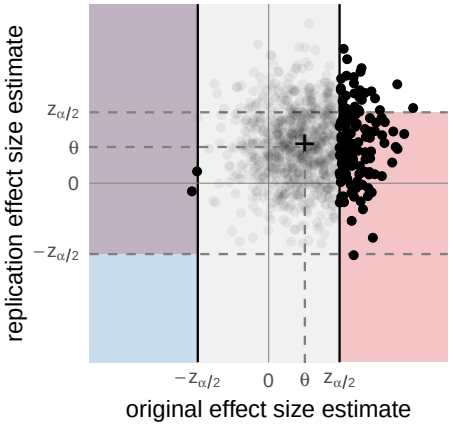
Figure 3.2a shows the expected fraction of replication studies which are not statistically significant in the same direction as the corresponding original studies, as a function of effect size θ, along with the true proportion of false directional claims; or type S errors. Even when the true error rate is low, e.g. at $\theta = 1$ as shown in Figure 3.2b, the proportion of replications reporting the same directional findings as the original studies can remain low.

Likewise, we simulate the expected fraction of 95% replication confidence intervals that fail to cover their original point estimates in Figure 3.3 and the expected fraction of effect sizes that decline in Figure 3.4. In both cases, we see that selection bias is more than sufficient to produce the metrics in RP:P, even in our idealized simulation with exact replications and relatively few type S errors.

Because selection bias could, in principle, provide a sufficient explanation for the metrics reported in RP:P, those metrics do not, in and of themselves, provide any evidence of any other problems. In particular, they shed no light on whether the FDR is actually high, or how much the effect sizes shifted, or whether effect sizes tend to decline. Nor do they provide evidence for any competing accounts of the replication crisis, such as QRPs like p-hacking, high between-study variability in effect sizes, or systematic biases in the original studies. To discern anything about other explanations, we must adjust for the pervasive effects of selection bias.

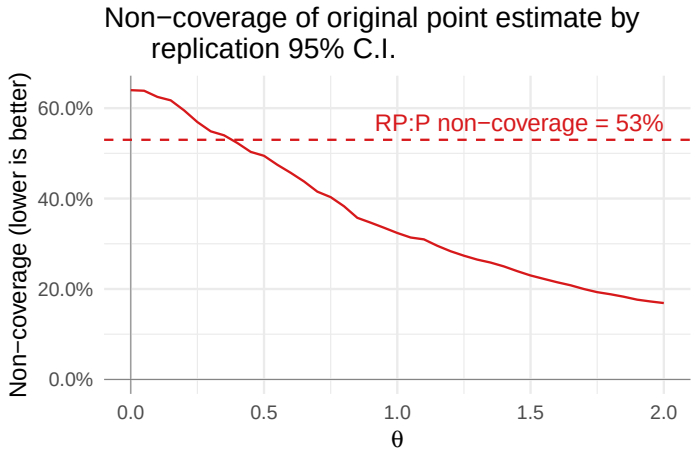


(a) The expected fraction of replications that do not confirm (at level 0.05) the original directional claim (red), and the proportion of false directional claims in the original studies (blue), as a function of effect size θ. For small θ, the fraction of replications that do not confirm the claims in the original studies may dramatically overestimate the fraction of false original claims.

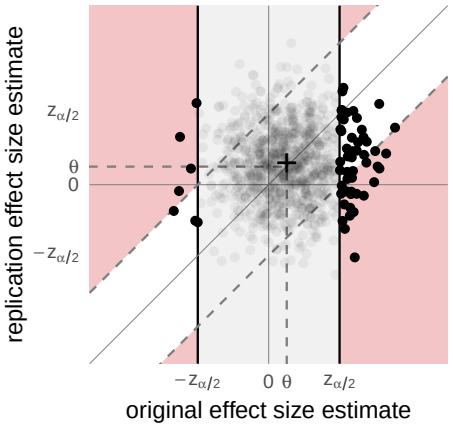


(b) $\theta = 1$. The gray region is unobserved. For points in the red region, the replication does not confirm the original directional claim, and for points in the blue region, the original claim is directionally false. The red and blue regions overlap in the purple region.

Figure 3.2

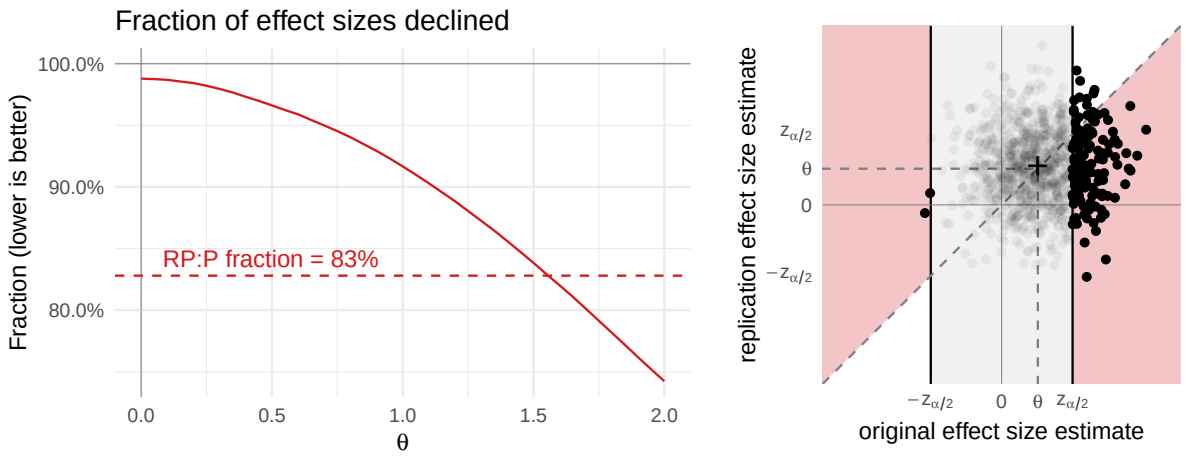


(a) Expected fraction of original point estimates falling outside the replication confidence interval, as a function of effect size θ. For small θ, the fraction of original point estimates falling outside the replication 95% confidence intervals can easily exceed the RP:P reported metric of 53%, even when all replications are perfectly exact.



(b) $\theta = 0.5$. The gray region is unobserved. For points in the red region, the original point estimate differs from the replication estimate by more than $z_{\alpha/2}$ and hence the original point estimate falls outside in the replication 95% confidence interval.

Figure 3.3



(a) Expected fraction of effect size point estimates that declined toward zero in replication, as a function of effect size of θ. For small θ, the fraction of effect size estimates declining from original to replication studies can easily exceed the RP:P reported metric of 83%, even when there is no decline in the true effect sizes.

(b) $\theta = 0.5$. The gray region is unobserved. Points in the red region represent declining point estimates in replications. When the original point estimate is positive, a decline is marked by a smaller replication estimate; on the other hand, if the original estimate is negative, a decline is indicated by a larger replication estimate.

Figure 3.4

Another good reason to disentangle selection bias from other sources of error is that the former is, in some sense, the most innocuous explanation for the phenomena observed by OSC while the others present much deeper scientific issues. The technical issues of selection bias can be addressed either retrospectively by statistical adjustments (e.g. Andrews and Kasy, 2018; Duval and Tweedie, 2000; Fithian, Sun, and Taylor, 2014; Hedges, 1992; Simonsohn, Nelson, and Simmons, 2014b), or prospectively with more preregistration or larger sample sizes. By contrast, it would be deeply worrying if psychologists were systematically unable to repeat their colleagues' experiments, or if most published claims about effect sizes were directionally incorrect.

Formalizing replicability

We now introduce a simple formal model for replication studies with selection bias. For study $i = 1, \ldots, m$, let $\theta_{i,O}$ and $\theta_{i,R}$ denote the true effect sizes in the original and the replication studies, respectively. Abstracting away experimental design details, assume that each study pair produces two normally distributed effect size estimators $\hat{\theta}_{i,O}$ and $\hat{\theta}_{i,R}$. Assume additionally that for the study pair to appear in our replication data, $\hat{\theta}_{i,O}$

must be statistically significant at level $\alpha = 0.05$;[4] then for some significance threshold $c > 0$ we have

$$\hat{\theta}_{i,O} \sim N\left(\theta_{i,O}, \sigma^2_{i,O}\right) 1_{\{|\hat{\theta}_{i,O}|>c\}} \qquad \text{and} \qquad \hat{\theta}_{i,R} \sim N\left(\theta_{i,R}, \sigma^2_{i,R}\right), \tag{3.1}$$

with all estimates assumed to be independent of each other. The indicator $1_{\{|\hat{\theta}_{i,O}|>c\}}$ beside the normal distribution in (3.1) means that the distribution of $\hat{\theta}_{i,O}$ has been truncated to the event where $|\hat{\theta}_{i,O}| > c$ and renormalized so that it integrates to 1. For the moment, we assume that the variances $\sigma^2_{i,O}$ and $\sigma^2_{i,R}$ are known; in that case $c = z_{0.05/2}\,\sigma_{i,O}$. We will relax this assumption in Section 3.2.

False directional claims To formalize false directional claims in terms of the parameters of model (3.1), we note that a type S error occurs when a statistically significant finding gets the sign of the parameter wrong:

$$H_i^{S,O} : \ \text{sign}(\theta_{i,O}) \neq \text{sign}(\hat{\theta}_{i,O}), \qquad \text{where } \text{sign}(x) = \begin{cases} +1, & x > 0 \\ -1, & x < 0 \\ 0, & x = 0 \end{cases}.$$

Note that $|\hat{\theta}_{i,O}|$ is always larger than c, so $\text{sign}(\hat{\theta}_{i,O}) \in \{-1, +1\}$. Letting $S_i = \text{sign}(\hat{\theta}_{i,O})$, we can rewrite the hypothesis as

$$H_i^{S,O} : \ S_i \cdot \theta_{i,O} \leq 0.$$

Here $H_i^{S,O}$ is fundamentally data-dependent as it is determined by S_i. Nonetheless it is a meaningful hypothesis: when $S_i = +1$, we want to test the null that $\theta_{i,O} \leq 0$; otherwise we want to test the null that $\theta_{i,O} \geq 0$. Our strategy is to condition on the value of S_i, since the null hypothesis is fixed again once we know S_i. We defer the discussion of valid testing of data-dependent hypotheses for now.

The question of false directional claims, then, boils down to asking how many $H_i^{S,O}$ are true: a multiple testing problem. Our estimand, the proportion of type S errors that occurred, is V/R, where V is the number of type S errors and R is the number of "discoveries," i.e. rejections. If we classify the hypotheses by whether $H_i^{S,O}$ is true and whether the test for $H_i^{S,O}$ is significant, then V and R correspond to the cell counts in Table 3.1.

In the multiple testing literature, V/R is called the *directional false discovery proportion* (directional FDP, or FDP_{dir}), the type S error analog of false discovery proportion (FDP; Benjamini and Hochberg, 2000). In addition to an estimate, we also provide an upper confidence bound for the directional FDP in Section 3.2. Both the estimator and the confidence bound are based on a "p-curve" analysis, i.e. an analysis of the distribution of significant p-values (Simonsohn, Nelson, and Simmons, 2014a). We further modify these methods to evaluate the proposal to lower the statistical significance threshold by Benjamin et al. (2018).

[4] We relax this assumption in Section 3.2.

Original p-value	$H_i^{S,O}$ is true	$H_i^{S,O}$ is false	Total
Significant	V	$*$	R
Not-significant	$*$	$*$	$*$
Total	$*$	$*$	$*$

Table 3.1: Classification of the hypotheses, in the style of Benjamini and Hochberg (1995). Only R is observed and we wish to infer on V.

Although $\hat{\theta}_{i,R}$ is irrelevant to testing $H_i^{S,O}$, it is informative for the closely related question of whether $\hat{\theta}_{i,O}$ incorrectly predicts the direction of the effect in a replication study, i.e.

$$H_i^{S,R} : \ S_i \cdot \theta_{i,R} \leq 0.$$

Note that S_i is computed from the original study, so this hypothesis is a measure of external validity as to the (claimed) directions of effects. If an experimental result has external validity, then any directional claim about the true effect should apply not only to the original study, but also to direct replications thereof. We provide analogous methods for multiple testing of the hypotheses $H_i^{S,R}$.

Effect shift To assess the effect shift in a specific replication attempt, we can test the hypothesis $H_i^E : \ \theta_{i,O} = \theta_{i,R}$ (an exact replication). As Anderson et al. (2016) noted, "there is no such thing as exact replication"; nevertheless, exactness serves usefully as an idealized null hypothesis. By inverting a test for H_i^E we can obtain a predictive interval for $\hat{\theta}_{i,R}$. Furthermore, by inverting tests for a related hypothesis $H_i^{E,\delta} : \ \theta_{i,O} - \theta_{i,R} = \delta$, we obtain a confidence interval for $\theta_{i,O} - \theta_{i,R}$, the effect shift in study i. Our methods explicitly take into account the truncation of $\hat{\theta}_{i,O}$.

Effect size decline The null hypothesis for effect size decline is closely related to effect shift, and can be formalized as the null hypothesis where the true effect size has declined by no more than a fraction $\rho \in [0, 1]$:

$$H_i^{D,\rho} : \ S_i \cdot \theta_{i,R} \geq S_i \cdot (1 - \rho)\theta_{i,O}.$$

If $S_i = +1$ and $\rho = 0.2$, for example, rejecting $H_i^{D,\rho}$ amounts to an assertion that $\theta_{i,R} < 0.8\,\theta_{i,O}$, i.e. the true effect declined by more than 20%, or is negative.

In particular, if $\rho = 0$ then $H_i^{D,0}$ is a one-sided version of H_i^E, and when $\rho = 1$, $H_i^{D,1}$ is equivalent to $H_i^{S,R}$. We can subsequently ask how many of $H_i^{D,\rho}$ are false: another multiple testing problem. We provide two estimators (one overestimate and one underestimate) and confidence interval for the proportion of false $H_i^{D,\rho}$.

Data-dependent hypotheses and conditional inference

Our hypotheses above, $H_i^{S,O}$, $H_i^{S,R}$ and $H_i^{D,\rho}$, are all innately data-dependent. Recall that a test of a data-dependent hypothesis is valid, so long as the type I error rate is controlled

conditioned on the portion of the data that generated the hypothesis. For our hypotheses here, S_i is the part of the data that determines the hypothesis: in effect, we can imagine ourselves in the position of having observed the signs of all the original estimators, but knowing nothing else about the data. At that stage, it is valid to formulate a hypothesis that depends on S_i, and plan to test it using the still-unobserved data: namely, $|\hat{\theta}_{i,O}|$ and $\hat{\theta}_{i,R}$.

After conditioning on S_i, each hypothesis discussed above amounts to testing a fixed linear hypotheses about $(\theta_{i,R}, \theta_{i,O})$, the natural parameter of the truncated bivariate normal model (3.1); as a result, they are all amenable to post-selection inference using the selective z-test built on the work of Lee et al. (2016). Section 3.2 discusses the methodology in detail.

Related work

There has been much commentary on how to define replicability for scientific experiments. Valentine et al. (2011) pointed out that the definition should depend on the scientific context. For example, sometimes one may wish to test the robustness of conclusions to subpopulation differences, but in other times, to changes in experimental conditions. Goodman, Fanelli, and Ioannidis (2016) expanded on this, and gave a few useful definitions for what replicability is, such as *methods reproducibility*, *results reproducibility*, *inferential reproducibility*, etc., but stopped short of an operational statistical criterion for replicability. False directional claims and effect shift can be loosely interpreted as inferential and results reproducibility, respectively.

Operationally, Valentine et al. (2011) and Nosek and Errington (2017) proposed the metrics used in RP:P and Camerer et al. (2018), a similar replication effort in experimental economics. These metrics however suffer the shortcomings discussed earlier, in that they do not answer a concrete statistical question and cannot disentangle selection bias from other explanations.

In this chapter, our definitions of replicability are inspired primarily by the statistical literature on multiple testing and meta-analysis, such as the estimator in Storey (2002), the FDP and directional FDP from Benjamini and Hochberg (2000) and Benjamini and Yekutieli (2005), and the partial conjunction testing framework of Benjamini and Heller (2008) and Heller et al. (2007). Related error rates have also been estimated before: Jager and Leek (2013) have modeled the p-value distributions under alternatives and the selection for statistical significance to estimate the FDR in the medical literature, accompanied by useful discussions from Gelman and O'Rourke (2013), Goodman (2013), and Ioannidis (2013); in addition, Camerer et al. (2018) used Bayesian methods to estimate the false positive rate, instead of the FDR, for published social science results in *Nature* and *Science*.

Furthermore, there are many past efforts to model and quantify selection bias, specifically using the RP:P dataset. For instance, Johnson et al. (2017) considered a publication bias model where the probability of publication is a step function of the p-value, which is generalized nonparametrically in Andrews and Kasy (2018). The two analyses estimated

that a statistically significant result was 200 (Johnson et al., 2017) or 30 (Andrews and Kasy, 2018) times as likely to be published as a statistically insignificant one.

Adjusting for selection, van Aert and van Assen (2017, 2018) have combined the evidences from both the original and replication experiments to provide estimates for the effect sizes. Specifically with a truncated Gaussian model, Etz and Vandekerckhove (2016) have also analyzed the RP:P dataset from a Bayesian perspective, and investigated the discrepancies between the original and replication studies. Our analysis provides a complementary point of view with frequentist hypothesis testing without any prior on the effect sizes, with the help of recent advances in post-selection inference, including primarily the selective z-test framework of Lee et al. (2016).

Outline

Section 3.2 details the methodology and assumptions used in this analysis, and is somewhat technical. Section 3.3 applies the developed methodology to the RP:P dataset, summarizes and interprets the results. Section 3.4 concludes.

3.2 Methodology

In this section we will construct an estimator for directional FDP, a test for the effect shift in replication i and an estimator for the proportion of effect sizes that declined. We also use $X \geq_{\text{st}} Y$ to denote that X is stochastically larger than Y. The index i is suppressed when there is no risk of ambiguity.

Since we need a well-defined notion of direction to consider the proportion of false directional claims, we restrict our attention to univariate tests, namely z-, t-, $F(1, \cdot)$-tests or correlations. Thus, studies that are not univariate or have p-values greater than $\alpha_0 = 0.05$ are discarded: our estimates and analyses below consider only the $m = 68$ remaining studies with univariate structure and conventionally significant original p-values.

Selection bias model

Model (3.1) assumes that results are only published if they achieved statistical significance at some conventional threshold level α_0, which is 0.05 in our data. While this assumption is not literally true in the case of RP:P since some original p-values are above 0.05, we note that the model can be relaxed to the following milder assumption:

Assumption 1. $p_O < \alpha_0$ is "significant enough": that is, not all results with significant p-values are necessarily published, but a result with $p_O < \alpha_0$ would have been equally likely to be published (or selected for replication), had the p-value taken on some other statistically significant value.

If Assumption 1 holds, then we can model the original test statistics as following their theoretical distribution, truncated to the event where the corresponding p-values are below α_0, as in Model 3.1.

Note that Assumption 1 contemplates a fairly straightforward mechanism for selection on statistical significance, which may not be adequate to describe the effects of more complex and difficult-to-model QRPs. In particular, p-hacking — the iterative tweaking of an analysis until the p-value drops below the researcher's desired significance level α_0 — is commonly suspected to produce a pileup of p-values just below the significance threshold (see e.g. Simonsohn, Nelson, and Simmons, 2014a). Because p-hacking is such a vaguely defined practice, it is unclear how we might incorporate it into our model, but in any case there is no evidence of a pileup just below 0.05 in the original RP:P studies (see Figure 3.6a).

False directional claims

We will adapt the method in Storey (2002) to estimate the directional FDP while accounting for selection bias. Furthermore, if we believe the chosen studies are representative of the publications in the journal or discipline (e.g. Stroebe, 2016), then this estimator can also be regarded as an estimator for the journal-wide or discipline-wide directional false discovery rate (FDR_{dir}), the expectation of the directional FDP (Benjamini and Yekutieli, 2005).

Adjusting for selection bias While dividing a post-selection p-value by α_0 intuitively adjusts for selection, it is not immediately valid when the null is one-sided with a true effect not on the boundary. We demonstrate below that this adjustment typically remains valid even in this case.

Recall that a valid p-value is a random variable that is stochastically larger than Uniform$[0, 1]$ (i.e. superuniform) under the null hypothesis. If we only observe the original p-value when it is significant, it is not superuniform after selection under $H^{S,O}$, and it is therefore not valid for testing the hypothesis of a false directional claim. To adjust these p-values for selection, we follow the principle in Fithian, Sun, and Taylor (2014) by conditioning on the event that the p-values are selected, and also on the variable $S = \text{sign}(\hat{\theta}_O)$ which determines the hypothesis $H^{S,O}$ that we test. We consider two cases: when the original study is a one-sided test and when it is a two-sided test. As we will see, the adjustment in either case is to divide by α_0.

First we consider the case where the original study was a one-sided test. Assume p_O is a p-value for a test of the hypothesis $H_0 : \ \theta_O \leq 0$, in which case $S = +1$ deterministically (the opposite case with $H_0 : \ \theta_O \geq 0$, and $S = -1$ deterministically, is directly analogous). Suppose p_O is the original p-value, which we observe only when it is significant at the conventional threshold, i.e. when $p_O < \alpha_0$. Under mild assumptions satisfied by both z-tests and t-tests,[5] $p_O \geq_{\text{st}} \text{Uniform}[0, \alpha_0]$ under $H^{S,O}$, in which case $p_O/\alpha_0 \geq_{\text{st}} \text{Uniform}[0, 1]$.

[5] namely, that the test statistic has monotone likelihood ratio in the parameter

Next we consider the case where p_O is a p-value for a two-sided test of $H_0 : \theta_O = 0$, and where $S = +1$ (the case with $S = -1$ is analogous). If p_O^+ was the original one-sided p-value for $H_0 : \theta_O \leq 0$, then $p_O = 2p_O^+$ when $S = +1$ ($p_O = 2 - 2p_O^+$ if $S = -1$). In our truncated model, under the same assumptions as above and conditional on $S = +1$, $p_O^+ \geq_{\text{st}} \text{Uniform}[0, \alpha_0/2]$ and therefore $p_O/\alpha_0 = 2p_O^+/\alpha_0 \geq_{\text{st}} \text{Uniform}[0, 1]$ under $H^{S,O}$. We write $p'_O = p_O/\alpha_0$ for the adjusted p-value.

Inference on FDP: estimate and upper confidence bound Using the adjusted original p-values, we can estimate the directional FDP in the original studies. Recall from Table 3.1 that

$$R = \#\{p_{i,O} \leq \alpha_0\} = m,$$
$$V = \#\{p_{i,O} \leq \alpha_0 \text{ and } H_i^{S,O} \text{ is true}\}.$$

Since all of the studies were deemed discoveries, $R = m$ is the total number of studies here. Table 3.2 classifies the m conventionally significant studies according to whether $H_i^{S,O}$ is true and whether the adjusted p-value is larger than some fixed value λ in $(0, 1)$, e.g. $\lambda = 0.5$.

Adjusted p-value	$H_i^{S,O}$ is true	$H_i^{S,O}$ is false	Total
$p'_{i,O} < \lambda$	$*$	$*$	$*$
$p'_{i,O} \geq \lambda$	U	$*$	B
Total	V	$*$	$R = m$

Table 3.2: Classification of the $R = m$ significant original studies. Here only R and B are observed, and we wish to infer on V.

Note that $B = \#\{\lambda\alpha_0 \leq p_{i,O} < \alpha_0\}$ from Table 3.2 is observable, while V and U are not. Under the one-sided null, the p-value is superuniform, and so

$$B \geq_{\text{st}} U \geq_{\text{st}} \text{Binomial}(V, 1 - \lambda). \tag{3.2}$$

As a result, $\mathbb{E}[B] \geq (1 - \lambda)V$ and a conservative (upwardly biased) estimator of the directional FDP is

$$\widehat{\text{FDP}}_{\text{dir}} = \frac{B}{(1 - \lambda)R}.$$

This estimate is conservative in the sense that it overestimates the type I error, and is equivalent to the estimator $\hat{\pi}_0$ of the true null proportion in Storey (2002). Using $\lambda = 0.5$ and $\alpha_0 = 0.05$, the estimate boils down to

$$\widehat{\text{FDP}}_{\text{dir}} = \frac{2}{m} \cdot \#\{0.025 \leq p_{i,O} < 0.05\}.$$

While the above is formally an estimator for the number of directional errors, it can be interpreted practically as an estimate of the fraction of directional claims where *either*

the direction is wrong *or* the effect has a negligible magnitude, cf. type M error from Gelman and Carlin (2014). This is because p-values whose effect sizes are very close to zero are nearly uniform and contribute to our estimator similarly as if the true effect were exactly zero.

Additionally, we can exploit (3.2) to obtain an upper confidence bound for the directional FDP, by testing the hypothesis $H_0 : V \geq v_0$, a partial conjunction hypothesis investigated in Heller et al. (2007). Here we combine only the coarse information of whether each p-value is greater than λ,[6] and reject for small values of B. We can compute the largest v_0 such that the test still accepts, which gives an upper confidence bound of V. Dividing this bound by R gives an upper confidence bound for the directional FDP.

Directional FDP at smaller thresholds One proposal to address the replicability crisis is to lower the conventional significance threshold from $\alpha_0 = 0.05$ to some smaller value α, such as 0.005 (Benjamin et al., 2018). As suggested by Goodman (2013), an empirical method to evaluate the hypothetical scenario with a smaller threshold can be helpful. We now discuss methods for inference on the directional FDP for those studies with $p_O < \alpha < \alpha_0$, based on comparing the number of adjusted p-values below α with the number above $\lambda\alpha_0$, for some $\lambda > \alpha/\alpha_0$. We call this method the *external comparison method* in contrast to the earlier *internal comparison method*. This method will be less conservative as we are not constrained to only using the p-values in $[0, \alpha)$.

Let $N \leq m$ denote the total number of original p-values in $[0, \alpha) \cup [\lambda\alpha_0, \alpha_0)$ (or equivalently, the number of adjusted p-values in $[0, \alpha') \cup [\lambda, 1)$ for $\alpha' = \alpha/\alpha_0$). Table 3.3 classifies these N studies according to whether $H_i^{S,O}$ is true and whether the adjusted p-value is larger than λ or smaller than α'. The numbers of false directional claims and all directional claims under the hypothetical threshold are V_α and R_α, respectively. Auxiliary counts, T_α and W, are defined according to Table 3.3 as well. The directional FDP, V_α/R_α, remains as our quantity of interest.

Adjusted p-value	$H_i^{S,O}$ is true	$H_i^{S,O}$ is false	Total
Small ($p'_{i,O} < \alpha'$)	V_α	T_α	R_α
Big ($p'_{i,O} \geq \lambda$)	U	W	B
Total	N_0	$*$	N

Table 3.3: Classification of the $N \leq m$ original studies with adjusted p-values in $[0, \alpha'] \cup [\lambda, 1]$. Only R_α, B and N are observed. Auxiliary unobserved quantities, N_0, T_α and R_α, are defined accordingly. Our goal is to infer on V_α.

Our method is inspired by the following stochastic inequality.

[6]More precisely, we count number of p-values that are greater than λ and consider its distribution under the partial conjunction null hypothesis

Lemma 8. *Conditional on N, T_α and W, we have*

$$B \mid N, T_\alpha, W \geq_{st} \text{Binomial}(N - T_\alpha, \beta). \tag{3.3}$$

Proof. All adjusted p-values are independent, and are either small ($p \leq \alpha'$) or big ($p \geq \lambda$). The adjusted p-values corresponding to a true null are big with probability at least $\beta = \frac{1-\lambda}{1-\lambda+\alpha'}$. We proceed to condition on T_α and W, so they are now considered deterministic. So the total number of big adjusted p-values, B, satisfies

$$B = U + W \geq_{\text{st}} \text{Binomial}(N - N_0, \beta) + W \geq_{\text{st}} \text{Binomial}(N - T_\alpha, \beta).$$

□

With (3.3), we can estimate $N - T_\alpha$ conservatively with B/β. Since $V_\alpha = N - T_\alpha - B$, a reasonable estimator for the directional FDP is

$$\widehat{\text{FDP}}_{\text{dir}} = \frac{1-\beta}{\beta} \cdot \frac{B}{R_\alpha}.$$

Furthermore (3.3) gives us a 95% upper confidence bound for the directional FDP:

$$\text{FDP}^*_{\text{dir}} = \frac{Q - B}{R_\alpha}, \qquad \text{where } Q = \max\{q : \mathbb{P}[\text{Binomial}(q, \beta) \geq B] \geq 0.95\}.$$

Proposition 9. *The expectation of $\widehat{FDP}_{dir}$ is at least the expectation of the true directional FDP, and FDP^*_{dir} is greater than the true directional FDP, with probability at least 95%.*

Proof. For the estimator, we start by taking the expectation of $\widehat{\text{FDP}}_{\text{dir}} - \text{FDP}_{\text{dir}}$, conditional on N, T_α and W:

$$\begin{aligned}
\mathbb{E}[\widehat{\text{FDP}}_{\text{dir}} - \text{FDP}_{\text{dir}} \mid N, T_\alpha, W] &= \mathbb{E}\left[\frac{\frac{1-\beta}{\beta}B - V_\alpha}{R_\alpha} \,\middle|\, N, T_\alpha, W\right] \\
&\geq \mathbb{E}\left[\frac{\frac{1-\beta}{\beta}(N_0 - V_\alpha) - V_\alpha}{V_\alpha + T_\alpha} \,\middle|\, N, T_\alpha, W\right] \\
&= \mathbb{E}\left[\frac{(1-\beta)N_0 - V_\alpha}{\beta(V_\alpha + T_\alpha)} \,\middle|\, N, T_\alpha, W\right] \\
&\geq \frac{(1-\beta)N_0 - \mathbb{E}[V_\alpha \mid N, T_\alpha, W]}{\beta(\mathbb{E}[V_\alpha \mid N, T_\alpha, W] + T_\alpha)} \qquad (3.4) \\
&\geq 0, \qquad (3.5)
\end{aligned}$$

where (3.4) follows from applying Jensen's inequality to the convex function $f(x) = \frac{(1-\beta)N_0 - x}{\beta(x+T_\alpha)}$, and (3.5) follows from $V_\alpha \mid N, T_\alpha, W \leq_{\text{st}} \text{Binomial}(N_0, 1-\beta)$. Taking expectation on both sides completes the proof.

For $\text{FDP}^*_{\text{dir}}$, we can directly compute the probability that it is greater than FDP_{dir}, conditional on N, T_α and W:

$$\begin{aligned}\mathbb{P}[\text{FDP}^*_{\text{dir}} \geq \text{FDP}_{\text{dir}} \mid N, T_\alpha, W] &= \mathbb{P}\left[\frac{Q-B}{R_\alpha} \geq \frac{V_\alpha}{R_\alpha} \,\middle|\, N, T_\alpha, W\right] \\ &= \mathbb{P}[Q \geq B + V_\alpha \mid N, T_\alpha, W] \\ &= \mathbb{P}[Q \geq N - T_\alpha \mid N, T_\alpha, W] \\ &\geq 0.95,\end{aligned}$$

from the construction of Q. Taking expectation on both sides hence yields the desired marginal coverage. □

Remark. This proof of conservativeness actually shows something stronger than marginal guarantees: the estimator and confidence upper bound are both conservative conditionally, even when we condition on the signs S_i.

Methods using replication p-values As mentioned in Section 3.1, we can use the replication p-values in lieu of the adjusted original p-values above, providing an estimate and confidence bound for the frequency of when the $\hat{\theta}_O$ incorrectly predicts the replication effect direction. While this approach requires potentially costly replications in future applications, it provides valuable additional information. In particular, the replication p-values are more likely to be free of QRPs or p-hacking that may violate our assumption that adjusted p-values are superuniform under the null, providing more robust evidence regarding replicability. The corresponding estimator for unadjusted replication p-values with $\lambda = 0.5$ is

$$\widehat{\text{FDP}}_{\text{dir}} = \frac{2}{m} \cdot \#\{p_{i,R} \geq 0.5\}.$$

Effect shift

We will derive a test for the hypothesis $H^E : \theta_O = \theta_R$ at level 0.05. Our test is based on a normal distribution, so we start by demonstrating that the effect size estimates of the univariate studies can be reasonably modeled by our truncated bivariate normal distribution in model (3.1). We classify these studies into two categories and provide a rough rationale in our definition of effect size in each category: (1) t-tests and $F(1, \cdot)$ ANOVAs, where all independent variables are categorical; and, (2) correlations and regressions, where one or more independent variables are continuous.

For a t-test or $F(1, \cdot)$ ANOVA, we can define the effect size as the noncentrality parameter, scaled for cell sizes. In other words, the t-statistic is distributed as $T \sim t_{df}(k\theta)$, for some real constant k chosen based on the study design. For example, $k = \sqrt{n}$ for a one-sample t-test. When df is sufficiently large, the t-statistic is approximated well by a z-statistic, and distributed approximately as

$$T \sim N(k\theta, 1).$$

For our analysis, we consider studies where the original and replication degrees of freedom are at least 30.[7]

For a (partial) correlation coefficient estimate, R, we can apply Fisher transformation (1921; 1924) to convert it into a z-statistic, which approximately follows

$$\sqrt{n-3-p}\tanh^{-1}(R) \sim N(\sqrt{n-3-p}\,\theta, 1),$$

where p is the number of controlled covariates and θ is a quantity that can be taken as the effect size.

In either case, the test statistic in 46 studies can be transformed to an approximate z-score $Z \sim N(k\theta, 1)$ for some real constant k. Additional considerations in certain studies are detailed in the supplement.

Adjusting for selection bias We turn next to address the issue of post-selection inference. Again, we condition on the event where the z-scores are observed, but we do not need to condition on S as the hypothesis H^E is no longer random. Since the statistic is only observed if it is statistically significant, the original and replication z-statistics follow a truncated bivariate normal joint distribution:

$$\begin{bmatrix} Z_O \\ Z_R \end{bmatrix} \sim N\left(\begin{bmatrix} k_O\theta_O \\ k_R\theta_R \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) 1_{\{Z_O \in A\}}.$$

Here A is the selection event, which contains the statistically significant values of Z_O. We are interested in testing $H^E : \theta_O = \theta_R$ and more generally the null hypothesis $H^{E,\delta} : \theta_O - \theta_R = \delta$, which can be inverted to yield a confidence interval.

We cast this as a more general testing problem here to benefit later derivations on effect decline. Suppose we have a truncated bivariate distribution

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N\left(\mu, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) 1_{\{Z_1 \in A\}}, \quad \text{where } \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},$$

and we want to test $\eta'\mu = \delta$ for some constant vector $\eta = (\eta_1, \eta_2)$ with $\eta_1 > 0$. Test for H^E and $H^{E,\delta}$ are special cases where $\eta = (1/k_O, -1/k_R)$.

We can perform this general testing problem with the *selective z-test*, based on the framework in Lee et al. (2016).

Definition 3 (Selective z-test)**.** *Let $\eta_\perp = (\eta_2, -\eta_1)$, $D = \eta' Z$ and $M = \eta_\perp' Z$. We now consider M as a constant and test $\eta'\mu = \delta$ using the test statistic D against the null distribution*

$$N(\delta, \|\eta\|^2) 1_{\left\{D \in \frac{\|\eta\|^2 A - \eta_2 M}{\eta_1}\right\}}.$$

Specifically, we reject $\eta'\mu = \delta$ when D is below the $\frac{0.05}{2}$-quantile or over the $(1 - \frac{0.05}{2})$-quantile of this null distribution.

[7] The choice of 30 complies with the analysis in Andrews and Kasy (2018). Further discussion on the approximation in available in Appendix A.

We proceed to show that this is a valid test by construction.

Proposition 10. *The selective z-test defined in Definition 3 has level* 0.05.

Proof. Leveraging the fact that $\eta'\eta_\perp = 0$, we reparametrize the joint distribution of (Z_1, Z_2) under the null such that δ is a parameter, i.e.

$$\begin{bmatrix} D \\ M \end{bmatrix} = \begin{bmatrix} \eta'\mu \\ \eta'_\perp\mu \end{bmatrix} \sim N\left(\begin{bmatrix} \delta \\ \eta'_\perp\mu \end{bmatrix}, \begin{bmatrix} \|\eta\|^2 & 0 \\ 0 & \|\eta\|^2 \end{bmatrix} \right) 1_{\{Z_1 \in A\}}.$$

In particular, the event $Z_1 \in A$ can be rewritten as

$$D \in \frac{\|\eta\|^2 A - \eta_2 M}{\eta_1}.$$

And so the distribution of D conditional on M under H_0^δ is a truncated Gaussian distribution,

$$[D \mid M] \sim N\left(\delta, \|\eta\|^2\right) 1_{\left\{D \in \frac{\|\eta\|^2 A - \eta_2 M}{\eta_1}\right\}}$$

and we obtain a valid test by rejecting when D is smaller than the $\frac{0.05}{2}$-quantile or larger than the $\left(1 - \frac{0.05}{2}\right)$-quantile. □

The construction above is represented graphically in Figure 3.5, in the style of Lee et al. (2016). We can represent the observation (Z_1, Z_2) as a point in $\mathbb{R}^2$. Conditioning on M is equivalent to conditioning on $M/\|\eta_\perp\|$, which means we are now considering the conditional distribution on the truncated line ℓ. The test statistic D, or equivalently $D/\|\eta\|$, indicates the position on ℓ. Under the null that $\eta'\mu = \delta$, the conditional distribution on ℓ is known and a valid p-value can be obtained, yielding the selective z-test.

Remark. It is not necessary to use $\frac{0.05}{2}$- and $(1 - \frac{0.05}{2})$-quantiles of the null distribution, as long as the desired significance level is achieved under the null distribution. For example, a uniformly most powerful unbiased test can be used in lieu of a test with equal tail cutoffs. Furthermore, if we are interested in a one-sided hypothesis, e.g. $\eta'\mu \leq 0$, we can reject on one tail only. This will be particularly useful for derivations about effect decline later.

Interval estimation Given a valid test $\phi(Z_O, Z_R)$ for testing $H^{E,\delta} : \theta_O - \theta_R = \delta$, we can obtain two intervals: a predictive interval for the replication effect size estimate, and a confidence interval for effect shifts.

Under the null hypothesis $H^E : \theta_O = \theta_R$, $\mathbb{P}[\phi(Z_O, Z_R) \text{ rejects}] = 0.05$, or equivalently,

$$\mathbb{P}[\{z_R : \phi(Z_O, z_R) \text{ accepts}\} \ni Z_R] = 0.95.$$

Hence $\{z_R : \phi(Z_O, z_R) \text{ accepts}\}$ is a predictive interval for Z_R, which translates to a predictive interval for the point estimate $\hat{\theta}_R$ of the replication effect size.

By the duality of hypothesis testing and confidence set, the set

$$\{\delta : H^{E,\delta} \text{ is rejected}\}$$

covers the difference of the original and replication effect sizes with probability 95%.

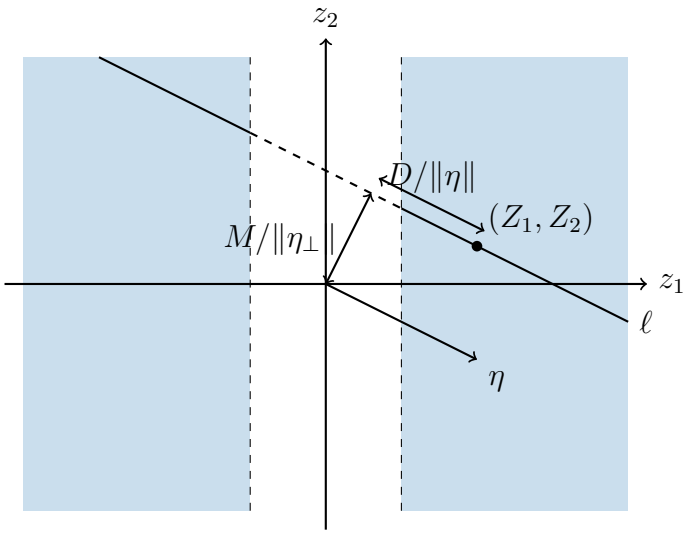


Figure 3.5: Graphical representation of the selective z-test. The observation (Z_1, Z_2) is a point and the truncation on Z_1 means that the shaded area is the support of the joint distribution (Z_1, Z_2). Conditioning on M is the same as conditioning on $M/\|\eta_\perp\|$, so we now consider the conditional distribution on the truncated line ℓ. The test statistic D indicates the position on ℓ. Under the null $H^{E,\delta} : \theta_1 - \theta_2 = \delta$, the conditional distribution on ℓ is known and a valid p-value can be obtained, yielding the selective z-test.

Effect decline

We will estimate the proportion of effect sizes that declined by at least a fraction of ρ. Our procedure consists of two parts: (1) for each study i, test and produce a p-value for the hypothesis $H_i^{D,\rho}$, and (2) adapt the method for the directional FDP to estimate the proportion of $H_i^{D,\rho}$ that are false.

Adjusting for selection bias As with the exactness test, we condition not only on the event where the z-scores are observed, but also on $S = \text{sign}(\hat{\theta}_O)$ as our hypothesis $H^{D,\rho}$ is determined by this random variable. In other words, we consider the z-statistic Z_O to be drawn from the set A_+, where A is the selection event from our test for effect shift and

$$A_+ = A \cap \mathbb{R}_+ = \{z_O : z_O \text{ is statistically significant}\} \cap \mathbb{R}_+.$$

Putting Z_O and Z_R together, they follow a truncated bivariate normal joint distribution:

$$\begin{bmatrix} Z_O \\ Z_R \end{bmatrix} \sim N\left(\begin{bmatrix} k_O\theta_O \\ k_R\theta_R \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) 1_{\{Z_O \in A_+\}}.$$

By convention RP:P chose $\hat{\theta}_O > 0$ so the hypothesis $H^{D,\rho}$ reduces to $\theta_{i,R} \geq (1-\rho)\theta_{i,O}$, or equivalently $\theta_{i,R} - (1-\rho)\theta_{i,O} \geq 0$. This can be tested using the selective z-test with $\eta = (1/k_O, -1/(1-\rho)k_R)$ and rejecting on one tail only.

Inference on effect decline: estimates and confidence bounds With the resulting p-values, our earlier methods on directional FDP can provide an overestimate and a upper confidence bound for the proportion of true $H^{D,\rho}$. Subtracting these from 1 yields an underestimate and a lower confidence bound for the proportion of false $H^{D,\rho}$. On the other hand, by considering the complement of the hypothesis $H^{D,\rho}$, we can also provide an overestimate and an upper confidence bound for the proportion of false $H^{D,\rho}$. These estimators and bounds together provide an overestimate, an underestimate and a 90% confidence interval for the proportion of effect sizes that at least declined by a fraction of ρ.

3.3 Re-analysis of RP:P

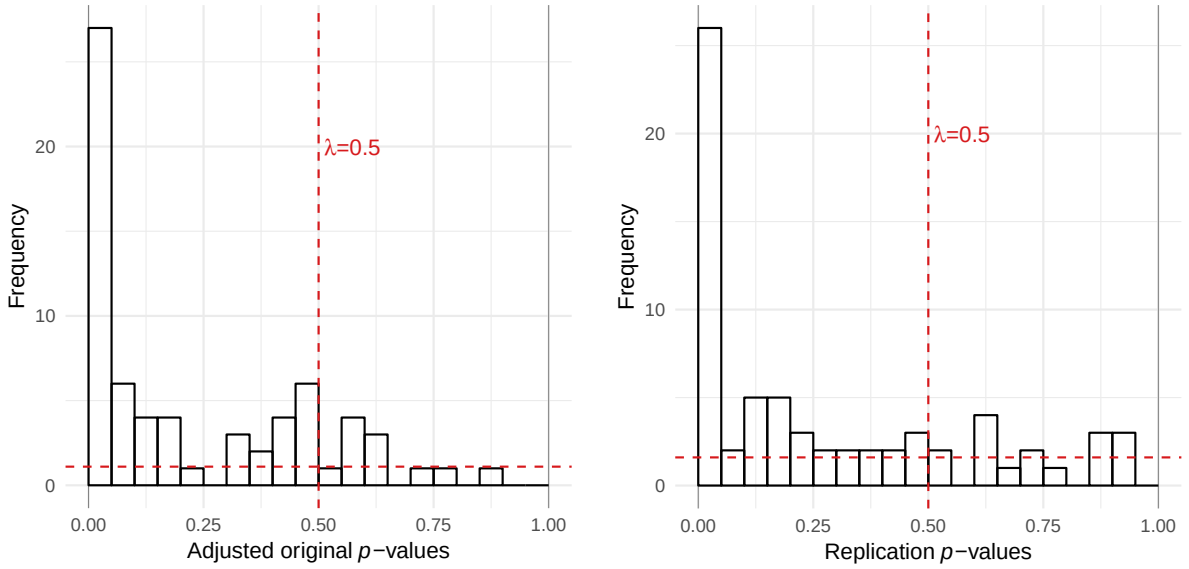
False directional claims

We implemented our method with $\lambda = 0.5$ to estimate the number of one-sided nulls and the directional FDP.[8] The adjusted original p-values and replication p-values are given in Figures 3.6a and 3.6b respectively. Using the original p-values, we estimate that 22 of the 68 (32%) original directional claims are false, with a 95% upper confidence bound of 47%. Using the replication p-values, we estimate that 32 of the 68 (47%) original directional claims incorrectly predict the direction of the replication effect, with a 95% upper confidence bound of 63%. In particular both of our FDP estimates are much lower than the 64% which could be suggested by a naive reading of RP:P (e.g. Baker, 2015). These numbers are summarized again in Table 3.4 later. Furthermore, while we can compute a lower confidence bound, it will always be 0% as the data is obviously consistent with many null hypotheses being slightly false.

We proceeded to evaluate the proposal to reduce the statistical significance threshold (Benjamin et al., 2018). We considered three candidates for the new threshold, 0.001, 0.005 and 0.01, using the external comparison method. The directional FDP estimates and upper confidence bounds are given in Table 3.4.

These estimates corroborate Benjamin et al. (2018)'s suggestion that reducing the statistical significance threshold may improve replicability, at least regarding the directional FDP of the original statistical hypotheses (of course, there is no way to account for potential change in researcher's behavior in response to the lowered threshold). Shall this be of interest, this method provides an empirical way to determine a better significance threshold, as no replications are needed. Nonetheless, potential effect heterogeneity is often a bigger concern. In this case, we are more concerned about the directional FDP for replications, which remains unacceptably high and requires replication experiments. Note, however, that a replication with low power could contribute to our estimates, even if there were no type S error.

[8] Choosing $\lambda = 0.5$ follows the convention in the multiple testing literature for a bias-variance trade off: if λ is too small, many true discoveries are counted as false; if λ is too big, the estimator can have large variance.



(a) Histogram of the adjusted original p-values. (b) Histogram of the replication p-values.

Figure 3.6: Histograms of p-values. We estimate the expected number of true nulls in each bin by the method from Storey (2002), shown by the horizontal red line. A net excess of p-values above this line means false directional claims.

α	Adjusted original		Replication	
	Est.	U.C.B.	Est.	U.C.B.
0.001	0.4/22 = 2%†	2/22 = 9%†	6/22 = 27%	12/22 = 55%
0.005	2.2/33 = 7%†	6/33 = 18%†	12/33 = 36%	20/33 = 61%
0.01	4.4/41 = 11%†	9/41 = 22%†	16/41 = 39%	25/41 = 61%
0.05	22/68 = 32%	32/68 = 47%	32/68 = 47%	43/68 = 63%

Table 3.4: The directional FDP estimates and 95% upper confidence bounds, using the adjusted original and replication p-values. The statistical significance level is α. The external comparison method was used for computing the directional FDP estimates and the upper confidence bounds marked with daggers(†) above, as information of p-values between α and 0.05 can improve the precision. The estimates and upper confidence bounds in the "Replication" column are relatively noisy due to the small number of p-values below the stricter rejection thresholds, and give little basis for any conclusions.

Effect shift

We performed the selective z-test for the hypothesis $H^E : \theta_O = \theta_R$ while adjusting for selection, where seven (15%) studies are rejected. In contrast, without adjusting for selection, 18 (39%) studies are rejected at 0.05 significance. If we wish to correct for multiplicity, we can apply Benjamini–Hochberg procedure (1995), which rules five (11%) replication studies as inconsistent with the original studies at false discovery rate 0.10.[9] Applying the more stringent Holm's method (1979) to control the familywise error rate rules only the replication of Farris et al. (2008) as inconsistent at familywise error rate 0.05.

We inverted the test for the hypothesis H^E, to yield a predictive interval for Z_R and hence a predictive interval for the replication effect size estimate $\hat{\theta}_R$, shown in Figure 3.7. By definition H^E is rejected when $\hat{\theta}_R$ is not included in the predictive interval. Adjusting for selection generally stretches the predictive intervals, resulting in fewer rejections.

We also inverted the test for $H^{E,\delta}$ and obtained a confidence interval for the effect shifts, $\theta_O - \theta_R$, given in Figure 3.8. By construction the null hypothesis $H^E : \theta_O = \theta_R$ is rejected when the confidence interval does not include 0. Adjusting for selection also generally lengthens the confidence intervals, resulting in fewer rejections.

If all procedures are replicated perfectly, we should expect to reject 5% of the tests on average, rather than the observed 15%, and after the Benjamini–Hochberg correction, there would be no rejection with 90% probability. In other words, while selection bias can partly explain the discrepancies between the original and replication studies, it does not explain all of it. Nevertheless, the RP:P data cannot be taken as strong evidence of widespread failure by replication teams to satisfactorily repeat the same experiment performed in the original study. The lack of strong evidence is hardly surprising: if the original study lacks power (Morey and Lakens, 2017) or $\hat{\theta}_O$ is closed to the rejection boundary, little can be said about θ_O and hence $\theta_O - \theta_R$. Furthermore, the replication sample sizes were determined based on the original effect size to achieve at least 80% in power. Selection bias inflated the original effect size, leading to lower test power and statistically insignificant replications (Camerer et al., 2018; Etz and Vandekerckhove, 2016). The lack of information about $\theta_O - \theta_R$ is evident in generally wider confidence intervals after adjustment in Figure 3.8.

Effect decline

Finally, we considered the proportion of effect sizes that declined. Using the selective z-test, we tested the hypothesis H^D, conditioning on the event where the z-scores are observed and the variable S. The resulting p-values are given in Figure 3.9. Our underestimate and overestimate are 35% (= 16/46) and 100% respectively, with a 90% confidence interval of $(11\%, 100\%)$.

More generally, we used the hypothesis $H^{D,\rho}$ to estimate the proportion of effect sizes that declined by at least a fraction of ρ. The underestimate, overestimate and the 90%

[9] The five rejected studies are Dodson, Darragh, and Williams (2008), Farris et al. (2008), Larsen and McKibban (2008), Purdie-Vaughns et al. (2008), and van Dijk et al. (2008).

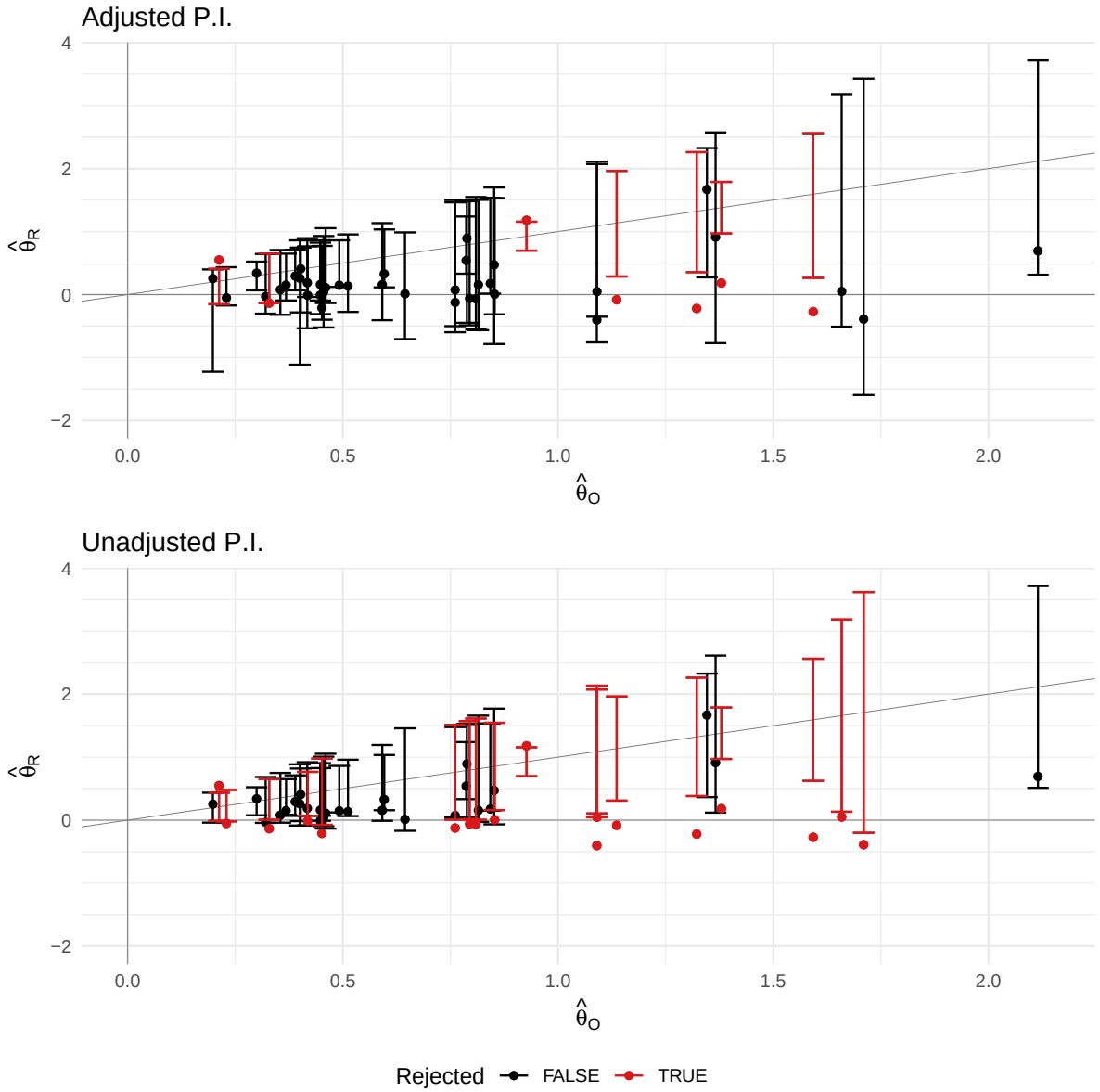


Figure 3.7: Predictive intervals for $\hat{\theta}_R$, both adjusted and unadjusted for selection, overlay with a plot of $\hat{\theta}_R$ against $\hat{\theta}_O$. Studies 36 and 145 are not shown here. By definition we reject $H_0 : \theta_O = \theta_R$ whenever the replication effect size estimate lies outside of the predictive interval. The intervals are generally longer after adjusting for selection.

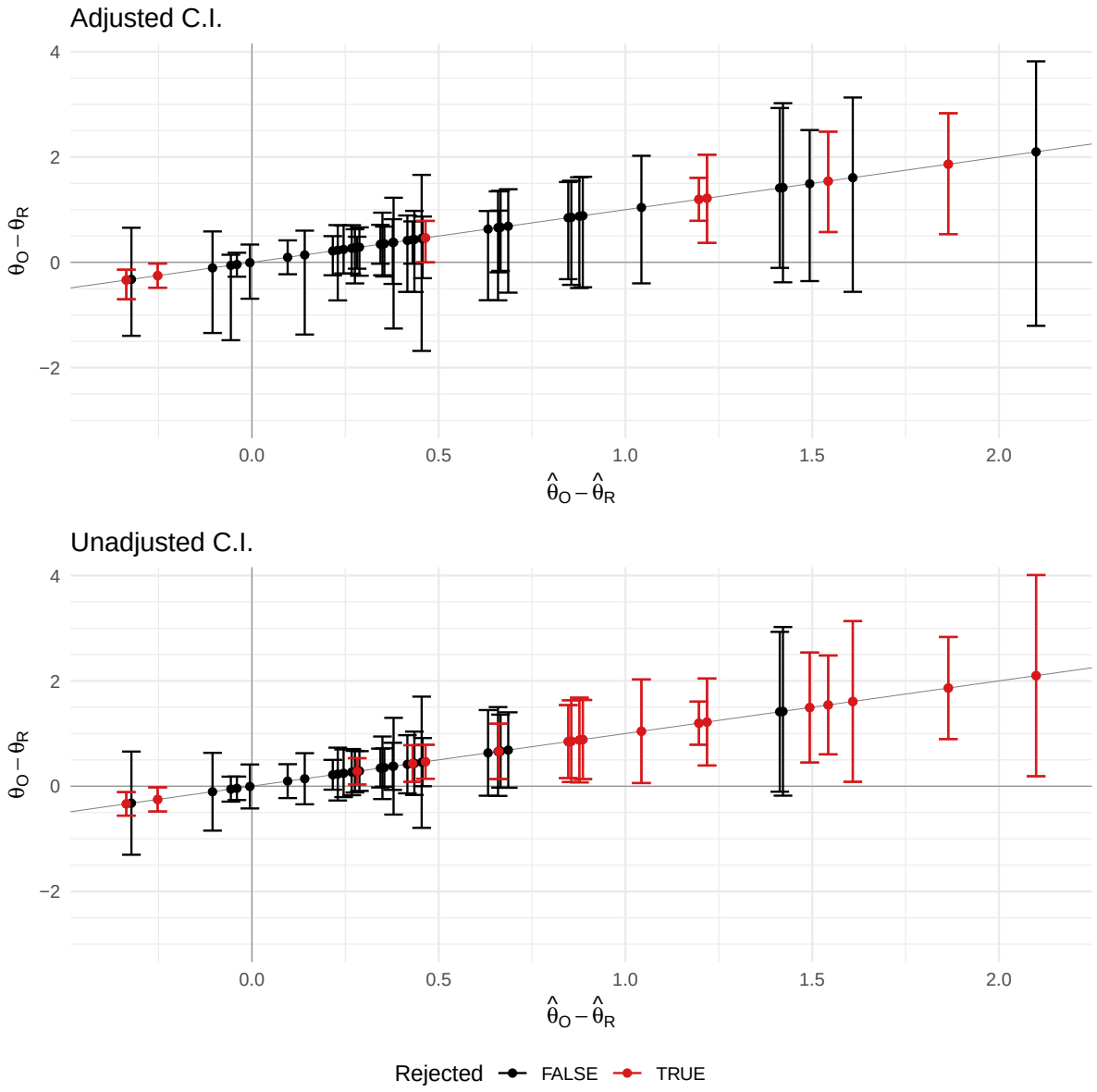


Figure 3.8: Confidence intervals for $\theta_O - \theta_R$, both adjusted and unadjusted for selection. By construction the null hypothesis $H_0 : \theta_O = \theta_R$ is rejected when the confidence interval does not include 0. Many of the adjusted intervals are fairly long as either the replication studies suffer low power or the original effect size estimate is near the rejection threshold. The intervals are generally longer after adjusting for selection.

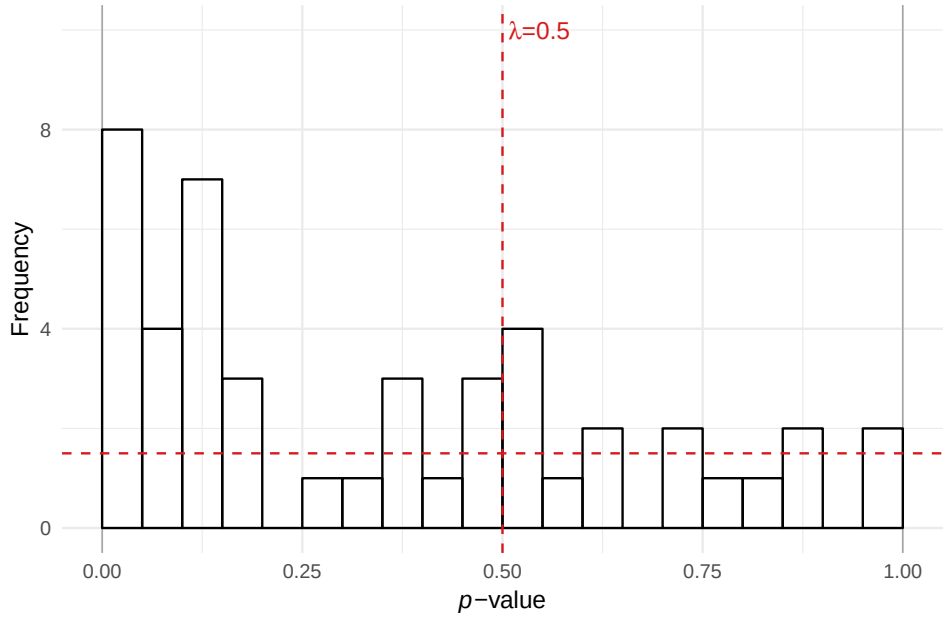


Figure 3.9: Histogram of the p-values for the null hypothesis $\theta_R \geq \theta_O$. p-values to the left gives more evidence for $\theta_R < \theta_O$ whereas p-values to the right gives more evidence for $\theta_R \geq \theta_O$. The estimate of the expected number of null p-values within each bin is given by the horizontal red line.

confidence interval are given in Figure 3.10. For example, we estimate that 16 of the 46 effect sizes (35% (with a 95% lower confidence bound 11%) decreased by at least 20%, even after adjusting for selection on measurement noise. Note that this does not exclude explanations by other forms of selection, e.g. selecting a large effect when there is a random effect.

3.4 Discussion

Importance of adjusting for selection bias

As we have seen, selection bias plays a powerful and pervasive role in shaping the data we observe in large-scale replication studies (and, by extension, the data we observe in published studies that have not yet been replicated!). It leads to many predictable pathologies and should be viewed as a proverbial "elephant in the room" whenever we discuss descriptive statistics computed from such studies. In particular, we should avoid leaping to any conclusions about how many false claims there were in the original studies, whether effect sizes declined or by how much, or which replication studies suffered from infidelities, until we have carefully ruled out the possibility that publication bias alone is to blame for whatever descriptive statistic we have computed.

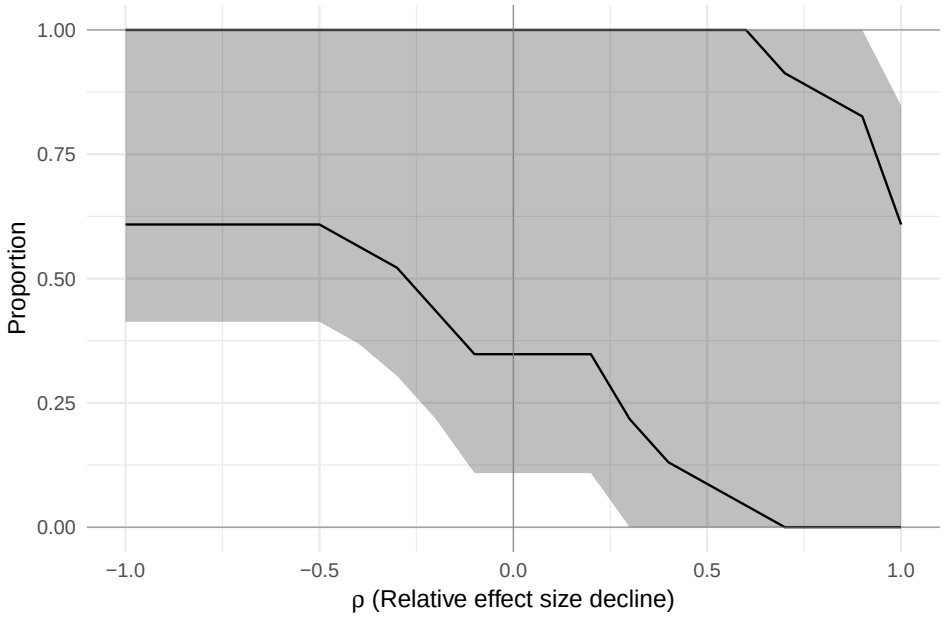


Figure 3.10: The underestimate, overestimate and the 90% confidence interval. The lower black line is the underestimate, the high black line is the overestimate and the gray band is the 90% confidence interval.

Fortunately, the truncated Gaussian model, properly combined with modern multiple testing and post-selection inference methods, opens many avenues for analyses that directly answer questions about true effect sizes with appropriate uncertainty quantification. We have explored several such avenues here (see also Andrews and Kasy, 2018) but many others are possible.

Importance of statistical formality

In addition, we hope this chapter serves to advocate for the benefits of careful formal statistical modeling in analyzing replication studies, in place of (or in addition to) descriptive statistics. In particular, using vaguely specified models or eschewing models altogether can lead to analyses from which it is difficult to draw firm conclusions. For example, in Open Science Collaboration (2015), McNemar's test was applied to a 2×2 contingency table of whether the original and replication studies are equally likely to be statistically significant. The very small p-value reported for this test establishes nothing more than that the original studies were selected to be statistically significant, a fact which is likely already known by most in the field. In fact, the test does not quite establish even that, because it is unclear whether this hypothesis would be true even without the effect of selection bias: The proportion of statistically significant p-values is a measure of the average power, which depends on the sample sizes, and the sample sizes often differed substantially between the original and replication studies.

Another example is RP:P's use of sample correlation coefficients between indepen-

dent and dependent variables as a standardized measure of effect size for comparison between the original and replication studies. This comparison implicitly assumes that the distribution of the independent variable is the same in the original and replication studies, an assumption that was violated by many of the replications. In an extreme case, an ANOVA in Purdie-Vaughns et al. (2008) with race as one of the factors used 40 African Americans and 37 Whites, but was replicated with 120 African Americans and 1370 Whites. With such a dramatic change in the distribution of an independent variable, there is no reason why the correlation coefficients should remain the same, as illustrated in the following example.

Example 5. A study with a two-sample t-test for some treatment condition is replicated. Suppose the treatment and control group are drawn from $N(1,1)$ and $N(0,1)$, respectively. If the ratio of the two group sizes changes from one study to another, the correlation coefficients may differ as well, even without any infidelities or hidden moderators. Borrowing the numbers from Purdie-Vaughns et al. (2008) for instance, if the original study contains 40 treatment and 37 control units, the true correlation coefficient is 0.45, whereas in a replication with 120 control and 1370 treatment units the true coefficient is 0.26 instead.

Replication projects similar to RP:P have since materialized, but few stated an explicit statistical hypothesis. For example, in economics, Camerer et al. (2016) used the same flawed metric of proportion of statistically significant results in the original direction. A statistical analysis with explicitly stated models and hypotheses will give us more meaningful estimates, particularly valuable given how costly these large scale replication efforts are.

Interpretation of effect shifts

While we have proposed several methods for quantifying discrepancies between the effect sizes in the original and replication studies, the data alone cannot tell us why they might differ. Several potential explanations include:

1. design failures, systematic biases or calculation errors in either the original or the replication study;
2. major differences in experimental conditions between the original and replication studies, which most researchers would recognize *a priori* as likely to affect the results; which Gilbert et al. (2016b) call *infidelities*; and
3. minor differences in experimental conditions between the studies — such as lighting, weather, or the passage of time — which cannot all be controlled but whose effects may nevertheless alter the true effect size in unforeseeable ways, often referred to as *hidden moderators* (e.g. Srivastava, 2015).

While there may be no sharp distinction in principle between infidelities and hidden moderators, there is a scientifically crucial difference between moderating factors that can be anticipated by experimenters and those that cannot. If we can anticipate in advance

when replications are likely to fail by carefully evaluating their designs, we might hope to solve the problem simply by being more careful in setting up experiments. By contrast, if hidden moderators confound most attempts to replicate most psychological studies, it would raise profound questions about the entire enterprise of experimental psychology. In the extreme case, if even trivial changes to those conditions have large and unpredictable effects on most phenomena of interest, we might begin to despair of gaining generalizable knowledge about psychology through laboratory experimentation.

Our analyses point to several conclusions regarding effect shifts: First, that there are a few studies where we can be confident the effect in the replication study was significantly different than in the original study; second, that in aggregate, when effects do shift, they tend to decline (shift toward zero) in replications rather than increase; and third, that there is insufficient evidence to conclude that the vast majority of experimental effects simply evaporated upon replication. In particular, 83% should not be treated as a reasonable estimator of the fraction of *true* effect sizes that declined; rather, it likely reflects that the estimates in the original studies overestimated their corresponding true effects due to selection bias.

One possible explanation for systematically declining effect involves a subtler form of selection bias, where every experiment's effect size is random, buffeted by hidden moderators, and those experiments whose moderators primarily magnify the effect size are more likely to be published. That is, in the same way that experimenters select studies whose sampling error is large, they also selects for studies whose true effect size is larger than usual. Further systematic replication studies may help to shed light on which factors are most often the culprits in moderating true effect sizes, possibly improving the reliability of experiments and leading to new scientific insights (Barrett, 2015; Klein et al., 2018).

Future work

As large-scale replicability studies are becoming more common in assessing the "well-being" of a scientific domain, this chapter serves as a stepping stone for improving methodologies in future replicability studies.

First, selection for significance is an inevitable consequence of the current scientific process. Our adjustments for selection is admittedly crude, but necessitated by the limitations in the given data. With more available information, a better model for selection can be used. For example, with the advancement of preregistration, we can use the external comparison method to produce less conservative estimates of the directional FDP at level $\alpha = 0.05$ if we have more information about statistically nonsignificant studies. With more replications carried out, we can estimate the publication bias model in Andrews and Kasy (2018) more precisely; together with higher powered design in replications (e.g. Camerer et al., 2018), we can enhance the precision of our estimators and power of our tests.

Second, we emphasized the importance of statistical formality. Our proposed criteria are based on clearly defined parameters. While these criteria may not suit all needs in

future replicability studies, additional formal hypotheses can also be analyzed under the post-selection inference framework similarly.

With our proposed criteria and procedures, researchers can perform more informative inferences than the current practice, and provide a clearer picture of the replicability crisis.

Reproducibility

A git repository containing with the code generating the images in this chapter is available at `https://github.com/kenhungkk/assessing-replicability.git`.

Supplement

The supplement is available in the git repository, or directly on `https://github.com/kenhungkk/assessing-replicability/raw/public/supplement.pdf`. The discussion of t-distribution approximations is also included in the appendix of this dissertation.

Chapter 4

Optimal Post-Selection Combined Inference

4.1 Introduction

There is a recent surge of interest in replicability research in many social science domains, such as psychology and economics. Experiments are selected and repeated in attempts to determine the validity of the original findings, e.g. Camerer et al. (2016), Klein et al. (2018), and Open Science Collaboration (2015). Statisticians developed to investigate the discrepancies between the original experiments and the replications, e.g. Andrews and Kasy (2018), Etz and Vandekerckhove (2016), Hung and Fithian (2019b), and Johnson et al. (2017).

These replication efforts provided a side benefit to the scientific community: with more replications, effect sizes can be estimated more precisely by combining the original experiments and replications. In some cases where many replications are performed (e.g. Klein et al., 2018), all of the replication experiments and the original experiment can be aggregated for a better estimate.

Historically there have been many methods for combining experiments to provide better inferences. A classical method is Fisher's combined test (1925), which combines the logarithms of the p-values of individual experiments. However, as the original experiments are selected in the publication process, or selected to be replicated based on its statistical significance, the p-values tend smaller and classical combination methods like Fisher's do not account for this bias.

To adjust the original p-value p_O, a truncated model is often used (Andrews and Kasy, 2018; Hung and Fithian, 2019b; van Aert and van Assen, 2018). In particular, the p-value can be conveniently adjusted by division by the statistical significance threshold, α_0, under mild conditions discussed in Hung and Fithian. van Aert and van Assen proposed to consider the sum of the adjusted original p-value, p'_O, and the replication p-value, p_R, as a test statistic, as $p'_O + p_R$ follows the Irwin–Hall distribution under the

null.[1]

Note that there are many potential ways to combine p'_O and p_R, e.g. Fisher's combined test. Different combination methods may have more power against different alternative hypotheses. To choose a combination method, there are two main considerations: (1) type of optimality in the test or estimator; (2) importance of different experiments. Since all experiments investigate a univariate parameter with two-sided alternatives, a uniformly minimal variance unbiased (UMVU) estimator is often considered optimal. For a univariate test, a uniformly most powerful unbiased (UMPU) test is commonly considered optimal, where for a simple null $H_0 : \theta = \theta_0$,

$$\beta(\theta_0) \leq \alpha, \qquad \beta(\theta) \geq \alpha \text{ for all } \theta \neq \theta_0,$$

α is the level of the test and β is the power function. The methods in van Aert and van Assen (2018), however, do not meet these optimality criteria.

Furthermore, van Aert and van Assen (2018) weighs the original experiment and replication equally, regardless of their sample sizes. Consider an extreme scenario where the replication has a sample size approaching infinity. The power of their test does not approach 1 as the adjusted original p-value p'_O has a non-vanishing probability of being bounded away from 0; meanwhile the asymptotic power using only the replication is 1.

Related work

Our method is inspired by the conditional inference framework proposed in Fithian, Sun, and Taylor (2014). In particular, we model the selection process are require the same assumption as in Hung and Fithian (2019b). We restate their assumption in Assumption 2 for readers' convenience.

Assumption 2. $p_O < \alpha_0$ is "significant enough": that is, not all results with significant p-values are necessarily published, but a result with $p_O < \alpha_0$ would have been equally likely to be published (or selected for replication), had the p-value taken on some other statistically significant value.

Furthermore, our methods are guided by ideas from mathematical statistics, such as sufficient statistics[2] and Rao–Blackwell theorem (1947). Finally, when selection bias is absent, our test and estimator coincide to the weighted inverse normal method by Lipták (1958).

Outline

We start by investigating a Gaussian model in Section 3.2, and extend it to sample correlation coefficients later. Section 4.3 simulates 1000 sample correlation coefficients

[1] More precisely, the sum should be stochastically larger than the Irwin–Hall distribution under the null, because the null does not need to be simple.

[2] For a brief introduction or review, see Keener (2010).

to illustrate the performance of our methods when compared to van Aert and van Assen (2018) and using only the replication, and to justify approximations involved in converting correlation coefficients to Gaussian variables. Section 3.4 concludes.

4.2 Methodology

We start with a Gaussian model. We set up our statistical model in the same way as Hung and Fithian (2019b): suppose the true parameter is θ, and we have noisy Gaussian observations, Z_O from the original study and Z_R from the replication, with variances σ_O^2 and σ_R^2 respectively. However, due to selection bias, the observation pair (Z_O, Z_R) is only observed when the original experiment is statistical significant, i.e. when $Z_O \geq c = \sigma_O z_{\alpha_0}$, where z_{α_0} is the upper α_0-quantile of a standard normal distribution. For simplicity we assume this selection is one-sided, and hence Z_O distributes according to

$$Z_O \sim N(\theta, \sigma_O^2) 1_{\{Z_O \geq c\}},$$

where the indicator function means truncation and renormalization of the distribution. Meanwhile, Z_R is free from selection bias, and distributes as

$$Z_R \sim N(\theta, \sigma_R^2).$$

For sample correlation coefficients, R_O and R_R, they can be Fisher tranformed (1921) into approximately Gaussian distributed statistics,

$$\begin{aligned} Z_O &= \tanh^{-1} R_O, \text{ and} \\ Z_R &= \tanh^{-1} R_R, \end{aligned}$$

with variances $\sigma_O^2 = 1/(n_O - 3)$ and $\sigma_R^2 = 1/(n_R - 3)$.

The joint distribution is thus a truncated bivariate Gaussian

$$\begin{pmatrix} Z_O \\ Z_R \end{pmatrix} \sim N\left(\begin{pmatrix} \theta \\ \theta \end{pmatrix}, \begin{pmatrix} \sigma_O^2 & 0 \\ 0 & \sigma_R^2 \end{pmatrix} \right) 1_{\{Z_O \geq c\}}. \tag{4.1}$$

Observe that this joint distribution can be viewed as an exponential family, as elucidated by the probability density function (p.d.f.):

$$\begin{aligned} p_\theta(z_O, z_R) &= \exp\left(-\frac{(z_O - \theta)^2}{2\sigma_O^2} - \frac{(z_R - \theta)^2}{2\sigma_R^2} \right) 1_{\{z_O \geq c\}} A(\theta) \\ &= \exp\left(\left(\frac{z_O}{\sigma_O^2} + \frac{z_R}{\sigma_R^2} \right) \theta \right) g(z_O, z_R) \tilde{A}(\theta), \end{aligned}$$

where some $A(\theta)$ and $\tilde{A}(\theta)$ are normalizing constants that depend only on θ, and g is a function of only z_O and z_R.

In fact, we can see that $S = Z_O/\sigma_O^2 + Z_R/\sigma_R^2$ is a sufficient statistic for the natural parameter θ, forming the groundwork of an optimal estimator and an optimal test.

Estimator Starting from an unbiased estimator, we can Rao–Blackwellize (1947) to obtain a UMVU estimator by taking its expectation conditioned on a sufficient statistic. Since Z_R is free from selection bias, it is an unbiased estimator of θ. Hence the UMVU estimator is given by

$$\hat{\theta}_{\text{UMVU}} = \mathbb{E}[Z_R \,|\, S] = \mathbb{E}\left[Z_R \,\middle|\, \frac{Z_O}{\sigma_O^2} + \frac{Z_R}{\sigma_R^2} \right]. \tag{4.2}$$

An explicit formula for the conditional expectation in $\hat{\theta}_{\text{UMVU}}$ can be obtained by an orthogonalization trick: we can reparametrize the joint distribution (4.1) as

$$\begin{bmatrix} Z_O - Z_R \\ S \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ (1/\sigma_O^2 + 1/\sigma_R^2)\,\theta \end{bmatrix}, \begin{bmatrix} \sigma_O^2 + \sigma_R^2 & 0 \\ 0 & 1/\sigma_O^2 + 1/\sigma_R^2 \end{bmatrix} \right) 1_{\{Z_O \geq c\}}, \tag{4.3}$$

where the truncation event can be further rewritten as $\{Z_O \geq c\} = \{Z_O - Z_R \geq a(S)\}$, where

$$a(S) = \frac{\sigma_O^2 + \sigma_R^2}{\sigma_O^2} c - S\sigma_R^2.$$

The conditional expectation in (4.2) is thus

$$\begin{aligned} \mathbb{E}[Z_R \,|\, S] &= \frac{\sigma_R^2}{\sigma_O^2 + \sigma_R^2} \mathbb{E}\left[\sigma_O^2 S - (Z_O - Z_R) \,\middle|\, S \right] \\ &= \frac{\sigma_O^2 \sigma_R^2}{\sigma_O^2 + \sigma_R^2} S - \frac{\sigma_R^2}{\sigma_O^2 + \sigma_R^2} \mathbb{E}[Z_O - Z_R \,|\, S] \\ &= \frac{\sigma_O^2 \sigma_R^2}{\sigma_O^2 + \sigma_R^2} S - \frac{\sigma_R^2}{\sqrt{\sigma_O^2 + \sigma_R^2}} \frac{\phi\left(a(S) \Big/ \sqrt{\sigma_O^2 + \sigma_R^2} \right)}{\bar{\Phi}\left(a(S) \Big/ \sqrt{\sigma_O^2 + \sigma_R^2} \right)}, \end{aligned}$$

where $\phi(\cdot)$ and $\bar{\Phi}(\cdot)$ are the p.d.f. and the survival function of a standard Gaussian, respectively. Note that the first term is in fact the estimate if there were no publication bias, and is the inverse-variance weighted average of the observations, Z_O and Z_R.

For investigating a true correlation, say r, a UMVU estimator for the Fisher transformed true correction, $\theta = \tanh^{-1} r$ can be computed, which can be transformed backwards to give an estimate for the true correlation. If the observed correlations are R_O and R_R, then the estimator is

$$\hat{r}_{\text{“UMVU”}} = \tanh \mathbb{E}\left[\tanh^{-1} R_R \,\middle|\, \frac{\tanh^{-1} R_O}{n_O - 3} + \frac{\tanh^{-1} R_R}{n_R - 3} \right].$$

Note that this is neither unbiased nor UMVU. Alternatively, we can Rao–Blackwellize an unbiased estimator, such as the estimator in Olkin and Pratt (2007). However, the conditional expectation is harder to compute. For most applications in social sciences where r is small, the bias in the estimator $\hat{r}_{\text{“UMVU”}}$ tends to be small, as illustrated in simulations in Section 4.3.

Test Since the distribution in (4.1) is a one-parameter exponential family, we can derive a UMPU test based on the sufficient statistic S, in the form of

$$\begin{cases} \text{reject} & \text{if } S \leq a \text{ or } S \geq b \\ \text{accept} & \text{otherwise,} \end{cases}$$

where a and b are threshold chosen such that

$$\mathbb{P}_{\theta=0}[S \leq a \text{ or } S \geq b] = \alpha, \tag{4.4}$$

$$\frac{\partial}{\partial\theta}\mathbb{P}_{\theta}[S \leq a \text{ or } S \geq b] = 0. \tag{4.5}$$

4.3 Simulation

We investigate the performance of our estimator and test by a Monte Carlo simulation. Each of the 1000 sample correlation coefficients is generated by passing Gaussian random numbers through Fisher transformation. Three estimators or tests are applied to the generated sample: (1) the *hybrid method* from van Aert and van Assen (2018), (2) testing or estimating using only the replication, and (3) the "UMVU" estimator or the UMPU test. For a fair comparison, we use the same range of original sample size n_O, the same range of replication sample size n_R, the same range of true correlation and the same significance threshold α_0 as in van Aert and van Assen. The biases and root-mean-square-errors (RMSEs) of the estimators are given in Figure 4.1 and Figure 4.2 respectively, and the type I error rates or powers of the tests are given in Figure 4.3.

Figure 4.1 shows a generally negative bias for both our "UMVU" estimator and using only the replication, stemming from the concavity of inverse Fisher transformation. However, our "UMVU" estimator is more concentrated, which alleviates this bias. Finally, Figure 4.2 shows that our "UMVU" estimator gives a smaller RMSE compared to the hybrid method and using only the replication.

For the test, Figure 4.3 demonstrates that our UMPU test achieves higher power than both the hybrid method and using only the replication data. In particular, when the original experiment has more samples than the replication ($n_O > n_R$), our method makes use the original study and thus performs better than using only the replication. On the other hand, when the replication has more samples than the original experiment ($n_R > n_O$), our method does not weigh the two experiments equally. Therefore it outperforms the hybrid method and performs more similar to using just the replication.

4.4 Discussion

With the high cost involved in social science experiments, every additional bit of information is valuable. We presented an optimal method for combining two experiments, that does not discard any information and always outperforms both the hybrid method in van Aert and van Assen (2018) and using only the replication data.

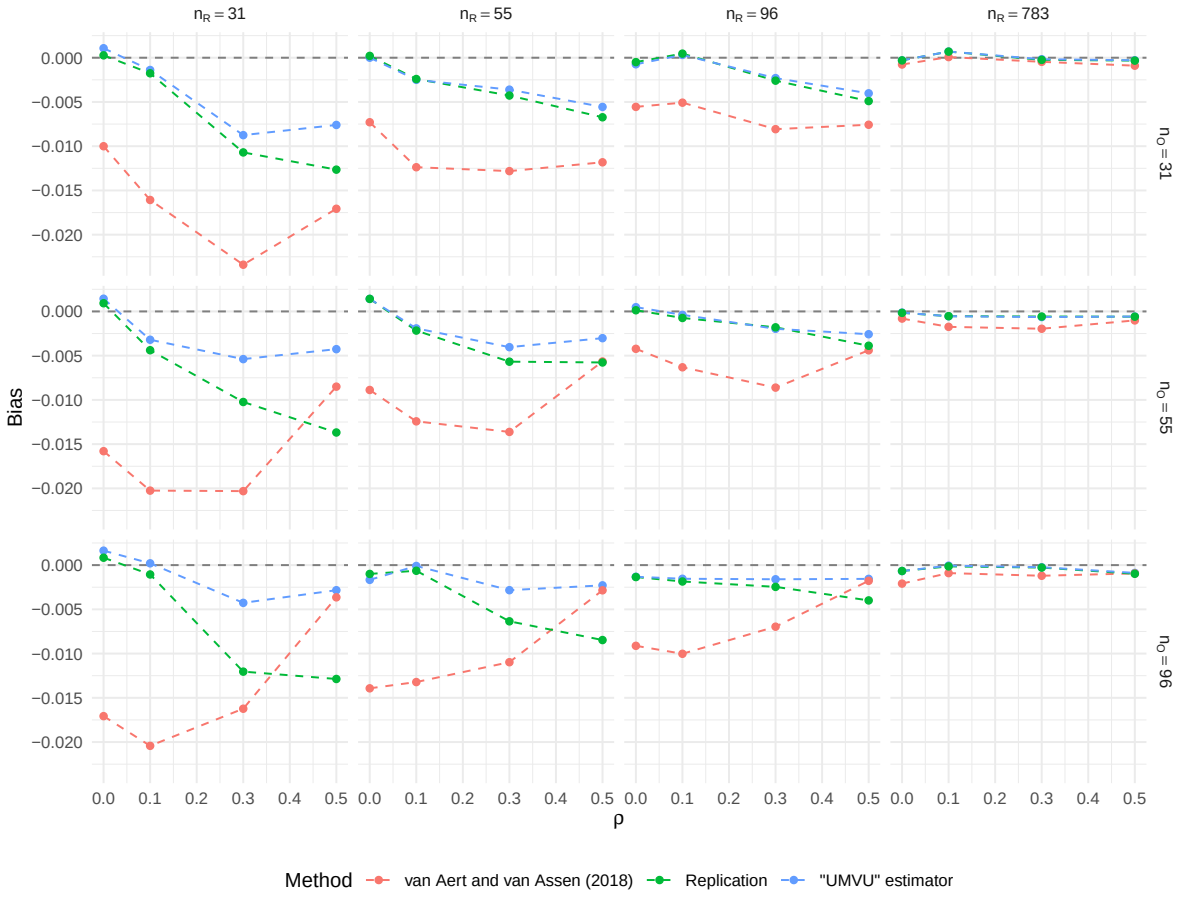


Figure 4.1: Bias for various settings of n_O, n_R and ρ, for an estimator using only the replication, the hybrid method and our "UMVU" estimator. Note that even using only the replication leads to a negative bias due to the concavity of Fisher transform for positive correlations. Our estimator has the least bias across various sample sizes and various correlation strengths.

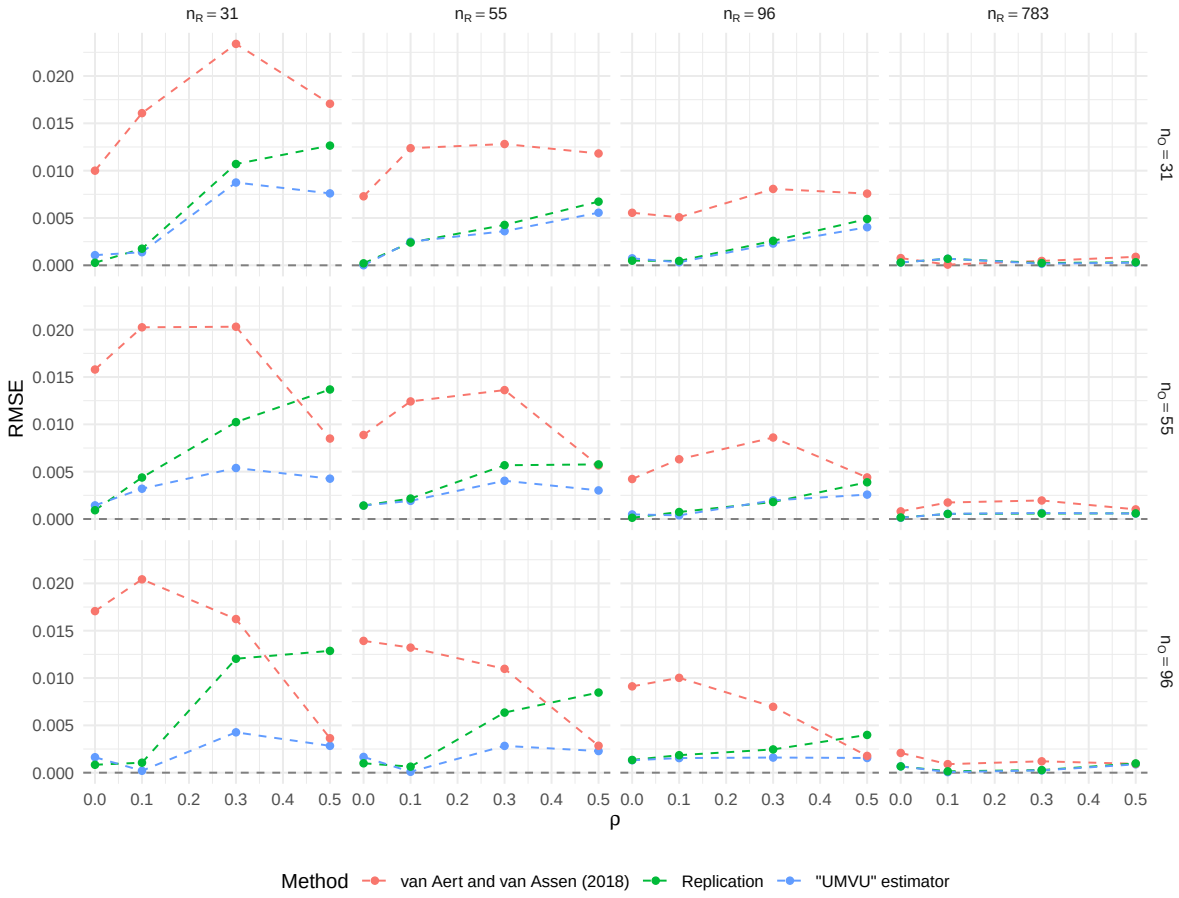


Figure 4.2: RMSEs for various settings of n_O, n_R and ρ, for an estimator using only the replication, the hybrid method and our "UMVU" estimator. Our estimator has smallest RMSE across various sample sizes and various correlation strengths.

The method can in fact be easily extended to setups with multiple replications (e.g. Klein et al., 2018), as the joint distribution of observations remains an exponential family. However, the implicit assumption that both experiments measured the same true parameter becomes less plausible as the number of replications increases, and a fixed effect model as used in this chapter may be less appropriate.

Finally, Assumption 2 requires that the original experiment to be statistical significant, but in some replicability studies, such as Open Science Collaboration (2015), studies close to significance (e.g. $p \in [0.05, 0.055]$) are included. Our method can be generalized to provide an estimate in such cases, if a more sophisticated selection model (e.g. Andrews and Kasy, 2018) is given.

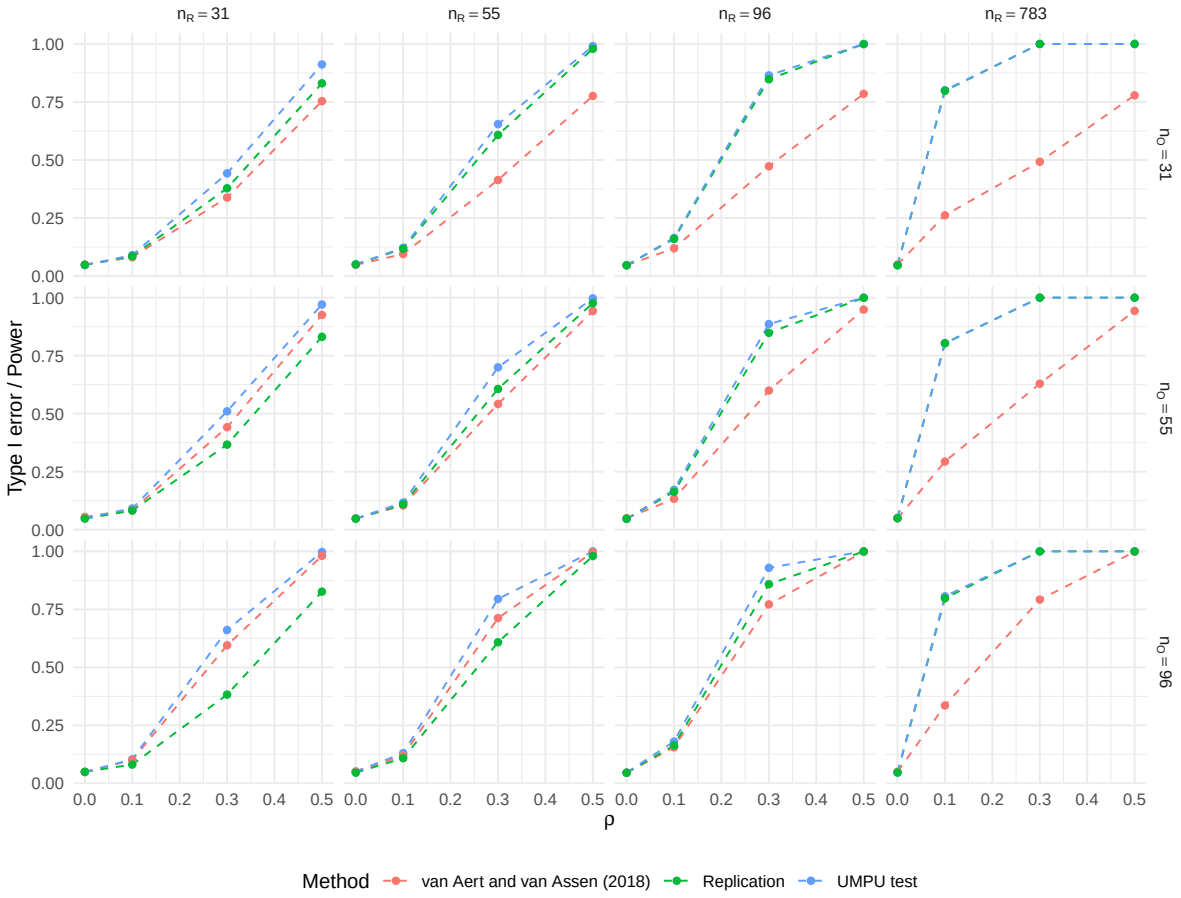


Figure 4.3: Type I error rate or power for various settings of n_O, n_R and ρ, for a test using only the replication, the hybrid method and our UMPU test. When the original experiment has more samples than the replication ($n_O > n_R$), our method makes use the original study and thus performs better that using only the replication. On the other hand, when the replication has more samples than the original experiment ($n_R > n_O$), our method does not weigh the two experiments equally. Therefore it outperforms the hybrid method and performs more similar to using just the replication information. Our method is the most powerful generally across various sample sizes and various correlation strengths.

Chapter 5

Discussion

This dissertation included a few examples of conditional inference in application to post-selection inference, arising from either data-dependent hypothesis or data screening. Conditional inference, and sometimes more generally post-selection inference, can be a powerful tool in statistics.

First, in the motivating example in Chapter 2, the hypotheses are innately dependent on the data. Deterministic hypotheses, such as $\pi_{\text{Cruz}} \leq \max_{i \neq \text{Cruz}} \pi_i$, can be uninteresting. At times, e.g. when Trump receives more votes than Cruz in the poll, considering these deterministic hypotheses can be even silly. Conditional inference allows rigorous testing of these hypotheses, even when they are traditionally considered ill-defined and untestable.

Second, an unintended benefit of conditional inference is that the data itself can play a role in choosing a test, and hence focus the statistical power against more "likely" alternatives. In Chapter 2, we observe the top two order statistics and their indices. This enables us to use that particular difference as the test statistic, as opposed to Gupta and Nagel (1967) that considers the maximum of all differences between the largest observation and other observations as the test statistic.

Third, with the abundant data available but one's limited attention, conditional inference let us screen the data before analysis. In Chapter 3, not all publications on the three psychology journals are replicated due to limited resources. In the end the selected experiments are mostly statistical significant, but we can nonetheless analyze the discrepancies between the original and replication experiments, free from the inevitable selection bias. A similar flavor of hypothesis testing after screening the given data can be found in Zhao, Small, and Su (2018) for global null testing.

Fourth, conditional inference allows us to make full use of the *leftover information* (Fithian, Sun, and Taylor, 2014). In Chapter 4, while part of the information is lost due to the selection bias on the original experiment, we do not need to forgo the original experiment altogether. With conditional inference, we are able to optimally combine the leftover information of the original experiment and the replication to give a test and an estimator that demonstrably outperform the existing methods.

As the typical dataset size skyrockets, the common data analysis process becomes more complex and more data-dependent. While the topics of conditional inference pre-

sented may seem idealized in comparison, the increasing computation power and theoretical advances can realize more complex versions of these ideas to provide more and better inferences.

Appendix A

Appendix for Chapter 3

To investigate how well the normal approximation works to t-distributions, we first consider the typical degrees of freedom for the t-distributions. Figure A.1 shows the degrees of freedom in the original and replication experiments where both are at least 30. All but one study pair falls in the blue region, and hence the grid points marked by "+" are generally representative.

For each grid point, we simulate a pair of one-sided t-tests with the same effect sizes. We generate

$$T_O \sim t_{df_O}(\text{ncp}_O)1_{\{|T_O|>t_{df_O,\alpha/2}\}} \qquad \text{and} \qquad T_R \sim t_{df_R}(\text{ncp}_R).$$

Since the original sample size is typically chosen to achieve a certain power, we assume ncp_O stays small. The type I error rate of the selective z-test is given in red in Figure A.2. The type I error rate can deviate from 0.05 as the noncentrality parameter grows.

This deviation is caused by inaccuracy of approximating a noncentral t-distribution with a location-shifted standard Gaussian. We propose a finite sample correction by approximating the distribution of T_O with

$$T_O \sim t_{df_O}(\text{ncp}_O)1_{\{|T_O|>t_{df_O,\alpha/2}\}} \approx N\left(\text{ncp}_O, 1+\frac{2\text{ncp}_O^2}{df_O}\right)1_{\{|T_O|>t_{df_O,\alpha/2}\}}$$

and approximating the distribution of T_R similarly. Note that the distribution of the test statistic relies on the unknown noncentrality parameter, which we replace with a plug-in estimator based on T_R: T_R can stand in for ncp_R, as well as ncp_O through the common effect size. The resulting type I error rate behaves better and is given in blue in Figure A.2.

With the finite sample correction, five (11%) studies are rejected. Controlling the false discovery rate at 0.10, we apply Benjamini–Hochberg procedure (1995) and rule four (9%) replication studies as inconsistent with the original studies, namely Dodson, Darragh, and Williams (2008), Farris et al. (2008), Purdie-Vaughns et al. (2008), and van Dijk et al. (2008), generally in line with our results without the finite sample correction. Farris et al. (2008) remains rejected at familywise error rate 0.05. To check if our assumptions still hold reasonably well, we recreate Figure A.2 with the effective p-value threshold of

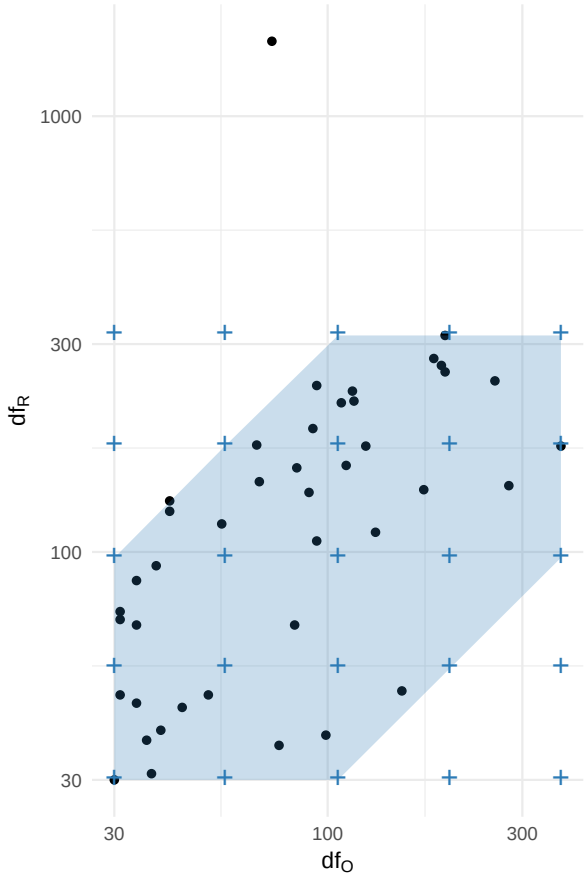


Figure A.1: Degrees of freedom in the original and replication experiments where both are at least 30, on log-scale. The blue region covers all but one study pair, and hence the grid points we choose, marked as "+", are generally representative.

0.004 ($= \frac{4}{46} \cdot 0.05$) used in Benjamini–Hochberg procedure and 0.001 ($= \frac{1}{46} \cdot 0.05$), given in Figure A.3 and Figure A.4 respectively.

For sake of completeness, we repeat the above plots specifically for the outlier (Study 97; Purdie-Vaughns et al., 2008) with exceptionally large replication degree of freedom, in Figure A.5.

Our overestimate, underestimate and confidence interval for the proportion of effect sizes that declined remain the same, but we now estimate conservatively that 14 (30%) of the effect sizes declined by at least 20% with a 95% lower confidence bound of three (7%).

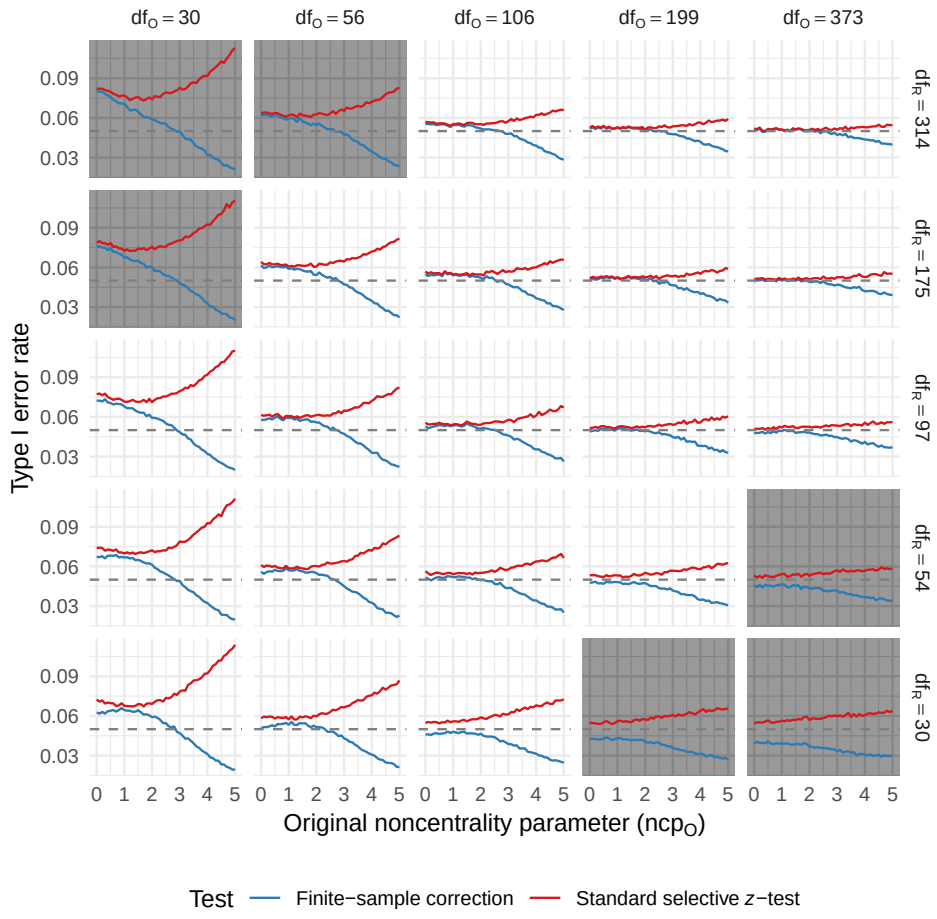


Figure A.2: The type I error rate as a function of the noncentrality parameter, based on a simulation. The type I error rate of the simple selective z-test is in red, which can deviate from 0.05 when the noncentrality parameter is large. The type I error rate of the selective z-test with our proposed finite sample correction is in blue, and stay mostly controlled. Extreme differences in degrees of freedom, as indicated by the gray background, is absent.

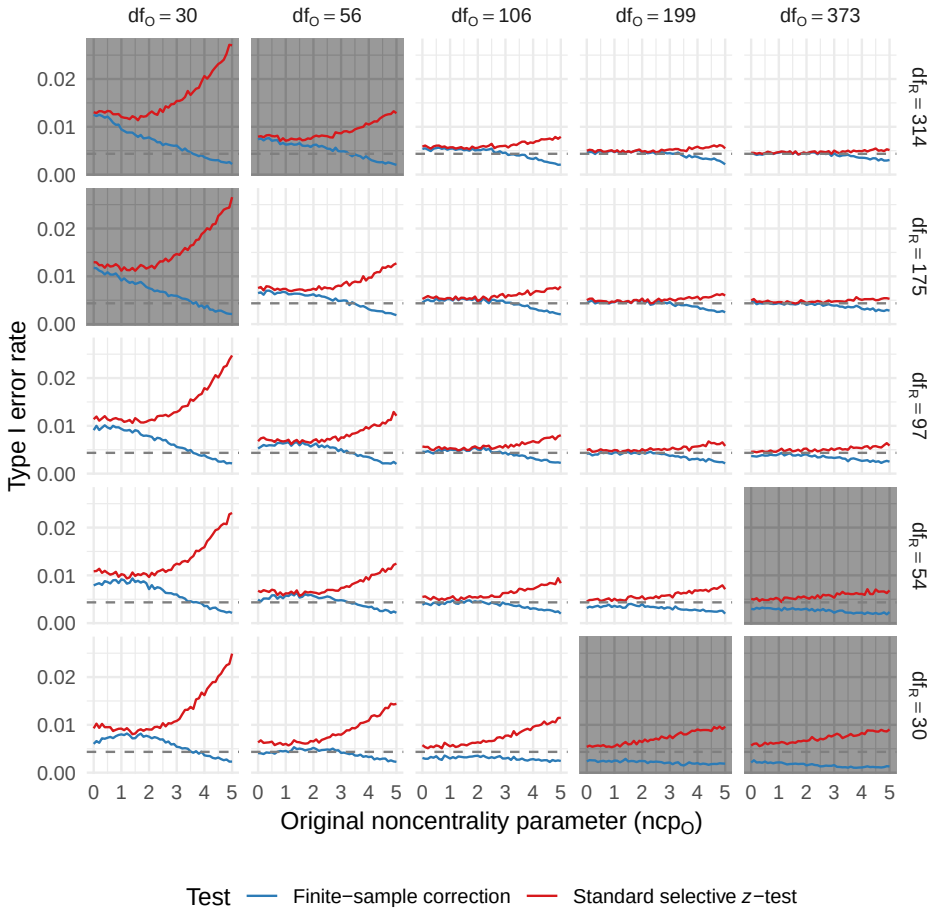


Figure A.3: The type I error rate as a function of the noncentrality parameter, based on a simulation. The type I error rate of the simple selective z-test is in red, which can deviate from 0.004 when the noncentrality parameter is large. The type I error rate of the selective z-test with our proposed finite sample correction is in blue. Extreme differences in degrees of freedom, as indicated by the gray background, is absent.

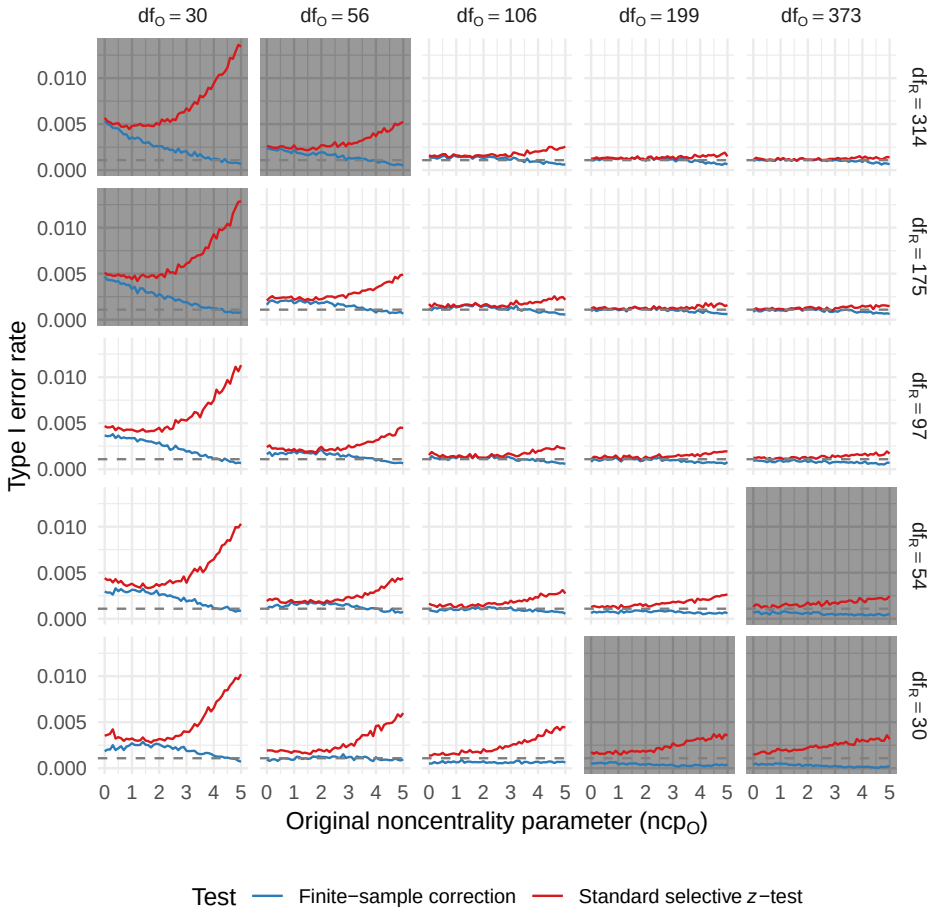


Figure A.4: The type I error rate as a function of the noncentrality parameter, based on a simulation. The type I error rate of the simple selective z-test is in red, which can deviate from 0.001 when the noncentrality parameter is large. The type I error rate of the selective z-test with our proposed finite sample correction is in blue. Extreme differences in degrees of freedom, as indicated by the gray background, is absent.

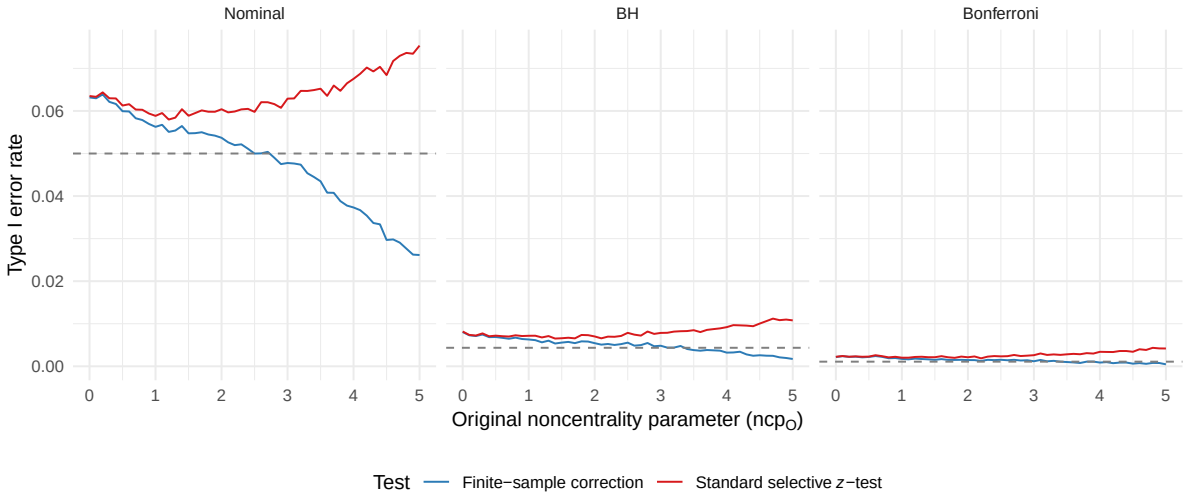


Figure A.5: The type I error rate as a function of the noncentrality parameter, based on a simulation. The type I error rate of the simple selective z-test is in red and the type I error rate of the selective z-test with our proposed finite sample correction is in blue. The error rate is evaluated for a test with the nominal level, the effective level from Benjamini–Hochberg procedure and from Bonferroni correction.

Appendix B

Appendix for Chapter 4

A useful function in implementing the UMPU test is the cumulative distribution function of S under θ,

$$F_\theta(s) = \mathbb{P}_\theta[S \leq s].$$

From (4.3), $F_\theta(s)$ can be expressed as

$$F_\theta(s) = \frac{\int_{-\infty}^{s} \bar{\Phi}_{0,\sigma_O^2+\sigma_R^2}(a(s))\phi_{(1/\sigma_O^2+1/\sigma_R^2)\theta,1/\sigma_O^2+1/\sigma_R^2}(s)\,ds}{\Phi(c/\sigma_O)},$$

where $\phi_{\mu,\sigma^2}(\cdot)$ and $\bar{\Phi}_{\mu,\sigma^2}(\cdot)$ are the p.d.f. and the survival function of $N(\mu,\sigma^2)$. The condition (4.4) can be rewritten as $F_\theta(a) + 1 - F_\theta(b) = \alpha$ and (4.5) as

$$\frac{\partial}{\partial\theta}(F_\theta(a) + 1 - F_\theta(b)) = 0$$
$$\frac{\partial}{\partial\theta}F_\theta(a) = \frac{\partial}{\partial\theta}F_\theta(b).$$

Since $\partial F_\theta(s)/\partial\theta$ is a unimodal function in s, we can perform a grid search for a and b such that the conditions above are satisfied in linear time.

Bibliography

Achenbach, Joel (Aug. 2015). "Many scientific studies can't be replicated. That's a problem". In: *The Washington Post*.

Alikhani, Lida (2011). *Study: Tween TV today is all about fame*. URL: `http://thechart.blogs.cnn.com/2011/08/05/study-tweens-aim-for-fame-above-all-else/` (visited on 07/26/2016).

Amrhein, Valentin, Fränzi Korner-Nievergelt, and Tobias Roth (2017). "The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research". In: *PeerJ* 5.2, e3544.

Anderson, Christopher J et al. (Mar. 2016). "Response to Comment on "Estimating the reproducibility of psychological science"". In: *Science* 351.6277, p. 1037c.

Andrews, Isaiah and Maximilian Kasy (May 2018). "Identification of and correction for publication bias". In: *GitHub*, pp. 1–85.

Baker, Monya (Aug. 2015). *Over half of psychology studies fail reproducibility test*. URL: `http://www.nature.com/doifinder/10.1038/nature.2015.17433`.

Barrett, Lisa Feldman (Sept. 2015). "Psychology Is Not in Crisis". In: *The New York Times*, A23.

Benjamin, Daniel J et al. (Jan. 2018). "Redefine statistical significance". In: *Nature Human Behaviour* 2, pp. 6–10.

Benjamini, Yoav and Ruth Heller (2008). "Screening for partial conjunction hypotheses". In: *Biometrics*.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 57.1, pp. 289–300.

— (2000). "On the adaptive control of the false discovery rate in multiple testing with independent statistics". In: *Journal of Educational and Behavioral Statistics* 25.1, pp. 60–83.

Benjamini, Yoav and Daniel Yekutieli (Mar. 2005). "False Discovery Rate-Adjusted Multiple Confidence Intervals for Selected Parameters". In: *Journal of the American Statistical Association* 100.469, pp. 71–81.

Berger, Roger L (1980). "Minimax subset selection for the multinomial distribution". In: *Journal of Statistical Planning and Inference* 4.4, pp. 391–402.

— (1982). "Multiparameter hypothesis testing and acceptance sampling". In: *Technometrics*.

Besag, Julian and Peter Clifford (1989). "Generalized monte carlo significance tests". In: *Biometrika* 76.4, pp. 633–642.

Blackwell, David (1947). "Conditional Expectation and Unbiased Sequential Estimation". In: *The Annals of Mathematical Statistics* 18.1, pp. 105–110. ISSN: 0003-4851. DOI: 10.1214/aoms/1177730497.

Bofinger, Eve (June 1991). "Selecting "Demonstrably best" or "Demonstrably worst" exponential population". In: *Australian Journal of Statistics* 33.2, pp. 183–190.

Camerer, Colin F et al. (Mar. 2016). "Evaluating replicability of laboratory experiments in economics". In: *Science* 351.6280, pp. 1433–1436.

Camerer, Colin F et al. (Aug. 2018). "Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015". In: *Nature Human Behaviour* 343, pp. 229–268.

Carey, Benedict (Aug. 2015). "Many psychology findings not as strong as claimed, study says". In: *The New York Times*, A1.

Dodson, Chad S, James Darragh, and Allison Williams (2008). "Stereotypes and retrieval-provoked illusory source recollections". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34.3, pp. 460–477.

Duval, Sue and Richard Tweedie (June 2000). "Trim and Fill: A Simple Funnel-Plot–Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis". In: *Biometrics* 56.2, pp. 455–463.

Edwards, Donald G and Jason C Hsu (1983). "Multiple comparisons with the best treatment". In: *Journal of the American Statistical Association* 78.384, pp. 965–971.

Etz, Alexander and Joachim Vandekerckhove (Feb. 2016). "A Bayesian Perspective on the Reproducibility Project: Psychology". In: *PLOS ONE* 11.2, e0149794–12.

Farris, Coreen et al. (Apr. 2008). "Perceptual mechanisms that characterize gender differences in decoding women's sexual intent". In: *Psychological Science* 19.4, pp. 348–354.

Finner, H and K Strassburger (2002). "The partitioning principle: a powerful tool in multiple decision theory". In: *The Annals of Statistics* 30.4, pp. 1194–1213.

Fisher, Ronald Aylmer (1921). "On the 'probable error' of a coefficient of correlation deduced from a small sample". In: *Metron* 1, pp. 3–32.

— (1924). "The distribution of the partial correlation coefficient". In: *Metron* 3, pp. 329–332.

— (1925). *Statistical methods for research workers*. Edinburgh Oliver & Boyd.

Fithian, William, Dennis L Sun, and Jonathan E Taylor (Oct. 2014). "Optimal Inference After Model Selection". In: *arXiv.org*. arXiv: 1410.2597v2 [math.ST].

Fithian, William, Jonathan E Taylor, and Ryan J Tibshirani (Dec. 2015). "Selective Sequential Model Selection". In: *arXiv.org*. arXiv: 1512.02565v1 [stat.ME].

Gelman, Andrew and John Carlin (Nov. 2014). "Beyond Power Calculations". In: *Perspectives on Psychological Science* 9.6, pp. 641–651.

Gelman, Andrew and Eric Loken (Nov. 2013). *The garden of forking paths*. URL: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Gelman, Andrew and Keith O'Rourke (Dec. 2013). "Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values". In: *Biostatistics* 15.1, pp. 18–23.

Gelman, Andrew and Francis Tuerlinckx (2000). "Type S error rates for classical and Bayesian single and multiple comparison procedures". In: *Computational Statistics* 15.3, pp. 373–390.

Gilbert, Daniel T et al. (Mar. 2016a). *A Response to the Reply to Our Technical Comment on "Estimating the Reproducibility of Psychological Science"*. URL: `https://projects.iq.harvard.edu/files/psychology-replications/files/gkpw_response_to_osc_rebutal.pdf`.

— (2016b). "Comment on "Estimating the reproducibility of psychological science"". In: *Science* 351.6277, 1037a.

— (Mar. 2016c). *More on "Estimating the Reproducibility of Psychological Science"*. URL: `https://projects.iq.harvard.edu/files/psychology-replications/files/gkpw_post_publication_response.pdf`.

Goodman, Steven N (Dec. 2013). "Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature". In: *Biostatistics* 15.1, pp. 23–27.

Goodman, Steven N, Daniele Fanelli, and John P A Ioannidis (June 2016). "What does research reproducibility mean?" In: *Science Translational Medicine* 8.341, 341ps12.

Gupta, Shanti Swarup, Deng-Yuan Huang, and S Panchapakesan (1984). "On some inequalities and monotonicity results in selection and ranking theory". In: *Inequalities in statistics and probability (Lincoln, Neb., 1982)*. Hayward, CA: Inst. Math. Statist., Hayward, CA, pp. 211–227.

Gupta, Shanti Swarup and TaChen Liang (1989). "Selecting the best binomial population: parametric empirical Bayes approach". In: *Journal of Statistical Planning and Inference* 23.1, pp. 21–31.

Gupta, Shanti Swarup and Klaus Nagel (1967). "On selection and ranking procedures and order statistics from the multinomial distribution". In: *Sankhyā: The Indian Journal of Statistics* 29.

Gupta, Shanti Swarup and S Panchapakesan (Dec. 1971). *On Multiple Decision (Subset Selection) Procedures*. Tech. rep. Purdue University.

— (Feb. 1985). "Subset Selection Procedures: Review and Assessment". In: *American Journal of Mathematical and Management Sciences* 5.3-4, pp. 235–311.

Gupta, Shanti Swarup and Wing-Yue Wong (July 1976). *On Subset Selection Procedures for Poisson Processes and Some Applications to the Binomial and Multinomial Problems*. Tech. rep.

Gutmann, Sam and Zakhar Maymin (1987). "Is the selected population the best?" In: *The Annals of Statistics* 15.1, pp. 456–461.

Hedges, Larry V (1992). "Modeling publication selection effects in meta-analysis". In: *Statistical Science* 7.2, pp. 246–255.

Heller, Ruth et al. (2007). "Conjunction group analysis: an alternative to mixed/random effect analysis". In: *Neuroimage* 37.4, pp. 1178–1185.

Holm, Sture (1979). "A simple sequentially rejective multiple test procedure". In: *Scandinavian Journal of Statistics* 6.2, pp. 65–70.

Hsu, Jason C (Sept. 1984). "Constrained Simultaneous Confidence Intervals for Multiple Comparisons with the Best". In: *The Annals of Statistics* 12.3, pp. 1136–1144.

— (1996). *Multiple comparisons: theory and methods*. CRC Press.

Hung, Kenneth and William Fithian (2019a). "Rank Verification for Exponential Families". In: *The Annals of Statistics* 47.2, pp. 758–782. URL: https://doi.org/10.1214/17-AOS1634.

— (2019b). "Statistical Methods for Replicability Assessment".

Ioannidis, John P A (Dec. 2013). "Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false". In: *Biostatistics* 15.1, pp. 28–36.

Jager, Leah R and Jeffrey T Leek (Dec. 2013). "An estimate of the science-wise false discovery rate and application to the top medical literature". In: *Biostatistics* 15.1, pp. 1–12.

Johnson, Valen E et al. (Mar. 2017). "On the Reproducibility of Psychological Science". In: *Journal of the American Statistical Association* 112.517, pp. 1–10.

Karnnan, Nandini and S Panchapakesan (2009). "Does the Selected Normal Population Have the Smallest Variance?" In: *American Journal of Mathematical and Management Sciences* 29.1-2, pp. 109–123.

Keener, Robert W (Jan. 2010). *Theoretical Statistics*. New York, NY, USA: Springer, pp. xviii–538. ISBN: 1431-875X. URL: http://dx.doi.org/10.1007/978-0-387-93839-4.

Klein, Richard A et al. (Oct. 2018). "Many Labs 2: Investigating Variation in Replicability Across Sample and Setting". URL: https://psyarxiv.com/9654g/.

Larsen, Jeff T and Amie R McKibban (2008). "Is Happiness Having What You Want, Wanting What You Have, or Both?" In: *Psychological Science* 19.4, pp. 371–377.

Lee, Jason D et al. (2016). "Exact post-selection inference, with application to the lasso". In: *The Annals of Statistics* 44.3, pp. 907–927.

Lipták, Tamás (1958). "On the Combination of Independent Tests". In: *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei* 3.1-2, pp. 171–197.

Marshall, Albert W, Ingram Olkin, and Barry Arnold (Dec. 2010). *Inequalities: Theory of Majorization and Its Applications*. Springer Series in Statistics. New York, NY: Springer.

Maymin, Zakhar and Sam Gutmann (1992). "Testing retrospective hypotheses". In: *The Canadian Journal of Statistics. La Revue Canadienne de Statistique* 20.3, pp. 335–345.

Morey, Richard D and Daniël Lakens (2017). "Why most of psychology is statistically unfalsifiable". URL: https://github.com/richarddmorey/psychology_resolution.

Nettleton, Dan (Sept. 2009). "Testing for the supremacy of a multinomial cell probability". In: *Journal of the American Statistical Association* 104.487, pp. 1052–1059.

Ng, H K T and S Panchapakesan (2007). "Is the selected multinomial cell the best?" In: *Sequential Analysis* 26, pp. 415–423.

Nosek, Brian A and Timothy M Errington (Jan. 2017). "Reproducibility in Cancer Biology: Making sense of replications". In: *eLife* 6, e23383.

Nosek, Brian A and Elizabeth Gilbert (Mar. 2016). *Let's not mischaracterize replication studies: authors*. Blog. URL: https://retractionwatch.com/2016/03/07/lets-not-mischaracterize-replication-studies-authors/.

Olkin, Ingram and John W. Pratt (2007). "Unbiased Estimation of Certain Correlation Coefficients". In: *The Annals of Mathematical Statistics* 29.1, pp. 201–211. ISSN: 0003-4851. DOI: 10.1214/aoms/1177706717.

Open Science Collaboration (2015). "Estimating the reproducibility of psychological science". In: *Science* 349.6251, p. 943.

Purdie-Vaughns, Valerie et al. (2008). "Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions". In: *Journal of Personality and Social Psychology* 94.4, pp. 615–630.

Quinnipiac University Poll Institute (2016). *First-Timers Put Trump Ahead In Iowa GOP Caucus, Quinnipiac University Poll Finds; Sanders Needs First-Timers To Tie Clinton In Dem Caucus*. URL: http://www.quinnipiac.edu/images/polling/ia/ia02012016_Ifsmb28.pdf (visited on 03/18/2016).

Sampson, Allan R and Michael W Sill (2005). "Drop-the-Losers Design: Normal Case". In: *Biometrical Journal* 47.3, pp. 257–268.

Simonsohn, Uri, Leif D Nelson, and Joseph P Simmons (2014a). "*P*-Curve: a Key to the File-Drawer". In: *Journal of Experimental Psychology: General* 143.2, pp. 534–547.

— (Nov. 2014b). "*p*-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results". In: *Perspectives on Psychological Science* 9.6, pp. 666–681.

Srivastava, Sanjay (Sept. 2015). *Moderator interpretations of the Reproducibility Project*. Blog. URL: https://thehardestscience.com/2015/09/02/moderator-interpretations-of-the-reproducibility-project/.

— (Mar. 2016). *Evaluating a new critique of the Reproducibility Project*. Blog. URL: https://thehardestscience.com/2016/03/03/evaluating-a-new-critique-of-the-reproducibility-project/.

Stefansson, Gunnar, Woo-Chul Kim, and Jason C Hsu (1988). "On confidence sets in multiple comparisons". In: *Statistical Decision Theory and Related Topics IV. ...* Decision Theory and ..., pp. 89–104.

Storey, John D (July 2002). "A direct approach to false discovery rates". In: *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 64.3, pp. 479–498.

Stroebe, Wolfgang (Sept. 2016). "Are most published social psychological findings false?" In: *Journal of Experimental Social Psychology* 66.C, pp. 134–144.

The Economist (Feb. 2016). "The scientific method". In: *The Economist*.

Uhls, Yalda T and Patricia M Greenfield (2012). "The value of fame: preadolescent perceptions of popular media and their relationship to future aspirations." In: *Developmental psychology*.

Valentine, Jeffrey C et al. (May 2011). "Replication in Prevention Science". In: *Prevention Science* 12.2, pp. 103–117.

Van Aert, Robbie C M and Marcel A L M van Assen (2017). "Bayesian evaluation of effect size after replicating an original study". In: *PLoS ONE* 12.4, e0175302–23. ISSN:

1932-6203. DOI: 10.1371/journal.pone.0175302. URL: http://dx.plos.org/10.1371/journal.pone.0175302.

Van Aert, Robbie C M and Marcel A L M van Assen (2018). "Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication". In: *Behavior Research Methods* 50.4, pp. 1515–1539. ISSN: 15543528. DOI: 10.3758/s13428-017-0967-6.

Van Dijk, Eric et al. (2008). "A social functional approach to emotions in bargaining: When communicating anger pays and when it backfires". In: *Journal of Personality and Social Psychology* 94.4, pp. 600–614.

Weinstein, Asaf, William Fithian, and Yoav Benjamini (2013). "Selection adjusted confidence intervals with more power to determine the sign". In: *Journal of the American Statistical Association* 108.501, pp. 165–176.

Yekutieli, Daniel (2012). "Adjusted Bayesian inference for selected parameters". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3, pp. 515–541.

Zhao, Qingyuan, Dylan S Small, and Weijie J Su (July 2018). "Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions". In: *Journal of the American Statistical Association*. URL: https://www.tandfonline.com/doi/full/10.1080/01621459.2018.1497499.

Zöllner, Sebastian and Jonathan K Pritchard (2007). "Overcoming the winner's curse: estimating penetrance parameters from case-control data". In: *The American Journal of Human Genetics* 80.4, pp. 605–615.