Multiple Hypothesis Testing in Experimental Economics*

John List
Department of Economics
University of Chicago
jlist@uchicago.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Yang Xu
Department of Economics
University of Chicago
yangxu@uchicago.edu

December 27, 2015

Abstract

Empiricism in the sciences allows us to test theories, formulate optimal policies, and learn how the world works. In this manner, it is critical that our empirical work provides accurate conclusions about underlying data patterns. False positives represent an especially important problem, as vast public and private resources can be misguided if we base decisions on false discovery. This study explores one especially pernicious influence on false positives—multiple hypothesis testing (MHT). While MHT potentially affects all types of empirical work, we consider three common scenarios where MHT influences inference within experimental economics: jointly identifying treatment effects for a set of outcomes, estimating heterogeneous treatment effects through subgroup analysis, and conducting hypothesis testing for multiple treatment conditions. Building upon the work of Romano and Wolf (2010), we present a correction procedure that incorporates the three scenarios, and illustrate the improvement in power by comparing our results with those obtained by the classic studies due to Bonferroni (1935) and Holm (1979). Importantly, under weak assumptions, our testing procedure asymptotically controls the familywise error rate – the probability of one false rejection – and is asymptotically balanced. We showcase our approach by revisiting the data reported in Karlan and List (2007), to deepen our understanding of why people give to charitable causes.

KEYWORDS: Experiments, Multiple Hypothesis Testing, Multiple Treatments, Multiple Outcomes, Multiple Subgroups, Randomized Controlled Trial, Bootstrap, Balance

*We would like to thank Joseph P. Romano for helpful comments on this paper. We also thank Joseph Seidel for his excellent research assistance. The research of the second author was supported by National Science Foundation Grants DMS-1308260 and SES-1227091. Documentation of our procedures and our Stata and Matlab code can be found here: https://github.com/seidelj/mht.

JEL classification codes: C12, C14

"What was observed by us in the third place is the nature or matter of the Milky Way itself, which, with the aid of the spyglass, may be observed so well that all the disputes that for so many generations have vexed philosophers are destroyed by visible certainty, and we are liberated from wordy arguments." - Galileo Galilei (1610)

"In general, we look for a new law by the following process. First, we guess it (audience laughter), no, don't laugh, that's really true. Then we compute the consequences of the guess, to see what, if this is right, if this law we guess is right, to see what it would imply and then we compare the computation results to nature, or we say compare to experiment or experience, compare it directly with observations to see if it works.

If it disagrees with the experiment, it's wrong. In that simple statement is the key to science. It doesn't make any difference how beautiful your guess is, it doesn't matter how smart you are who made the guess, or what his name is... If it disagrees with experiment, it's wrong. That's all there is to it." - Richard Feyman's lectures

1 Introduction

As the world continually provides deeper and more informative data, decision-makers are increasingly relying on the scientific approach to inform their decisions. In those cases where false discovery drives resource allocation and future streams of thought, the private and social costs can be quite high. Consider drug trials. In 2011, the success rate of Phase 2 trials was 18 percent and in 2006 that rate was 28 percent (Arrowsmith, 2011a). The same trend for Phase 3 trials was 50 percent in 2011 and 80 percent in 2001 (Arrowsmith, 2011b). These failed trials are quite expensive with the cost of running one ranging from $5 to $25 million depending on the number of patients involved. Examples abound for individuals as well: a recent study due to Ong and Mandl (2015) estimates that, just for U.S. women aged 40-59, we spend $4 billion a year on unnecessary medical costs due to mammograms that generate false alarms, and on unnecessary treatment of tumors.

The stakes are raised considerably when empirical methods are further used to test the foundations of theoretical models. In these cases, whether the theory is rejected critically affects future research directions, potentially leading to profound paradigmatic changes. This is especially true when initial results are difficult to overturn because the culture of the field is to "file drawer" results at odds with received findings or when failed replications are difficult to publish.

In this study, we consider one key reason for false positives—multiple hypothesis testing (MHT). MHT refers to any instance in which a family of hypotheses is carried out simultaneously and the analyst must decide which hypotheses to reject. Within the broad class of empiricism, there are three common scenarios that involve MHT. The first is when the analyst jointly identifies treatment effects for a set of outcomes. In such cases, theory provides several dependent variables that are potentially affected by treatment, and the analyst explores each in turn. As an example, consider the Perry Pre-School intervention, which was carried out from 1962-67, and provided pre-school education to three and four year old African American children living in poverty. By now, the outcomes measured run in the dozens and range from educational attainment to social outcomes, such as crime rates (see, e.g., Heckman et al., 2011). While this is but one example, the problem is ubiquitous. Anderson (2008) reports that 84% of experimental papers published from 2004 to 2006 in a set of social sciences fields jointly test five or more outcomes, and 61% have ten or more outcomes simultaneously tested. Yet, only 7% of these papers conduct any multiplicity correction.

Second, in many cases the analyst is interested in whether the treatment has heterogeneous effects. For instance, it might be the case that the intervention affects Hispanics differently than African Americans, or women differently than men. One set of studies in this spirit is the work on gender and competition (see, e.g., Gneezy et al., 2003; Niederle and Vesterlund, 2007; Flory et al., 2014). In this case, the researcher begins by testing for an overall treatment effect, then explores whether the treatment affects men differently than women (or perhaps explores effects across different subgroups, such as old and young (see, e.g., Sutter and Glätzle-Rützler, 2014; Flory et al., 2015)).

In this way, subgroup analysis is a key contribution of the study. How pervasive is this particular problem? Fink et al. (2011) review all field-experiment based articles published in premier economics journals and find that 76% of the 34 articles that they study involve subgroup analysis, and 29% estimate treatment effects for ten or more subgroups. Yet, few of them properly correct for MHT in their subsequent statistical analyses.

Third, hypothesis testing is often conducted for multiple treatment groups. In particular, the third scenario may typically include two cases - assessing treatment effects for multiple treatment conditions and making all pairwise comparisons across multiple treatment conditions and a control condition. In testing any single hypothesis, we typically conduct a t-test where the Type I error rate is set at α. That is, for each single hypothesis, the probability of rejecting the null hypothesis when it is true is α. When multiple hypotheses are considered together, however, the probability that at least some Type I errors are committed would often increase dramatically with the number of hypotheses. For instance, consider the case in which N null hypotheses being tested at the same time and all the null hypotheses are true. If all the tests are mutually independent, then the probability of at least one true null hypothesis being rejected would equal

to $1-(1-\alpha)^N$. Setting $\alpha = 0.05$, the chance of rejecting at least one true hypothesis would be 0.226 for 5 hypotheses, 0.401 for 10 hypotheses, and 0.994 for 100 hypotheses.

If an experimental economist focuses on a particular hypothesis, then the conventional approaches such as a t-test would still be valid. However, one often needs to consider multiple hypotheses simultaneously and search for any rejection from a group of tests. Without taking into account the statistical inference problem arising from multiple testing, experimentalists would be quite likely to undertake a substantially large false discovery risk and draw ungrounded conclusions. This third case presents itself in nearly every experiment that is published today.

The multiple testing issue is well known in the econometric theory field (e.g. Romano and Wolf, 2005a,b; Romano and Shaikh, 2006b; Romano et al., 2008b) and the biostatistics literature (e.g. Hochberg, 1988; O'Brien, 1984; Marcus et al., 1976). MHT has also been pervasive in the experimental economics community. Indeed, a growing body of literature in experimental economics has begun to address the multiple testing problem. In some of the previous literature, adjustments for multiple testing have been made when making inferences for only multiple outcomes (e.g. Anderson, 2008; Casey et al., 2012; Heckman et al., 2011; Kling et al., 2007), only multiple subgroups (e.g. Lee and Shaikh, 2014), and for both multiple outcomes and subgroups (e.g. Heckman et al. 2010).

Nevertheless, the potential solutions have yet to be systematically discussed. The goal of our study is to fill this gap, and provide the necessary code for others to easily use our approach. To achieve control of the familywise error rate in finite samples, Romano and Wolf (2005a) develop a stepdown procedure for multiple testing based upon permutation and randomization tests under the monotonicity assumption of estimated critical values.[1] However, multiple testing involving multiple treatment conditions may result in violation of the monotonicity assumption so that this finite-sample procedure may not be valid.[2] Similarly, Romano and Wolf (2005a) introduce asymptotically valid procedures based upon bootstrap and subsampling methods. Building upon previous work, Romano and Wolf (2010) demonstrate how to construct asymptotically balanced procedures by choosing appropriate test statistics so that all marginal probabilities of rejecting any true null hypothesis are approximately equal.

In this paper, we exploit results on bootstrap multiple testing procedures discussed in Romano and Wolf (2010), and construct a asymptotically balanced testing procedure that incorporates the three common scenarios for multiple testing in experimental economics. Under weak assumptions, the resulting testing

[1] See, e.g., Heckman et al. 2011; Lee and Shaikh 2014, for applications of this testing procedure.

[2] As discussed in Romano and Wolf (2005a), a simple solution would be to use the maximum of the critical values of all the subset hypotheses. However, this approach can be computationally prohibitive. Furthermore, the above modification may result in a dramatic decline in power due to larger critical values so that the resulting procedure may be inferior to classical multiple testing procedures - like Bonferroni (1935) and Holm (1979) - in terms of power.

procedure asymptotically controls the familywise error rate – the probability of even one false rejection. The methodology that we introduce differs from classical multiple testing procedures , such as Bonferroni (1935) and Holm (1979), in that it incorporates information about the joint dependence structure of the test statistics when determining which null hypotheses to reject. This leads to important gains in power.

We showcase our approach by applying it to various hypotheses of interest using data originally presented in Karlan and List (2007). The Karlan and List (2007) study explored the economics of charity by measuring, amongst other things, the effectiveness of a matching grant on charitable giving. We illustrate the improvement in power by comparing our results with those obtained by Bonferroni and Holm. Notably, we also demonstrate that further improvements could be made by exploiting transitivity and imposing smaller critical values when there are multiple treatment conditions.

The remainder of our study proceeds as follows. Section 2 describes our setup and notation. Section 3 introduces the stepwise multiple testing procedure. Section 4 demonstrates various applications of our methodology to the large-scale natural field experiment discussed in Karlan and List (2007). Section 5 concludes.

2 Setup and Notation

For $k \in \mathcal{K}$, let $Y_{i,k}$ denote the (observed) kth outcome of interest for the ith unit, D_i denote treatment status for the ith unit, and Z_i denote observed, baseline covariates for the ith unit. Further denote by $\mathcal{D}$ and $\mathcal{Z}$ the supports of D_i and Z_i, respectively. For $d \in \mathcal{D}$, let $Y_{i,k}(d)$ be the kth potential outcome for the ith unit if treatment status were (possibly counterfactually) set equal to d. As usual, the kth observed outcome and kth potential outcome are related to treatment status by the relationship

$$Y_{i,k} = \sum_{d \in \mathcal{D}} Y_{i,k}(d) I\{D_i = d\} \ .$$

It is useful to introduce the shorthand notation $Y_i = (Y_{i,k} : k \in \mathcal{K})$ and $Y_i(d) = (Y_{i,k}(d) : k \in \mathcal{K})$. We assume that $((Y_i(d) : d \in \mathcal{D}), D_i, Z_i), i = 1, \ldots, n$ are i.i.d. with distribution $Q \in \Omega$, where our requirements on Ω are specified below. It follows that the observed data $(Y_i, D_i, Z_i), i = 1, \ldots, n$ are i.i.d. with distribution $P = P(Q)$. Denote by $\hat{P}_n$ the empirical distribution of the observed data.

The family of null hypotheses of interest is indexed by

$$s \in \mathcal{S} \subseteq \{(d, d', z, k) : d \in \mathcal{D}, d' \in \mathcal{D}, z \in \mathcal{Z}, k \in \mathcal{K}\} \ .$$

For each $s \in \mathcal{S}$, define

$$\omega_s = \{Q \in \Omega : E_Q[Y_{i,k}(d) - Y_{i,k}(d')|Z_i = z] = 0\} \ .$$

Using this notation, the family of null hypotheses of interest is given by

$$H_s : Q \in \omega_s \text{ for } s \in \mathcal{S} \ . \tag{1}$$

In other words, the sth null hypothesis specifies that the average effect of treatment d on the kth outcome of interest for the subpopulation where $Z_i = z$ equals the average effect of treatment d' on the kth outcome of interest for the subpopulation where $Z_i = z$. For later use, let

$$\mathcal{S}_0(Q) = \{s \in \mathcal{S} : Q \in \omega_s\} \ .$$

Our goal is to construct a procedure for testing these null hypotheses in a way that ensures asymptotic control of the familywise error rate uniformly over $Q \in \Omega$. More precisely, we require for each $Q \in \Omega$ that

$$\limsup_{n\to\infty} FWER_Q \leq \alpha \tag{2}$$

for a pre-specified value of $\alpha \in (0, 1)$, where

$$FWER_Q = Q\{\text{reject any } H_s \text{ with } s \in \mathcal{S}_0(Q)\} \ .$$

We additionally require that the testing procedure is "balanced" in that for each $Q \in \Omega$,

$$\lim_{n\to\infty} Q\{\text{reject } H_s\} = \lim_{n\to\infty} Q\{\text{reject } H_{s'}\} \text{ for any } s \text{ and } s' \text{ in } \mathcal{S}_0(Q) \ . \tag{3}$$

We now describe our main requirements on Ω_Q. The assumptions make use of the notation

$$\begin{aligned} \mu_{k|d,z}(Q) &= E_Q[Y_{i,k}(d)|D_i = d, Z_i = z] \\ \sigma^2_{k|d,z}(Q) &= \text{Var}_Q[Y_{i,k}(d)|D_i = d, Z_i = z] \ . \end{aligned}$$

Assumption 2.1 *For each $Q \in \Omega$,*

$$(Y_i(d) : d \in \mathcal{D}) \perp\!\!\!\perp D_i|Z_i$$

under Q.

Assumption 2.2 *For each $Q \in \Omega$, $k \in \mathcal{K}$, $d \in \mathcal{D}$ and $z \in \mathcal{Z}$,*

$$0 < \sigma^2_{k|d,z}(Q) = Var_Q[Y_{i,k}(d)|D_i = d, Z_i = z] < \infty \ .$$

Assumption 2.3 *For each $Q \in \Omega$, there is $\epsilon > 0$ such that*

$$Q\{D_i = d, Z_i = z\} > \epsilon \tag{4}$$

for all $d \in \mathcal{D}$ and $z \in \mathcal{Z}$.

Assumption 2.1 simply requires that treatment status was randomly assigned. Assumption 2.2 is a mild non-degeneracy requirement. Assumption 2.3 simply requires that both D_i and Z_i are discrete random variables (with finite supports).

Remark 2.1 Note that we have assumed in particular that treatment status $D_i, i = 1, \ldots, n$ is i.i.d. While this assumption accommodates situations in which treatment status is assigned according to simple random sampling, it does not accommodate more complicated treatment assignment rules, such as those in which treatment status is assigned in order to "balance" baseline covariates among the subsets of individuals with different treatment status. For a discussion of such treatment assignment rules and the implications for inference about the average treatment effect, see Bugni et al. (2015).

Remark 2.2 When $\mathcal{S}$ is very large, requiring control of the familywise error rate may significantly limit the ability to detect genuinely false null hypotheses. For this reason, it may be desirable in such situations to relax control of the familywise error rate in favor of generalized error rates that penalize false rejections less severely. Examples of such error rates include : the m-familywise error rate, defined to be the probability of m or more false rejections; the tail probability of the false discovery proportion, defined to be the fraction of false rejections (understood to be zero if there are no rejections at all); and the false discovery rate, defined to be the expected value of false discovery proportion. Control of the m-familywise error rate and the tail probability of the false discovery proportion using resampling are discussed in Romano et al. (2008b), Romano and Wolf (2010). For procedures based only on (multiplicity-unadjusted) p-values, see Lehmann and Romano (2005), Romano and Shaikh (2006a), and Romano and Shaikh (2006b). For resampling-based control of the false discovery rate, see Romano et al. (2008a).

3 A Stepwise Multiple Testing Procedure

In this section, we describe a stepwise multiple testing procedure for testing (1) in a way that satisfies (2) and (3) for any $Q \in \Omega$. In order to do so, we first require some additional notation. To this end, first define the "unbalanced" test statistic for H_s,

$$T_{s,n} = \sqrt{n} \left| \frac{1}{n_{d,z}} \sum_{1 \leq i \leq n: D_i = d, Z_i = z} Y_{i,k} - \frac{1}{n_{d',z}} \sum_{1 \leq i \leq n: D_i = d', Z_i = z} Y_{i,k} \right| , \tag{5}$$

and its re-centered version

$$\tilde{T}_{s,n}(P) = \sqrt{n} \left| \frac{1}{n_{d,z}} \sum_{1 \leq i \leq n: D_i = d, Z_i = z} (Y_{i,k} - \tilde{\mu}_{k|d,z}(P)) - \frac{1}{n_{d',z}} \sum_{1 \leq i \leq n: D_i = d', Z_i = z} (Y_{i,k} - \tilde{\mu}_{k|d',z}(P)) \right| , \tag{6}$$

where

$$\tilde{\mu}_{k|d,z}(P) = E_P[Y_{i,k} | D_i = d, Z_i = z] .$$

Next, for $s \in \mathcal{S}$, define

$$J_n(x, s, P) = P\left\{ \tilde{T}_{s,n}(P) \leq x \right\} .$$

In order to achieve "balance," rather than reject H_s for large values of $T_{n,s}$, we reject H_s for large values of

$$J_n(T_{s,n}, s, \hat{P}_n) . \tag{7}$$

Note that (7) is simply one minus a (multiplicity-unadjusted) bootstrap p-value for testing H_s based on $T_{s,n}$. Finally, for $\mathcal{S}' \subseteq \mathcal{S}$, let

$$L_n(x, \mathcal{S}', P) = P\left\{ \max_{s \in \mathcal{S}'} J_n(\tilde{T}_{s,n}(P), s, P) \leq x \right\} .$$

Using this notation, we may describe our proposed stepwise multiple testing procedure as follows:

Algorithm 3.1

Step 0. *Set* $\mathcal{S}_1 = \mathcal{S}$.

$\vdots$

Step j. *If* $S_j = \emptyset$ *or*

$$\max_{s \in \mathcal{S}_j} J_n(T_{s,n}, s, \hat{P}_n) \leq L_n^{-1}(1 - \alpha, \mathcal{S}_j, \hat{P}_n) ,$$

then stop. Otherwise, reject any H_s *with* $J_n(T_{s,n}, s, \hat{P}_n) > L_n^{-1}(1-\alpha, \mathcal{S}_j, \hat{P}_n)$*, set*

$$\mathcal{S}_{j+1} = \{s \in \mathcal{S}_j : J_n(T_{s,n}, s, \hat{P}_n) \leq L_n^{-1}(1-\alpha, \mathcal{S}_j, \hat{P}_n)\} \ ,$$

and continue to the next step.

$\vdots$

The following theorem describes the asymptotic behavior of our proposed multiple testing procedure.

Theorem 3.1 *Consider the procedure for testing (1) given by Algorithm 3.1. Under Assumptions 2.1 – 2.3, Algorithm 3.1 satisfies (2) and (3) for any* $Q \in \Omega$*.*

Remark 3.1 If $\mathcal{S} = s$, i.e., $\mathcal{S}$ is a singleton, then the familywise error rate is simply the usual probability of a Type I error. Hence, Algorithm 3.1 provides asymptotic control of the probability of a Type I error. In this case, Algorithm 3.1 is equivalent to the usual bootstrap test of H_s, i.e., the test that rejects H_s whenever $T_{s,n} > J_n^{-1}(1-\alpha, s, \hat{P}_n)$.

Remark 3.2 As noted above, $\hat{p}_{s,n} = 1 - J_n(T_{s,n}, s, \hat{P}_n)$ may be interpreted as a bootstrap p-value for testing H_s. Indeed, for any $Q \in \omega_s$, it is possible to show that

$$\limsup_{n\to\infty} Q\{\hat{p}_{s,n} \leq u\} \leq u$$

for any $0 < u < 1$. A crude solution to the multiplicity problem would therefore be to apply a Bonferroni or Holm correction to these p-values. Such an approach would indeed satisfy (2), as desired, but implicitly relies upon a "least favorable" dependence structure among the p-values. To the extent that the true dependence structure differs from this "least favorable" one, improvements may be possible. Algorithm 3.1 uses the bootstrap to incorporate implicitly information about the dependence structure when deciding which null hypotheses to reject. In fact, Algorithm 3.1 will always reject at least as many null hypotheses as these procedures.

Remark 3.3 Implementation of Algorithm 3.1 typically requires approximating the quantities $J_n(x, s, \hat{P}_n)$ and $L_n(x, \mathcal{S}', \hat{P}_n)$ using simulation. As noted by Romano and Wolf (2010), doing so does not require nested bootstrap simulations. To explain further, for $b = 1, \ldots, B$, draw a sample of size n from $\hat{P}_n$ and denote by

$\tilde{T}^{*,b}_{s,n}(\hat{P}_n)$ the quantity $\tilde{T}_{s,n}(P)$ using the bth resample and $\hat{P}_n$ as an estimate of P. Then, $J_n(x, s, \hat{P}_n)$ may be approximated as

$$\hat{J}_n(x, s, \hat{P}_n) = \frac{1}{B} \sum_{1 \le b \le B} I\{\tilde{T}^{*,b}_{s,n}(\hat{P}_n) \le x\}$$

and $L_n(x, \mathcal{S}', \hat{P}_n)$ may be approximated as

$$\hat{L}_n(x, \mathcal{S}', \hat{P}_n) = \frac{1}{B} \sum_{1 \le b \le B} I\{\hat{J}_n(T^{*,b}_{s,n}(\hat{P}_n), s, \hat{P}_n) \le x\} \ .$$

In particular, the same set of bootstrap resamples may be used in the two approximations.

Remark 3.4 In terms of higher-order asymptotic properties, it is often desirable to studentize, i.e., to replace $T_{s,n}$ and $\tilde{T}_{s,n}(P)$, respectively, with

$$\begin{aligned} T^{\text{stud}}_{s,n} &= \frac{T_{s,n}}{\sqrt{n \cdot \left(\frac{\tilde{\sigma}^2_{k|d,z}(\hat{P}_n)}{n_{d,z}} + \frac{\tilde{\sigma}^2_{k|d',z}(\hat{P}_n)}{n_{d',z}} \right)}} \\ \tilde{T}^{\text{stud}}_{s,n}(P) &= \frac{\tilde{T}_{s,n}(P)}{\sqrt{n \cdot \left(\frac{\tilde{\sigma}^2_{k|d,z}(\hat{P}_n)}{n_{d,z}} + \frac{\tilde{\sigma}^2_{k|d',z}(\hat{P}_n)}{n_{d',z}} \right)}} \ , \end{aligned}$$

where

$$\tilde{\sigma}^2_{k|d,z}(P) = \text{Var}_P[Y_{i,k}|D_i = d, Z_i = z] \ .$$

Theorem 3.1 continues to hold with these changes.

Remark 3.5 In some cases, it may be of interest to consider one-sided null hypotheses, e.g., $H^-_s : P \in \omega^-_s$, where

$$\omega^-_s = \{Q \in \Omega : E_Q[Y_{i,k}(d) - Y_{i,k}(d')|Z_i = z] \le 0\} \tag{8}$$

In this case, it suffices simply to replace T_{s_n} and $\tilde{T}_{s_n}(P)$, respectively, with $T^-_{s,n}$ and $\tilde{T}^-_{s,n}(P)$, which are, respectively, defined as in (5) and (6), but without the absolute values. An analogous modification can be made for null hypotheses $H^+_s : P \in \omega^+_s$, where ω^+_s is defined as in (8), but with the inequality reversed.

Remark 3.6 Note that a multiplicity-adjusted p-value for H_s, $\hat{p}^{\text{adj}}_{s,n}$, may be computed simply as the smallest value of α for which H_s is rejected in Algorithm 3.1.

Remark 3.7 It is possible to improve Algorithm 3.1 by exploiting transitivity (i.e., $\mu_{k|d,z}(Q) = \mu_{k|d',z}(Q)$

and $\mu_{k|d',z}(Q) = \mu_{k|d'',z}(Q)$ implies that $\mu_{k|d,z}(Q) = \mu_{k|d'',z}(Q)$). To this end, for $\mathcal{S}' \subseteq \mathcal{S}$, define

$$\mathbb{S}(\mathcal{S}') = \{\mathcal{S}'' \subseteq \mathcal{S}' : \exists\ Q \in \Omega \text{ s.t. } \mathcal{S}'' = \mathcal{S}_0(Q)\}$$

and replace $L_n^{-1}(1-\alpha, \mathcal{S}_j, \hat{P}_n)$ in Algorithm 3.1 with

$$\max_{\tilde{\mathcal{S}} \in \mathbb{S}(\mathcal{S}_j)} L_n^{-1}(1-\alpha, \tilde{\mathcal{S}}, \hat{P}_n)\ .$$

With this modification to Algorithm 3.1, Theorem 3.1 remains valid. Note that this modification is only non-trivial when there are more than two treatments and may be computationally prohibitive when there are more than a few treatments.

Remark 3.8 Note that we only require that the familywise error rate is asymptotically no greater than α for each $Q \in \Omega$. By appropriately strengthening the assumptions of Theorem 3.1, it is possible to show that Algorithm 3.1 satisfies

$$\limsup_{n\to\infty} \sup_{Q\in\Omega} FWER_Q \leq \alpha\ .$$

In particular, it suffices to replace Assumption 2.2 with a mild uniform integrability requirement and require in Assumption 2.3 that there exists $\epsilon > 0$ for which (4) holds for all $Q \in \Omega$, $d \in \mathcal{D}$ and $z \in \mathcal{Z}$. Relevant results for establishing this claim can be found in Romano et al. (2012), Bhattacharya et al. (2012), and Machado et al. (2013).

4 Empirical Applications

In this section, we apply our testing methodology to the large-scale natural field experiment presented in Karlan and List (2007). As a quick summary, using direct mail solicitations targeted to previous donors of a nonprofit organization, Karlan and List (2007) focus on the effectiveness of a matching grant on charitable giving. Their sample includes all 50,083 individuals who had given to the organization at least once since 1991. These individuals were randomly assigned to two groups: a treatment "match" group (33,396, or 67 percent of the sample) and a control group (16,687 subjects, or 33 percent of the sample). The specifics of the match offer were then randomized along three dimensions: the price ratio of the match, the maximum size of the matching gift across all donations, and the example donation amount suggested to the donor.

Each of these sub-treatments (price ratio, maximum size of match, and example amount) was assigned with equal probability. There were three treatments for the price ratio of the match, \$1:\$1, \$2:\$1, and \$3:\$1.

An \$X:\$1 ratio means that for every dollar the individual donates, the matching donor also contributes \$X; hence, the charity receives \$X+1 (subject to the maximum amount across all donations). There were four treatments for the maximum matching grant amount: \$25,000, \$50,000, \$100,000, and unstated. The organization also offered three individual-specific suggested amounts equal to the individual's highest previous contribution, 1.25 times the highest previous contribution, and 1.50 times the highest previous contribution, respectively.

In the following subsections, we discuss four common scenarios of multiple hypothesis testing in experimental economics that manifest themselves in the Karlan and List (2007) data. First, we consider the MHT issue of jointly identifying treatment effects for a set of outcomes. Second, we show how to use our procedure to identify heterogeneous treatment effects through subgroup analysis. Third, we apply our methodology to settings with multiple treatment conditions. In particular, we consider two cases - assessing treatment effects for multiple treatment conditions, and making all pairwise comparisons across multiple treatment conditions and a control condition.

Notably, we demonstrate that further improvements can be made by exploiting transitivity when making all pairwise comparisons across multiple treatment conditions and a control condition. Lastly, we apply our testing procedure to the general MHT issue in experimental economics with multiple outcomes, multiple subgroups, and multiple treatments all at once. Among all four scenarios, we consider the studentized test statistics described in Remark 3.4 and compare our results with those obtained by the classic studies of Bonferroni and Holm.[3]

4.1 Multiple Outcomes

We simultaneously assess the effects of the "match" treatment for the four outcome variables: response rate, dollars given not including match, dollars given including match, and amount change. Table 1 displays for each of the four outcomes of interest, the following five quantities: column 2 displays the difference in means between the treated and the untreated subjects for the four outcomes. Column 3 displays a (multiplicity-unadjusted) p-value computed using Remark 3.1; column 4 displays a (multiplicity-adjusted) p-value computed using Theorem 3.1. Column 5 displays a (multiplicity-adjusted) p-value obtained by applying a Bonferroni adjustment to the p-values in column 3; column 6 displays a (multiplicity-adjusted) p-value obtained by applying a Holm adjustment to the p-values in column 3.

In column 3, the single testing procedure indicates that the "match" treatment has a significant effect

[3]While all the following empirical results are based upon our MATLAB code, our Stata code generates very similar results. Our Stata and Matlab code can be found here: https://github.com/seidelj/mht.

Table 1: Multiple Outcomes

Outcome	DI	*p*-values			
		Unadj.	Multiplicity Adj.		
		Remark 3.1	Thm. 3.1	Bonf.	Holm
Response Rate	0.0042	0.0003***	0.0003***	0.0013***	0.0013***
Dollars Given Not Including Match	0.1536	0.0500*	0.0967*	0.2000	0.1000
Dollars Given Including Match	1.9340	0.0003***	0.0003***	0.0013***	0.0010***
Amount Change	6.3306	0.7200	0.7200	1.0000	0.7200

Table 1: Note: DI refers to "difference in means." * and *** indicate the corresponding p-values less than 10% and 1%, respectively.

on response rate, dollars given not including match, and dollars given including match. However, by taking into account the multiplicity of the comparisons, column 4 suggests that the effect of the "match" treatment on dollars given not including match becomes only marginally significant at the 0.10 significance level.

Importantly, the p-values from Theorem 3.1 are an improvement upon those obtained by applying Bonferroni or Holm adjustments because of the incorporation of information about the joint dependence structure of the test statistics when determining which null hypotheses to reject. This feature is evident in Table 1, as the p-values in column 4 are always weakly smaller than the p-values in columns 5 and 6.

4.2 Multiple Subgroups

In this subsection, we demonstrate how to apply Remark 3.1 and Theorem 3.1 to identify heterogeneous treatment effects for multiple subgroups. We consider four subgroups contained in Karlan and List (2007): red county in a red state, blue county in a red state, red county in a blue state, and blue county in a blue state. We focus on the outcome variable response rate. Table 2 displays for each of the four subgroups of interest, the five quantities similar to those in Table 1.

Compared to the single testing procedure shown in column 3, column 4 indicates that after accounting for multiple testing the "match" treatment has an effect among a smaller number of subgroups. In particular, single hypothesis testing in column 3 indicates that the "match" treatment has a significant effect on response rate among the subgroups "Red County in a Red State" and "Blue County in a Red State." After taking into account the multiplicity issue, column 4 suggests that the treatment effect for the subgroup "Blue County in a Red State" becomes insignificant. Importantly, the procedure described in Theorem 3.1 is more powerful than the Bonferroni and Holm procedures, as the p-values in column 4 are always smaller than the p-values in columns 5 and 6.[4]

[4]105 out of the 50,083 individuals in our dataset do not have complete subgroup information. In Sections 4.2 and 4.4, we regard these 105 individuals as a subgroup of no interest for our analysis.

Table 2: Multiple Subgroups

Subgroup	DI	*p*-values			
		Unadj.	Multiplicity Adj.		
		Remark 3.1	Thm. 3.1	Bonf.	Holm
Red County in a Red State	0.0095	0.0003***	0.0003***	0.0013***	0.0013***
Blue County in a Red State	0.0070	0.0503*	0.1427	0.2013	0.1510
Red County in a Blue State	0.0016	0.4560	0.7017	1.0000	0.9120
Blue County in a Blue State	0.0000	0.9920	0.9920	1.0000	0.9920

Table 2: Note: DI refers to "difference in means." * and *** indicate the corresponding p-values less than 10% and 1%, respectively.

Table 3: Multiple Treatments (Comparing Multiple Treatments with a Control)

Treatment/Control Groups	DI	*p*-values			
		Unadj.	Multiplicity Adj.		
		Remark 3.1	Thm. 3.1	Bonf.	Holm
Control vs 1:1	0.1234	0.2627	0.2627	0.7880	0.2627
Control vs 2:1	0.2129	0.0477**	0.1297	0.1430	0.1430
Control vs 3:1	0.1245	0.2060	0.3537	0.6180	0.4120

Table 3: Note: DI refers to "difference in means." ** indicates the corresponding p-values less than 5%.

4.3 Multiple Treatments

In this empirical example, we apply our testing methodology to the multiple testing issue with multiple treatment conditions. We focus on the three treatments on matching-ratio dimension: 1:1, 2:1, and 3:1. We consider "dollars given not including match" as our outcome of interest. In particular, we address two typical scenarios in experimental economics: assessing treatment effects for multiple treatments, and making all pairwise comparisons across multiple treatments and a control condition.

For each of the three treatment effects, Table 3 displays the five quantities as described previously. By applying Remark 3.1, single testing in column 3 suggests that the match ratio 2:1 has a significant effect on the outcome "dollars given not including match." Nonetheless, as shown in column 4, the treatment effect vanishes after applying Theorem 3.1 to this multiple testing problem. Again, the empirical results confirm that the p-values from Theorem 3.1 improve upon those obtained by applying the Bonferroni or Holm procedure.

For each of the six pairwise comparisons among the treatment and control groups, Table 4 presents the corresponding five quantities as well as the p-values from Remark 3.7. Unlike all the other empirical applications, Remark 3.7 becomes non-trivial under this scenario. As shown in column 4, the treatment effect based on the pairwise comparison between the control and the match ratio 2:1 becomes negligible after accounting for the multiple testing issue.

Table 4: Multiple Treatments (All Pairwise Comparisons across Multiple Treatments and a Control)

Treatment/Control Groups	DI	*p*-values				
		Unadj.		Multiplicity Adj.		
		Remark 3.1	Thm. 3.1	Remark 3.7	Bonf.	Holm
Control vs 1:1	0.1234	0.2627	0.5810	0.4973	1.0000	1.0000
Control vs 2:1	0.2129	0.0477**	0.1930	0.1930	0.2860	0.2860
Control vs 3:1	0.1245	0.2060	0.5533	0.4167	1.0000	1.0000
1:1 vs 2:1	0.0895	0.4627	0.7467	0.7467	1.0000	1.0000
1:1 vs 3:1	0.0011	0.9920	0.9920	0.9920	1.0000	0.9920
2:1 vs 3:1	0.0883	0.4633	0.6963	0.4633	1.0000	0.9267

Table 4: Note: DI refers to "difference in means." ** indicates the corresponding p-values less than 5%.

Notice that the (multiplicity-adjusted) p-values in columns 4, 5, and 6 of Table 3 are always smaller than their counterparts in Table 4, suggesting that the multiple testing problem would often become more severe with a larger number of hypotheses. Among all of the multiplicity adjustments considered in Table 4, the procedure described in Remark 3.7 appears to be the most powerful approach. In particular, Remark 3.7 may improve upon Theorem 3.1 by exploiting transitivity and imposing smaller critical values when there are multiple treatment conditions.

4.4 Multiple Outcomes, Subgroups, and Treatments

More often than not, experimentalists wish to conduct hypothesis testing that involves multiple outcomes, multiple subgroups, and multiple treatments simultaneously (as in Karlan and List (2007)). In this subsection, we simultaneously consider the four outcome variables described in Section 4.1, the four subgroups described in Section 4.2, and the three treatment conditions described in Section 4.3. For each outcome and subgroup, we compare all of the treatments to the control group.

For each of the 48 hypotheses, Table 5 displays the corresponding five quantities. Similar to our previous discussion, after accounting for MHT we find that many of the treatment effects are no longer significant. Given such a large number of hypotheses, we can see that ignoring the multiplicity of the comparisons being made would deflate the p-values by a considerable margin. For instance, single testing by applying Remark 3.1 suggests that the null hypothesis for "response rate," "red county in a red state," and "control vs 1:1" could be rejected at the 0.05 significance level with the p-value being 0.0217. Yet, by taking multiple testing into account, Theorem 3.1 yields a much larger p-value of 0.4597. Without accounting for multiple testing, 21 null hypotheses are rejected at $p < 0.10$ level. In contrast, Theorem 3.1 indicates that with multiplicity adjustment, only 13 null hypotheses are rejected. Notice that Theorem 3.1 always gives weakly smaller p-values than the Bonferroni and Holm procedures. For some of the hypotheses, Theorem 3.1 generates strictly smaller p-values than the other two procedures.

Table 5: Multiple Outcomes, Subgroups, and Treatments

Outcome	Subgroup	Treatment/Control Groups	DI	*p*-values			
				Unadj.	Multiplicity Adj.		
				Remark 3.1	Thm. 3.1	Bonf.	Holm
Response Rate	**Red County in a Red State**	**Control vs 1:1**	0.0079	0.0217**	0.4597	1.0000	0.7367
Response Rate	**Red County in a Red State**	**Control vs 2:1**	0.0100	0.0017***	0.0453**	0.0800*	0.0650*
Response Rate	**Red County in a Red State**	**Control vs 3:1**	0.0107	0.0017***	0.0443**	0.0800*	0.0633*
Response Rate	**Blue County in a Red State**	**Control vs 1:1**	0.0024	0.5973	1.0000	1.0000	1.0000
Response Rate	**Blue County in a Red State**	**Control vs 2:1**	0.0080	0.0987*	0.8997	1.0000	1.0000
Response Rate	**Blue County in a Red State**	**Control vs 3:1**	0.0108	0.0247**	0.4950	1.0000	0.8140
Response Rate	**Red County in a Blue State**	**Control vs 1:1**	0.0003	0.9060	0.9990	1.0000	1.0000
Response Rate	**Red County in a Blue State**	**Control vs 2:1**	0.0010	0.7190	1.0000	1.0000	1.0000
Response Rate	**Red County in a Blue State**	**Control vs 3:1**	0.0034	0.2290	0.9953	1.0000	1.0000
Response Rate	**Blue County in a Blue State**	**Control vs 1:1**	0.0006	0.8667	1.0000	1.0000	1.0000
Response Rate	**Blue County in a Blue State**	**Control vs 2:1**	0.0026	0.5033	1.0000	1.0000	1.0000
Response Rate	**Blue County in a Blue State**	**Control vs 3:1**	0.0032	0.3740	1.0000	1.0000	1.0000
Dollars Given Not Including Match	**Red County in a Red State**	**Control vs 1:1**	0.4260	0.0903*	0.9027	1.0000	1.0000
Dollars Given Not Including Match	**Red County in a Red State**	**Control vs 2:1**	0.4097	0.0557*	0.7813	1.0000	1.0000
Dollars Given Not Including Match	**Red County in a Red State**	**Control vs 3:1**	0.3214	0.0710*	0.8483	1.0000	1.0000
Dollars Given Not Including Match	**Blue County in a Red State**	**Control vs 1:1**	0.0374	0.8950	1.0000	1.0000	1.0000
Dollars Given Not Including Match	**Blue County in a Red State**	**Control vs 2:1**	0.4325	0.1853	0.9853	1.0000	1.0000
Dollars Given Not Including Match	**Blue County in a Red State**	**Control vs 3:1**	0.5728	0.0933*	0.8983	1.0000	1.0000
Dollars Given Not Including Match	**Red County in a Blue State**	**Control vs 1:1**	0.0256	0.8683	1.0000	1.0000	1.0000
Dollars Given Not Including Match	**Red County in a Blue State**	**Control vs 2:1**	0.0928	0.5893	1.0000	1.0000	1.0000
Dollars Given Not Including Match	**Red County in a Blue State**	**Control vs 3:1**	0.0243	0.8847	1.0000	1.0000	1.0000
Dollars Given Not Including Match	**Blue County in a Blue State**	**Control vs 1:1**	0.0074	0.9747	0.9747	1.0000	0.9747
Dollars Given Not Including Match	**Blue County in a Blue State**	**Control vs 2:1**	0.0380	0.8650	1.0000	1.0000	1.0000
Dollars Given Not Including Match	**Blue County in a Blue State**	**Control vs 3:1**	0.2173	0.2847	0.9997	1.0000	1.0000
Dollars Given Including Match	**Red County in a Red State**	**Control vs 1:1**	1.0782	0.0020***	0.0533*	0.0960*	0.0720*
Dollars Given Including Match	**Red County in a Red State**	**Control vs 2:1**	2.1238	0.0007***	0.0123**	0.0320**	0.0273**
Dollars Given Including Match	**Red County in a Red State**	**Control vs 3:1**	2.9206	0.0003***	0.0003***	0.0160**	0.0160**
Dollars Given Including Match	**Blue County in a Red State**	**Control vs 1:1**	0.7996	0.0043***	0.1173	0.2080	0.1517
Dollars Given Including Match	**Blue County in a Red State**	**Control vs 2:1**	2.3892	0.0017***	0.0437**	0.0800*	0.0617*
Dollars Given Including Match	**Blue County in a Red State**	**Control vs 3:1**	4.0048	0.0013***	0.0347**	0.0640*	0.0533*
Dollars Given Including Match	**Red County in a Blue State**	**Control vs 1:1**	0.7993	0.0003***	0.0003***	0.0160**	0.0157**
Dollars Given Including Match	**Red County in a Blue State**	**Control vs 2:1**	1.8355	0.0003***	0.0003***	0.0160**	0.0153**
Dollars Given Including Match	**Red County in a Blue State**	**Control vs 3:1**	2.5479	0.0003***	0.0003***	0.0160**	0.0150**
Dollars Given Including Match	**Blue County in a Blue State**	**Control vs 1:1**	1.0042	0.0003***	0.0003***	0.0160**	0.0147**
Dollars Given Including Match	**Blue County in a Blue State**	**Control vs 2:1**	2.0991	0.0003***	0.0003***	0.0160**	0.0143**
Dollars Given Including Match	**Blue County in a Blue State**	**Control vs 3:1**	2.3830	0.0003***	0.0003***	0.0160**	0.0140**
Amount Change	**Red County in a Red State**	**Control vs 1:1**	1.8253	0.1310	0.9497	1.0000	1.0000
Amount Change	**Red County in a Red State**	**Control vs 2:1**	0.5491	0.6443	1.0000	1.0000	1.0000
Amount Change	**Red County in a Red State**	**Control vs 3:1**	0.0681	0.9593	0.9987	1.0000	1.0000
Amount Change	**Blue County in a Red State**	**Control vs 1:1**	92.3221	0.4410	1.0000	1.0000	1.0000
Amount Change	**Blue County in a Red State**	**Control vs 2:1**	93.7227	0.4410	1.0000	1.0000	1.0000
Amount Change	**Blue County in a Red State**	**Control vs 3:1**	94.2640	0.4410	1.0000	1.0000	1.0000
Amount Change	**Red County in a Blue State**	**Control vs 1:1**	51.9652	0.4530	1.0000	1.0000	1.0000
Amount Change	**Red County in a Blue State**	**Control vs 2:1**	0.4450	0.6817	1.0000	1.0000	1.0000
Amount Change	**Red County in a Blue State**	**Control vs 3:1**	1.1372	0.2593	0.9973	1.0000	1.0000
Amount Change	**Blue County in a Blue State**	**Control vs 1:1**	0.9294	0.4617	1.0000	1.0000	1.0000
Amount Change	**Blue County in a Blue State**	**Control vs 2:1**	0.2938	0.8277	1.0000	1.0000	1.0000
Amount Change	**Blue County in a Blue State**	**Control vs 3:1**	0.5147	0.6577	1.0000	1.0000	1.0000

Table 5: Note: DI refers to "difference in means." *, **, and *** indicate the corresponding p-values less than 10%, 5%, and 1%, respectively.

5 Conclusion

As the 'credibility crises' begins to take shape and gain momentum in the empirical sciences, policymakers and academics are increasingly taking note of the assumptions underlying appropriate inference. For example, in the biological and human sciences as well as within economics it has been claimed that there is a 'reproducibility crisis', whereby established results fail to be replicated (Jennions and Moller, 2002; Ioannidis, 2005; Nosek et al., 2012; Bettis, 2012; and in economics, Maniadis et al., 2014). One reason for these failed replications is false positives, whereby many ungrounded conclusions are drawn due to a large false discovery rate. As modern societies gradually demand evidence based policies and make decisions based on science, the scientific community is relied on more heavily to provide sound advice.

In this paper, we consider three common threats to appropriate inference that involve multiple hypothesis testing within the field of experimental economics: jointly identifying treatment effects for a set of outcomes, estimating heterogeneous treatment effects through subgroup analysis, and conducting hypothesis testing for multiple treatment groups. These three core areas present themselves in nearly every empirical exercise that we are aware, and are particularly acute problems within experimental economics.

Building on stepwise multiple testing procedures discussed in Romano and Wolf (2010), we introduce a testing procedure that may be applied to any combination of these three common scenarios. The testing procedure incorporates information about the joint dependence structure of the test statistics so that it improves upon the classical multiple testing procedures such as Bonferroni and Holm. Furthermore, the testing procedure is asymptotically balanced in the sense that all marginal probabilities of rejecting any true null hypothesis are approximately equal.

Notably, we also demonstrate that further improvements could be made by exploiting transitivity and imposing smaller critical values when there are multiple treatment conditions. We highlight our methodology by exploring various plausible hypotheses of interest in the original Karlan and List (2007) data set. These data are interesting in their own right, as they help us to understand the economics of charity, which represents an important area of study in its own right.

In sum, as could be anticipated, we find that a smaller number of hypotheses are rejected when compared to results that do not adjust for multiple testing. Furthermore, and importantly from a methodological perspective, the p-values for all hypotheses by applying the testing procedure that we construct are always weakly smaller than the corresponding p-values from the classic studies due to Bonferroni and Holm. This result suggests the considerable power of our approach. Lastly, we show empirically that exploiting transitivity yields further improvements in power. In this way, our framework should have broad appeal to

empiricists, and more narrowly experimental economists.

A Appendix

A.1 Proof of Theorem 3.1

First note that under Assumption 2.1, $Q \in \omega_s$ if and only if $P \in \tilde{\omega}_s$, where

$$\tilde{\omega}_s = \{P(Q) : Q \in \Omega, E_P[Y_{i,k}|D_i = d, Z_i = z] = E_P[Y_{i,k}|D_i = d', Z_i = z]\} \ .$$

The proof of this result now follows by verifying the conditions of Corollary 5.1 in Romano and Wolf (2010). In particular, we verify Assumptions B.1 – B.4 in Romano and Wolf (2010).

In order to verify Assumption B.1 in Romano and Wolf (2010), let

$$T^*_{s,n}(P) = \sqrt{n}\left(\frac{1}{n_{d,z}} \sum_{1 \le i \le n : D_i = d, Z_i = z} (Y_{i,k} - \tilde{\mu}_{k|d,z}(P)) - \frac{1}{n_{d',z}} \sum_{1 \le i \le n : D_i = d', Z_i = z} (Y_{i,k} - \tilde{\mu}_{k|d',z}(P))\right) \ ,$$

and note that

$$T^*_n(P) = (T^*_{s,n}(P) : s \in \mathcal{S}) = f(A_n(P), B_n) \ ,$$

where

$$A_n(P) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} A_{n,i}(P) \ ,$$

with $A_{n,i}(P)$ equal to the $2|\mathcal{S}|$-dimensional vector formed by stacking vertically for $s \in \mathcal{S}$ the terms

$$\begin{pmatrix} (Y_{i,k} - \tilde{\mu}_{k|d,z}(P)) I\{D_i = d, Z_i = z\} \\ (Y_{i,k} - \tilde{\mu}_{k|d',z}(P)) I\{D_i = d', Z_i = z\} \end{pmatrix} \ , \tag{9}$$

and B_n is the $2|\mathcal{S}|$-dimensional vector formed by stacking vertically for $s \in \mathcal{S}$ the terms

$$\begin{pmatrix} \frac{1}{\frac{1}{n}\sum_{1 \le i \le n} I\{D_i = d, Z_i = z\}} \\ -\frac{1}{\frac{1}{n}\sum_{1 \le i \le n} I\{D_i = d', Z_i = z\}} \end{pmatrix} \ . \tag{10}$$

and $f : \mathbf{R}^{2|\mathcal{S}|} \times \mathbf{R}^{2|\mathcal{S}|} \to \mathbf{R}^{2|\mathcal{S}|}$ is the function of $A_n(P)$ and B_n whose sth argument for $s \in \mathcal{S}$ is given by the inner product of the sth pair of terms in $A_n(P)$ and the sth pair of terms in B_n, i.e., the inner product of (9) and (10). The weak law of large numbers and central limit theorem imply that

$$B_n \xrightarrow{P} B(P) \ ,$$

where $B(P)$ is is the $2|\mathcal{S}|$-dimensional vector formed by stacking vertically for $s \in \mathcal{S}$ the terms

$$\begin{pmatrix} \frac{1}{P\{D_i=d, Z_i=z\}} \\ -\frac{1}{P\{D_i=d', Z_i=z\}} \end{pmatrix} .$$

Next, note that $E_P[A_{n,i}(P)] = 0$. Assumption 2.3 and the central limit theorem therefore imply that

$$A_n(P) \xrightarrow{d} N(0, V_A(P))$$

for an appropriate choice of $V_A(P)$. In particular, the diagonal elements of $V_A(P)$ are of the form

$$\tilde{\sigma}^2_{k|d,z}(P) P\{D_i = d, Z_i = z\} .$$

The continuous mapping theorem thus implies that

$$T^*_n(P) \xrightarrow{d} N(0, V(P))$$

for an appropriate variance matrix $V(P)$. In particular, the sth diagonal element of $V(P)$ is given by

$$\frac{\tilde{\sigma}^2_{k|d,z}(P)}{P\{D_i = d, Z_i = z\}} + \frac{\tilde{\sigma}^2_{k|d',z}(P)}{P\{D_i = d', Z_i = z\}} . \tag{11}$$

In order to verify Assumptions B.2–B.3 in Romano and Wolf (2010), it suffices to note that (11) is strictly greater than zero under our assumptions. Note that it is not required that $V(P)$ be non-singular for these assumptions to be satisfied.

In order to verify Assumption B.4 in Romano and Wolf (2010), we first argue that

$$T^*_n(P_n) \xrightarrow{d} N(0, V(P)) \tag{12}$$

under P_n for an appropriate sequence of distributions P_n for (Y_i, D_i, Z_i). To this end, assume that

(a) $P_n \xrightarrow{d} P$.

(b) $\tilde{\mu}_{k|d,z}(P_n) \to \tilde{\mu}_{k|d,z}(P)$.

(c) $B_n \xrightarrow{P_n} B(P)$.

(d) $\mathrm{Var}_{P_n}[A_{n,i}(P_n)] \to \mathrm{Var}_P[A_{n,i}(P)]$.

Under (a) and (b), it follows that $A_{n,i}(P_n) \xrightarrow{d} A_{n,i}(P)$ under P_n. By arguing as in Theorem 15.4.3 in Lehmann and Romano (2006) and using (d), it follows from the Lindeberg-Feller central limit theorem that

$$A_n(P_n) \xrightarrow{d} N(0, V_A(P))$$

under P_n. It thus follows from (c) and the continuous mapping theorem that (12) holds under P_n. Assumption B.4 in Romano and Wolf (2010) now follows simply by nothing that the Glivenko-Cantelli theorem, strong law of large numbers and continuous mapping theorem ensure that $\hat{P}_n$ satisfies (a) – (d) with probability one under P.

References

Anderson, M. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association*, 103(484):1481–1495.

Arrowsmith, J. (2011a). Trial watch: phase ii failures: 2008–2010. *Nature Reviews Drug Discovery*, 10(5):328–329.

Arrowsmith, J. (2011b). Trial watch: phase iii and submission failures: 2007–2010. *Nature Reviews Drug Discovery*, 10(2):87–87.

Bettis, R. A. (2012). The search for asterisks: compromised statistical tests and flawed theories. *Strategic Management Journal*, 33(1):108–113.

Bhattacharya, J., Shaikh, A. M., and Vytlacil, E. (2012). Treatment effect bounds: An application to swan–ganz catheterization. *Journal of Econometrics*, 168(2):223–243.

Bonferroni, C. E. (1935). *Il calcolo delle assicurazioni su gruppi di teste*. Tipografia del Senato.

Bugni, F., Canay, I., and Shaikh, A. (2015). Inference under covariate-adaptive randomization.

Casey, K., Glennerster, R., and Miguel, E. (2012). Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *The Quarterly Journal of Economics*, 127(4):1755–1812.

Fink, G., McConnell, M., and Vollmer, S. (2011). Testing for heterogeneous treatment effects in experimental data: False discovery risks and correction procedures.

Flory, J. A., Gneezy, U., Leonard, K. L., and List, J. A. (2015). Gender, age, and competition: the disappearing gap. *Under Review*.

Flory, J. A., Leibbrandt, A., and List, J. A. (2014). Do competitive workplaces deter female workers? a large-scale natural field experiment on job-entry decisions. *The Review of Economic Studies*, page rdu030.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, 118(3):1049–1074.

Heckman, J., Moon, S., Pinto, R., Savelyev, P., and Yavitz, A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the highscope perry preschool program. Technical report, National Bureau of Economic Research.

Heckman, J., Pinto, R., Shaikh, A., and Yavitz, A. (2011). Inference with imperfect randomization: The case of the perry preschool program. Technical report, National Bureau of Economic Research.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Med*, 2(8):e124.

Jennions, M. D. and Moller, A. P. (2002). Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill'method. *Biological Reviews of the Cambridge Philosophical Society*, 77(02):211–222.

Karlan, D. and List, J. (2007). Does price matter in charitable giving? evidence from a large-scale natural field experiment. *The American economic review*, 97(5):1774–1793.

Kling, J., Liebman, J., and Katz, L. (2007). Experimental analysis of neighborhood effects. *Econometrica*, 75(1):83–119.

Lee, S. and Shaikh, A. M. (2014). Multiple testing and heterogeneous treatment effects: Re-evaluating the effect of progresa on school enrollment. *Journal of Applied Econometrics*, 29(4):612–626.

Lehmann, E. and Romano, J. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3):1138–1154.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Machado, C., Shaikh, A., and Vytlacil, E. (2013). Instrumental variables, and the sign of the average treatment effect. Technical report, Citeseer.

Maniadis, Z., Tufano, F., and List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *The American Economic Review*, 104(1):277–290.

Marcus, R., Eric, P., and Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

Nosek, B. A., Spies, J. R., and Motyl, M. (2012). Scientific utopia ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631.

O'Brien, P. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, pages 1079–1087.

Ong, M.-S. and Mandl, K. D. (2015). National expenditure for false-positive mammograms and breast cancer overdiagnoses estimated at $4 billion a year. *Health Affairs*, 34(4):576–583.

Romano, J. and Wolf, M. (2005a). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Romano, J. and Wolf, M. (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Romano, J. P. and Shaikh, A. M. (2006a). On stepdown control of the false discovery proportion. *Lecture Notes-Monograph Series*, pages 33–50.

Romano, J. P. and Shaikh, A. M. (2006b). Stepup procedures for control of generalizations of the familywise error rate. *The Annals of Statistics*, pages 1850–1873.

Romano, J. P., Shaikh, A. M., et al. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *The Annals of Statistics*, 40(6):2798–2822.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008a). Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, 17(3):417–442.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008b). Formalized data snooping based on generalized error rates. *Econometric Theory*, 24(02):404–447.

Romano, J. P. and Wolf, M. (2010). Balanced control of generalized error rates. *The Annals of Statistics*, pages 598–633.

Sutter, M. and Glätzle-Rützler, D. (2014). Gender differences in the willingness to compete emerge early in life and persist. *Management Science*.