Selective inference is easier with p-values

Anav Sood
Stanford University

November 12, 2024

Abstract

Selective inference is a subfield of statistics that enables valid inferences after selection of a data-dependent question. In this paper, we introduce selectively dominant p-values, a class of p-values that allow practitioners to easily perform selective inference after arbitrary selection procedures. Unlike a traditional p-value, whose distribution must stochastically dominate the uniform distribution under the null, a selectively dominant p-value must have a post-selection distribution that stochastically dominates that of a uniform having undergone the same selection process; moreover, this property must hold simultaneously for all possible selection processes. Despite the strength of this condition, we show that all commonly used p-values (e.g., p-values from two-sided testing in parametric families, one-sided testing in monotone likelihood ratio and exponential families, and permutation tests) are selectively dominant. By recasting two canonical selective inference problems—inference on winners and rank verification—in our selective dominance framework, we provide simpler derivations, a deeper conceptual understanding, and new generalizations and variations of these methods. Additionally, we use our insights to introduce selective variants of methods that combine p-values, such as Fisher's combination test.

1 Introduction

Selective inference is a subfield of statistics that allows practioners to make valid inferences even when statistical question at hand was chosen by a data-driven selection process. Most selective methods however, which operate by conditioning on the selection event, can be difficult to derive, hard to implement, and exhibit counterintuitive behaviors. To statisticians outside of the field, each selective procedure may seem to come from a different argument or approach.

In this paper, we provide a unifying framework for selective inference centered around p-values. So long as a statistician knows how to construct a p-value for their inferential question at hand, our framework provides an algorithmic approach for delivering hypothesis testing procedures and confidence intervals that are valid even after selection. Our framework (1) can greatly simplify the process of designing new selective methods and (2) results in more natural and general derivations of some existing selective methods, allowing for a deeper understanding of their behavior as well as new variations and extensions.

1.1 Motivation

To more concretely motivate the need for our framework, we recall two canonical problems in selective inference. In both problems, we imagine observing n independent samples $X_i \sim N(\mu_i, 1)$.

First, we consider the inference on winners problem. We want to provide a level α confidence lower bound (CLB) for the mean $\mu_{(n)}$ corresponding to the largest observed value $X_{(n)}$. Note that we have abused notation, and $\mu_{(n)}$ is not the largest of the μ_i but rather it is the random variable whose value is the true mean of the largest observation. The classical approach to this problem is to deliver a Sidak corrected level $1-(1-\alpha)^n$ lower bound. These lower bounds

These simultaneous lower bounds suffer from, what seems like, a necessary curse of dimensionality - as n grows, the distance from $X_{(n)}$ to the lower bound grows as well.

The conditional approach has very strange behavior ... seemingly avoiding the curse of dimensionality faced by the classical approach.

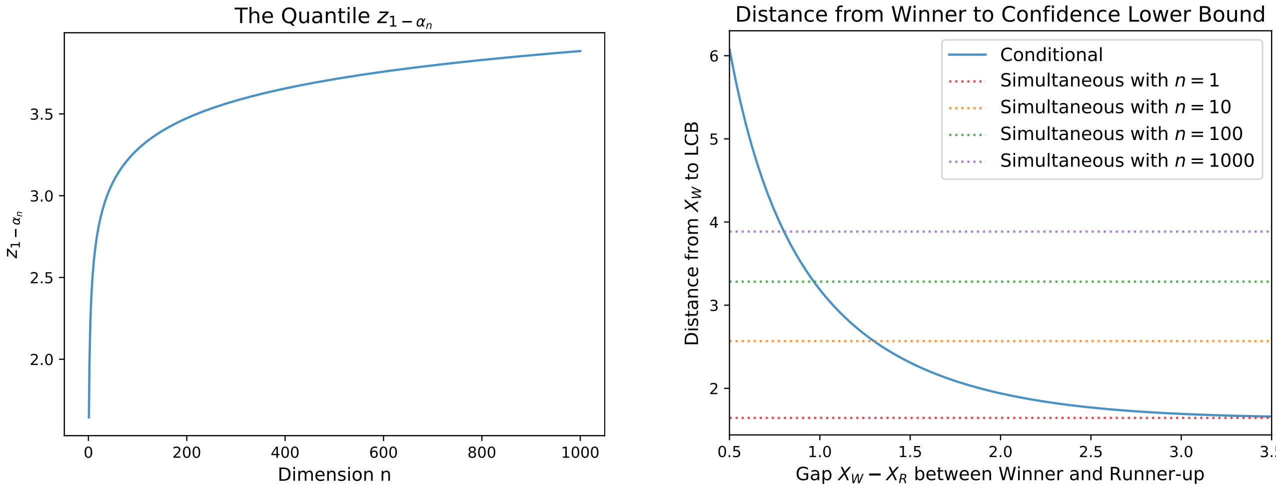


Figure 1: The first panel (left) shows the growth of the quantile $z_{1-\alpha_n}$ as a function of the dimension n. The second panel (right) gives the distance between the level $\alpha = 0.05$ conditional LCB and the winner X_W as a function of the gap $X_W - X_R$ between winner and the runner-up. For dimensions $n = 1, 10, 100, 1000$, it also gives the distance from X_W to the level $\alpha = 0.05$ simultaneous LCB.

Second, we consider the rank verification problem. In particular, we want to verify that the mean $\mu_{(n)}$ corresponding to the largest observation $X_{(n)}$ is actually largest than the rest. We can accomplish this by testing the null $\cap_{i=1}^{n-1} \mu_{(n)} - \mu_{(i)} \geq 0$. The seminal work Hung and Fithian [2019] provides a seemingly magical result.

After observing the largest value $X_{(n)}$ we want to provide a confidence lower bound (CLB)

We derive a framework that not only establishes this to be true, but turns out to be an all encompassing framework for selective inference.

1.2 Our Contributions

In this paper we introduce the **selective dominance** framework, which we summarize here. In the framework, we imagine using a p-value p to test the null hypothesis H_0.

The sufficient statisics can vary quite heavily from problem to problem, making many selective inference methods appear quite different. By focusing on the p-value...

The remainder of the paper illustrates our framework's utility via a number of applications.

1.3 Related Work

O

Related work (Andrews and Fithian on selection). (Adapt for increasing p-values). Other works on selecting the winner (Tijana, Benjamini)

Sequential selective hypothesis testing.

2 Selectively Dominant p-Values

In this section we define selectively dominant p-values, a class of p-values that enable us to easily do inference after selection. We give a precise characterization of when p-values are selectively dominant and illustrate that the most commonly used p-values are all selectively dominant. Finally, we discuss how to use this machinery to conduct data dependent hypothesis tests and construct selective confidence regions.

2.1 Selective Dominance

In classical statistics, a p-value is a random variable p supported on $[0, 1]$ that stochastically dominates the uniform distribution $U \sim \text{Unif}([0,1])$ under the null, i.e., $p \succeq_{H_0} U$. When working with p-values, we maintain Type I error control if we reject the null H_0 when p is small:

$$P_{H_0}(\text{reject null}) = P_{H_0}(p \leq \alpha) \leq P(U \leq \alpha) = \alpha.$$

In essence, stochastic dominance allows us to use the uniform distribution as a reference distribution. We control the probability of p being small under the null by comparing it to the probability of a uniform being small.

In the problems we consider, we use p to test the null H_0 only after it has been "selected". In full generality, we consider a p-value p for testing the null H_0 that is conditionally valid given some random vector Z, i.e., $p \mid Z = z \succeq_{H_0} U$:

$$P_{H_0}(p \leq \alpha \mid Z = z) \leq P(U \leq \alpha) = \alpha. \tag{1}$$

Additionally, we consider a binary selection random variable $S \in \{0, 1\}$ that takes value one when p is selected and zero otherwise. The relationship between p, Z, and S is governed by a **selection function**,

$$s(x, z) = P(S = 1 \mid p = x, Z = z).$$

Intuitively, we imagine observing $p = x$ and $Z = z$ and then flipping a biased coin that comes up heads with probability $s(x, z)$. We only use p to test the null H_0 when this coin comes up heads, and otherwise do not perform inference. This process turns out to capture what happens in a wide range of selective inference problems. We decide the selection procedure in the problems we consider, so $s(x, z)$ is known. In cases where no Z is present, we can imagine $Z = 0$ and write the selection function $s(x)$ purely in terms of x.

Because we only perform inference when the coin comes up heads, our goal should be to design a procedure that controls Type I error conditional on this selection event:

$$P_{H_0}(\text{reject } H_0 \mid S = 1) \leq \alpha \tag{2}$$

As illustrated by our next example, the classical approach of rejecting when $p \leq \alpha$ does not maintain selective Type I error control as in (2).

Example 1 (Publication bias and the failure of classical inference)**.** *Consider a p-value p that is uniform on $[0, 1]$ under the null H_0. If we use the selection function $s(x) = I_{p \leq \alpha}$, i.e., we select p when it is at most α, then $p \mid S = 1 \sim$ Unif$([0, \alpha])$. Our classical procedure will clearly fail to control Type I error conditional on selection:*

$$P_{H_0}(\textit{reject } H_0 \mid S = 1) = P_{H_0}(p \leq \alpha \mid S = 1) = 1 > \alpha.$$

Example 1 is a standard example in the literature on publication bias. If researchers publish studies testing the null H_0 only when their p-values are below α, the reader only gets to observe inferences made after this selection. As a consequence, the reader's observed type I error rate can be as high as one.

Essentially, after selection, the uniform distribution no longer suffices as a reference distribution. Naturally, we may instead try and use the distribution of a uniform after it has been selected by the same selection function. Formally, suppose that U has uniform distribution conditional on Z, i..e., $U \mid Z = z \sim \text{Unif}([0, 1])$, and let $S' \in \{0, 1\}$ be a different binary selection random variable whose joint distribution with U and Z is governed by the same selection function

$$P(S' = 1 \mid U = x, Z = z) = s(x, z).$$

Then, we could use the conditional distribution $U \mid Z, S' = 1$ of U given selection as our reference distribution. This approach is valid exactly when our p-value is selectively dominant, as we define below [1].

[1] For the sake of simplicity, we require selective dominance to hold point-wise for *every* z rather than almost all, and our later results similarly hold point-wise. Our definition and results, however, can be modified to accomodate more general case.

Definition 1 (Selective dominance)**.** *Considering a p-value p for the null H_0 that is valid given Z as in* (1)*, we say that p is* ***selectively dominant given*** *Z if, under the null H_0, it has a conditional probability density function (PDF) given $Z = z$ and satisfies*

$$p \mid Z = z, S = 1 \succeq_{H_0} U \mid Z = z, S' = 1 \tag{3}$$

for every selection function $s(x, z)$ under which p and U both have a positive probability of being selected given $Z = z$.

As we will soon see, the majority of p-values that practitioners use are selectively dominant as described in (3). In Definition 1, we restrict to p-values with conditional PDFs under the null because it makes our theory and methods simpler to state. Because we can always make a p-value both have a conditional PDF and be more powerful via randomization, this restriction is never a practical issue. Also, after applying our machinery with randomized p-values, the user can always derandomize the resulting method if they would like.

To perform valid post-selection inference using a selectively dominant p-value, we can transform it so that it remains a p-value after selection. As Theorem 1 explains, we can "undo" the effects of selection by applying the conditional cumulative distribution function (CDF) $F_{U|Z,S'=1}(\cdot)$ of U given selection to p. In line with prior literature, we refer to this transformed p-value as a **selective p-value**. For simple selection functions, this selective p-value is often computable in closed form.

Theorem 1. *Let $F_{U|Z,S'=1}(u)$ denote the CDF of U conditional on selection. Then, under the null, the selective p-value*

$$p_{sel} = F_{U|Z,S'=1}(p) = \frac{\int_0^p s(x, Z)dx}{\int_0^1 s(x, Z)dx} \tag{4}$$

stochastically dominates the uniform distribution conditional on Z and selection,

$$P_{H_0}(p_{sel} \leq \alpha \mid Z = z, S = 1) \leq \alpha, \tag{5}$$

for any selection function $s(x, z)$ under which p and U both have a positive probability of being selected given $Z = z$. Further, for any distribution in H_0 under which p has an exact uniform distribution given $Z = z$, (5) *holds with equality.*

Essentially, Theorem 1 tells us that if we want selective Type I error control as in (2), then we should reject H_0 when p is less than the α quantile of $U \mid Z = z, S' = 1$ rather than the α quantile of U.

2.2 Characterizing Selectively Dominant p-Values and Examples

Theorem 2 tells us that p-values are selectively dominant precisely when their conditional PDF is non-decreasing under the null.

Theorem 2 (Selective dominance and increasing density)**.** *If the conditional PDF of the p-value p given $Z = z$ is always non-decreasing under the null, then it is selectively dominant given Z as described in Definition 1. Conversely, if ever under the null, the conditional PDF of p given $Z = z$ is everywhere continuous and not non-decreasing, then p is not selectively dominant given Z.*

In what follows, we give a number of examples of selectively dominant p-values. Our examples include all the common p-values that practitioners use in real life. We recommend that the unfamiliar reader review uniformly most powerful (UMP) and uniformly most powerful unbiased (UMPU) testing [Lehmann et al., 1986, Chapter 3 and Chapter 4] prior to proceeding.

Example 2 (Two-sided testing in parametric families)**.** *Consider observing data from a parametric family P_θ and testing the null $H_0 : \theta = \theta_0$. Because the null is a point null, most p-values we construct will have an exact Unif$([0, 1])$ distribution under the null and are therefore trivially selectively dominant.*

Example 3 (One-sided testing in monotone likelihood ratio families). *Consider observing one-dimensional data from a parametric family $X \sim P_\theta$ that admits density $p_\theta(x)$ with respect to some carrier measure μ. We say that P_θ has a monotone likeliihood ratio (MLR) in the real valued function $T(x)$ if, the densities $p_\theta(x)$ share a common support and, for any $\theta < \theta'$, the ratio $p_{\theta'}(x)/p_\theta(x)$ is a non-decreasing function of $T(x)$. In this case, the UMP test for the null $H_0 : \theta \leq \theta_0$ rejects when $T(X)$ is large. The associated randomized p-value for this test (see Appendix D) is selectively dominant.*

Example 4 (Testing in in exponential families). *Suppose we observe data $X \in \mathbb{R}^m$ from an exponential family P_θ parameterized by $\theta \in \mathbb{R}^n$ i.e., under P_θ the data X has density*

$$g_\theta(x) = \exp(\theta_1 T_1(X) + \cdots + \theta_n T_n(X) - \psi(\theta))g(x)$$

with respect to some carrier measure μ. In both the case of testing the two-sided null $H_0 : \theta_1 \neq \theta_{0,i}$ or one-sided null $H_0 : \theta_i \leq \theta_{0,i}$, the UMPU test conditions on the nuisance statistics $T_{-i}(X)$. The p-value associated with the UMPU test for $H_0 : \theta_1 \neq \theta_{0,i}$ has an exact Unif($[0,1]$) distribution conditional on $T_{-i}(X)$, so it is trivially selectively dominant given $Z = T_{-i}(X)$. For testing $H_0 : \theta_1 \leq \theta_{0,i}$, we are in the setting of an MLR family once we condition on $T_{-i}(X)$, so Example 3 implies that the p-value associated with the UMPU test is also selectively dominant given $Z = T_{-i}(X)$.

Example 5 (Permutation testing). *In a permutation test we observe data $X \in \mathcal{X}$ and compute a test statistic $T(X)$ that, under the null H_0, has a distribution that is invariant under a finite group of transformations $G : \mathcal{X} \to \mathcal{X}$. That is, $T(X) \stackrel{d}{=}_{H_0} T(g(X))$ for all $g \in G$. To run the test, we consider a collection of group elements $g_1, g_2, \ldots, g_w$ where $g_1 = id$ is fixed to be the identity transformation and $g_2, \ldots, g_w$ are either a random sample from G with replacement or a random sample from $G \setminus \{id\}$ without replacement. The test then rejects when $T(X)$ is large compared to the $T(g_j(X))$. Specifically, the randomized permutation test from Proposition 3 of Hemerik and Goeman [2018] uses the p-value*

$$p = \frac{\#\{1 \leq j \leq w : T(g_j(X)) > T(X)\}}{w} + U_{aux}\frac{\#\{1 \leq j \leq w : T(g_j(X)) = T(X)\}}{w},$$

where $U_{aux} \sim$ Unif($[0,1]$) adds auxiliary randomness that is independent of X. This p-value always has an exact Unif($[0,1]$) distribution under H_0 and is therefore is trivially selectively dominant.

Etablishing Example 3 and Example 4 is non-trivial, and the bulk of Appendix D is spent doing so.

2.3 Example Applications of Selective Dominance

Having developed our machinery, we provide a few examples that illustrate how to use it.

As an introductory example, we show how to correct for Example 1's publication bias. Using our selective dominance machinery, we can provide a one-line derivation of the p-value adjustment from Hung and Fithian [2020]. Hung and Fithian [2020] derive this correction specifically for p-values coming from z- and t-tests, but our machinery applies for all selectively dominant p-values.

Example 6 (Correcting for publication bias). *Suppose we have a selectively dominant p-value p for the null hypothesis H_0, and we choose to test H_0 only after observing that $p \leq \alpha$. We can apply our framework with $p = p$, $Z = 0$, and $s(x,z) = I_{x \leq \alpha}$. The selective p-value from (4) is p/α, so Theorem 1 tells us that rejecting when $p \leq \alpha^2$ controls selective Type I error:*

$$P_{H_0}(p \leq \alpha^2 | S = 1) = P_{H_0}(p/\alpha \leq \alpha | S = 1) \leq \alpha$$

As we have learned that essentially all the p-values researchers typically use are selectively dominant, Example 6 gives a simple way for readers to make valid inferences in the presence of publication bias: declare a studies' result significant when the associated p-value is at most α^2.

Our rule of thumb of rejecting when $p \leq \alpha^2$ should also deliver valid inferences in the presence of p-hacking. Rather discarding an experiment after observing a p-value larger than α, researches more typically tweak their analysis until the p-value crosses the significance threshold. This process, known as p-hacking, is difficult to study theoretically (hence Hung and Fithian [2020] do not study it). But it has been emprirically well established that, under the null, p-values resulting from p-hacking have left-skewed distributions, i.e., null p-hacked p-values can be reasonably modeled as having an increasing density on $[0, \alpha]$ [Simonsohn et al., 2013]. The transformed p-value p/α then has a density that is increasing on $[0, 1]$ under the null, so Theorem 2 guarantees that it is indeed a valid p-value.

Our second example shows how to use Theorem 1 to perform inference using the "winning" p-value. It illustrates how our selective dominance machinery enables us to test data dependent hypotheses, the core problem of selective inference.

Example 7 (Inference on the winning p-value)**.** *Suppose we have n independent and selectively dominant p-values p_i for the null hypotheses $H_{0,i}$, and we choose to test only the jth null $H_{0,j}$ after observing that p_j is the smallest of the p_i. We will assume that under $H_{0,i}$ each p_i has density that is positive on all of $(0, 1)$. Applying our framework with $p = p_j$, $Z = p_{-j}$, and the selection function $s(x, z) = I_{x<\min_k z_k}$, it is straightforward to compute that the adjusted p-value p_{adj} from (4) is $p_j / \min_{i \neq j} p_i$, so Theorem 1 tells us that rejecting when $p_j \leq \alpha \min_{i \neq j} p_i$ controls selective Type I error:*

$$P_{H_{0,j}}(p_j \leq \alpha \min_{i \neq j} p_i \mid p_{-j}, S = 1) \leq \alpha. \tag{6}$$

If we let W be the index of the smallest p-value, it is now easy to see that rejecting the data-dependent "winning" null $H_{0,W}$ when $p_{(1)} \leq \alpha p_{(2)}$ is controls Type I error both conditionally on W and marginally. Consider only the set of indices $j \in \mathcal{J}$ for which p_j has a positive probability of being the smallest. Conditional error control is immediate: If $H_{0,j}$ is not true, then trivially $P(\text{falsely reject } H_{0,W} \mid W = j) = 0 \leq \alpha$. For the case that $H_{0,j}$ is true, the event $W = j$ is the same event as selecting p_j for inference in (6), so

$$\begin{aligned} P(\textit{falsely reject } H_{0,W} \mid W = j) &= P(p_{(1)} \leq \alpha p_{(2)} \mid W = j) \\ &= P(p_j \leq \alpha \min_{i \neq j} p_i \mid W = j) \\ &\leq \alpha. \end{aligned}$$

Marginal error control follows from the law of total probability.

$$\begin{aligned} P(\textit{falsely reject } H_{0,W}) &= \sum_{j \in \mathcal{J}} P(\textit{falsely reject } H_{0,j} \mid W = j) P(W = j) \\ &\leq \alpha \sum_{j \in \mathcal{J}} P(W = j) \\ &\leq \alpha. \end{aligned}$$

If the nulls are all true and the p_i are exactly uniform, then the inequalities become equalities and our error control is tight.

Rejecting the null $H_{0,W}$ when $p_{(1)} \leq \alpha p_{(2)}$ may seem like a strange procedure, but we will see that doing so is central to the conditional inference for winners method that arises from Fithian et al. [2017]. In fact, Fithian et al. [2017] implies that, in many settings, this test is UMP amongst tests that are valid conditional on W.

Lastly, we show how our framework also applies to data-carving. Specifically we consider the file-drawer problem from Fithian et al. [2017]. Fithian et al. [2017] argue that data-splitting, which involves using a chunk of the data for selection and an independent chunk of data for inference, is often an inadmissible approach in selective inference problems. In such settings, data-carving, as we describe below, results in strictly more powerful procedures. Although it initially appears that data-carving's selection procedure does not involve selecting a p-value as we describe in Section 2, we show via a coupling argument that it can be viewed in this way. This both serves to illustrate the breadth of our framework's applicability, as well as provide a new perspective on data carving.

Example 8 (Data carving and the file-drawer problem)**.** *In the file-drawer problem we observe two independent samples $X_1, X_2 \sim N(\mu, 2)$ (e.g., X_1 comes from the first half of the data and X_2 from the second). We test the null $H_0 : \mu \leq 0$, but only when we observe that $X_1 > t$ for some $t \in \mathbb{R}$.*

Data splitting ignores the first observation, which was used for selection, and simply tests the null with the p-value $p_{split} = 1 - \Phi(X_2/\sqrt{2})$. Intuitively, because this p-value is independent of the selection process, we should maintain Type I error control without performing any correction. Applying our framework with $p = p_{split}$, $Z = X_1$, and the selection function $s(x,z) = I(z > t)$, it is not hard to see that the selective p-value indeed offers no correction.

Data-carving, however, attempts to still use the more powerful p-value $p_{full} = 1 - \Phi((X_1 + X_2)/2)$, which leverages information from both samples. How can we apply our framework to this data-carving problem? In how we have stated the problem, it is <u>not</u> the case that we observe p_{full} and then decide whether or not to use it for inference. Instead, we decide based on X_1, and unlike in data-splitting, p is not a valid p-value given X_1. Noting that

$$X_1 | \frac{X_1 + X_2}{2} = y \sim N(y, 1),$$

however, we can compute the probability that selection happened given that p took a particular value:

$$P(X_1 > t | p_{full} = x) = 1 - \Phi(t - \Phi^{-1}(1 - x)).$$

We may as well therefore imagine that we observed that $p_{full} = x$, and then decided to use it to test the null H_0 with probability $1 - \Phi(t - \Phi^{-1}(1-x))$. Although this is not what happens in the original problem (in our new characterization we may test H_0 even when $X_1 \leq 3$), the conditional distribution of p given selection is the same in both cases. Hence we can apply our framework with $p = p_{full}$, $Z = 0$ and $s(x) = 1 - \Phi(t - \Phi^{-1}(1-x))$. Fithian et al. [2017] argue that the resulting selective p-value

$$p_{carve} = \frac{\int_0^{p_{full}} 1 - \Phi(t - \Phi^{-1}(1-x))dx}{\int_0^1 1 - \Phi(t - \Phi^{-1}(1-x))dx} = \frac{\int_0^{p_{full}} 1 - \Phi(t - \Phi^{-1}(1-x))dx}{1 - \Phi(t/\sqrt{2})}$$

will result in strictly more rejections than p_{split}. Appendix A.10 provides more details on the specific calculations done in this example.

Crucially, in Example 8, the conditional distribution of the random variable X_1 used for selection given the p-value p_{full} did not depend on the unknown parameter μ. Hence, the selection function $s(x)$ had no dependence on μ, and we were able to correct our p-value without any issues. This did not happen by accident, and is actually a consequence of more general and interesting fact regarding the relationship between between data splitting, data carving, data fission [Leiner et al., 2023], and data thinning [Dharamshi et al., 2024, Neufeld et al., 2024].

In the most basic version of data fission, we add and subtract independent normal noise $Z \sim N(0,1)$ to a normal sample $X \sim N(\mu, 1)$ to get two independent samples $X_1, X_2 \sim N(\mu, 2)$ centered at the same mean. This is meant to mimic data splitting: the first sample can be used for selection and the second for inference. Data thinning generalizes this idea to the setting where X is a random vector from a parametric family, and we add noise make k new random vectors $X_1, \ldots, X_k$ that (1) are independent and (2) can be used to recover X via a deterministic function $X = T(X_1, \ldots, X_k)$. Vanilla data thinning would involve using some of the X_i to perform selection and then the rest to do inference. Data carving, however, suggests using a p-value for inference that is a function of all the data $X = T(X_1, \ldots, X_k)$, despite some of the X_i being used for selection. Because the noise we add to X has no dependence on the unknown parameter, the joint distribution of the X_i given $T(X_1, \ldots, X_k)$ also has no dependence on the unknown paramter. Therefore, contrary to the what the language in Leiner et al. [2023] suggests, the selection function $s(x)$ is always <u>known</u>, and we can always apply our framework to data carve and get more power. If the selection process is highly complicated, it is true that $s(x)$ may be very difficult to compute, but in theory it is always accessible to us via extensive simulations.

Our framework also applies to regression problems, including Lee et al. [2016]'s foundational problem of doing inference on LASSO selected regression coefficients. For sake of brevity, we have moved this discussion to Appendix A.9.

Example 6, Example 7, Example 8 all share a common theme. In all three examples, the practioner cheats. They peak at the p-value and, to varying degrees, they only test the null when the p-value looks promising. The purpose of selective procedures is to adjust the p-value in a way that accounts for this cheating. The harsher the cheating is, the more this adjustment inflates the original p-value.

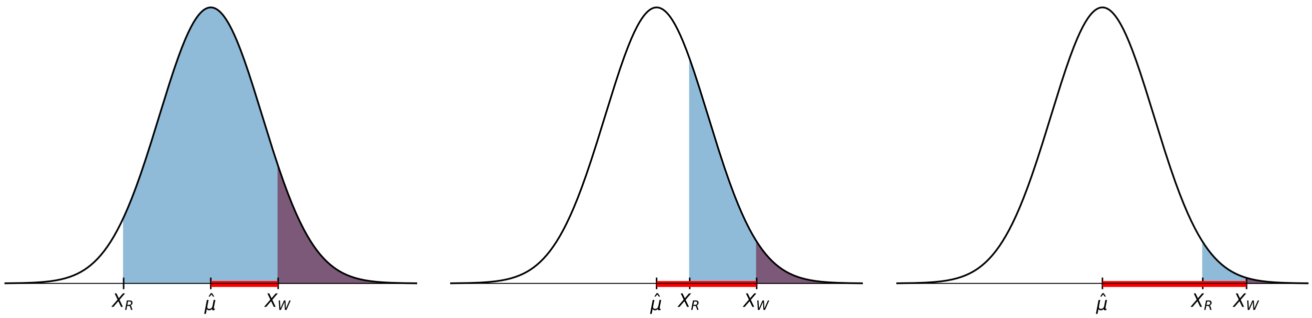


Figure 2: We plot the level $\alpha = 0.1$ conditional LCB $\hat{\mu}$ for different gaps between the winning value X_W and the runner up value X_R and highlight the distance between $\hat{\mu}$ and X_W in red. The LCB $\hat{\mu}$ is chosen exactly so that the tail probability $P(N(\hat{\mu}, 1) > X_R)$, shaded in blue, is $1/\alpha = 10$ times the tail probability $P(N(\hat{\mu}, 1) > X_W)$, shaded in red (the overlap appears purple). As X_W and X_R get closer, we need to take $\hat{\mu}$ further and further back for this condition to be satisfied.

3 Inference on Winners

In this section we use our framework to study the inference on winners problem. Along with providing further discussion about Example 7, we also discuss how hybrid inference [Andrews et al., 2023], which offers a solution to the exploding interval problem, also naturally arises in our framework. For both appproaches, our discussion results in novel interpretations, generalizations, and methods. To be concrete, we will imagine performing inference at level $\alpha = 0.1$ throughout the section.

For now, we focus on the independent data setting. The tools we develop in the next section, however, enable us to do inference on winners when data is generated from a multi-parameter exponential family, which encompasses many correlated data settings.

3.1 Conditional Inference

In this sub-section we discuss Fithian et al. [2017]'s conditional approach for performing inference on winners. This conditional approach turns out to be highly related to the testing procedure that we derived in Example 7.

Corollary 1 (Testing the winner)**.** *Suppose that p_i are n independent and selectively dominant p-values under the nulls $H_{0,i}$, and let W be the index of the smallest p-value. Rejecting $H_{0,W}$ when $p_{(1)} \leq \alpha p_{(2)}$ controls Type I error at level α conditionally on W, and therefore also marginally.*

Unlike Sidak's simultaneous approach, which rejects the winning null when the smallest p-value is small in absolute terms, the conditional approach rejects the winning null when the smallest p-value is small relative to the second smallest p-value. This procedure is strange, but fairly easy to interpret: we reject the winning null when the most extreme observation is $1/\alpha = 10$ more extreme under its null than the second most extreme observation. This can be quite a stringent requirement!

Once written in terms of p-values, its easy to mathematically see the merits and pitfalls of the conditional approach. If all the p-values except the smallest provide essentially no evidence against the null, then $p_{(2)} \approx 1$ and we reject when $p_{(1)} < \alpha$, the same p-value cutoff as a one-dimensional problem. On the other hand, if even just one other p-value provides a similar amount of evidence against the null as the smallest p-value, then then $p_{(2)} \approx p_{(1)}$ and we will never reject because $p_{(1)} \not< \alpha p_{(1)} \approx \alpha p_{(2)}$.

The p-value viewpoint also makes it clear that the conditional approach can only outperform the simultaneous approach when some null p-values are conservative (i.e., they are super-uniform). We give a heuristic argument here, and a more formal statement in Appendix A.1. Suppose one of our p-values p_1 is a very strong signal (so it is very small with high probability) but the remaining p-values $p_2, \ldots, p_n$ are null p-values that are uniform (i.e., they are not conservative). Our conditional procedure will reject when our smallest p-value, likely p_1, is less than α times the smallest of $p_2, \ldots, p_n$. The minimum of these $n-1$

uniform p-values is $1/n$ on average. Hence, roughly speaking, the conditional approach also rejects when $p_{(1)}$ is less than α/n, which is essentially the same as Sidak's simultaneous approach when n is large.

3.1.1 Confidence regions for the winner

In parametric problems, we can invert Corollary 1's test to get selective confidence intervals for the winning parameter. Consider observing independent data $X_i \sim P_{\theta_i}$ from an MLR family P_θ parametrized by $\theta \in \mathbb{R}$. Let $p_i^{\theta_0}$ (which is a function of X_i) be the UMP p-value for testing the null $H_0 : \theta_i \leq \theta_0$. Details regarding these p-values can be found in Appendix D.1. We can define the winner $W = \text{argmin}_{j\in[n]}\, p_j^{\theta_0}$ to be the index of the smallest and most promising p-value. This winning index will be the same irrespsective of θ_0 [2]. By inverting Corollary 1's test, we get an LCB

$$\{\theta_0 : p_{(1)}^{\theta_0}/p_{(2)}^{\theta_0} > \alpha\} \tag{7}$$

for the winning parameter θ_W that holds conditionally on W with probability exactly $1-\alpha$:

$$P(\theta_W \in \{\theta_0 : p_{(1)}^{\theta_0}/p_{(2)}^{\theta_0} > \alpha\}|W) = 1-\alpha.$$

The fact that the confidence region (7) is actually an LCB is a consequence of the selective p-value $p_{(1)}^{\theta_0}/p_{(2)}^{\theta^0}$ being monotone non-decreasing in null parameter θ_0. Appendix D.3 provides general conditions under which selective p-values like $p_{(1)}^{\theta_0}/p_{(2)}^{\theta^0}$ are monotone in the null parameter. We show these conditions apply to the winner problem, and also argue that (7) has exact $1-\alpha$ coverage in **??**. We also show in **??** how to invert Corollary 1's test to get a CI (rather than LCB), and that both our CI and LCB match Fithian et al. [2017]'s approach in the Gaussian case.

Writing the conditional inference in terms of p-values helps us better understand Fithian et al. [2017]'s conditional LCB (7). In particular, we learn that Figure 1's LCB stretches back exactly to the $\hat{\theta}$ under which it is $1/\alpha = 10$ times less likely to see something as extreme as the winner than something as extreme as the runner-up. Figure 2 provides an illustration for the Gaussian case. It demonstrates why why the LCB diverges to $-\infty$ as the winner and runner-up get closer. If the winner and runner-up are very close, we will need the LCB $\hat{\mu}$ to be very far back for the winner to be ten times more extreme than the runner-up. Thanks to the decay of the Gaussian tail, however, we can always find such a mean if we go far back enough.

For non-Gaussian data, the amount the conditional LCB (7) stretches back depends on the tail decay of P_θ. The faster the tail decays, the larger the first $\hat{\theta}$ for which the winner is $1/\alpha = 10$ times as extreme as the runner-up. Seeing as the Gaussian distribution, whose tail shrinks as e^{-x^2}, still often results in very low lower bounds, we should expect that the distance from the winning observation to the lower bound will often be quite large in many settings.

As an example, consider observing independent exponential random variables $X_i \sim \text{Exp}(\lambda_i)$. The exponential distribution has a tail e^{-x} that decays fast, but not as fast as the Gaussian tail. Crucially, it also has a parameter space $\lambda \in (0,\infty)$ that is bounded below. The exponential distribution has an MLR in $T(x) = 1/x$, so the UMP test for $H_{0,i}^{\lambda_0} : \lambda_i \leq \lambda_0$ uses a p-value $p_i^{\lambda_0}$ that rejects when X_i is small. It turns out that

$$\lim_{\lambda^0 \downarrow 0} p_{(1)}^{\lambda_0}/p_{(2)}^{\lambda_0} = X_{(1)}/X_{(2)},$$

so the conditional LCB for the winning parameter λ_W,

$$\{\lambda_0 : p_{(1)}^{\lambda_0}/p_{(2)}^{\lambda_0} > \alpha\}, \tag{8}$$

is vacuous whenever $X_{(1)}/X_{(2)} > \alpha$, i.e., with positive probability the confidence region (7) spans the whole parameter space $(0,\infty)$. A careful derviation of this test and result can be found in Appendix A.5.

The failure of the conditional LCB (8) often manifests in real data examples. On a dataset of car engine failure times [Molotaliev, 2024], we find that the conditional LCB (8) is always vacuous. The dataset has the failure times of one-hundred car engines, which we model as independent exponentials. Taking a random

[2] If we use the same auxiliary randomness to compute $p_i^{\theta_0}$ for every θ_0, then one index will result in the smallest p-value for every θ_0 and W is well-defined. The discussion in Appendix D.1 implies that this will be the case.

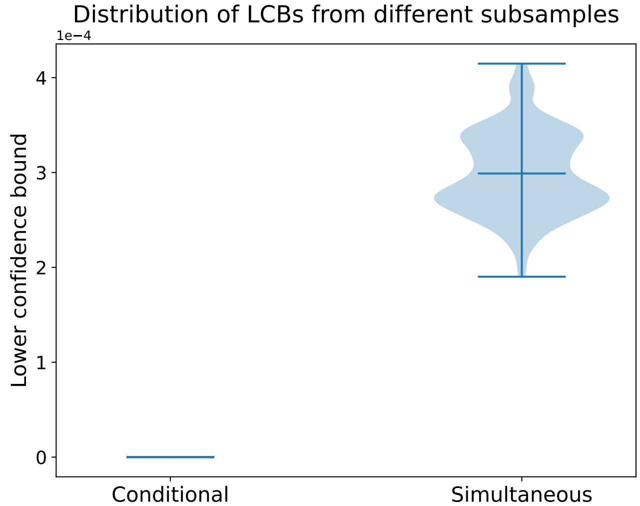


Figure 3: Over $B = 1000$ different subsamples of $n = 2$ failure times from the dataset Molotaliev [2024], the distribution of the LCB for the "winning" parameter resulting from the conditional and simulatneous approaches. The conditional LCB is always vacuous.

subsample of just $n = 2$ failure times, the LCB (7) does inference on the worse of the two engines, but always gives a vacuous lower bound of zero. In contrast, the simultaneous approach always gives a non-vacuous lower bound. Figure 3 depicts the results.

3.1.2 More discoveries via the closure principle

Once written in terms of p-values, it is natural to treat Corollary 1's test as a test of the global null and try to close it (as in Marcus et al. [1976]). Closing a global null test precludes us from making confidence regions, but it allows us to make more individual discoveries. Closed global null testing procedures are often computationally intractable to implement, so it is interesting that Corollary 1's global null test admits a tractable closure.

Corollary 2 (Closed testing for winners)**.** *Suppose that p_i are n independent and selectively dominant p-values under the nulls $H_{0,i}$. As shorthand, let $H_{0,(j)}$ denote the null corresponding to the jth smallest p-value (ties broken randomly) and define $p_{(n+1)} = 1$. Rejecting the null hypotheses $H_{0,(k)}$ for which $p_{(j)} \leq \alpha p_{(j-1)}$ for every $j \leq k$ controls FWER error at level α.*

As is often the case for closed procedures, Corollary 2 procedure is best understood sequentially. We reject $H_{0,(1)}$ if $p_{(1)} \leq \alpha p_{(2)}$. Then, if we rejected $H_{0,(1)}$, we reject $H_{0,(2)}$ if $p_{(2)} \leq \alpha p_{(3)}$, so on and so forth until we fail to reject.

3.2 Hybrid Inference

Hybrid inference, originally proposed by Andrews et al. [2023], is an inference on winners procedure that attempts to balance the benefits of the simultaneous and conditional approaches. It is a very elegant idea, but it currently only applies to Gaussian data and can be difficult to parse and implement. Using our selective dominance framework, we give a simpler exposition of hybrid inference that enables its application in more general settings, provided that the data is independent. As a bonus, our new procedure is very easy to undestand and implement.

3.2.1 Hybrid Test for the Winner

Corollary 3 presents our hybrid testing procedure. We provide a proof sketch and defer a detailed proof to **??**.

Corollary 3 (Hybrid test for the winner)**.** *Suppose that p_i are n independent and selectively dominant p-values under the nulls $H_{0,i}$, let W be the index of the smallest p-value, and fix some $\beta < \alpha$. Rejecting $H_{0,W}$ when*

$$p_{(1)} \leq \frac{\alpha - \beta}{1 - \beta} p_{(2)} + \left(1 - \frac{\alpha - \beta}{1 - \beta}\right) \beta_n \tag{9}$$

controls Type I error at level α. The same is thus true for rejecting the global null $\cap_{i=1}^{n} H_{0,i}$ whenever (9) holds.

Proof sketch. Let B be the event that the smallest p-value comes from a null and is at most β_n. We know from Sidak's procedure that $P(B) \leq \beta$. Hence, on the complementary event B^c, which has probability $\geq 1 - \beta$, it suffices to ensure that we fail to falsely reject $H_{0,W}$ with probability at least $(1-\alpha)/(1-\beta)$. Supposing $H_{0,j}$ is true, imagine testing $H_{0,j}$ using p_j only when B^c happens and $W = j$. This is exactly like selecting p_j to use for inference when it is between β_n and $\max_{i \neq j} p_i$. For this selection, Theorem 1's selective p-value is given by $(p_j - \beta_n)/(\max_{i \neq j} p_i - \beta_n)$, so we can ensure that we fail to reject $H_{0,j}$ with probability at least $(1-\alpha)/(1-\beta)$ if we fail to reject whenever

$$\frac{p_j - \beta_n}{\max_{i \neq j} p_i - \beta_n} > 1 - \frac{1 - \alpha}{1 - \beta} = \frac{\alpha - \beta}{1 - \beta}.$$

Our hybrid inference procedure turns out to fail to reject exactly in this case. □

Written in terms of p-values, it is easy to see how the hybrid approach balances the benefits of the simultaneous and conditional approaches. It will reject both when the smallest p-value is small in absolute terms or when it small relative to the second smallest p-value. When the other p-values provide essentially no evidence against the null (i.e., $p_{(2)} \approx 1$), hybrid rejects when $p_{(1)}$ is less than a cutoff that is bounded away from zero. In this situation, it performs at least on par with the conditional procedure run at level $(1-\alpha)/(1-\beta)$. On the other hand, even if some other p-value provides as much evidence against the null as the smallest, hybrid still rejects whenever the level β simultaneous approach does. This is because when $p_{(1)} \leq \beta_n$, the hybrid cutoff is a mixture of two things that are larger than $p_{(1)}$, and we will reject.

The parameter β allows hybrid inference to interpolate between the simultaneous and conditional approaches. When we set $\beta = 0$ then the hybrid cutoff (9) becomes $\alpha p_{(2)}$ and we recover the conditional method, and if we set $\beta = \alpha$ it becomes α_n and we recover the classical method.

3.2.2 Confidence Regions

Recalling Example 11's setting where we had independent samples from an MLR family, we can invert Corollary 3's test to get confidence lower bounds

$$\left\{ \theta_0 \in \mathbb{R} : \frac{p_{(1)}^{\theta_0} - \beta_n}{p_{(2)}^{\theta_0} - \beta_n} > \frac{\alpha - \beta}{1 - \beta} \right\}, \tag{10}$$

as well as confidence intervals,

$$\left\{ \theta_0 \in \mathbb{R} : 1 - \frac{\alpha - \beta}{2(1 - \beta)} > \frac{p_{(1)}^{\theta_0} - \beta_n}{p_{(2)}^{\theta_0} - \beta_n} > \frac{\alpha - \beta}{2(1 - \beta)} \right\}. \tag{11}$$

In hybrid inference, the selection event actually does depend on the specific θ_0 of the nulls $H_{0,i}^{\theta_0} : \theta_i \leq \theta_0$ that we are testing. Thus our discussion from **??** does not apply, and it is not immediate that these confidence regions are actually lower bounds and intervals. In **??**, however, we use the monotonicity of the conditional selective p-values to show that this is indeed the case.

3.2.3 Comparison to the Union Bound

Another way to balance the benefits of the conditional and simultaneous approaches would be to apply a union bound. Naively, we can reject the winning null whenever the level β simultaneous approach rejects or the level $\alpha - \beta$ conditional approach rejects, i.e., whenever

$$p_{(1)} \leq \max\{(\alpha - \beta)p_{(2)}, \beta_n\}. \tag{12}$$

The union bound harshly switches between the simultaneous and conditional approaches, whereas the hybrid approach smoothly interpolates between them. This is illustrated in Figure 1, which compares the LCBs resulting from the hybrid versus union bound approaches in the $n = 10$ dimensional Gaussian case.

Written in terms of p-values, we easily see that the hybrid approach dominates the union bound approach. Both methods reject when $p_{(1)} \leq \beta_n$, and when $p_{(1)} > \beta_n$ the hybrid cutoff (9) is clearly strictly larger than the union bound cutoff (12), meaning hybrid will reject whenever the union bound does and more [3].

There are two reasons hybrid inference out performs the union bound. First, union bounds are not tight and waste some of the Type I error budget. This is captured by the fact that the coefficient on $p_{(2)}$ is larger in the hybrid cutoff (9) than the union bound cutoff (12). Second, conditioning on the "good event" in the hybrid approach causes an upper truncation in (29), which results in $\left(1 - \frac{\alpha-\beta}{1-\beta}\right)\beta_n$ being added to the cutoff [4].

Practically speaking, however, hybrid inference typically results in very little improvement over the union bound. This is already somewhat evident in Figure 1's rightmost panel, where we see that the hybrid LCB and union bound LCB are very similar. Even when the hybrid cutoff is larger than that of the union bound, it is provably not much larger. We detail why in Appendix A.4, where we also run a number of simulations comparing the power of the hybrid and union bound approaches. We are hard pressed to find a setting where the hybrid approach results in a meaningful power gain.

Hence, we suggest adopting the view that hybrid inference's main contribution is squeezing out the little power that is left from the union bound approach. As it is not computationally more expensive and our p-value viewpoint makes it equally easy to implement, we feel it is still worth using.

3.2.4 Applying the Closure Principal

Like was true in the conditional case, treating Corollary 3's test as a global null test and closing it results in a tractable procedure that allows us to make more discoveries. As we allow β to range from 0 to α, this closed procedure interpolates between Corollary 2's closed procedure and the Holm-Sidak procedure, which is the closure of Sidak's global null test.

Corollary 4 (Closed hybrid testing for winners)**.** *Suppose that p_i are n independent and selectively dominant p-values under the nulls $H_{0,i}$. As shorthand, let $H_{0,(j)}$ denote the null corresponding to the jth smallest p-value and define $p_{(n+1)} = 1$. Fixing some $\beta < \alpha$, rejecting the null hypotheses $H_{0,(k)}$ for which*

$$p_{(j)} \leq \frac{\alpha - \beta}{1 - \beta} p_{(j-1)} + \left(1 - \frac{\alpha - \beta}{1 - \beta}\right) \beta_{n-j+1}$$

for every $j \leq k$ controls FWER error at level α.

This closed procedure is also best understood sequentially. We reject $H_{0,(1)}$ if $p_{(1)} \leq \frac{\alpha-\beta}{1-\beta}p_{(2)} + (1 - \frac{\alpha-\beta}{1-\beta})\beta_n$. Then, if we rejected $H_{0,(1)}$, we reject $H_{0,(2)}$ if $p_{(2)} \leq \frac{\alpha-\beta}{1-\beta}p_{(3)} + (1 - \frac{\alpha-\beta}{1-\beta})\beta_{n-1}$, so on and so forth until we fail to reject.

[3] the authors Andrews et al. [2023] only point out that hybrid dominates the level β classical approach, which is weaker than our statement

[4] the authors Andrews et al. [2023] mention these differences between the union bound and hybrid inference in a footnote, but they only point out that hybrid dominates the level β classical approach, which is weaker than our statement.

4 Rank Verification in Exponential Families

In this section we consider the problem of verifying that that the winning parameter is actually larger than the other parameters, i.e., rather than doing inference on the winning parameter, we do inference on the gap between the winning and remaining parameters. Mainly, we give an account of the seminal work Hung and Fithian [2019] in our selective dominance framework. We show, however, that Hung and Fithian [2019] do not correctly handle cases where there can be ties between the winner. This is a subtle point, but it is a mistake that is easy to avoid when using our selective dominance framework.

Overall, the section serves to illustrate how our selective dominance machinery provides a straightforward way to correctly design intricate and counter-intuitive selective procedures. For examples of how one may apply these methods, we refer the reader to the original article Hung and Fithian [2019], where there are many.

4.1 Warm-up: Rank Verification and Type III Error Control

To motivate the rank verification problem and shed some light on its relationship with selective dominance, we consider a seemingly unrelated classical statistical question about Type III errors.

A researcher wants to test if the unknown means of two univariate Gaussian samples, $X_1 \sim N(\mu_1, 1/\sqrt{2})$ and $X_2 \sim N(\mu_2, 1/\sqrt{2})$, are different. They end up rejecting the null hypothesis $H_0 : \mu_1 = \mu_2$ because the two-sided p-value $2(1 - \Phi(|X_1 - X_2|))$ they learned in introductory statistics is at most α. After rejecting, they note $X_1 > X_2$, and claim "not only are the two means are different, but they must be different because μ_1 is bigger than μ_2". Your friend, however, only rejected the null that the means are equal. Can they make a claim about the direction of inequality? This is a quesiton of Type III error, and we can use our selective dominance framework to show that the researcher's claim is actually statistically valid.

Based on the claim, it seems that what the researcher really wants to do is test the one-sided null $H_{0,12} : \mu_1 \leq \mu_2$ whenever they observe that $X_1 > X_2$, and test the complementary one-sided null $H_{0,21} : \mu_2 \leq \mu_1$ whenever they observe that $X_2 < X_1$. To test the null $H_{0,ij} : \mu_i \leq \mu_j$ we normally use the UMP p-value $p_{ij} = 1 - \Phi(X_i - X_j)$. In the researcher's case, however, they only select this p-value to use for inference when they observe that $X_j > X_{-j}$, or equivalently that $p_{ij} < 1/2$. Since the p_{ij} are selectively dominant, Theorem 1 tells us that, when using p_{ij} to test $H_{0,ij}$, they should correct for this selection when $2p_{ij} \leq \alpha \iff p_{ij} \leq \alpha/2$.

The procedure we describe above verifies the rank of the winning mean with Type I error control, i.e., it affirms not just that the means are different, but that the mean of the winning observation is the larger of the two. The procedure, which rejects when the smaller of the two one-sided p-valuesis at most $\alpha/2$, is identical to the procedure that rejects when the above two-sided p-value is at most α! Hence, for reasons likely unbeknowst to them, the researcher's original claim is indeed statistically valid. We walk through deriving this procedure much more carefully in Appendix A.7, where we give a treatment whos level of detail is similar to Example 7

In what follows, we generalize (10)'s rank verification procedure to the exponential family setting. Although this generalization appears quite a bit more complicated, the core idea remains the same.

4.2 Rank Verification in Exponential Families

In this sub-section we illustrate how to do rank verification when we observe data $X \in \mathbb{R}^m$ from a multiparameter exponential family P_θ with density

$$g_\theta(x) = \exp(\theta_1 T_1(x) + \cdots + \theta_n T_n(x) - \psi(\theta))g(x) \tag{13}$$

We typically care about doing inference on some subset $\mathcal{S} \subseteq [p]$ of the θ_j (e.g., in the case of a multivariate Gaussian with unknown covariance $\sigma^2 I_n$, we usually are only interested in θ_j that determine the means and not the θ_j that determines the variance). Since $E_\theta[T] = \theta$, it makes sense to let the winner W be the index in $\mathcal{S}$ corresponding to the largest sufficient statistic T_i, with ties broken randomly.

Considering the nulls $H^\delta_{0,ij} : \theta_i - \theta_j \leq \delta$, we want to reject the data-dependent global null $\cap_{j\in\mathcal{S}-W} H^\delta_{0,Wj}$ and affirm that θ_W is more than δ larger than any other parameter in $\mathcal{S}$. Then we can make an LCB for the difference $\theta_W - \max_{j\in\mathcal{S}-W} \theta_j$ between the winning and next largest parameter in $\mathcal{S}$ by inverting this test.

Considering some $i \neq j$, we start by constructing the UMPU p-value $p_{ij}^{\delta}(X)$ for testing $H_{0,ij}^{\delta} : \theta_i - \theta_j \leq \delta$. Defining the transformed sufficient statistics $\widetilde{T} \in \mathbb{R}^n$ by

$$\widetilde{T}_i = \frac{T_i - T_j}{2}, \qquad \widetilde{T}_j = \frac{T_i + T_j}{2}, \qquad \widetilde{T}_\ell = T_\ell \text{ for } \ell \neq i, j \tag{14}$$

we can reparametrize the density (15) as

$$g_\theta(x) = \exp\left((\theta_i - \theta_j)\widetilde{T}_i + (\theta_i + \theta_j)\widetilde{T}_j + \sum_{\ell \neq i,j} \widetilde{T}_\ell \theta_\ell\right) g(x). \tag{15}$$

From this reparameterization, it is clear that the conditional left-continuous survival function of $\widetilde{T}_i$ and its righthand limit

$$G_{ij}^{\delta}(\tilde{t}_i|\tilde{t}_{-i}) = P_{\theta_i - \theta_j = \delta}(\widetilde{T}_i \geq \tilde{t}_i|\widetilde{T}_{-i} = \tilde{t}_{-i}) \qquad G_{ij}^{\delta,+}(\tilde{t}_i|\tilde{t}_{-i}) = \lim_{u \downarrow \tilde{t}_i} G_{ij}^{\delta}(u|\tilde{t}_{-i}), \tag{16}$$

have no dependence on θ, and the UMPU p-value p_{ij}^{δ} for testing $H_{0,ij}^{\delta} : \theta_i - \theta_j \leq \delta$ is given by

$$p_{ij}^{\delta}(X) = G_{ij}^{\delta,+}(\widetilde{T}_i|\widetilde{T}_{-i}) + U_{ij,aux}(G_{ij}^{\theta_0}(\widetilde{T}_i|\widetilde{T}_{-i}) - G_{ij}^{\theta_0,+}(\widetilde{T}_i|\widetilde{T}_{-i})), \tag{17}$$

where $U_{ij,aux}$ are Unif([0, 1]) random variables that are independent from each other and X (see Chapter Four of Lehmann et al. [1986] and **??** for a justification).

Crucially, we can tell if $i \in \mathcal{S}$ is a winner by examining p_{ij}^{δ}. It is straightforward to confirm that $i \in \mathcal{S}$ is the sole winner exactly when $\widetilde{T}_i > \max_{k \in \mathcal{S}-i} \widetilde{T}_k - \widetilde{T}_j$. Equivalently, this happens when p_{ij}^{δ} is strictly smaller than

$$q_{ij}^{\delta,+}(\widetilde{T}_{-i}) = G_{ij}^{\delta,+}(\max_{k \in \mathcal{S}-i} \widetilde{T}_k - \widetilde{T}_j|\widetilde{T}_{-i}). \tag{18}$$

Likewise, one can confirm that $i \in \mathcal{S}$ is one of multiple winners in $\mathcal{S}$ exactly when $\widetilde{T}_i = \max_{k \in \mathcal{S}-i} \widetilde{T}_k - \widetilde{T}_j$, or equivalently when p_{ij}^{δ} is at least $q_{ij}^{\delta,+}$ but at most

$$q_{ij}^{\delta}(\widetilde{T}_{-i}) = G_{ij}^{\delta}(\max_{k \in \mathcal{S}-i} \widetilde{T}_k - \widetilde{T}_j|\widetilde{T}_{-i}), \tag{19}$$

Moreover, in the case that there are multiple winners, the number of winners is also a deterministic function of $\widetilde{T}_{-i}$:

$$N_i(\widetilde{T}_{-i}) = 1 + \{k \in \mathcal{S} - i : \widetilde{T}_k = \max_{\ell \in \mathcal{S}-i} \widetilde{T}_\ell\}. \tag{20}$$

Note that $N_i(\widetilde{T}_{-i})$, which is always at least two, is <u>not</u> the same as the number of winners, which can be one. Rather, it is the number of winners there will be if $\widetilde{T}_i$ is a winner and at least one other $\widetilde{T}_j$ is as well.

Leveraging these facts, Example 9 shows how we can use our selective dominance framework to come up with valid tests for both the winning null H_{Wj}^{δ} and the winning global null $\cap_{j \in \mathcal{S}-W} H_{0,Wj}^{\delta}$.

Example 9. *Suppose p is a p-value for the null H_0 that is selectively dominant given Z. If we select p to use for inference according to the selection function*

$$s(x, z) = \begin{cases} 1 & \textit{if } x < q^+(z), \\ \frac{1}{N(z)} & \textit{if } x \in [q^+(z), q(z)], \\ 0 & \textit{otherwise}, \end{cases},$$

where $N(z) > 1$ and $0 \leq q^+(z) \leq q(z) \leq 1$ are known functions of z, then the adjusted p-value from (4) turns out to be (see Appendix A.6 for computations)

$$p_{adj} = f(p, q^+(Z), q(Z), N(Z)) \qquad f(a, b, c, d) = \frac{a - (1 - \frac{1}{d})(a - b)_+}{b + \frac{1}{d}(c - b)}. \tag{21}$$

Therefore, Theorem 1 tells us that rejecting when (21) is at most α is a selective Type I error controlling procedure:

$$P_{H_0}\left(f(p, q^+(Z), q(Z), N(Z)) \leq \alpha \mid Z, S = 1\right) \leq \alpha. \tag{22}$$

Now, suppose we observe X drawn from the exponential family (15) and let W be the index $i \in \mathcal{S}$ of the largest sufficient satistic (with ties broken randomly). For $i \neq j$, Example 4 tells us that the UMPU p-value $p = p_{ij}^{\delta}$ from (17) for the null $H_{0,ij}^{\delta} : \theta_i - \theta_j \leq \delta$ is selectively dominant given the transformed nuisance statistics $Z = \widetilde{T}_{-i}$ from (14). Taking $q^+(Z) = q_{ij}^{\delta,+}(\widetilde{T}_{-i})$, $q(Z) = q_{ij}^{\delta}(\widetilde{T}_{-i})$, $N(Z) = N_i(\widetilde{T}_{-i})$ and f from (18), (19), (20), and (21), it is now easy to show that rejecting the data-dependent null $H_{0,Wj}^{\delta}$ when $f(p_{Wj}^{\delta}, q_{Wj}^{\delta,+}, q_{Wj}^{\delta}, N_W) \leq \alpha$ controls Type I error both conditional on W and marginally. Again it suffices to restrict our attention to indices $i \in \mathcal{I}$, $i \neq j$ that have a positive probability of being the winner (if $i = j$, then we know H_{ii}^0 is true and we simply do not reject). If $H_{0,ij}^{\delta}$ is not true, then trivially $P($falsely reject $H_{0,Wj}^{\delta} | W = i) = 0 \leq \alpha$. For the case that $H_{0,Wj}^{\delta}$ is true (and $i \neq j$), the event $W = i$ is the same event as selecting $p_{0,ij}^{\delta}$ for inference in (21), so

$$\begin{aligned} P(\text{falsely reject } H_{0,Wj}^{\delta} | W = i) &= P(f(p_{Wj}^{\delta}, q_{Wj}^{\delta,+}, q_{Wj}^{\delta}, N_{Wj}) \leq \alpha | W = i) \\ &= P(f(p_{ij}^{\delta}, q_{ij}^{\delta,+}, q_{ij}^{\delta}, N_i) \leq \alpha | W = i) \\ &\leq \alpha. \end{aligned}$$

Marginal error control follows from the usual law of total probability argument.

We can reject the data-dependent global null when $\cap_{j \in \mathcal{S}-W} H_{0,Wj}^{\delta}$ when we reject all of the individual nulls $H_{0,Wj}^{\delta}$ for $j \in \mathcal{S} - W$. It is straightforward to see that this will control Type I error both conditionally on W and marginally: If there is an $i \in \mathcal{I}$ for which $\cap_{j \in \mathcal{S}-i} H_{0,ij}^{\delta}$ is false, then trivially $P($falsely reject $\cap_{j \in \mathcal{S}-W}$ $H_{0,Wj}^{\delta} | W = i) = 0 \leq \alpha$ for this i. Otherwise, there exists a $k \in \mathcal{S} - i$ for which $\theta_i \leq \theta_k$, and

$$P(\text{falsely reject } \cap_{j \in \mathcal{S}-W} H_{0,Wj}^{\delta} | W = i) \leq P(\text{falsely reject } H_{0,Wk}^{\delta} | W = i) \leq \alpha.$$

Again, marginal error control follows from the usual law of total probability argument.

By inverting Example 9's test for the data-dependent global null $\cap_{j \in \mathcal{S}-W} H_{0,Wj}^{\delta}$, we can get our desired LCB for $\theta_W - \max_{j \in \mathcal{S}-W} \theta_j$. We give a precise statement in Theorem 3 below.

Theorem 3. *Let X be drawn from the exponential family (15) and let W be the index $i \in \mathcal{S}$ corresponding to the largest sufficient statistic T_i (with ties broken randomly). If p_{ij}^{δ} is the UMPU p-value (17) for testing $H_{0,ij}^{\delta} : \theta_i - \theta_j \leq \delta$ and $q_{ij}^{\delta,+}$, q_{ij}^{δ}, and N_i are as in (18), (19), and (37), then*

$$\left\{ \delta : \max_{j \in \mathcal{S}-W} \frac{p_{Wj}^{\delta} - \left(1 - \frac{1}{N_i}\right)(p_{Wj}^{\delta} - q_{Wj}^{\delta,+})_+}{q_{Wj}^{\delta,+} + \frac{1}{N_i}(q_{Wj}^{\delta} - q_{Wj}^{\delta,+})} > \alpha \right\} \tag{23}$$

is a LCB for θ_W that holds with probability at least $1-\alpha$ conditionally on W (and therefore also marginally).

Our derivation of the test for $H_{0,Wj}$ differs from that given in Hung and Fithian [2019] in two important ways. First, the derivation in Hung and Fithian [2019] only implies Type I error control at the boundary of the one-sided null, although they assume it to be true for the entire composite null. Due to selection, establishing validity for entire composite null is non-trivial. Thankfully, it is immediately implied by our selective dominance machinery and Example 4. Second, Hung and Fithian [2019] incorrectly use the same adjusted p-value when there are and are not ties amongst the T_j. If there cannot be ties amongst the T_j, then $q_{ij}^{\delta} = q_{ij}^{\delta,+}$ and the adjusted p-value (21) we use to test $H_{0,ij}^{\delta}$ conditional on $W = i$ simplifies to $p_{ij}^{\delta}/q_{ij}^{\delta}$. If we use $p_{ij}^{\delta}/q_{ij}^{\delta}$ as our adjusted p-value when ties are possible, we will not achieve conditional error control (as is claimed in Fithian et al. [2017]). In the case that $X \in \mathbb{R}^3$ is three independent binomials, the left panel of Figure 4 depicts the conditional distribution of p_{12}^0 given $W = 1$ and a specific setting of the nusiances statistics Z. It makes it clear that rejecting when $p_{ij}^{\delta}/q_{ij}^{\delta} < \alpha$ does not maintain error control, as affirmed in Figure 4's right panel.

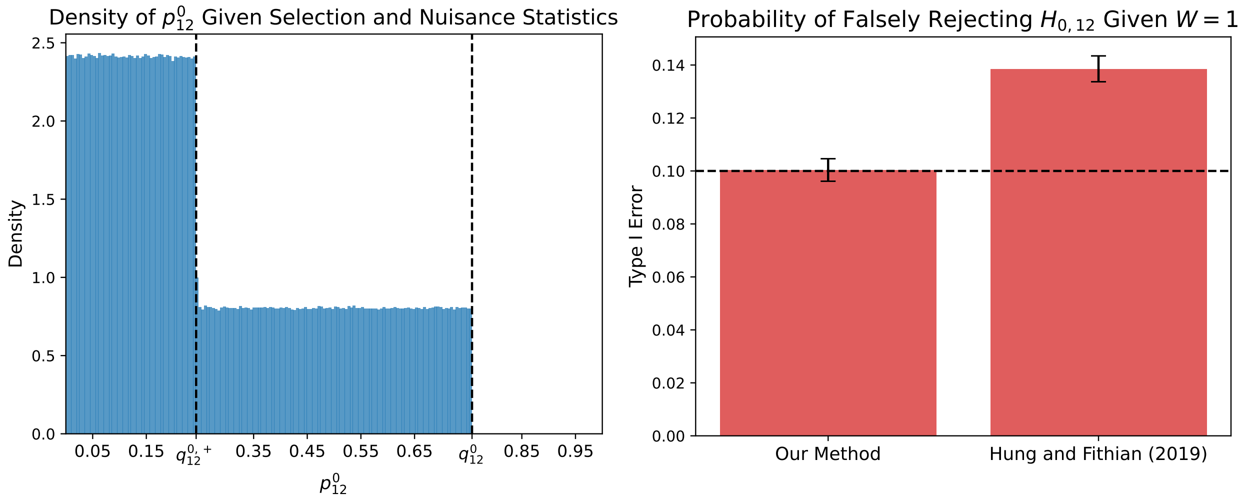


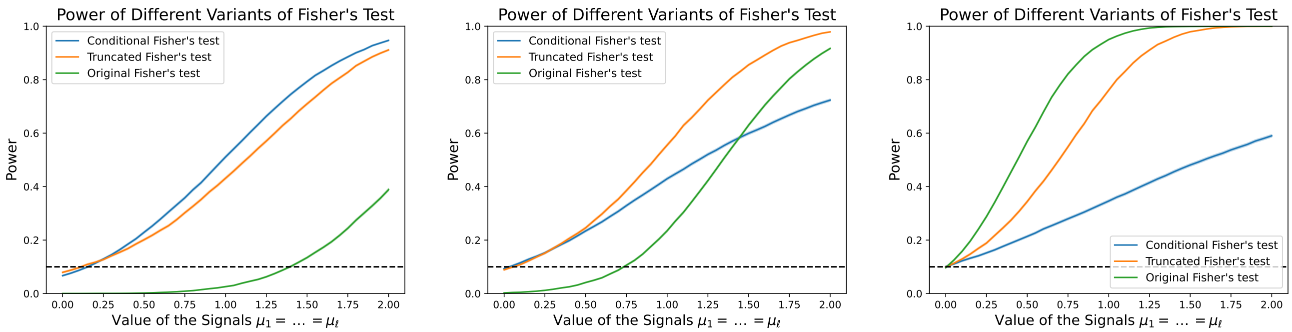
Figure 4: Considering three independent binomial random variables $X_i \sim \text{Bin}(n, s_i)$ with random variables with $n = 4$ and $s_i = 1/2$, the first panel (left) depicts $N = 10^6$ draws from the conditional distribution of the p-value p^0_{12} (used for testing $H^0_{0,12} : s_1 \leq s_2$) given selection $W = 1$ and the nuisance statistics $(X_1 + X_2)/2 = 4$, $X = 3$. When $p^0_{12} < q^{0,+}_{12}$, then X_1 is the sole winner and the p-value is selected for inference with probability one, but when $p^0_{12} \in [q^{0,+}_{12}, q^0_{12}]$ there is a three-way tie and it is selected for inference with probability 1/3. Hence, the p-value's conditional distribution is not uniform on $[0, q^0_{12}]$ as Hung and Fithian [2019] implicitly assume. The next panel (right) displays the consequence. Conditional on $W = 1$, Hung and Fithian [2019] do not maintain Type I error control when testing H^0_{W2} at level $\alpha = 0.1$ (denoted by horizontal dashed line), whereas our method does. Error bars denote a 99% confidence interval.

In closing, we mention that the the final result (Theorem 3) is perhaps not as surprising as Hung and Fithian [2019] originally suggest. In Hung and Fithian [2019], there are further computations showing that if the exponential family density (15) is such that $T(X) = X$, the function $g(x)$ is Schur concave (see [Hung and Fithian, 2019, Definition 2]), and there are no ties, then (1) the maximum in (23) is achieved by the runner up R (the index $j \in S$ with second largest sufficient statistic T_j) and (2) q^0_{WR} is equal to 1/2. Hence, in this restricted setting, we can claim that the winning parameter is indeed the largest whenever $p^0_{WR} \leq \alpha/2$, i.e., whenever the two-sided test comparing the winner and runner-up rejects. Since this test has non-vanishing level even as n grows, the authors claim that it seems to surprisingly circumvent a multiple comparisons issue. However, Example 10 already illustrates why verifying rank by running a two-sided test comparing the winner and loser gives Type I error control in the $n = 2$ dimensional case. It should be obvious then that running a two-sided test to compare the winner to the best of $n-1$ losers in the $n > 2$ dimensional case is a more conservative procedure. In other words, the multiplicty correction does not appear in the level of the test, but it is baked in by the fact that we compare the winner to the largest of $n - 1$ different losers, which scales with n.

5 New Selective Methods

In this section we illustrate how our selective dominance framework allows us to easily design new methods for selective inference, global null testing, and dealing with publication bias.

Rather than just selecting just one p-value as in Section 2, some of this section's methods select multiple p-values to use for inference. As such, we need to slightly generalize our framework from Section 2. The generalization is intuitive, and for sake of brevity we have deferred a formal account of it to Appendix C. The proofs of validity for this section's methods, which are in the appendix, often rely on Appendix C.



(a) $\ell = 3$ (b) $\ell = 5$ (c) $\ell = 10$

Figure 5: For $\ell = 3$ (left), $\ell = 5$ (middle), and $\ell = 10$ (right), power of the top $k = 3$ conditional, $\tau = 0.5$ truncated, and original Fisher's combination test for data drawn from $N(\mu, I_{10})$ with $\mu_1 = \cdots = \mu_\ell$ varying according to the x-axis and $\mu_{\ell+1} = \cdots = \mu_n = -2$. Power results from an average over $N = 10^4$ trials, bands denote one standard error (barely visible), and the level $\alpha = 0.1$ is denoted by the dashed line.

5.1 Adaptive Versions of Fisher's Combination Test

If we are just interested in verifying the existence of a signal, we can try and combine inferences on the top k effects rather than trying to simultaneous inference on them individually. This is made possible by Corollary 5, which gives a conditional version of Fisher's combination test. Recall that Fisher's original combination test considers n independent p-values p_i for the nulls $H_{0,i}$ and rejects the global null $\cap_{i=1}^n H_{i,0}$ when the test statistic $-2\sum_{i=1}^n \log(p_i)$ is at least as large as the $1-\alpha$ quantile of the χ^2_{2n} distribution.

Corollary 5. *Suppose that p_i are n independent and selectively dominant p-values under the nulls $H_{0,i}$ and let $H_{0,(j)}$ denote the null corresponding to the jth smallest p-value. Rejecting the data-dependent global null $\cap_{j=1}^k H_{0,(j)}$ (and therefore also the global null $\cap_{i=1}^n H_{0,i}$) when*

$$-2\sum_{j=1}^{k} \log(p_{(j)}/p_{(k+1)}) \geq \mathit{Quantile}(1-\alpha, \chi^2_{2k}) \tag{24}$$

controls Type I error at level α conditional on the indicies of the smallest k p-values (and therefore also marginally).

Examining (24), we see that when the $(k+1)$st p-value provides essentially no evidence against the null (so $p_{(k+1)} \approx 1$), running Corollary 5's test is like running Fisher's test using just the top k p-values, completely ignoring that we selected from n of them. In this case, our test statistic will have an essentially identical value to Fisher's original test statistic, but the critical value required for rejection will be much smaller. On the flipside, if many of the $p_{(j)}$ for $j \leq k$ are not sufficiently smaller than $p_{(k+1)}$, then Corollary 5's test statistic will be small and the test will fail to reject.

Another approach to improving Fisher's combination test is truncation: only add a p-value to Fisher's statistic if it is below some fixed threshold $\tau \in \mathbb{R}$ [Zaykin et al., 2002]. Past and previous work, however, only establishes validity of this test when the null p-values have exact Unif[0, 1] distributions, excluding the possbility of one-sided null hypotheses [Zhang et al., 2020]. Corollary 6, however, gives a generalized version of Fisher's truncated combination test that is still valid whenever the p-values are independent and selectively dominant.

Corollary 6. *Suppose that p_i are n independent and selectively dominant p-values under the nulls $H_{0,i}$ and fix n thresholds $\tau_i \in [0,1]$. Letting $j \in J$ denote the random set of indices for which $p_j \leq \tau_j$, rejecting the data-dependent global null $\cap_{j\in J} H_{0,j}$ (and therefore also the global null $\cap_{i=1}^n H_{0,i}$) when*

$$-2\sum_{j=1}^{k} \log(p_j/\tau_j) \geq \mathit{Quantile}(1-\alpha, \chi^2_{2|J|})$$

controls type I error at level α conditional on which indicies are in J (and therefore also marginally).

If some p_j are substantially lower than their truncation point τ_j but most are above it, compared to Fisher's original combination test, Corollary 6's test will use a slightly smaller test statistic but a much smaller critical value. Hence, the truncated Fisher test is most powerful when some p-values come from strong alternatives but many p-values come from conservative nulls. (i.e, the null p-values do not have exact Unif([0, 1]) distributions). As such, Corollary 6 actually generalizes the truncated Fisher test to the settings where it is most useful. On the flipside, if most of the p_j are below τ_j, the truncated test statistic pays a penalty due to selection while, compared to Fisher's original test, the test's critical value remains essentially unchanged.

To illustrate the benefits and drawbacks of these methods, we display their power alongside that of Fisher's original test for a simple $n = 10$ dimensional Gaussian problem, where we sample $X \sim N(\mu, I_n)$ and use the p-values $p_i = 1 - \Phi(X_i)$ try and detect the existence of a positive mean. For $\ell \in \{3, 5, 10\}$, we vary the strength $\mu_1 = \cdots = \mu_\ell > 0$ of our signals and set $\mu_{\ell+1} = \cdots = \mu_n = -2$ to be conservative nulls. We do inference on the top $k = 3$ p-values using the conditional version of Fisher's method and set the truncation $\tau = 0.5$ for the truncated version (i.e., we include p_i for which $X_i > 0$). The results are displayed in Figure 5

As expected, the new methods outperform Fisher's original method when conservative nulls are present. When $\ell = 3$ and the bottom three p-values are much smaller than the rest, the conditional method does incredibly well. But its performance quickly degrades when $\ell = 5, 10$ and the fourth smallest p-value becomes often close to the bottom three. By focusing inference on just the top five p-values, the truncated method still considerably improves power when $\ell = 5$. Unsurprisingly, both methods perform worse than Fisher's original method when $\ell = 10$ and every μ_i is a signal.

Acknowledgements

We thank John Cherian, Yash Nair, and Will Hartog for helpful discussions.

References

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on Winners*. *The Quarterly Journal of Economics*, 139(1):305–358, 09 2023. ISSN 0033-5533. doi: 10.1093/qje/qjad043. URL https://doi.org/10.1093/qje/qjad043.

Yuval Benjamini, Jonathan Taylor, and Rafael A Irizarry. Selection-corrected statistical inference for region detection with high-throughput assays. *Journal of the American Statistical Association*, 114(527):1351–1365, 2019.

Ameer Dharamshi, Anna Neufeld, Keshav Motwani, Lucy L Gao, Daniela Witten, and Jacob Bien. Generalized data thinning using sufficient statistics. *Journal of the American Statistical Association*, (just-accepted):1–26, 2024.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection, 2017. URL https://arxiv.org/abs/1410.2597.

Charles J Geyer and Glen D Meeden. Fuzzy and randomized confidence intervals and p-values. *Statistical Science*, pages 358–366, 2005.

Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *TEST*, 27(4):811–825, 2018. doi: 10.1007/s11749-017-0571-1. URL https://doi.org/10.1007/s11749-017-0571-1.

Kenneth Hung and William Fithian. Rank verification for exponential families. *The Annals of Statistics*, 47 (2):758 – 782, 2019. doi: 10.1214/17-AOS1634. URL https://doi.org/10.1214/17-AOS1634.

Kenneth Hung and William Fithian. Statistical methods for replicability assessment. *The Annals of Applied Statistics*, 14(3):1063 – 1087, 2020. doi: 10.1214/20-AOAS1336. URL https://doi.org/10.1214/20-AOAS1336.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. 2016.

Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 3. Springer, 1986.

Lihua Lei and William Fithian. AdaPT: An Interactive Procedure for Multiple Testing with Side Information. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):649–679, 06 2018. ISSN 1369-7412. doi: 10.1111/rssb.12274. URL https://doi.org/10.1111/rssb.12274.

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: splitting a single data point. *Journal of the American Statistical Association*, pages 1–12, 2023.

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

Alexander Molotaliev. Engine time to failure dataset, 2024. URL https://www.kaggle.com/datasets/m0ntecarl0/engine-time-to-failure.

Anna Neufeld, Ameer Dharamshi, Lucy L Gao, and Daniela Witten. Data thinning for convolution-closed distributions. *Journal of Machine Learning Research*, 25(57):1–35, 2024.

Souhardya Sengupta and Lucas Janson. The ℓ-test: leveraging sparsity in the gaussian linear model for improved inference. *arXiv preprint arXiv:2406.18390*, 2024.

Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons. P-curve: A key to the file drawer. *Cognitive Linguistics: Cognition*, 2013. URL https://api.semanticscholar.org/CorpusID:8505270.

Dmitri V Zaykin, Lev A Zhivotovsky, Peter H Westfall, and Bruce S Weir. Truncated product method for combining p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 22(2):170–185, 2002.

Hong Zhang, Tiejun Tong, John Landers, and Zheyang Wu. Tfisher: A powerful truncation and weighting procedure for combining p-values. 2020.

A Additional Derivations, Details, and Comments

A.1 Comparing Conditional and Sidak Global Null Testing

We consider a setting where we have n independent and selectively dominant p-values $p_1, \dots, p_n$ that are all anti-conservative, i.e., $p_j \preceq \text{Unif}([0,1])$. At worst, these p-values are exact uniforms (e.g., they come from the boundary of the null).

We will show that, on an event with probability at least $1-\epsilon$, the conditional procedure, which rejects when $p_{(1)} \leq \alpha p_{(2)}$, can only reject if $p_{(1)} \leq C_\epsilon/n$ for some constant $C_\epsilon > 0$. Hence, without conservative nulls, the conditional approach behaves roughly on the same order as the classical approach (Sidak).

Letting $U_1, \dots, U_n$ be independent Unif([0, 1]) random variables, two facts are clear. First that $U_{(2)} \sim \text{Beta}(2, n-1)$ has mean $\frac{2}{n+1} < \frac{2}{n}$ and standard deviation $\sqrt{\frac{2(n-1)}{(n+1)^2(n+2)}} \leq \frac{2}{n}$, and second that $p_{(2)} \preceq U_{(2)}$.

Fix any $\epsilon > 0$. We have by Chebyshev's inequality that

$$
\begin{aligned}
P\left(p_{(2)} \leq \frac{2}{n}(1+\epsilon^{-\frac{1}{2}})\right) &\geq P(p_{(2)} \leq E[U_{(2)}] + \sqrt{\text{Var}(U_{(2)})}/\sqrt{\epsilon}) \\
&\geq P(U_{(2)} \leq E[U_{(2)}] + \sqrt{\text{Var}(U_{(2)})}/\sqrt{\epsilon}) \\
&\geq P(|U_{(2)} - E[U_{(2)}]| < \sqrt{\text{Var}(U_{(2)})}/\sqrt{\epsilon}) \\
&\geq 1-\epsilon
\end{aligned}
$$

Define $C_\epsilon = 2(1+\epsilon^{-\frac{1}{2}})\alpha$ and the event $A_\epsilon = \{p_{(2)} \leq \frac{2}{n}(1+\epsilon^{-\frac{1}{2}})\}$. The event A_ϵ has probability at least $1-\epsilon$, and on this event a the conditional procedure never rejects when the procedure $p_{(1)} > C_\epsilon/n$, completing our claim.

In other words, the conditional procedure can only reject when $p_{(1)} \leq C_\epsilon/n$ (except for a small probability event) and hence suffers the same curse of dimensionality as the classical method:

$$P(p_{(1)} \leq \alpha p_2 \text{ and } p_{(1)} > C_\epsilon/n) \leq P(A_\epsilon^c) \leq \epsilon.$$

A.2 Conditional Inference on Winners

A.2.1 Standard Derivation

Suppose we observe independent Gaussian data $X \sim N(\mu, I_n)$ and want to make a LCB for the mean μ_W of the winner $W = \operatorname{argmax}_{i\in[n]} X_i$.

The conditional approach tries to avoid the classical approach's curse of dimensionality by providing inferences for only the winning mean. It does so by delivering an LCB for μ_W that is valid conditionally on W. To construct the conditional LCB, we follow the framework of Fithian et al. [2017]. Letting R be the index of the runner-up (second largest observation), we note that the deviation of X_W from μ_W has a truncated normal distribution once we condition on W and X_{-W}:

$$X_W - \mu_W \mid W, X_{-W} \sim TN(0, 1, X_R - \mu_W, \infty). \tag{25}$$

Letting

$$q_{1-\alpha}(w, x_{-w}, \mu_w) = \text{Quantile}_\mu(1-\alpha, X_W - \mu_W \mid W = w, X_{-w} = x_{-w}) \tag{26}$$

denote the $1-\alpha$ quantile of this conditional distribution (25), it is straightforward to show that

$$C_{cond}(X) = \{\eta : \eta > X_W - q_{1-\alpha}(W, X_{-W}, \eta)\} \tag{27}$$

is a $1-\alpha$ confidence region for μ_W that is valid conditional on W and X_{-W}:

$$P_\mu(\mu_W \in C_{cond}(X)|W, X_{-W}) = P_\mu(X_W - \mu_W < q_{1-\alpha}(W, X_{-W}, \mu_W)|W, X_W) = 1-\alpha.$$

Since η appears on both sides of the membership condition in (27), it is not clear that $C_{cond}(X)$ actually provides a lower bound for μ_W. We will see, however, that we can rewrite $C_{cond}(X)$ as

$$C_{cond}(X) = \{\eta : \eta > X_W - L(X_W - X_R)\}. \tag{28}$$

where $L : \mathbb{R} \to \mathbb{R}$ is a complicated "length" function that determines the distance from the winner X_W to the lower confidence bound. We will de-mistify $L(\cdot)$ in the next sub-section. For now, we provide a plot of it in the middle panel of Figure 1.

As can be seen in Figure 1, the conditional LCB has very interesting behavior. In some situations, it indeed seems to avoid the curse of dimensionality. If the gap between X_W and X_R is large, then the conditional LCB for μ_W will be roughly $X_W - z_{1-\alpha}$, i.e, what we expect in a one-dimensional inference problem. But the cost of using the conditional approach can be tremendous. As the runner-up gets close to the winner, the conditional LCB goes quickly to $-\infty$ and can give much worse inferences than the classical approach. For example, if the runner-up is within just one standard deviation of the winner, we get inferences similar to those in the $n = 100$ dimensional classical problem.

Compared to the classical approach, which provides a lower bound that is a fixed distance from the winner, the conditional approach is much harder to interpret. Its quality depends on a somewhat opaque relationship between the runner-up and the winner. How separated does the winner need to be from the runner-up for the conditional approach to be useful? And how close can the winner and runner-up be before the classical approach is better? We shed light on these questions by rewriting the conditional procedure in terms of p-values.

A.2.2 p-Value Viewpoint

Considering data $X \sim N(\mu, I_n)$ with unknown mean μ and the p-values $p_i^{\mu_0} = 1 - \Phi(X_i - \mu_0)$, we want to characterize when the conditional LCB is at least $\mu_0 \in \mathbb{R}$. This happens exactly when μ_0 is not included in the set (27). Examining (25), (26), and (27), this happens when $X_W - \mu_0$ is at least as large as the $1-\alpha$ quantile Q of a standard normal truncated to be larger than $X_R - \mu_0$. This quantile satisfies

$$\alpha = \frac{1 - \Phi(Q)}{1 - \Phi(X_R - \mu_0)}.$$

Solving for Q gives $Q = \Phi^{-1}(1 - \alpha(1 - \Phi(X_R - \mu_0)))$, meaning we reject exactly when

$$\begin{aligned} X_W - \mu_0 \geq \Phi^{-1}(1 - \alpha(1 - \Phi(X_R - \mu_0))) &\iff 1 - \Phi(X_W - \mu_0) \leq \alpha(1 - \Phi(X_R - \mu_0)) \\ &\iff p_{(1)}^{\mu_0} \leq \alpha p_{(2)}^{\mu_0}. \end{aligned}$$

A.2.3 Distance Between X_W and Conditional LCB

Consider observing Gaussian data $X \sim N(\mu, I_n)$ and let W and R be the indicies of the winner and runner up respectively. We briefly justify that the distance between X_W and the conditional LCB for μ_W depends only on the gap between $X_W - X_R$. Define $D = X_W - \mu_0$ to be distance between X_W and the conditional LCB $\hat{\mu}$. The conditional LCB $\hat{\mu}$ satisfies

$$\frac{p^{\hat{\mu}}(X_W)}{p^{\hat{\mu}}(X_R)} = \alpha \iff \frac{1 - \Phi(X_W - \hat{\mu})}{1 - \Phi(X_R - \hat{\mu})} = \alpha \iff \frac{1 - \Phi(D)}{1 - \Phi(D - (X_W - X_R))} = \alpha.$$

Clearly then D is a function of $X_W - X_R$.

A.2.4 The selective p-value

coming soon

A.2.5 Confidence lower bound

coming soon

A.2.6 Confidence interval

coming soon

A.3 Hybrid Inference on Winners

A.3.1 Standard Deviation

Like before, we want to make a LCB for the mean μ_W of the winner $W = \text{argmax}_{i \in [n]} X_i$ in the case of independent Gaussian data $X \sim N(\mu, I_n)$ with unknown mean $\mu \in \mathbb{R}^n$.

The core idea behind hybrid inference is giving a confidence region $C_{hyb}(X)$ that has a very high probability of containing μ_W on a "good" event G_μ. Oddly, this good event depends on the unknown parameter. For some $\beta < \alpha$, we need G_μ to happen with probability at least $1 - \beta$. Then, if we ensure that $C_{hyb}(X)$ has at least $(1-\alpha)/(1-\beta)$ coverage on the G_μ, it will achieve $1 - \alpha$ coverage overall:

$$\begin{aligned} P_\mu(\mu_W \in C_{hyb}(X)) &= P_\mu(G_\mu) P_\mu(\mu_W \in C_{hyb}(X) | G_\mu) + P_\mu(G_\mu^c) P_\mu(\mu_W \in C_{hyb}(X) | G_\mu^c) \\ &\geq (1 - \beta) \left(\frac{1 - \alpha}{1 - \beta} \right) \\ &= 1 - \alpha. \end{aligned}$$

Considering some $\beta < \alpha$ and defining $\beta_n = 1 - (1-\beta)^{\frac{1}{n}}$ as in (**??**), our good event is that the confidence lower bounds $X_i - z_{1-\beta_n}$ for the means μ_i all simultaneously hold:

$$G_\mu = \{X_i < \mu_i + z_{1-\beta_n} \text{ for all } i \in [n]\}.$$

From our earlier reasoning, we know that this good event happens with probability exactly $1 - \beta$.

Now, we can make a confidence region that contains the mean with probability at least $(1-\alpha)/(1-\beta)$ on this good event. If we condition on G_μ along with W and X_{-W}, the deviation of X_W from μ_W has a truncated normal distribution like (25) that is further truncated from above:

$$X_W - \mu_W \mid W, X_{-W}, G_\mu \sim TN(0, 1, X_R - \mu_W, z_{1-\beta_n}). \tag{29}$$

On the good event G_μ, we always have that $X_R - \mu_W < X_W - \mu_W < z_{1-\beta_n}$, so the lower truncation is indeed below the upper one. Let

$$q^h_{\frac{1-\alpha}{1-\beta}}(w, x_{-w}, \mu_w) = \text{Quantile}_\mu\left(\frac{1-\alpha}{1-\beta}, X_W - \mu_W \mid W = w, X_{-w} = x_{-w}, G_\mu\right) \tag{30}$$

denote the $(1-\alpha)/(1-\beta)$ quantile of the conditional distribution (29). Per the prior discussion, the function (30) only makes sense if the largest value in x_{-w} at most $z_{1-\beta_n}$, and we will take the quantile (30) to be $-\infty$ if it is not. It is then straightforward to show that

$$C_{hyb}(X) = \{\eta : \eta > X_W - q^h_{\frac{1-\alpha}{1-\beta}}(W, X_{-W}, \eta)\} \tag{31}$$

contains μ_W with high probability conditional on W, X_{-W}, and the event G_μ:

$$P_\mu(\mu_W \in C_{hyb}(X) | W, X_{-W}, G_\mu) = P_\mu(X_W - \mu_W < q^h_{\frac{1-\alpha}{1-\beta}}(W, X_{-W}, \mu_W) | W, X_W, G_\mu) = \frac{1-\alpha}{1-\beta}.$$

Based on our earlier discussions, this is sufficient to imply that $C_{hyb}(X)$ from (31) will contain μ_W with probability at least $1-\alpha$.

As was the case for our conditional confidence region (27), the hybrid confidence region (31) is a little hard to interpret at first. We will get a clearer sense of its benefits when we write it in terms of p-values. As a teaser, the rightmost panel of Figure 1 plots distance between the winner X_W and the hybrid LCB for a $n = 10$ dimensional problem. We argue (using our upcoming p-value viewpoint) in Appendix A.3.3 that, once we fix α and β, this distance is a function of just the gap $X_W - X_R$ between the winner and runner up and the problem dimension n.

A.3.2 p-Value Viewpoint

Considering data $X \sim N(\mu, I_n)$ with unknown mean μ and the p-values $p_i^{\mu_0} = 1 - \Phi(X_i - \mu_0)$, we want to characterize when the hybrid LCB is at least $\mu_0 \in \mathbb{R}$. Examining (29), (30), and (31), we can consider two cases to figure out when this happens.

Case One - $X_R - \mu_0 \geq z_{1-\beta_n}$: If $X_R - \mu_0 \geq z_{1-\beta_n}$, then $q^h_{\frac{1-\alpha}{1-\beta}}(W, X_{-W}, \mu_0) = -\infty$, so μ_0 cannot be in (31). This case happens precisely when

$$X_R - \mu_0 \geq z_{1-\beta_n} \iff 1 - \Phi(X_R - \mu_0) \leq 1 - \Phi(z_{1-\beta_n}) \iff p_{(2)}^{\mu_0} \leq \beta_n.$$

Case Two - $X_R - \mu_0 < z_{1-\beta_n}$: If $X_R - \mu_0 < z_{1-\beta_n}$, then μ_0 is not in (31) exactly when $X_W - \mu_0$ is at least as large as the $\frac{1-\alpha}{1-\beta}$ quantile Q of a standard normal truncated to be larger than $X_R - \mu_0$ but smaller than $z_{1-\beta_n}$. This quantile satisfies

$$\frac{\alpha - \beta}{1-\beta} = \frac{\Phi(z_{1-\beta_n}) - \Phi(Q)}{\Phi(z_{1-\beta_n}) - \Phi(X_R - \mu_0)} = \frac{1 - \beta_n - \Phi(Q)}{1 - \beta_n - \Phi(X_R - \mu_0)}$$

Solving for Q gives

$$Q = \Phi^{-1}\left(1-\beta_n - \frac{\alpha-\beta}{1-\beta}\left(1-\beta_n - \Phi(X_R-\mu_0)\right)\right) = \Phi^{-1}\left(\left(1-\frac{\alpha-\beta}{1-\beta}\right)(1-\beta_n) + \frac{\alpha-\beta}{1-\beta}\Phi(X_R-\mu_0)\right),$$

so we reject exactly when

$$\begin{aligned}
X_W - \mu_0 &\geq \Phi^{-1}\left(\left(1-\frac{\alpha-\beta}{1-\beta}\right)(1-\beta_n) + \frac{\alpha-\beta}{1-\beta}\Phi(X_R-\mu_0)\right) \\
\iff 1-\Phi(X_W-\mu_0) &\leq \left(1-\frac{\alpha-\beta}{1-\beta}\right)\beta_n + \frac{\alpha-\beta}{1-\beta}(1-\Phi(X_R-\mu_0)) \\
\iff p_{(1)}^{\mu_0} &\leq \frac{\alpha-\beta}{1-\beta}p_{(2)}^{\mu_0} + \left(1-\frac{\alpha-\beta}{1-\beta}\right)\beta_n
\end{aligned}$$

It turns out we can combine these two cases. Because $p_{(1)}^{\mu_0} \leq p_{(2)}^{\mu_0}$, the fact that $p_{(2)}^{\mu_0} \leq \beta_n$ in Case One implies that $p_{(1)}^{\mu_0} \leq \beta_n$ also. Therefore in Case One, $p_{(1)}^{\mu_0}$ must be strictly smaller than a mixture of $p_{(2)}^{\mu_0}$ and β_n:

$$p_{(1)}^{\mu_0} \leq \frac{\alpha-\beta}{1-\beta}p_{(2)}^{\mu_0} + \left(1-\frac{\alpha-\beta}{1-\beta}\right)\beta_n. \tag{32}$$

Therefore, if we reject according to (32), we will always reject in Case One, and we will reject at the appropriate times in Case Two.

A.3.3 Distance Between X_W and Hybrid LCB

Consider observing Gaussian data $X \sim N(\mu, I_n)$ and let W and R be the indicies of the winner and runner up respectively. We briefly justify that the distance between X_W and the chybrid LCB for μ_W depends only on the gap between $X_W - X_R$ and the dimension n. Define $D = X_W - \mu_0$ to be distance between X_W and the hybrid LCB $\hat{\mu}$. The hybrid LCB $\hat{\mu}$ satisfies

$$\begin{aligned}
p^{\hat{\mu}}(X_W) &= \frac{\alpha-\beta}{1-\beta}p^{\hat{\mu}}(X_R) + \left(1-\frac{\alpha-\beta}{1-\beta}\right)\beta_n \\
1-\Phi(X_W-\hat{\mu}) &= \frac{\alpha-\beta}{1-\beta}(1-\Phi(X_R-\hat{\mu})) + \left(1-\frac{\alpha-\beta}{1-\beta}\right)\beta_n \\
\iff 1-\Phi(D) &= \frac{\alpha-\beta}{1-\beta}(1-\Phi(D-(X_W-X_R))) + \left(1-\frac{\alpha-\beta}{1-\beta}\right)\beta_n.
\end{aligned}$$

Clearly then D is a function of $X_W - X_R$ and n.

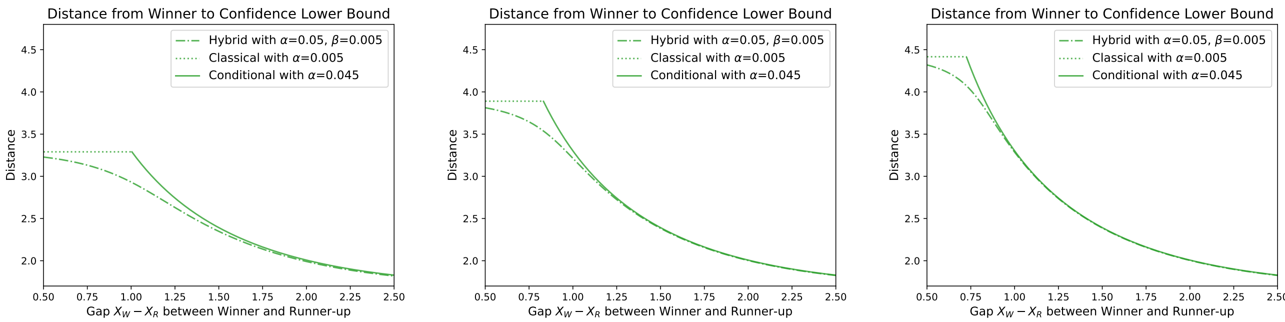
A.4 Comparing Hybrid Inference to the Union Bound

As discussed earlier the hybrid cutoff (9) is strictly larger than the union bound cutoff (12) when $p_{(1)} > \beta_n$. Thus, both procedures reject when $p_{(1)} \leq \beta_n$. When $p_{(1)} > \beta_n$, the hybrid procedure rejects but the union bound does not whenever

$$p_{(1)} \in \left((\alpha-\beta)p_{(2)}, \frac{(\alpha-\beta)}{1-\beta}p_{(2)} + \left(1-\frac{(\alpha-\beta)}{1-\beta}\right)\beta_n\right]$$

When $p_{(2)} = 1$ and $n = 1$, the length of the interval is β, which is the largest it can possible be. Thus, in the case where we can have additional rejections, the hybrid cutoff is never more than β plus the union bound cutoff. Andrews et al. [2023] suggests taking $\beta = \alpha/10$, so when $\alpha = 0.05$, for example, $\beta = 0.005$ is quite small.

Still, this is not a precise statement about power gain. The computations required to compute the power gain analytically are messy, so instead we gauge the power gain via simulation. We sample data $X \sim N(\mu, I_n)$



(a) $n = 10$ (b) $n = 100$ (c) $n = 1000$

Figure 6: For $n = 10$ (left), $n = 100$ (middle), and $n = 1000$ (right) the distance between the hybrid LCB to the winner (dash-dot line) and union bound LCB to the winner (dotted and solid line) with $\alpha = 0.05$ and $\beta = 0.005$ plotted as a function of the gap between the winning and runner-up observation.

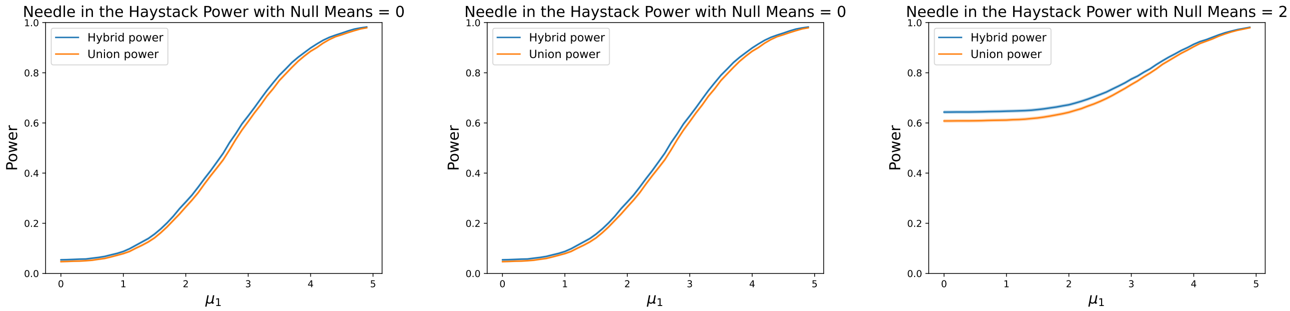
for $n = 10$ and attempt to reject the winning null $H_W : \mu_W \leq 0$ where $W = \text{argmin}_{i\in[n]} 1 - \Phi(X_i)$ is the index of the smallest p-value $p_i = 1 - \Phi(X_i)$.

We choose $n = 10$ because it is a reasonably small dimension size where one may apply hybrid inference (e.g., the main example from Andrews et al. [2023] has $n = 13$). Let R denote the index of the runner-up (second largest observation). For the dimensions $n = 10, 100, 1000$, Figure 6 compares the distance from the winner X_W the hybrid LCB and the union bound LCB. As illustrated in the plot, the (small) benefit of hybrid inference dissapates as the dimension of the problem increases. This is because conditioning on the good event has less and less of an effect as n grows.

Our simulation results indicate that hybrid inference typically results in a fairly small power gain. We consider two simulated settings:

Needle in a haystack: First, we consider a needle in the haystack setting where $\mu_1 > 0$ and all the other μ_i for $i \neq 1$ are set to μ_2. We try $\mu_2 = -2, 0, 2$. The power comparison is ploted in Figure 7. Whenever we truly have a needle in the haystack problem, i.e., $\mu_1 > \mu_2$, hybrid inference results in essentially no power gain. The only setting where we see some gain (up to around 0.05) is when $\mu_2 > \mu_1$. In this setting we actually have a dense alternative (many small signals). We expect the top two p-values to be close to each other, so conditional methods should perform poorly. The union bound approach indeed performs essentially identically to the level β classical test (not pictured). Hybrid, however, manages to eke out some additional power. Both methods pale in comparison to the level α classical test however, which would achieve power > 0.95 throughout the whole plot (not pictured).For various values of σ_1 and σ_2 which we assume are known, we also tried re-running the experiments with $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_i \sim N(\mu_2, \sigma_2^2)$ when $i > 1$. The results were not appreciably different.

Two possible signals: Seeing as the hybrid and union bound approaches both reject based on the winning and running-up p-value, we ran a simulation for all pairs $\mu_1, \mu_2 \in \{-3, -2.9, \ldots, 2.9, 3\}$ with $\mu_1 > \mu_2$ and $\mu_1 > 0$. We forcibly set $\mu_i = -\infty$ for $i > 2$. For each setting we ran $N = 10000$ to get an empirical estimate of power for each method. Across the 1492 simulated settings, the median emprical power increase from hybrid was ≈ 0.003, the 90th percentile empirical power increase was ≈ 0.023 and the maximum empirical power increase was ≈ 0.042. As the results indicate, the power increases from hybrid were negligble for most settings. We also re-ran the same simulations but sampled $X_1 \sim N(\mu, \sigma_1^2)$ and $X_2 \sim N(\mu, \sigma_2^2)$ for various values of σ_1 and σ_2, which we assume are known. The results were not appreciably different, and, if anything, the difference in power was notably smaller for some values of σ_1 and σ_2.



(a) $\mu_2 = -2$ (b) $\mu_2 = 0$ (c) $\mu_2 = 2$

Figure 7: For $\mu_2 = 0$ (left), $\mu_2 = -0.5$ (middle), and $\mu_2 = -1$ (right) the empirical power over $N = 10^4$ trials of the hybrid inference approach versus the union bound approach for the needle in the haystack alternative. One standard error bands are also plotted. For the most part, they are so small that they are hardly visible.

A.5 Conditional inference on winners for exponentials

Recall the exponential distribution $X \sim \text{Exp}(\lambda_i)$ which has PDF

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0, \\ 0 & x \leq 0 \end{cases}.$$

Defining $T(x) = 1/x$, we see for $x > 0$ and $\lambda_2 \geq \lambda_1$, the ratio

$$f_{\lambda_2}(x)/f_{\lambda_1}(x) = \frac{\lambda_2}{\lambda_1} \exp\left(-\frac{\lambda_2 - \lambda_1}{T(x)}\right)$$

is monotone non-decreasing in $T(x)$. Thus this family has an MLR in $T(x)$, and the UMP test for $H_0 : \lambda \leq \lambda_0$ thus rejects when $T(X)$ is large, or correspondingly when X is small. In particular, noting that the CDF of X is given by

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x > 0, \\ 0 & x \leq 0 \end{cases}$$

it rejects according to the p-value $p^{\lambda^0} = 1 - e^{-\lambda_0 X}$ (see Appendix D.1 for details).

Now consider observing some independent data $X_i \sim \text{Exp}(\lambda_i)$. We know from Appendix A.2.2 that the selective p-value $p_{(1)}^{\lambda_0}/p_{(2)}^{\lambda_0}$ is monotone non-decreasing in λ_0. Using L'hopital's rule, we can compute that

$$\begin{aligned} \lim_{\lambda_0 \downarrow 0} \frac{p_{(1)}^{\lambda_0}}{p_{(2)}^{\lambda_0}} &= \lim_{\lambda_0 \downarrow 0} \frac{1 - e^{-\lambda_0 X_{(1)}}}{1 - e^{-\lambda_0 X_{(2)}}} \\ &= \lim_{\lambda_0 \downarrow 0} \frac{X_{(1)} e^{-\lambda_0 X_{(1)}}}{X_{(2)} e^{-\lambda_0 X_{(2)}}} \\ &= \frac{X_{(1)}}{X_{(2)}}, \end{aligned}$$

which suffices to imply our claims in the main text.

A.6 p-value Adjustment for Exponential Families

Suppose p is a p-value for the null H_0 that is selectively dominant given Z and we select p to use for inference according to the selection function

$$s(x, z) = \begin{cases} 1 & \text{if } x < q^+(z), \\ \frac{1}{N(z)} & \text{if } x \in [q^+(z), q(z)], \\ 0 & \text{otherwise}, \end{cases},$$

Then the adjusted p-value (4) is given by

$$p_{adj} = \frac{\int_0^p s(x,Z)dx}{\int_0^1 s(x,Z)dx} = \begin{cases} \frac{p}{q^+(Z)+\frac{1}{N(Z)}(q(Z)-q^+(Z))} & \text{if } p < q^+(Z), \\ \frac{q^+(Z)+\frac{1}{N(Z)}(p-q^+(Z))}{q^+(Z)+\frac{1}{N(Z)}(q(Z)-q^+(Z))} & \text{if } p \in [q^+(Z), q(Z)], \end{cases},$$

which can be re-written as

$$p_{adj} = \frac{p - (1-\frac{1}{N(Z)})(p-q^+(Z))_+}{q^+(Z) + \frac{1}{N(Z)}(q^+(Z) - q(Z))}.$$

This is sufficient to imply the claim from the main text.

A.7 Rank verification warm-up additional details

Example 10 (Rank verification in a simple case)**.** *Suppose that p is a selectively dominant p-value for testing the null H_0, but we only choose to test H_0 when $p < 1/2$. Applying our framework with the p-value p and selection function $s(x) = I_{x<1/2}$, Theorem 1 tells us that we control selective Type I error if we reject according to the adjusted p-value from (4) is $p_{adj} = 2p$:*

$$P_{H_0}\left(p_j \le \frac{\alpha}{2}|S=1\right) \le \alpha. \tag{33}$$

Now, consider data $X_1 \sim N(\mu_1, 1/\sqrt{2})$ and $X_2 \sim N(\mu_2, 1/\sqrt{2})$, the one-sided nulls $H_{0,j} : \mu_j \le \mu_{-j}$, and their corresponding selectively dominant p-values $p_j = 1 - \Phi(X_j - X_{-j})$ (selective dominance follows from Example 3). Denoting the winner $W = \operatorname{argmax}_{j=1,2} X_j$, it is now clear rejecting the data-dependent null $H_{0,W} : \mu_W \le \mu_{-W}$ when $p_W \le \alpha/2$ maintains Type I error control both conditionally on W and marginally. If $H_{0,j}$ is not true, then trivially $P(\text{falsely reject } H_{0,W} \mid W = j) = 0 \le \alpha$. For the case that $H_{0,j}$ is true, the event $W = j$ is identical to the event $p_j < 1/2$, and hence is the same event as selecting p_j for inference in (33). Therefore,

$$\begin{aligned} P(\textit{falsely reject } H_{0,W} \mid W = j) &= P\left(p_W \le \frac{\alpha}{2} \mid W = j\right) \\ &= P\left(p_j \le \frac{\alpha}{2} \mid W = j\right) \\ &\le \alpha, \end{aligned}$$

implying error control conditional on W. Marginal error control then follows from the law of total probability:

$$\begin{aligned} P(\textit{falsely reject } H_{0,W}) &= \sum_{j=1,2} P(\textit{falsely reject } H_{0,j} \mid W = j) P(W = j) \\ &\le \alpha \sum_{j=1,2} P(W = j) \\ &= \alpha. \end{aligned}$$

If $\mu_1 = \mu_2$ then the inequalities become equalities and our error control is tight.

A.8 Inference on Winners in Exponential Families

We can use the same tools from the previous sub-section to do inference on the winner in multiparameter exponential families. Again consider data X drawn from the exponential family (15) and let the winner W be the index $i \in \mathcal{S} \subseteq [p]$ corresponding to the largest sufficient statistic T_i, with ties broken randomly.

The p-value

$$p_i^{\theta_0}(T(X)) = G_i^{\theta_0,+}(T_i|T_{-i}) + U_{i,aux}(G_i^{\theta_0}(T_i|T_{-i}) - G_i^{\theta_0,+}(T_i|T_{-i})), \tag{34}$$

corresponds to the UMP test for $H_{0,i}^{\theta_0} : \theta_i \le \theta_0$, where, like before,

$$G_i^{\theta_0}(t_i|t_{-i}) = P_{\theta_0}(T_i \ge t|T_{-i} = t_{-i}) \qquad G_i^{\theta_0,+}(t) = \lim_{u \downarrow t}(T_i \ge t|T_{-i} = t_{-i}),$$

denote the conditional left-continuous survival function of T_i and its right-hand limit, and $U_{i,aux}$ are Unif([0, 1]) random variables that are independent from each other and X.

To come up with a test for the data dependent null $H_{0,W}^{\theta_0}$ (which we will then invert to get an LCB for the winning parameter), we need to apply a similar adjustment to our p-value as in Example 9. Rather than re-work through essentially identical arguments, we simply define the analogous quantities to those from the previous sub-section,

$$q_i^{\theta_0,+}(T_{-i}) = G_i^{\theta_0,+}(\max_{k\in\mathcal{S}-i} T_k|T_{-i}), \tag{35}$$

$$q_i^{\theta_0}(T_{-i}) = G_i^{\theta_0}(\max_{k\in\mathcal{S}-i} T_k|T_{-i}), \tag{36}$$

$$N_i(T_{-i}) = 1 + |\{k \neq i : T_k = \max_{\ell\in\mathcal{S}-i} T_\ell\}|, \tag{37}$$

and state the final result in Corollary 7. In ?? we use a modification of the arguments in Fithian et al. [2017] to show that our test $H_{0,W}^{\theta_0}$ is selectively UMPU, meaning it is UMPU amongst all tests that are are valid conditional on W. Hence, as stated in Corollary 7, its inversion is selectively uniformly most accurate unbaised (UMAU) in the analogous sense.

Corollary 7 (Conditional inference for multiparameter exponential families)**.** *Let X be drawn from the exponential family (15) and let W be the index $i \in \mathcal{S}$ corresponding to the largest sufficient statistic T_i (with ties broken randomly). If $p_i^{\theta_0}$ is the UMP p-value (47) for testing $H_{0,i}^{\theta_0} : \theta_i \leq \theta_0$ and $q_i^{\theta_0,+}$, $q_i^{\theta_0}$, and N_i are as in (35), (36), and (37), then*

$$\left\{\theta_0 : \frac{p_W^{\theta_0} - \left(1 - \frac{1}{N_i}\right)(p_W^{\theta_0} - q_W^{\theta_0,+})_+}{q_W^{\theta_0,+} + \frac{1}{N_i}(q_W^{\theta_0} - q_W^{\theta_0,+})} > \alpha\right\} \tag{38}$$

is an LCB for θ_W that holds with probability exactly $1-\alpha$ conditional on W (and therefore also marginally), and it is selectively UMAU conditional on W.

Briefly, we point out that selecting the largest sufficient statistic does not always correspond to selecting the "most promising" effect. Even though $T_i > T_j$, it might be the case that $p_j^{\theta_0}$ is smaller than $p_i^{\theta_0}$. Unlike in Section 3, we cannot simply have the winner correspond to the smallest p-value for two reasons. First, in this more general setting, the index resulting in the smallest p-value changes depending on θ_0 (see ?? for an example). Hence, it is not well-defined to determine the winner by looking at the smallest p-value (and we thus have to deal with ties). Second, even if we settled on a θ_0 to prioritize, because the p-values have intricate correlation strucutre, the selection event $p_i^{\theta_0} < \max_{j\neq i} p_j^{\theta_0}$ corresponds to a very complicated selection function that we would have to try and characterize on a case-by-case basis. Indeed, this discussion applies to the rank verification problem for exponential families we discussed previously as well.

A.9 Post selection inference for the LASSO

Coming soon.

A.10 Data carving for Gaussian file-drawer

We have two data samples $X_1 \sim N(\mu, 2)$ and $X_2 \sim N(\mu, 2)$ that are independent and want to test $H_0 : \mu \leq 0$. Suppose we only do inference because we observed that $X_1 > t$ for some threshold t. If we consider the p-value $p_{full} = 1 - \Phi((X_1 + X_2)/2)$, then our selection function is given by

$$\begin{aligned} s(x) &= P(X_1 > t | p_{full} = x) \\ &= P(X_1 > t | \frac{X_1 + X_2}{2} = \Phi^{-1}(1 - x)) \\ &= 1 - \Phi(t - \Phi^{-1}(1 - x)) \end{aligned}$$

where we have used that

$$\begin{bmatrix} X_1 \\ \frac{X_1+X_2}{2} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}\right)$$

so

$$X_1|\frac{X_1+X_2}{2} = y \sim N(y, 1)$$

Thus our corrected p-value is given by

$$p_{carve} = \frac{\int_0^{p_{full}} 1 - \Phi(t - \Phi^{-1}(1-x))dx}{\int_0^1 1 - \Phi(t - \Phi^{-1}(1-x))dx}$$

$$p_{carve} = \frac{\int_{\bar{X}}^{\infty} \phi(z)(1 - \Phi(t-z))dz}{\int_{-\infty}^{\infty} \phi(z)(1 - \Phi(t-z))dz} = \frac{\int_{\bar{X}}^{\infty} \phi(z)(1 - \Phi(t-z))dz}{1 - \Phi(t/\sqrt{2})}$$

We now show that p_{carve} is monotone non-decreasing in t. Letting Z and Y be independent standard normal random variables and fixing some constant a, the selective p-value is given by

$$p_{carve} = \frac{P(Z+Y > t, Z > a)}{P(Z+Y > t)} = P(Z > a|Z+Y > t)$$

for $a = \bar{X}$. Letting $W = Z + Y$ we can write $Z = \frac{1}{2}W + \epsilon$ where ϵ is independent of W. This gives us

$$p_{carve} = P(\frac{1}{2}W + \epsilon > a|W > t) = E[P(W > 2(a-\epsilon)|W > t, \epsilon)|W > t] = E[P(W > 2(a-\epsilon)|W > t, \epsilon)]$$

Then the fact that p_{carve} is monotone non-decreasing in t follows from the fact that $P(W > c|W > t)$ is monotone non-decreasing in t for every constant c:

$$P(W > c|W > t) = \begin{cases} \frac{P(W>c)}{P(W>t)} & \text{if } t \leq c, \\ 1 & \text{if } t > c. \end{cases}$$

B Inverting selective tests to get selective confidence regions

In parametric settings, we can invert tests of data dependent null hypotheses to get selective confidence regions. These confidence regions cover a random parameter. Example 11 gives the beginnings of an illustrative example.

Example 11 (Confidence region for the winner)**.** *Suppose the $X_i \sim P_{\theta_i}$ are n independent samples from an MLR family P_θ, and $p_j^{\theta_0} = p^{\theta_0}(X_j)$ are the UMP p-values for testing $H_{0,j} : \theta_j \leq \theta_0$. Define the parameter vector $\Theta = (\theta_1, \ldots, \theta_n)$ and let W be the index of the smallest p-value, so that $X_W = \max_{i \in [n]} X_i$ is a winner. We can get a $1-\alpha$ confidence region for the winning parameter θ_W that is valid conditional on W by inverting Example 7's test of the winning null $H_{0,W} : \theta_W \leq \theta_0$:*

$$\begin{aligned} P_\Theta\left(\theta_W \in \left\{\theta_0 : \frac{p_{(1)}^{\theta_0}}{p_{(2)}^{\theta_0}} > \alpha\right\} |W = j\right) &= P_\Theta\left(\theta_j \in \left\{\theta_0 : \frac{p_j^{\theta_0}}{\min_{i \neq j} p_i^{\theta_0}} > \alpha\right\} |W = j\right) \\ &= P_\Theta\left(\frac{p_j^{\theta_j}}{\min_{i \neq j} p_i^{\theta_j}} > \alpha |W = j\right) \\ &= 1 - \alpha, \end{aligned}$$

where we have equality because $p_j^{\theta_j}$ has an exact uniform distribution given $p_{-j}^{\theta_j}$ under P_Θ.

Since the region in Example 11 results from inverting a one sided test, we would hope that it is not just a confidence region, but a confidence lower bound. This is not obvious, however. While the p-values $p_j^{\theta_0}$ are known to be monotone in θ_0, it is less clear whether the selective p-values $p_j^{\theta_0} / \min_{i \neq j} p_i^{\theta_0}$ resulting from

Theorem 1 are. Indeed confirming that selective confidence regions are actually invervals or lower bounds can be tricky in selective problems, and proving that they are is sometimes done on a case by case basis [Benjamini et al., 2019, Lee et al., 2016] or not at all [Sengupta and Janson, 2024].

It turns out that selective p-values resulting from UMP or UMPU one-sided testing are essentially always monotone in the null parameter θ_0. Suppose we observe data (X, Z) where the conditional distribution $X|Z = z \sim P_{\theta,z}$ is parameterized by $\theta \in \mathbb{R}$ and has an MLR in $T(x)$. This setting is general enough to encompass both when we have independent samples from an MLR family as in Example 11 and also when we have data from an exponential family. Conditional on $Z = z$, the UMP p-value p^{θ_0} for testing $H_0 : \theta \leq \theta_0$ is a deterministic function of $T(X)$ and an independent uniform U_{aux} (see Appendix D):

$$p^{\theta_0} = m^{\theta_0}(T(X), U_{aux}, z). \tag{39}$$

We usually get the selective p-value $p^{\theta_0}_{sel}$ by applying Theorem 1's adjustment with the selection function $s^{\theta_0}(x, z)$. The resulting selective p-value $p^{\theta_0}_{sel}$ from Theorem 1 will be monotone non-decreasing in θ_0 whenever $s^{\theta_0}(x, z)$ depends on θ_0 only through m^{θ_0}, i.e., if

$$\text{there exists } \tilde{s} \text{ such that } s^{\theta_0}(x, z) = \tilde{s}(t, u, z) \text{ for all } t, u \text{ with } x = m^{\theta_0}(t, u, z).$$

Essentially, this condition requires that the selection event, once written in terms of the data, has no dependence on the specific parameter θ_0 that we are testing. Admittedly, this is the one case where writing our problem in terms of p-values instead of the original data adds an additional complication. But, since this condition holds for essentially all selective problems, a practioner seldom has to worry about this complication.

In **??**, we carefully show that the selective p-values from Example 11 satisfy the above condition. This allows us to give a more refined version of our above example.

Example 12 (Confidence lower bound and interval for the winner)**.** *Recall the setting of Example 11. Then*

$$\left\{\theta_0 : \frac{p^{\theta_0}_{(1)}}{p^{\theta_0}_{(2)}} > \alpha\right\} \tag{40}$$

and

$$\left\{\theta_0 : 1 - \frac{\alpha}{2} > \frac{p^{\theta_0}_{(1)}}{p^{\theta_0}_{(2)}} > \frac{\alpha}{2}\right\} \tag{41}$$

are a CLB and CI that cover the winning parameter θ_W with probability exactly $1 - \alpha$ conditional on W. In the CI, we can replace the left bound with $1 - \alpha_1$ and the right bound with α_2 for any α_1, α_2 such that $\alpha_1 + \alpha_2 = \alpha$. The choice $\alpha_1 = \alpha_2 = \alpha/2$ ensures we miscover because the CI lies above and below the parameter with equal probability.

The CI and CLB from Example 12 immediately generalize the conditional CI and CLB for the winner that one gets by following Fithian et al. [2017] to any MLR family. These confidence regions are normally only presented in the Gaussian case, and we show explicitly in **??** that our CI and CLB match those in the existing literature in the Gaussian case.

Something about p-values makes it easy to see that it only depends on the gap...

C Selecting Multiple p-Values for Inference

In this appendix, we generalize our selective dominance framework to allow us to select multiple p-values for inference. We only do this generilization

Suppose we have n ndependent and selectively dominant p-values for the nulls $H_{0,i}$. We imagine conditioning on some collection of them, which, without loss of generality, we can take to be $Z = (p_{k+1}, \ldots, p_n)$. Note that, due to independence, conditioning on Z does not change the distribution of p_j. Thus, the p_j remain p-values under the nulls $H_{0,j}$ that are independent. Now, for $1 \leq j \leq k$, we consider k binary selection random variables $S_j \in \{0, 1\}$, where $S_j = 1$ when p_j is selected. The relationship between p_j, Z, and S_j is governed by the selection function

$$s_j(x, z) = p(S_j = 1 | p_j = x, Z = z)$$

Supposing that U is a uniform random variable that has a uniform distribution given Z, we can imagine selecting U using the same selection functions. I.e., we can imagine that a binary selection variable $S_j \in 0, 1$ whose joint distribution with U is governed by

$$P(S_j = 1 | U = x, Z = z) = s(x, z)$$

Then the machinery from Section 2 tells us that

$$p_{adj,j} = F_{U|Z,S_j=1}(p_j) = \frac{\int_0^{p_j} s_j(x, Z)dx}{\int_0^1 s_j(x, Z)dx}$$

is p-value (it stochastically dominates the uniform distribution under the null) conditional on Z and selection.

We will further assume that the selection happens independently, i.e.,

$$P(S_1 = 1, \dots, S_k = 1 \mid p_1, \dots, p_k, Z) = \prod_{j=1}^{k} P(S_j = 1 \mid p_j, Z)$$

By taking expectations conditional on Z with respect to both sides, we find that the S_j are independent given Z:

$$\begin{aligned}
P(S_1 = 1, \dots, S_k = 1 \mid Z) &= E[P(S_1 = 1, \dots, S_k = 1 | p_1, \dots, p_k, Z) \mid Z] \\
&= E\left[\prod_{j=1}^{k} P(S_j = 1 | p_j, Z) \,\middle|\, Z \right] \\
&= \prod_{j=1}^{k} E[P(S_j = 1 | p_j, Z)] \\
&= \prod_{j=1}^{k} P(S_j = 1 | Z)
\end{aligned}$$

where we have used that the p_j are conditionally independent given Z to move the expectation inside the product. Finally, conditional on Z and all the selections $S_j = 1$, the adjusted p-values $p_{adj,j}$ are all independent of one another.

$$P(p_{adj,1} \in A_1, \dots, p_{adj,k} \in A_k \mid Z, S_1 = 1, \dots, S_k = 1) = \prod_{j=1}^{k} P(p_{adj,j} \in A_j \mid Z, S_j = 1)$$

This can be confirmed via Bayes rule:

$$\begin{aligned}
&P(p_{adj,1} \in A_1, \dots, p_{adj,k} \in A_k \mid Z, S_1 = 1, \dots, S_k = 1) \\
&= \frac{p(p_{adj,1} \in A_1, \dots, p_{adj,k} \in A_k, |Z) P(S_1 = 1, \dots, S_k = 1 \mid Z, p_{adj,1} \in A_1, \dots, p_{adj,k} \in A_k)}{P(S_1 = 1, \dots, S_k = 1 | Z)} \\
&= \frac{a + b}{c}
\end{aligned}$$

D Selective Dominance and One-Sided Testing

In this appendix, we establish the selective dominance property for UMP p-values in MLR familes and UMPU p-values in exponential families. We also show that in these cases, the adjusted p-value from Theorem 1 is monotone in the parameter, under suitable conditions. Throughout the appendix, we draw from the discussion and proof strategy in Appendix B.1 of Lei and Fithian [2018].

D.1 MLR Families

We consider a parametric family P_θ parameterized by a real parameter $\theta \in R$ such that each P_θ has density $p_\theta(x)$ with respect to some carrier measure μ. We will suppose that these densities share support (i.e., for any two θ and θ' we have $p_\theta(x) > 0 \iff p_{\theta'}(x) > 0$). Further, we suppose that for any $\theta < \theta'$, the likelihood ratio $p_{\theta'}(x)/p_\theta(x)$ is a monotone non-decreasing function of some real-valued function $T(x)$ on this support (i.e, for any $x_1 \leq x_2$ with $p_\theta(x_1), p_\theta(x_2) > 0$, we have $p_{\theta'}(x_1)/p_\theta(x_1) \leq p_{\theta'}(x_2)/p_\theta(x_2)$).

Recall that for a testing problem, the critical function $\phi(x)$ (see [Lehmann et al., 1986, Section 3.1]) tells us the probability of rejecting the null having observed data x, so $\phi(X) = P(\text{reject } H_0|X)$. From Theorem 3.4.1 of Lehmann et al. [1986] we know the test governed by the critical function

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > C \\ \gamma & \text{if } T(x) = C \\ 0 & \text{otherwise} \end{cases} \tag{42}$$

is UMP for testing $H_0 : \theta \leq \theta_0$ against the alternatives $H_a : \theta > \theta_0$ so long C and γ satisfy

$$P_{\theta_0}(T(X) > C) + \gamma P_{\theta_0}(T(X) = C) = \alpha. \tag{43}$$

Denote the left-continuous survival function of $T(X)$ and its righthand limit under P_{θ_0} as

$$G(t) = P_{\theta_0}(T(X) \geq t) \qquad G^+(t) = \lim_{u \downarrow t} G(u).$$

Since $G(t)$ is a monotone non-increasing function, we can also define its generalized inverse

$$G^{-1}(z) = \inf\{t : G(t) \leq z\},$$

Lemma D.1 gives a natural way to set C and γ in Equation (42) to get an UMP test.

Lemma D.1 (An UMP test)**.** *Adopting the convention that $0/0 = 0$, taking $C = G^{-1}(\alpha)$ and $\gamma = (\alpha - G^+(C))/(G(C) - G^+(C))$ in Equation (42) gives an UMP test.*

Proof. At continuity points t of $G(\cdot)$, we have $G(G^{-1}(t)) = t$ and also $P(T(X) = t) = 0$. Thus, if $C = G^{-1}(\alpha)$ is a continuity point of $G(\cdot)$, then $G^+(C) = P_{\theta_0}(T(X) > C) = P_{\theta_0}(T(X) \geq C) = G(C) = \alpha$ and $\gamma = 0$, so the constraint (43) is immediately satisfied.

If $C = G^{-1}(\alpha)$ is not a continuity point of $G(\cdot)$, then $G(C) - G^+(C) > 0$ and the constraint (43) is also immediately satisfied. To ensure that we still have a valid test, however, we need $\gamma \in [0, 1]$. This is true so long as $\alpha \in [G^+(C), G(C)]$. We know that $G(t) \leq \alpha$ for any $t > C$, so $G^+(C) \leq \alpha$. If $G(C) < \alpha$, then we could find some $t^- < C$ such that $G(C) < \alpha$ by left-continuity, but this would contradict that $C = G^{-1}(\alpha)$, finishing the proof. □

Letting $U_{aux} \sim \text{Unif}([0, 1])$ be auxiliary randomness that is independent of X, a simple way to instantiate the test from Lemma D.1 is to reject whenever $T(X) > C$ or when $T(X) = C$ and $U_{aux} \leq \gamma$. Lemma D.2 explains how this is the same as rejecting when the p-value, termed as a fuzzy p-value in Geyer and Meeden [2005],

$$p = G^+(T(X)) + U_{aux}(G(T(X)) - G^+(T(X))) \tag{44}$$

is at most α.

Lemma D.2 (Fuzzy p-value is UMP)**.** *Rejecting $H_0 : \theta \leq \theta_0$ when the fuzzy p-value (44) is at most α instantiates the test from Lemma D.1, and is therefore UMP.*

Proof. We rewrite

$$p = (1 - U_{aux})G^+(T(X)) + U_{aux}G(T(X))$$

and consider four cases.

- If $t < G^{-1}(\alpha)$ then we can find some $t^+ > t$ such that $G(t^+) > \alpha$. Thus $G(t) \geq G^+(t) > \alpha$. So $p > \alpha$ whenever $T(X) < G^{-1}(\alpha)$

- If $t = G^{-1}(\alpha)$ and $G^{-1}(\alpha)$ is a continuity point of $G(\cdot)$, then $G^+(t) = G(t) = \alpha$. Thus, in this case $p \leq \alpha$ whenever $T(X) = G^{-1}(\alpha)$ and $U_{aux} \leq \frac{\alpha - G^+(G^{-1}(\alpha))}{G(G^{-1}(\alpha)) - G^+(G^{-1}(\alpha))} = \infty$.
- If $t = G^{-1}(\alpha)$ and $G^{-1}(\alpha)$ is not a continuity point of $G(\cdot)$, then we must have that $G(t) - G^+(t) > 0$. Also by right continuity we have $G(t) \geq \alpha$ and by how $G^{-1}(\cdot)$ is defined we have $G^+(t) \leq \alpha$. In this case $p \leq \alpha$ also whenever $T(X) = G^{-1}(\alpha)$ and $U_{aux} \leq \frac{\alpha - G^+(G^{-1}(\alpha))}{G(G^{-1}(\alpha)) - G^+(G^{-1}(\alpha))}$.
- If $t > G^{-1}(\alpha)$ then $G^+(t) \leq G(t) \leq \alpha$. So $p \leq \alpha$ whenever $T(X) > G^{-1}(\alpha)$.

This implies the following set equalities:

$$\begin{aligned}
\{p \leq \alpha\} &= \{T(X) > G^{-1}(\alpha)\} \cup \left\{T(X) = G^{-1}(\alpha), U_{aux} \leq \frac{\alpha - G^+(G^{-1}(\alpha))}{G(G^{-1}(\alpha)) - G^+(G^{-1}(\alpha))}\right\} \\
&= \{T(X) > C\} \cup \{T(X) = C, U_{aux} \leq \gamma\}
\end{aligned}$$

□

Lemma D.3 shows that $p \sim \text{Unif}([0,1])$ under P_{θ_0}, which will be useful for us later.

Lemma D.3 (Fuzzy p-value is uniform at null boundary)**.** *Under P_{θ_0}, the p-value* (44) *has a Unif*$([0,1])$ *distribution.*

Proof. Using the set equality from Lemma D.2 but replacing α with $z \in (0,1)$, we find

$$\begin{aligned}
&P_{\theta_0}(G^+(T(X)) + U_{aux}(G(T(X)) - G^+(T(X))) \leq z) \\
&= P_{\theta_0}(T(X) > G^{-1}(z)) + P_{\theta_0}\left(T(X) = G^{-1}(z), U \leq \frac{z - G^+(G^{-1}(z))}{G(G^{-1}(z)) - G^+(G^{-1}(z))}\right) \\
&= P_{\theta_0}(T(X) > G^{-1}(z)) + P_{\theta_0}(T(X) = G^{-1}(z)) P_{\theta_0}\left(U \leq \frac{z - G^+(G^{-1}(z))}{G(G^{-1}(z)) - G^+(G^{-1}(z))}\right) \\
&= G^+(G^{-1}(z)) + (G(G^{-1}(z)) - G^+(G^{-1}(z))) \cdot \frac{z - G^+(G^{-1}(z))}{G(G^{-1}(z)) - G^+(G^{-1}(z))} \\
&= z.
\end{aligned}$$

□

Now we can show that p is a selectively dominant p-value for testing the null $H_0 : \theta \leq \theta_0$. In what follows, we consider some fixed $\theta \leq \theta_0$ and prove some facts that allow us to relate the distribution of $T(X)$ under P_{θ_0} to its distribution under P_θ.

Lemma D.4 (Distribution of $T(X)$)**.** *Let $g_\theta(T(x))$ be a non-increasing function that equals the likelihood ratio $p_\theta(x)/p_{\theta_0}(x)$ on the support and*

$$\nu(A) = \int I(T(x) \in A) p_{\theta_0}(x) \mu(dx)$$

be the meausre of $T(X)$ under $X \sim P_{\theta_0}$. Then

$$P_\theta(T(X) \in A) = \int_A g_\theta(t) \nu(dt)$$

Proof. We know that

$$P_\theta(T(X) \in A) = \int I(T(x) \in A) p_{\theta_0}(x) g_\theta(T(x)) \mu(dx),$$

so we need to show that

$$\int I(T(x) \in A) p_{\theta_0}(x) g_\theta(T(x)) \mu(dx) = \int_A g_\theta(t) \nu(dt) \tag{45}$$

If $g_\theta(T(x)) = I(T(x) \in A')$ happens to be an indicator then (45) holds. Therefore, we can apply the standard machine (see the discussion after Equation 42 in Lei and Fithian [2018]) to show that (45) holds for all non-negative functions $g_\theta(\cdot)$. □

Lemma D.5 (Distribution of $(T(X), U_{aux})$)**.** *If ω denotes the product measure of ν and Lesbesgue measure λ on $[0,1]$, i.e., the distribution of $(T(X), U_{aux})$ under P_{θ_0}, then*

$$P_\theta((T(X), U) \in B) = \int_B g_\theta(t) \omega(dt, du) \tag{46}$$

Proof. We will first argue that (46) holds for any B which is a product set $A_1 \times A_2$. We can futher reduce to the case that $g_\theta(T(x)) = I(T(x) \in A_1')$ is an indicator. Then we see using our previous lemma that

$$\begin{aligned}
P_\theta((T(X), U_{aux}) \in A_1 \times A_2) &= P_\theta(T(X) \in A_1) P(U_{aux} \in A_2) \\
&= \int_{A_1} g_\theta(t) \nu(dt) \cdot \int_{A_2} \lambda(du) \\
&= \int_{A_1 \cap A_1'} \nu(dt) \cdot \int_{A_2} \lambda(du) \\
&= \int_{A_1 \cap A_1' \times A_2} \omega(dt, du) \\
&= \int_{A_1 \times A_2} g_\theta(t) \omega(dt, du)
\end{aligned}$$

To handle the case of general $g_\theta(\cdot)$ we can again simply apply the standard machine.

The full result then follows from an application of the $\pi - \lambda$ theorem: the set of B for which (46) holds is a λ-system, and (46) holds for every set in the π system of all product sets $B = A_1 \times A_2$. □

Note that our p-value p is a deterministic function of $T(X)$ and U_{aux}:

$$p = m(T(X), U_{aux}) \qquad m(t, u) = G^+(t) + u(G(t) - G^+(t)).$$

As such, we sometimes write our selection function as a function of $T(X)$ and U_{aux}:

$$s(t, u) = s(m(t, u)).$$

We use this abuse of notation in our next lemma, which characterizes the conditional distribution of $T(X)$ given selection.

Lemma D.6 (Distribution of $(T(X), U_{aux})$ given selection)**.** *For any selection function $s(x)$ under which p has a positive probability of selection under P_θ,*

$$P_\theta((T(X), U_{aux}) \in B | S = 1) = \frac{\int_B g_\theta(t) s(t, u) \omega(dt, du)}{\int g_\theta(t) s(t, u) \omega(dt, du)}$$

Proof. First note that

$$P_\theta((T(X), U_{aux}) \in B | S = 1) = \frac{P_\theta((T(X), U_{aux}) \in B, S = 1)}{P_\theta(S = 1)}.$$

Thus it suffices to show for any set B that

$$P_\theta((T(X), U_{aux}) \in B, S = 1) = \int_B g_\theta(t) s(t, u) \omega(dt, du).$$

By the definition of conditional expecation

$$\begin{aligned} P_\theta((T(X), U_{aux}) \in B, S = 1) &= E_\theta[E_\theta[I(S = 1) \mid T(X), U_{aux}] I((T(X), U_{aux}) \in B)] \\ &= E_\theta[s(T(X), U_{aux}) I((T(X), U_{aux}) \in B)] \end{aligned}$$

If $s(t, u) = I_{(t,u)\in B}$ is an indicator function, then the result is implied by our previous lemma. We again get the result for general selection functions $s(t, u)$ by applying the standard machine. $\square$

With these lemmas under our belt, we can show Proposition D.1, the main result of this sub-section. Since $p \sim_{P_{\theta_0}} \text{Unif}([0, 1])$ by Lemma D.3, this proposition is sufficient to imply selective dominance.

Proposition D.1. *For any selection function $s(x)$ for which p has positive probability of selection under both θ and θ_0,*

$$P_\theta(p \le z|S = 1) \le P_{\theta_0}(p \le z|S = 1).$$

Proof. Fix $z \in (0, 1)$. If z is such that $P_\theta(p \le z|S = 1) = 0$ then the desired inequality is trivial. To handle the non-trival case, we note three facts from the proof of Lemma D.2:

- If $(t, u) \in m^{-1}([0, z])$ then $t \ge G^{-1}(z)$,
- If $(t, u) \in m^{-1}((z, 1])$ then $t \le G^{-1}(z)$,
- The sets $m^{-1}([0, z])$ and $m^{-1}((z, 1])$ are disjoint.

Thus,

$$\begin{aligned} \frac{1}{P_\theta(p \le z|S = 1)} &= \frac{\int_{m^{-1}([0,1])} g_\theta(t)s(t, u)\omega(dt, du)}{\int_{m^{-1}([0,z])} g_\theta(t)s(t, u)\omega(dt, du)} \\ &= \frac{\int_{m^{-1}([0,z])} g_\theta(t)s(t, u)\omega(dt, du) + \int_{m^{-1}((z,1])} g_\theta(t)s(t, u)\omega(dt, du)}{\int_{m^{-1}([0,z])} g_\theta(t)s(t, u)\omega(dt, du)} \\ &= 1 + \frac{\int_{m^{-1}((z,1])} g_\theta(t)s(t, u)\omega(dt, du)}{\int_{m^{-1}([0,z])} g_\theta(t)s(t, u)\omega(dt, du)} \\ &\ge 1 + \frac{g_\theta(G^{-1}(z)) \int_{m^{-1}((z,1])} s(t, u)\omega(dt, du)}{g_\theta(G^{-1}(z)) \int_{m^{-1}([0,z])} s(t, u)\omega(dt, du)} \\ &= 1 + \frac{\int_{m^{-1}((z,1])} s(t, u)\omega(dt, du)}{\int_{m^{-1}([0,z])} s(t, u)\omega(dt, du)} \\ &= \frac{1}{P_{\theta_0}(p \le z|S = 1)}, \end{aligned}$$

where to finish we have noted that $g_{\theta_0}(t) = 1$ almost eveywhere in the measure ω. $\square$

D.2 Exponential Families

Suppose we observe data $X \in \mathbb{R}^m$ from an exponential family P_θ parameterized by $\theta \in \mathbb{R}^n$ i.e., under P_θ the data X has density

$$g_\theta(x) = \exp(\theta_1 T_1(x) + \cdots + \theta_n T_n(x) - \psi(\theta))g(x)$$

with respect to some carrier measure μ. We consider the problem of testing $H_0 : \theta_i \le \theta_{0,i}$.

The UMPU test for $H_0 : \theta_i \le \theta_{0,i}$ is valid conditional on $T_{-i}(X)$. More specifically, Theorem 4.4.1 of Lehmann et al. [1986] tells us that any test of the form

$$\phi(t_i, t_{-i}) = \begin{cases} 1 & \text{if } t_i > C_0(t_{-i}) \\ \gamma(t_{-i}) & \text{if } t_i = C_0(t_{-i}) \\ 0 & \text{otherwise} \end{cases}$$

where the functions $\gamma(\cdot)$ and $C_0(\cdot)$ satisfy

$$E_{\theta_{0,i}}[\phi(T_i(X), t_{-i})|T_{-i}(X) = t_{-i}] = \alpha$$

is UMPU for testing $H_0 : \theta_i \leq \theta_{0,i}$. Lemma 2.7.2 of Lehmann et al. [1986] tells us that the conditional distribuition of $T_i(X)$ given $T_{-i}(X) = t_{-i}$ admits a density

$$g_{\theta_i, t_{-i}}(t_i) = \exp(\theta_i t_i - \tilde{\psi}(\theta_i))$$

with respect to some base measure $\mu_{t_{-i}}$. This density has an MLR in t_i (to be specific, we are imagining observing $T_i(X)$ from its conditional distribution $T_{-i}(X)$, and the map $T(\cdot)$ from the previous sub-section is actually the identity). Hence, a concrete UMPU test is to just run our UMP test from the previous section using the conditional distribution given $T_{-i}(X) = t_{-i}$. In particular, our work from the previous section implies that it is UMPU to reject when the p-value

$$p = G^+(T_i(X)|T_{-i}(X)) + U_{aux}(G(T_i(X)|T_{-i}(X)) - G_i^+(T_i(X)|T_{-i}(X))), \tag{47}$$

where U_{aux} is an uniform random variable independent of the data and

$$G(t_i|t_{-i}) = P_{\theta_0}(T_i(X) \geq t|T_{-i}(X) = t_{-i}) \qquad G^+(t_i|t_{-i}) = \lim_{u \downarrow t_i} G(u|t_{-i}),$$

is at most α. Our work from the previous section also implies that this p-value is selectively dominant given $T_{-i}(X)$.

D.3 Montonicty of Selective MLR p-Values

In this sub-section, we consider data (X, Z) where the conditional distribution $X|Z = z \sim P_{\theta,z}$ is parameterized by $\theta \in \mathbb{R}$ and has an MLR in $T(x)$. Without loss of generality however, we can fix $Z = z$ and pretend we just observe X. Letting

$$G^{\theta_0}(t) = P_{\theta_0}(T(X) \geq t) \qquad G^{\theta_0,+}(t) = \lim_{u \downarrow t} G^{\theta_0}(u)$$

we let

$$p^{\theta_0} = G^+(T(X)) + U_{aux}(G(T(X)) - G^+(T(X)))$$

be the UMP p-value for testing $H_0 : \theta \leq \theta^0$. Again, U_{aux} is an uniform random variable independent of the data. Let

$$m^{\theta_0}(t, u) = G^{\theta_0,+}(t) + u(G^{\theta_0}(t) - G^{\theta_0,+}(t))$$

be the map such that $m^{\theta_0}(T(X), U) = p^{\theta_0}$.

We consider a class of selection functions $s^\theta(x)$ under which p^θ has a positive probability of being selected under P_θ:

$$\int_0^1 s^{\theta_0}(z)dx > 0.$$

Using them, we define the selective p-values

$$p_{sel}^\theta = \frac{\int_0^{p^\theta} s^\theta(x)dx}{\int_0^1 s^\theta(x)dx}.$$

Our goal is to show that p_{sel}^θ is monotone non-decreasing in θ.

Fixing some $\theta \leq \theta_0$, define $r_\theta(x, z) = s_\theta(x, z)/s_{\theta_0}(x, z)$ to be the selection ratio, and assume that it is monotone ..... in x.

Lemma D.7.

Let p^{θ_0} be the UMP p-value for testing $H_0 : \theta \leq \theta_0$. If $s^{\theta_0}(x, z)$ is a selection function for which p^{θ_0} has a positive probability of being selected under P_{θ_0}, then the selective p-value

$$p_{sel}^{\theta_0} = \frac{\int_0^{p^{\theta_0}} s^{\theta_0}(x, Z)dx}{\int_0^1 s^{\theta_0}(x, Z)dx}$$

is monotone non-decreasing in θ_0 so long as, for $\theta \leq \theta'$, $s^{\theta}(x, z) = 0 \implies s^{\theta'}(x, z) = 0$ and the **selection ratio** $s_{\theta'}(x, z)/s_{\theta'}(x, z)$ is monotone non-decreasing in x.

Proposition D.2. *Let $s_\theta(x)$ be a selection function such that for $\theta \leq \theta_0$, $s_{\theta_0}(x) = 0 \implies s_\theta(x) = 0$ and the ratio $s_{\theta_0}(x)/s_\theta(x)$ is non-decreasing in x. Then, if p^θ_{adj} is the adjustment from $F_{U|S_\theta=1}(p^\theta)$ then p^θ_{adj} is monotone non-decreasing increasing in θ.*

Proof. Fix a t and u and let $A_{r,s} = \{(t, u) : t > r \text{ or } t = r \text{ and } u \leq s\}$. Then it suffices to show that

$$\frac{\int_{A_{t,u}} g_\theta(t)s_\theta(t, u)\omega(dt, du)}{\int g_\theta(t)s_\theta(t, u)\omega(dt, du)}$$

is monotone non-decreasing. Equivalently that

$$\frac{\int_{A_{t,u}} g_\theta(t)s_\theta(t, u)\omega(dt, du)}{\int g_\theta(t)s_\theta(t, u)\omega(dt, du)} \leq \frac{\int_{A_{r,s}} s_{\theta_0}(t, u)\omega(dt, du)}{\int s_{\theta_0}(t, u)\omega(dt, du)}$$

If $\int_{A_{r,s}} g_\theta(t)s_\theta(t, u)\omega(dt, du) = 0$ then the inequality is trivial. So it suffices to consider the other case and show that

$$\frac{\int g_\theta(t)s_\theta(t, u)\omega(dt, du)}{\int_{A_{r,s}} g_\theta(t)s_\theta(t, u)\omega(dt, du)} \geq \frac{\int s_{\theta_0}(t, u)\omega(dt, du)}{\int_{A_{r,s}} s_{\theta_0}(t, u)\omega(dt, du)}$$

$$1 + \frac{\int_{A^c_{r,s}} g_\theta(t)s_\theta(t, u)\omega(dt, du)}{\int_{A_{r,s}} g_\theta(t)s_\theta(t, u)\omega(dt, du)} \geq 1 + \frac{\int_{A^c_{r,s}} s_{\theta_0}(t, u)\omega(dt, du)}{\int_{A_{r,s}} s_{\theta_0}(t, u)\omega(dt, du)}$$

Definining $0/0 = 0$, our assumptions allow us to do the following:

$$\begin{aligned}
\frac{\int_{A^c_{r,s}} g_\theta(t)s_\theta(t, u)\omega(dt, du)}{\int_{A_{r,s}} g_\theta(t)s_\theta(t, u)\omega(dt, du)} &= \frac{\int_{A^c_{r,s}} g_\theta(t)s_{\theta_0}(t, u)\left(\frac{s_\theta(t,u)}{s_{\theta_0}(t,u)}\right)\omega(dt, du)}{\int_{A_{r,s}} g_\theta(t)s_{\theta_0}(t, u)\left(\frac{s_\theta(t,u)}{s_{\theta_0}(t,u)}\right)\omega(dt, du)} \\
&\geq \frac{g_\theta(G_\theta^{-1}(m(r, s)))\left(\frac{s_\theta(r,s)}{s_{\theta_0}(r,s)}\right)\int_{A^c_{r,s}} s_{\theta_0}(t, u)\omega(dt, du)}{g_\theta(G_\theta^{-1}(m(r, s)))\left(\frac{s_\theta(r,s)}{s_{\theta_0}(r,s)}\right)\int_{A_{r,s}} s_{\theta_0}(t, u)\omega(dt, du)} \\
&= \frac{\int_{A^c_{r,s}} s_{\theta_0}(t, u)\omega(dt, du)}{\int_{A_{r,s}} s_{\theta_0}(t, u)\omega(dt, du)}
\end{aligned}$$

□

E Proofs and Derivations

E.1 Proof of Theorem 1

Recall we are considering a selection function such that the probability that U is selected $P(S' = 1|Z = z) = \int_0^1 s(x, z)dx$ is positive. The CDF of U given selection is continuous because it cannot have any point masses:

$$P(U = x|Z = z, S' = 1) = \frac{P(U = x, S' = 1|Z = z)}{P(S' = 1|Z = z)} \leq \frac{P(U = x|Z = z)}{P(S' = 1|Z = z)} = 0.$$

Therefore, defining

$$F^{-1}_{U|Z=z,S'=1}(t) = \inf\{x : F_{U|Z=z,S'=1}(x) > t\}$$

continuity implies that $F_{U|Z=z,S=1}(F^{-1}_{U|Z=z,S=1}(t)) = t$ and $F_{U|Z=z,S=1}(x) \leq t \iff x \leq F^{-1}_{U|Z=z,S=1}(t)$. Then

$$\begin{aligned}
P_{H_0}(F_{U|Z=z,S'=1}(p) \leq t|Z = z, S = 1) &= P_{H_0}(p \leq F^{-1}_{U|Z=z,S'=1}(t)|Z = z, S = 1) \\
&\leq P(U \leq F^{-1}_{U|Z=z,S'=1}(t)|Z = z, S' = 1) \\
&= P(F_{U|Z=z,S'=1}(U) \leq t|Z = z, S' = 1)
\end{aligned}$$

where we have used that $p|Z = z, S = 1 \succeq_{H_0} U|Z = z, S' = 1$ to get the middle inequality. Finally

$$\begin{aligned}
P(F_{U|Z=z,S'=1}(U) \leq t|Z = z, S' = 1) &= P(U \leq F^{-1}_{U|S'=1}(t)|Z = z, S' = 1) \\
&= F_{U|Z=z,S'=1}(F^{-1}_{U|Z=z,S'=1}(t)) = t
\end{aligned}$$

so $F_{U|Z=z,S'=1}(U)|Z = z, S' = 1 \sim \text{Unif}([0, 1])$, which finishes the proof.

Further, if for some distribution under the null, p has an exact uniform distribution given $Z = z$, then the distributions $U|Z = z, S = 1'$ and $p|Z = z, S = 1$ are indentical, so the fact that $F_{U|Z=z,S'=1}(U)|Z = z, S' = 1 \sim \text{Unif}([0, 1])$ also implies that $F_{U|Z=z,S'=1}(p)|Z = z, S = 1 \sim \text{Unif}([0, 1])$, and therefore (5) holds with equality in this case.

E.2 Proof of Theorem 2

Let f_z be the density of $p|Z = z$ under a distribution in the null H_0. We start by showing that, if f_z is non-decreasing, then $p|Z = z, S = 1$ dominates $U|Z = z, S' = 1$. Fixing a selection function $s(x, z)$, it suffices to show that for any $t \in [0, 1]$.

$$\begin{aligned}
&P(p \leq t|Z = z, S = 1) \leq P(U \leq t|Z = z, S = 1) \\
\iff &\frac{\int_0^t s(x, z)f_z(x)dx}{\int_0^1 s(x, z)f_z(x)dx} \leq \frac{\int_0^t s(x, z)dx}{\int_0^1 s(x, z)dx}
\end{aligned}$$

If $P(p \leq t|Z = z, S = 1)$ is zero then this trivially holds. Otherwise $P(p \leq t|Z = z, S = 1) = \int_0^t s(x, z)f_z(x)dx > 0$ and we see that,

$$\begin{aligned}
\frac{\int_0^1 s(x, z)f_z(x)dx}{\int_0^t s(x, z)f_z(x)dx} &= 1 + \frac{\int_t^1 s(x, z)f_z(x)dx}{\int_0^t s(x, z)f_z(x)dx} \\
&\geq 1 + \frac{f_z(t)\int_t^1 s(x, z)dx}{f_z(t)\int_0^t s(x, z)dx} \\
&= 1 + \frac{\int_t^1 s(x, z)dx}{\int_0^t s(x, z)dx} \\
&= \frac{\int_0^1 s(x, z)dx}{\int_0^t s(x, z)dx},
\end{aligned}$$

which is sufficient to imply the claim.

Now assuming that $f_{z'}$ is continuous and not non-decreasing for some z', we can show that p is not selectively dominant. In general, it suffices for there to be two points $y_1 < y_2$ such that $f_{z'}$ is strictly larger in a neighborhood around y_1 than in a neighborhood around y_2, where these neighborhoods are disjoint. In particular for $\epsilon > 0$ let $N_\epsilon(y) = (y - \epsilon, y + \epsilon)$ be a ball around y. Then we need there to be y_1, y_2, $\epsilon > 0$, and some $\eta > 0$ such that, for all $w_1 \in N_\epsilon(y_1)$ and $w_2 \in N_\epsilon(y_2)$, $w_1 < w_2$ but $f_{z'}(w_2) + \eta < f_{z'}(w_1)$. If f is continuous and not non-decreasing, then this must be true. First define $B_{high} = \inf\{f_{z'}(w_1) : w_1 \in N_\epsilon(y_1)\}$ and $B_{low} = \sup\{f_{z'}(w_2) : w_2 \in N_\epsilon(y_2)\}$ so $B_{high} > B_{low}$. Then consider the selection function

$$s(x, z') = \begin{cases} 1 & \text{if } x \in N_\epsilon(y_1) \cup N_\epsilon(y_2) \text{ and } z = z' \\ 0 & \text{otherwise} \end{cases}$$

and let t be a value such that $t > w_1$ for all $w_1 \in N_\epsilon(y_1)$ and $t < w_2$ for all $w_2 \in N_\epsilon(y_2)$. Trivially,

$$P(U \leq t | Z = z', S' = 1) = \frac{1}{2}.$$

But the fact that

$$\begin{aligned} \frac{1}{P(p \leq t | Z = z', S = 1)} &= \frac{\int_{y_1-\epsilon}^{y_1+\epsilon} f_{z'}(x)dx + \int_{y_2-\epsilon}^{y_2+\epsilon} f_{z'}(x)dx}{\int_{y_1-\epsilon}^{y_1+\epsilon} f_{z'}(x)dx} \\ &= 1 + \frac{\int_{y_2-\epsilon}^{y_2+\epsilon} f_{z'}(x)dx}{\int_{y_1-\epsilon}^{y_1+\epsilon} f_{z'}(x)dx} \\ &\leq 1 + \frac{2\epsilon B_{low}}{2\epsilon B_{high}} \\ &< 2 \end{aligned}$$

implies that $P(p \leq t | Z = z', S = 1) > \frac{1}{2}$, which means that p is not selectively dominant.

E.3 Proof of ??

It suffices to argue that closing Sidak's rejects $H_{0,(k)}$ if and only if $p_{(j)} \leq \alpha_{n-j+1}$ for every $1 \leq j \leq k$. For a subset $I \subseteq [p]$, Sidak's global null test rejects $H_{0,I}$ when the smallest p-value in I is at most $\alpha_{|I|}$.

Necessity: For $1 \leq j \leq k$, let I_{n-j+1} be the size $n - j + 1$ subset that excludes the $j - 1$ smallest p-values for (when $j = 1$ then $I = [p]$). This subset includes the index of the kth smallest p-value, so we must reject $H_{0,I}$ to reject $H_{0,(k)}$. It rejects exactly when $p_{(j)} \leq \alpha_{n-j+1}$, so our conditions are neccesary.

Sufficiency: Consider a size $n-j+1$ size subset I that has the kth smallest p-value. The smallest p-value in this subset is the ℓth smallest p-value for some $\ell \leq k$ and also $\ell \leq j$. We reject $H_{I,0}$ because

$$p_{(\ell)} \leq \alpha_{n-\ell+1} \leq \alpha_{n-j+1} = \alpha_I.$$

E.4 Proof of Corollary 2

It suffices to argue that closing our conditional global null testing procedure rejects $H_{0,(k)}$ if and only if $p_{(j)} \leq \alpha p_{(j+1)}$ for every $1 \leq j \leq k$. We will define $p_{(n+1)} = 1$ and, for subsets $I \subseteq [p]$ of size one, we define the conditional procedure to reject the global null $H_{I,0}$ when the lone p-value is at most α. For subsets I of size strictly more than one, the procedure rejects the global null $H_{I,0}$ when the smallest p-value in I is at most alpha times the second smallest p-value in I.

Necessity: For $1 \leq j \leq k$, let I_{n-j+1} be the size $n - j + 1$ subset that excludes the $j - 1$ smallest p-values for (when $j = 1$ then $I = [p]$). This subset includes the index of the kth smallest p-value, so we must reject $H_{0,I}$ to reject $H_{0,(k)}$. It rejects exactly when $p_{(j)} \leq \alpha p_{(j+1)}$, so our conditions are neccesary.

Sufficiency: Consider a subset I that contains the index of the kth smallest p-value. If it is size one, then we reject because

$$p_{(k)} \leq \alpha p_{(k+1)} \leq \alpha$$

If it is of size at least two, suppose the smallest p-value in the set is then the ℓth smallest p-value for $\ell \leq k$ and the econd smallest p-value in I is mth smallest p-value for some $m > \ell$. We will reject $H_{0,I}$ because

$$p_{(\ell)} \leq \alpha p_{(\ell+1)} \leq \alpha p_{(m)}.$$

E.5 Proof of ??

Suppose we have n independent and selectively dominant p-values p_i for the null hypotheses $H_{0,i}$. We restrict our attention to $j \in \mathcal{J}$ for which p_j has positive probability of being the smallest. Suppose we use p_j to test $H_{0,j}$ only when we observe that p_j is strictly larger than β_n but still the smallest of all the p-values. We can apply Section 2's framework with $p = p_j$, $Z = p_{-j}$ and the selection function $s(x, z) = I_{\beta_n < p_j < \min_{i \neq j} p_i}$. It is straightforward to see that our adjusted p-value p_{adj} is $(p_j - \beta_n)/(\min_{i \neq j} p_i - \beta_n)$, and Theorem 1 therefore tells us that

$$P_{H_{0,j}} \left(\frac{p_j - \beta_n}{\min_{i \neq j} p_i - \beta_n} \leq \frac{\alpha - \beta}{1 - \beta} \,\middle|\, p_{-j}, S = 1 \right) \leq \frac{\alpha - \beta}{1 - \beta}.$$

Re-arranging things we get

$$P_{H_{0,j}} \left(p_j \leq \frac{\alpha - \beta}{1 - \beta} \min_{i \neq j} p_i + \left(1 - \frac{\alpha - \beta}{1 - \beta}\right) \beta_n \,\middle|\, p_{-j}, S = 1 \right) \leq \frac{\alpha - \beta}{1 - \beta}. \tag{48}$$

Letting W be the index of the smallest p-value, we can now prove the claim that rejecting $H_{0,W}$ when

$$p_{(1)} \leq \frac{\alpha - \beta}{1 - \alpha} p_{(2)} + \left(1 - \frac{\alpha - \beta}{1 - \alpha}\right) \beta_n$$

controls Type I error at level α. Let $\widetilde{W}$ be the index of the smallest p-value if all the p-values are strictly larger than β_n. If some p-value is at most β_n, then force $\widetilde{W} = 0$. Let G_P be the event that a p-value corresponding to a true null is at most β_n (note the event G depends on the data generating process P). We know from Sidak's procedure (**??**) that $P(G_P) \leq \beta$. Three facts are immediate:

$$G_P \subseteq \{\widetilde{W} = 0\} \implies \{\widetilde{W} > 0\} \subseteq G_P^c \implies P(\widetilde{W} > 0) \leq 1 - \beta$$

$$P(\text{falsely reject } H_{0,W}, G_P, \widetilde{W} = 0) \leq P(G_P) \leq \beta,$$

$$P(\text{falsely reject } H_{0,W}, G_P^c, \widetilde{W} = 0) = 0.$$

If $H_{0,j}$ is true, the event $\widetilde{W} = j$ is the same event as selecting p_j for inference in (48), so

$$\begin{aligned}
P(\text{falsely reject } H_{0,W} | \widetilde{W} = j) &= P\left(p_{(1)} \leq \frac{\alpha - \beta}{1 - \alpha} p_{(2)} + \left(1 - \frac{\alpha - \beta}{1 - \alpha}\right) \beta_n | \widetilde{W} = j\right) \\
&= P\left(p_j \leq \frac{\alpha - \beta}{1 - \alpha} \min_{i \neq j} p_i + \left(1 - \frac{\alpha - \beta}{1 - \alpha}\right) \beta_n | \widetilde{W} = j\right) \\
&\leq \frac{\alpha - \beta}{1 - \alpha},
\end{aligned}$$

and if $H_{0,j}$ is not true then trivially $P(\text{falsely reject } H_{0,W} | \widetilde{W} = j) = 0 \leq \frac{\alpha - \beta}{1 - \alpha}$. Our result then follows from law of total probability:

$$\begin{aligned}
&P(\text{falsely reject} H_{0,W}) \\
&= P(\text{falsely reject } H_{0,W}, \widetilde{W} = 0, G_P) + P(\text{falsely reject } H_{0,W}, \widetilde{W} = 0, G_P^c) \\
&\quad + \sum_{j \in \mathcal{J}} P(\text{falsely reject } H_{0,W} | \widetilde{W} = j) P(\widetilde{W} = j) \\
&\leq \beta + \frac{\alpha - \beta}{1 - \beta} \sum_{j \in \mathcal{J}} P(\widetilde{W} = j) \\
&= \beta + \frac{\alpha - \beta}{1 - \beta} P(\widetilde{W} > 0) \\
&\leq \alpha.
\end{aligned}$$

E.6 Proof of Corollary 4

It suffices to argue that closing our hybrid global null testing procedure rejects $H_{0,(k)}$ if and only if

$$p_{(j)} \leq \frac{\alpha-\beta}{1-\beta} p_{(j+1)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{n-j+1}$$

for every $1 \leq j \leq k$. We will define $p_{(n+1)} = \alpha$ so that the right-hand side of the above equals α when $j = n$. Correspondingly, for subsets $I \subseteq [p]$ of size one, we define the hybrid procedure to reject the global null $H_{I,0}$ when the lone p-value is at most α. For subsets I of size strictly more than one, supposing that the smallest p-value in I is the ℓth smallest p-value and the second smallest p-value in I is the mth smallest p-value, the hybrid procedure rejects the global null $H_{I,0}$ when

$$p_{(\ell)} \leq \frac{\alpha-\beta}{1-\beta} p_{(m)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{|I|}.$$

Necessity: For $1 \leq j \leq k$, let I_{n-j+1} be the size $n-j+1$ subset that excludes the $j-1$ smallest p-values (when $j = 1$ then $I = [p]$). This subset includes the index of the kth smallest p-value, so we must reject $H_{0,I}$ to reject $H_{0,(k)}$. It rejects exactly when

$$p_{(j)} \leq \frac{\alpha-\beta}{1-\beta} p_{(j+1)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{n-j+1}$$

so our conditions are neccesary.

Sufficiency: Consider a subset I that contains the index of the kth smallest p-value. If it is size-one, then we reject because

$$p_{(k)} \leq \frac{\alpha-\beta}{1-\beta} p_{(k+1)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{n-k+1} \leq \frac{\alpha-\beta}{1-\beta} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta = \alpha.$$

Now suppose that I is size $n-j+1$. Its smallest p-value is the ℓth smallest p-value for some $\ell \leq k$ and $\ell \leq j$, and its second smallest p-value is the mth smallest p-value for some $m > \ell$. We reject because

$$\begin{aligned} p_{(\ell)} &\leq \frac{\alpha-\beta}{1-\beta} p_{(\ell+1)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{n-\ell+1} \\ &\leq \frac{\alpha-\beta}{1-\beta} p_{(m)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{n-j+1} \\ &= \frac{\alpha-\beta}{1-\beta} p_{(m)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{|I|}. \end{aligned}$$