ENSAI

ATPA Track
Academic Year 2024–2025

Methodological Project Report

Selective Inference for Hierarchical Clustering

Student: Jasmin Neveu

Supervisor: Javier González-Delgado

Analyzed article:
Lucy L. Gao, Jacob Bien, and Daniela Witten,
Selective Inference for Hierarchical Clustering,
Journal of the American Statistical Association,
119(545), 332–342, 2024.

Submission date: February 13, 2026

Contents

1 Introduction

Introduction text...

2 Classical versus selective testing

2.1 Hypothesis testing and *p-values*

To do:

- Number (almost) all the equations.
- Use $\mathcal{SU}$ everywhere instead of SU (search+replace).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(E, \mathcal{E})$ a topological space and $\mathcal{T}$ the σ-algebra generated by $\mathcal{E}$. A *random variable* is a measurable function $X : (\Omega, \mathcal{F}) \to (E, \mathcal{T})$. The *(probability) distribution* of X is the mapping $P : \mathcal{T} \to [0, 1]$ such that $P(O) = (\mathbb{P} \circ X^{-1})(O)$ for all $O \in \mathcal{T}$. We say that P is *supported* on E and denote by $\mathcal{M}_1(E)$ the set of all probability distributions supported on $\mathcal{E}$. From now on, we will set $E = \mathbb{R}$ and simply write $\mathcal{M}(\mathbb{R}) = \mathcal{M}$.

Definition 1 (Hypothesis, Definition 1.1 in [4])**.** *A hypothesis $\mathcal{H}_0$ is a set of probability distributions in $\mathcal{M}$. A hypothesis is simple if it is a singleton, such as $\{P\}$ or $\{Q\}$, and composite otherwise. The complementary set $\mathcal{M} \setminus \mathcal{H}_0$ is called the alternative hypothesis of $\mathcal{H}_0$.*

Definition 2 (Test)**.** *A test for $\mathcal{H}_0$ is a binary partition of the sample space, defined by a mapping~~: function.~~Mapping = application in french.*

$$\begin{array}{rcl} \pi_{\mathcal{H}_0} : \Omega & \to & \{0, 1\} \\ \omega & \mapsto & \pi_{\mathcal{H}_0}(\omega). \end{array} \tag{1}$$

For $\omega \in \Omega$, we say that the test rejects $\mathcal{H}_0$ based on ω if $\pi_{\mathcal{H}_0}(\omega) = 1$, and does not reject $\mathcal{H}_0$ based on ω if $\pi_{\mathcal{H}_0}(\omega) = 0$.

In what follows~~From now~~, we will adopt the more standard~~use the following~~ notation:

$$\mathbb{P}\left(\text{Reject } \mathcal{H}_0\right) = \mathbb{P}\left(\{\omega \in \Omega \,:\, \pi_{\mathcal{H}_0}(\omega) = 1\}\right),$$

making the~~The~~ dependence on π ~~is~~ implicit.

Definition 3 (Type I error)**.** *We say that a test controls the type I error at level α if*

$$\mathbb{P}(\text{Reject } \mathcal{H}_0) \leq \alpha, \quad \textit{for } \alpha \in (0, 1).$$

We say that it controls the type I error exactly at level α if

$$\mathbb{P}(\text{Reject } \mathcal{H}_0) = \alpha, \quad \textit{for } \alpha \in (0, 1).$$

Definition 4 (Stochastic dominance)**.** *Let X and Y be real-valued random variables. We say that Y stochastically dominates X, and write*

$$X \preceq_{\text{st}} Y,$$

if

$$\mathbb{P}(X \leq u) \geq \mathbb{P}(Y \leq u), \quad \forall\, u \in \mathbb{R}.$$

Definition 5 (Super-uniform random variable)**.** *Let X be a real-valued random variable. We say that X is super-uniform on [0,1], and write $X \sim \mathcal{SU}(0,1)$, if X stochastically dominates a uniform random variable on $[0,1]$, that is, if*

$$\mathbb{P}(X \leq u) \leq u, \quad \forall\, u \in [0,1].$$

Proposition 1. *Let X and Y be real-valued random variables, with cumulative distribution functions F_X and F_Y, respectively. If $X \preceq_{\mathrm{st}} Y$, then $F_X(Y) \sim SU(0,1)$.*

Remark 1. *If $X = Y$, then $F_X(X)$ is super-uniform on $[0,1]$. Moreover, if X has a continuous distribution function F_X, then*

$$F_X(X) \sim \mathcal{U}(0,1).$$

Definition 6 (p-value, Definition 1.1 in [4])**.** *Let $\mathcal{H}_0$ be a hypothesis. A p-value for $\mathcal{H}_0$ is a super-uniform random variable under $\mathcal{H}_0$.*

We often build tests using p-values as $\pi = \mathbb{1}\{p \leq \alpha\}$.

Proposition 2. *Let $\mathcal{H}_0$ be a hypothesis and~~,~~ p a p-value for $\mathcal{H}_0$.~~ and $\mathcal{R}$ the rejection rule defined by $\pi = \mathbb{1}\{p \leq \alpha\}, \quad \alpha \in (0,1)$.~~ Then, the test $\pi = \mathbb{1}\{p \leq \alpha\}$~~$\mathcal{R}$~~ controls the type I error at level α, for $\alpha \in (0,1)$.*

Definition 7 (Test statistic)**.** *A test statistic is a measurable function $T : \Omega \to \mathbb{R}$.*

From now on, we will consider the case of unilateral tests. In this setting, to define a test based on the information given by a real-valued random variable X, p-values are often built in the form:

$$p(x) = \mathbb{P}_{H_0}(T(X) \geq T(x)) \tag{2}$$

where T is a test statistic and x a realization of X.

The p-value (2) follows a uniform distribution~~law~~ The french term *loi* doesn't translate well to law in this context... under $\mathcal{H}_0$. However, we have defined p-values as being super-uniform random variables under the null~~$\mathcal{SU}$~~, which corresponds to~~is~~ a more general family of distributions. To adapt the classical form (2)~~of the p-value~~ to that setting, we introduce~~adopt the construction of~~ the following proposition.

Proposition 3. *Let T and T' be two test statistics. Let X be a random variable and x a realization of X. Define*

$$p(x) = \mathbb{P}_{\mathcal{H}_0}\big(T'(X) \geq T(x)\big).$$

If

$$T'(X) \preceq_{\mathrm{st}} T(X) \quad \textit{under } \mathcal{H}_0,$$

then p is a p-value for $\mathcal{H}_0$.

Remark 2. *If $T' = T$ and the cumulative distribution function F_T of $T(X)$ is continuous under $\mathcal{H}_0$, then*

$$p \overset{\mathcal{H}_0}{\sim} \mathcal{U}(0,1).$$

In the classical setting, the null hypothesis is independent of the data. However, in many practical applications $\mathcal{H}_0$ is chosen *after seeing the data*. In this framework~~setting~~ to avoid repeating *setting*, the classical testing approaches built for type I error control are unsuitable. Instead, statistical guarantees need to be provided via *selective inference*. In particular, the theory of statistical testing of data-driven null hypotheses is known as *selective testing* [2]. The next section provides some examples that motivate this framework.

2.2 Examples motivating selective testing

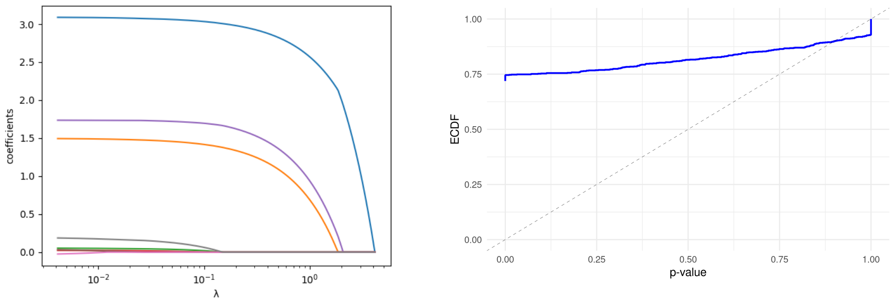
2.2.1 Lasso

To determine whether an explanatory variable helps explain a response variable through a linear regression model, a common practice is to test whether its associated coefficient is significantly different from zero. In the context of classical linear regression, this falls within the framework of non-selective inference, because all coeficients are fixed *a priori*. In contrast, Lasso regression uses an ℓ_1-penalty to perform variable selection cite the Lasso paper: https://academic.oup.com/jrsssb/article/58/1/267/7027929. As only the coefficients selected by the Lasso can be tested, the null hypothesis depends on the outcome of the Lasso regression, and is therefore data-driven. This corresponds to a setting of selective inference, where the classical control of type I error fails.

To illustrate the unsuitability of non-selective inference, consider a centered Gaussian vector $X = (X_1, \ldots, X_8)$ and set the model~~. The response is modeled as~~

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1), \quad \beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top.$$

We generated $n = 100$ independent realizations of (X, Y) and implemented the Lasso algorithm, obtaining the regularization path presented in Figure 1(a). Then, we tested whether a randomly chosen coefficient selected by Lasso regression equals 0. Modify this sentence to say that we are only testing coefficients that are trully equal to zero. After repeating this pipeline $M = 2000$ times, we obtained the empirical p-value distribution depicted in Figure 1(b). We clearly see that the p-values are not super-uniform, motivating the use of~~therefore~~ a selective testing~~inference~~ approach ~~should be used~~ instead of naive testing after selection.



(a) (b)

Figure 1: (a) Regularization path of the coefficients obtained using the Lasso algorithm. This is for a single realization of (X, Y), right? Say it! Some coefficients are zero, and three converge toward the theoretical values (3,1.5,2). (b) Empirical cumulative distribution function (ECDF) of the p-values obtained from tesing wheter a randomly chosen coefficient is null after selecting the coefficient by a Lasso regression. The ECDF was computed from $M = 2000$ simulations.

Make the axes labels, ticks and names bigger (they are hard to read).

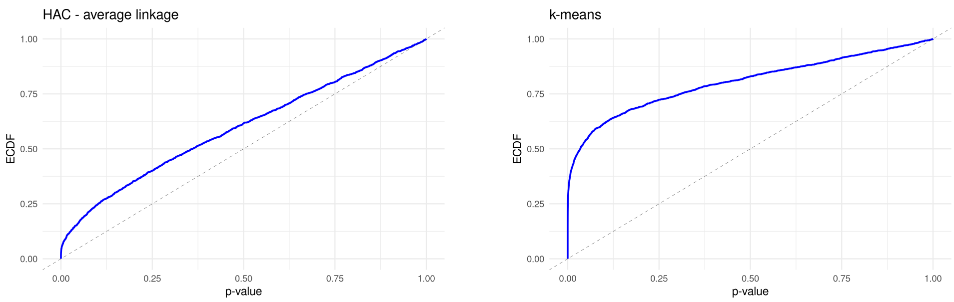
2.2.2 Publication bias

Most published studies gain publication due to their demonstrated significance. Only studies presentating major results are published by scientifics. Thus, there is a selection process. The previous sentences are not very clear. I propose a reformulation below:

In many areas of research, scientists tend to test for signficance only when the associated effect is found to be substantial. In other words, testing is performed after a selection process that filters out small effects. In that setting, controlling type I error at a fixed level α yields inflated false positive rates when considering the ensemble of all published studies. If $Y_i \sim \mathcal{N}(\mu_i, 1)$ represents the effect size in~~of~~ a scientific study, whose significance is tested only if ~~and only those with~~ $|Y_i| > 1$~~, are published, denoted~~ $\cancel{\hat{I} = \{i : |Y_i| > 1\}}$~~, then~~ a naive level α test $H_{0,i} : \mu_i = 0$ ~~for~~ $\cancel{i \in \hat{I}}$ is invalid. Indeed, Fithian [2] demonstrates that the false positive rate among ~~all studies~~true nulls reaches approximately 0.16, far exceeding the nominal 0.05 level. Valid inference requires thresholding $|Y_i|$ at 2.41 rather than 1.96, the 0.95 quantile of the standard normal, imposing a more stringent criterion.

2.2.3 Clustering

Another remarkable example of selective inference appears when evaluating the performance of clustering algorithms by testing for the equality of cluster means. To illustrate the need of using appropriate tests in the context of selective inference, we simulated $n = 1000$ samples of a one-dimensional standard~~centered~~ Gaussian random variable. Each sample ~~that~~ was classified into $K = 3$ groups using hierarchical agglomerative clustering and k-means. Then~~For each sample~~, the equality of cluster means was tested using a classical t-test, for two randomly selected clusters. If the test controlled the type I error in this setting, the resulting p-value would be super-uniformly distributed under the null hypothesis. However, the 'clustering + post-selection testing' procedure leads to a deviation from super-uniformity, as shown in Figure 2.



(a) (b)

Figure 2: Empirical cumulative distribution functions of the p-values obtained after testing the~~from tests of~~ equality of cluster means for~~between clusters after~~ (a) hierarchical agglomerative clustering (HAC) and (b) k-means algorithms. The distribution functions were computed using $M = 2000$ simulations from a univariate centered normal distribution.

Make labels, ticks and text in the plot bigger.

2.3 Addressing selective testing

Since the type I error is no longer controlled, alternative approaches are required. Accordingly, we present three methods as described by Yoav Benjamini [1].

2.3.1 Simultaneous Inference

This approach controls the family-wise error rate across all hypotheses:

$$\mathbb{P}(\text{At least 1 false positive among all hypotheses}) \leq \alpha.$$

This strategy proves highly conservative, ensuring that for every possible set of hypotheses, the probability of at least one false positive remains below α.

2.3.2 Sample splitting

This approach consists in splitting the dataset into a training set X and a test set Y. The training set X is used to choose which hypotheses to test, denoted $H_0(X)$. Then, the tests are performed on the test set Y.

Although relatively simple to implement, this method raises several issues. First, statistical guarantees on the tests hold only if X and Y are independent, which is rarely the case in practice. In addition, comparing cluster means on the observations in Y requires assigning each test point to one of the clusters obtained from the clustering performed on X, a step that compromises validity. As discussed in [3], this strategy does not yield valid post-clustering inference in general.

2.3.3 Conditional Inference

This approach constitutes the most extensively studied framework for post-clustering inference. It controls the false positive rate conditional on hypothesis selection:

$$\mathbb{P}(\text{Reject } H_0(X) \mid H_0 \text{ selected}) \leq \alpha.$$

In the remainder of the article, we employ this method to develop a statistical procedure for testing equality of cluster means following a clustering algorithm.

Why new page?

3 Post clustering inference

As we are focusing on one article, I think it is better if we start right away with their approach. Here we can add a sentence saying that this article was the first to propose a feasible solution to post-clustering inference, and that we focus on the setting and approach introduced by the autorhs. Later in the discussion, we will mention its limitations and speak about other articles.

3.1 Gao *et al.*'s approach

Use bold for matrix notation but not for matrix coefficients: $\mathbf{X} = (X_{ij})_{ij}$. For vectors in $\mathbb{R}^p$ don't use bold either, write $\bar{\mu}_G$ or $\bar{X}_G$, to be consistent with Gao *et al.*'s notation.

Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the design matrix. A cluster is an element of a partition of the samples. We note C a clustering algorithm and $C_1 \in C(\mathbf{X})$ a cluster obtained by the algorithm C on $\mathbf{X}$.

We note $\boldsymbol{\mu} = (\boldsymbol{\mu}_{ij})_{ij}$ such as $\boldsymbol{\mu}_{ij} = \mathbb{E}[\mathbf{X}_{ij}]$, with $\mathbf{X}_{ij}$ the element in row i and column j of the matrix $\mathbf{X}$. For a subset G of $\{1, \dots, n\}$, we note $\overline{\boldsymbol{\mu}}_G = \frac{1}{|G|}\sum\limits_{i\in G} \mu_i \in \mathbb{R}^p$ and $\overline{\mathbf{X}}_G = \frac{1}{|G|}\sum\limits_{i\in G} X_i \in \mathbb{R}^p$.

Here, introduce the vector ν and the compact notation $\nu^T\mathbf{X}$, $\nu^T\boldsymbol{\mu}$.

Definition 8 (Null Hypothesis)**.**

$$H_0^{\{C_1,C_2\}} : \overline{\boldsymbol{\mu}}_{C_1} = \overline{\boldsymbol{\mu}}_{C_2} \tag{H0}$$

The previous equation is not really a definition. Say it rather in the text, also helping the reader understand the flow: 'The goal is to test for the equality of cluster means, that is, testing the following null hypothesis:' and then add the equation (with a number).

Definition 9 (Type I selective error for clustering)**.** *We say that a test controls the type I selective error for clustering at leval α if*

$$\mathbb{P}_{H_0^{\{C_1,C_2\}}}(\text{Reject } H_0^{\{C_1,C_2\}} \mid C_1, C_2 \in C(\mathbf{X})) \leq \alpha, \quad \alpha \in (0,1).$$

We say that it controls exactly the type I selective error for clustering at level α if

$$\mathbb{P}_{H_0^{\{C_1,C_2\}}}(\text{Reject } H_0^{\{C_1,C_2\}} \mid C_1, C_2 \in C(\mathbf{X})) = \alpha, \quad \alpha \in (0,1).$$

I think we should introduce the model here. The idea should be something like: To define a p-value that controls the selective type I error, model assumptions need to be imposed on $\mathbf{X}$. The authors in [3] adopt the following matrix normal model:

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbb{I}_n, \sigma^2\mathbb{I}_p), \tag{3}$$

which means that the lines of $\mathbf{X}$ are independent p-dimensional random vectors distributed as $\mathcal{N}_p(\mu_i, \sigma^2 \mathbb{I}_p)$, for every line $i \in \{1, \ldots, n\}$. Note that this model imposes an indepedence assumption between the features (columns) of $\mathbf{X}$. Then, we mention the p-value as you did below:

To run a proper test, we need to control this error at level α. Ideally~~In the ideal~~ so french..., we would like to define a p-value as follows~~following~~:

$$p_{ideal}(x) = \mathbb{P}_{H_0^{\{C_1, C_2\}}}(T(\mathbf{X}) \geq T(x) \mid C_1, C_2 \in C(\mathbf{X})),$$

with T being a test statistic.

Now we mention the choice of T. Say that in [3] they set $T(\mathbf{X}) = ||\nu^T \mathbf{X}||_2$ because we know its distribution under the null, which is $\mathcal{N}_p(\mathbf{0}_p, ||\nu||_2^2 \sigma^2 \mathbb{I}_p)$, explaining why. Then, continue:

However, p_{ideal} cannot be evaluated in practice as it depends on parameters that are unknown [3]. Whenever you make a statement that you don't really justify, cite the source. To~~Thus, to~~ address this issue, the authors in [3] propose~~we need~~ to add technical events to the conditioning set, considering the following quantity:~~and consider:~~ $\cancel{p(x) = \mathbb{P}_{H_0^{\{C_1, C_2\}}}(T(\mathbf{X}) \geq T(x) \mid C_1, C_2 \in C(\mathbf{X}), E[\mathbf{X}])}$ Here add the p-value defined by Gao, that is, Equation 8 in [3]. as a p-value for (H0).

Now we haved defined the model, the hypothesis to test and the candidate p-value. The goal now is to prove that the candidate p-value (i) can be computed under an analitically tractable form and (ii) controls the selective type I error for clustering. Now, we say that to prove all that we first need a technical lemma that we state now (the Lemma about the independence that we proved). Add the lemma here and its proof to the appendix.

A Proofs

A.1 Proofs of Section 2

Proof of Proposition 1. Let G_X denote the generalized inverse (quantile function) of F_X, defined for $u \in [0, 1]$ as

$$G_X(u) = \inf\{x \in \mathbb{R} : F_X(x) \geq u\}.$$

By definition of the generalized inverse,

$$\{F_X(Y) \leq u\} = \{Y < G_X(u)\}.$$

Therefore,

$$\mathbb{P}\big(F_X(Y) \leq u\big) = \mathbb{P}\big(Y < G_X(u)\big) = F_Y\big(G_X(u)^-\big),$$

where $G_X(u)^-$ denotes the left limit at $G_X(u)$.

Since $X \preceq_{\text{st}} Y$, we have $F_Y \leq F_X$ pointwise, and thus

$$F_Y\big(G_X(u)^-\big) \leq F_X\big(G_X(u)^-\big).$$

By the defining property of the generalized inverse,

$$F_X\big(G_X(u)^-\big) \leq u.$$

Combining these inequalities yields

$$\mathbb{P}\big(F_X(Y) \leq u\big) \leq u, \quad \forall\, u \in [0, 1],$$

which proves that $F_X(Y)$ is super-uniform.

□

Proof of Remark 1. By the proposition 1, we have that $F_X(X) \sim SU(0, 1)$. If F_X is continuous, then $F_X(G_X(u)) = u$ for all $u \in [0, 1]$, and hence

$$\mathbb{P}\big(F_X(X) \leq u\big) = F_X\big(G_X(u)\big) = u, \quad \forall\, u \in [0, 1],$$

which concludes the proof.

□

Proof of Proposition 2. By definition of the rejection rule,

$$\mathbb{P}_{\mathcal{H}_0}(\text{Reject } \mathcal{H}_0) = \mathbb{P}_{\mathcal{H}_0}(p \leq \alpha).$$

Since p is a p-value for $\mathcal{H}_0$, it is super-uniform under $\mathcal{H}_0$, hence

$$\mathbb{P}_{\mathcal{H}_0}(p \leq \alpha) \leq \alpha.$$

This establishes control of the type I error at level α. □

Proof of Proposition 3. Let $F_{T'(X)}$ denote the distribution function of $T'(X)$ under $\mathcal{H}_0$. By Proposition 1, the stochastic dominance $T'(X) \preceq_{\mathrm{st}} T(X)$ implies that

$$F_{T'(X)}\big(T(X)\big) \sim \mathrm{SU}(0,1).$$

By definition,

$$p(x) = \mathbb{P}_{\mathcal{H}_0}\big(T'(X) \geq T(x)\big) = 1 - F_{T'(X)}\big(T(x)\big).$$

Let $u \in [0,1]$. Then

$$\begin{aligned}\mathbb{P}_{\mathcal{H}_0}\big(p(X) \leq u\big) &= \mathbb{P}_{\mathcal{H}_0}\big(1 - F_{T'(X)}(T(X)) \leq u\big)\\ &= \mathbb{P}_{\mathcal{H}_0}\big(F_{T'(X)}(T(X)) \geq 1-u\big)\\ &= 1 - \mathbb{P}_{\mathcal{H}_0}\big(F_{T'(X)}(T(X)) \leq 1-u\big).\end{aligned}$$

Since $F_{T'(X)}(T(X))$ is super-uniform,

$$\mathbb{P}_{\mathcal{H}_0}\big(F_{T'(X)}(T(X)) \leq 1-u\big) \leq 1-u,$$

and therefore

$$\mathbb{P}_{\mathcal{H}_0}\big(p(X) \leq u\big) \leq u.$$

Thus, p is super-uniform under $\mathcal{H}_0$, and hence a p-value. □

Proof of Remark 2. When $T' = T$

$$p(X) = \mathbb{P}_{\mathcal{H}_0}\big(T(X') \geq T(X)\big) = 1 - F_T\big(T(X)\big),$$

By Remark 1, if F_T is continuous,

$$F_T\big(T(X)\big) \sim \mathcal{U}(0,1).$$

Consequently, for any $u \in [0,1]$,

$$\begin{aligned}\mathbb{P}_{\mathcal{H}_0}\big(p(X) \leq u\big) &= \mathbb{P}_{\mathcal{H}_0}\big(1 - F_T(T(X)) \leq u\big)\\ &= \mathbb{P}_{\mathcal{H}_0}\big(F_T(T(X)) \geq 1-u\big)\\ &= 1 - \mathbb{P}_{\mathcal{H}_0}\big(F_T(T(X)) \leq 1-u\big)\\ &= 1-(1-u)\\ &= u.\end{aligned}$$

Thus p is uniformly distributed on $[0,1]$ under $\mathcal{H}_0$. □

A.2 Proofs of Section 3

References

[1] Y. Benjamini. Selective Inference: The Silent Killer of Replicability. *Harvard Data Science Review*, 2(4), dec 16 2020. https://hdsr.mitpress.mit.edu/pub/l39rpgyc.

[2] W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection, 2017. arXiv:1410.2597.

[3] L. L. Gao, J. Bien, and D. Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 119(545):332–342, 2024.

[4] A. Ramdas and R. Wang. Hypothesis testing with e-values. *arXiv*, Oct. 2024.