The Multiple-Comparison Trap and the Raven's Paradox – Perils of Using Null Hypothesis Testing

Song S. Qian[*] Thomas F. Cuffney[†]

October 5, 2017

Summary

1. The hidden multiple comparison trap (leading to a much higher risk of committing a type I error than the declared significance level) and the Raven's paradox (rendering the rejection of the null hypothesis meaningless) are two frequently encountered problems when using statistical null hypothesis testing.

2. These problems are often hidden when hypothesis testing is used as part of a more elaborate model, especially models based on computer-intensive methods.

3. Hidden multiple comparison trap can be exposed using computer simulation to estimate the real type I error probability.

4. The Ravens paradox can be avoided by clearly stating the null and alternative hypotheses using scientific terms to justify that the rejection of the null is equivalent to proving the alternative is true.

5. The hidden multiple comparison trap suggests that a null hypothesis testing based on computer-intensive method should be used with caution.

6. To avoid the Raven's paradox, we should focus on providing evidence to support the proposed hypothesis or model, rather than seeking evidence against the irrelevant null hypothesis.

7. With these two problems and many others related to null hypothesis test, statistical hypothesis testing should be used only as a component of the body of evidence, perhaps, as the devil's advocate.

[*] Department of Environmental Sciences, The University of Toledo, Toledo, OH, USA, Email: song.qian@utoledo.edu

[†] USGS South Atlantic Water Science Center, Raleigh, NC, USA

Keywords: deviance reduction methods, null hypothesis testing, statistical inference, TITAN, threshold

1 Introduction

Null hypothesis testing (NHT) is widely used in ecological studies. Open any issue of this journal and we can find examples where NHT is used as the main analytic tool for ecological inference. Anytime we see a p-value, an NHT is implied. The use of NHT is, however, controversial. The controversy is often centered around the use and interpretation of the key statistic, the p-value [Ellison et al., 2014]. Because we often use the p-value as the only criterion for accepting or rejecting the finding of a study, misuse and misinterpretation of the p-value often lead to misleading conclusions. In this paper, we discuss two issues that are rarely mentioned – the hidden multiple-comparison trap and the Raven's paradox. These issues are of particular concern because of the increased use of NHT in conjunction with computer-intensive methods. We introduce these issues with two examples from published papers and discuss how to detect and avoid them.

1.1 Beyond the Misuse of p-values

The misinterpretation and misuse of the p-value are the most commonly seen statistical errors in applications of NHT. With increased awareness, we are now better at discovering and avoiding such errors. However, many statistical issues of NHT are not quite obvious. We discuss two related issues.

1.1.1 The Multiple Comparison Trap

The issue of multiple comparisons concerns the increased likelihood of incorrectly rejecting the null hypothesis when multiple tests are carried out. In a one-way ANOVA problem, the simplest multiple comparison scenario is to carry out pairwise t-tests. The problem arises when the ANOVA null hypothesis (i.e., all means are equal) is not rejected and the pairwise t-tests use the same significance level of $\alpha = 0.05$. For each t-test, we have a 5% chance of erroneously rejecting the null hypothesis ($H_0 : \mu_i = \mu_j$). The chance of at least one erroneous rejection (among the multiple t-tests) is, however, much larger than 0.05. In other words, we are more likely to find a statistically significant difference when we make multiple comparisons even when all the means are the same. This problem is often masked when only the largest difference among all pairwise comparisons is tested. Although recognized in the 1940s [Tukey, 1949], the multiple comparison problems received wider attention in the 1970s, marked by the publication of the Holm-Bonferroni method [Holm, 1979]. The issue came back in the 1980s and 1990s as part of the study on false discovery rate (e.g., Benjamini and Hochberg [1995]). In ecological applications, the issue of multiple comparisons can often be hidden, especially when computer-intensive methods are used. Specifically, when a test is used after an automatic search of a maximum "difference." Such a search can be hidden in the computer program which exhuasts all possible comparisons. But the consequence is the same – the stated type I error probability of $\alpha = 0.05$ is an underestimate of the true type I error probability.

1.1.2 The Raven's Paradox

The Raven's paradox, also known as the Hampel's paradox of confirmation, is a paradox about what constitutes as evidence for confirming a hypothesis. The logical contrapositive of the statement "all ravens are black" is "everything that is not black is not a raven." In mathematics, proof by contrapositive is an oft-used inference method. That is, you can show "If not B then not A" to support "If A then B." The Raven's paradox is to show that the proof by contrapositive method can be paradoxical because it suggests that observing, for example, a white shoe is evidence supporting the claim that all ravens are black, because a white shoe is a non-white and non-raven object. I.J. Good proposed a Bayesian explanation (or solution) of the paradox [Good, 1983]. The weight of evidence provided by observing a white shoe (or any non-raven and non-black object) is positive, but small if the number of ravens is small compared to the number of all non-black objects. But how is the paradox relevant to statistical hypothesis testing? The NHT is actually based on the logic of proof by contrapositive. For example, in a one-sample t-test problem, we are interested in showing that the alternative (e.g., $\mu \neq 0$) is true. But the test provides evidence against the null hypothesis, $\mu = 0$. If all we want to know is whether or not μ is equal to 0, NHT is perfectly logical. However, we often want to know, for example, whether μ is equal to a specific value μ_a. By showing that $\mu = 0$ is not true does not provide much evidence to support $\mu = \mu_a$. In a simple t-test situation, we can provide the confidence interval of the estimated mean to suggest where the mean is likely located. This step goes beyond NHT and narrows down the range of the alternative mean. In a more complicated situation where we want to show a specific response pattern (e.g., threshold response, subsidy-stress response), rejecting a null hypothesis of no response is akin to showing a white shoe as proof of the blackness of a raven.

2 Methods

Many statistical results are counterintuitive because we are not accustomed to thinking in terms of probability or average behavior [Kahneman and Tversky, 1972]. Because the concept of probability rests on the long-run frequency, Monte Carlo simulation is often the best way to expose hidden problems of a statistical procedure. A Monte Carlo simulation can be designed to replicate the theoretical characteristics of a method or model, such that we can examine the outcome under known scenarios. For NHTs, we often use simulations to verify the test's prescribed behavior. Specifically, a test's type I error probability should be at the commonly accepted significance level of 0.05. In fact, testing an NHT's type I error probability is a standard method of verifying a test's robustness against departures from various assumptions, a practice started well before the advent of high-performing personal computers (see, for example, Box [1976]). Using a simulation, we repeatedly generate data from the distribution defined by the null hypothesis and carry out the test. The test should reject the null hypothesis about 5% of the time, the predetermined acceptable type I error probability. When a test is subject to the multiple-comparison trap, we will reject the null hypothesis more frequently. Using Monte Carlo simulation, we can approximate the type I error probability of a test.

The Raven's paradox is about the use of irrelevant evidence to support a claim. The best practice to avoid the paradox is to clearly state the null and alternative hypotheses of a test. When using the NHT, the p-value is a conditional probability – the probability of observing data as extreme as the current data when *the null hypothesis is true*. The null hypothesis is a specific hypothesis with a specific probabilistic model. This probabilistic model enables the calculation of a p-value and is often a straw man to be shot down. The question we must ask is whether the rejection of the null hypothesis leads to the validation of the alternative hypothesis in which we are interested. Without clearly stating the null and alternative hypotheses, we cannot readily answer this question. This is why Qian [2014b] recommended to verbally describe a statistical model, even for a simple t-test.

2.1 Examples

We use two examples to illustrate the use of simulation for assessing an NHT's type I error probability, as well as to explain why the paradoxical nature of the results from such applications can be exposed when the null and alternative hypotheses are verbally described. Both examples are from published papers with more than 100 citations.

2.1.1 Using CART-based Method for Detecting and Quantifying Ecological Threshold

One of the two methods for detecting an environmental threshold in Qian et al. [2003], the deviance reduction method, uses a classification and regression tree (CART) model. The basic idea is to find a break point along an environmental gradient to reduce the total deviance of the response variable. Specifically, the deviance reduction method is a CART model with one predictor variable (the gradient) and the estimated threshold is the first split point. Qian et al. [2003] included a generic χ^2-test for testing the statistical significance of the estimated threshold. The null hypothesis of this test is that the response variable distribution is the same across the entire gradient and the alternative hypothesis is that the response variable has two distributions, separated by the split point. The -2 log likelihood ratio (null over alternative) has a χ^2 distribution if the null model is true. The degrees of freedom (df) of the χ^2 distribution is the difference between the df of the null model and the df of the alternative model. For example, when the response variable is a count variable approximated by the Poisson distribution, the null hypothesis describes the response variable using a single Poisson distribution with one parameter, and the alternative hypothesis describes the response variable using two Poisson distributions (two parameters). The χ^2 test df is then 1. If the response variable distribution is normal or log-normal, χ^2 test df can be either 2 (assuming both mean and variance are different before and after crossing the split point) or 1 (assuming a constant variance). When assuming constant variance, the test is reduced to a two-sample t-test for normal response variables.

The test can, however, be deceiving. The CART model selected split point is the point that results in the largest reduction in deviance, among all possible split points. This process is analogous to selecting and testing the largest difference among all pairwise comparisons in an ANOVA problem. Because the CART model does the selection "behind the scene," the fact that the resulting split point is selected among all possible split points (i.e., multiple comparisons) is "hidden" from most of us. Although we set the test significance level at 0.05, the actual probability of making a type I error (erroneously rejecting the null hypothesis and claiming a threshold response) can be much higher, resulting in a test that rejects the null hypothesis far too frequent than we expected.

We use two Monte Carlo simulation approaches for this example, assuming that the response variable has a normal distribution. For this example, the null hypothesis assumes the response variable has the same normal distribution along the environmental gradient. We can draw random numbers from this normal distribution and pair them to randomly drawn gradient values. This step produces one realization of the data under the null hypothesis. We then apply the CART model to find the split point and conduct a two-sample t-test to compare the means of the two groups created by the split point. The process is then repeated many times, each time we record whether the null hypothesis is rejected (or the p-value is below the predetermined significance level of, e.g., 0.05). The proportion of times we reject the null hypothesis is an estimate of the true type I error. Alternatively, we can use an existing data set to carry out the simulation by randomly permuting the observed response variable data along the gradient, thereby removing the gradient effect. In the online supporting materials, we provided a simple R function to carry out both simulations. In general, the true type I error is much higher than the declared significance level of 0.05, and the larger the sample size is the higher the type I error probability is.

2.1.2 Permutation Test for Species Aggregation

The second example is a component of a computer program using an indicator of species aggregation to detect a threshold along an environmental gradient [Baker and King, 2010]. At the species (or taxon) level, the program measures species aggregation using an indicator developed by Dufrêne and Legendre [1997]. Specifically, the computer program finds all possible split points along the gradient to divide the data into two groups. For each possible split, an indicator value (IV) is calculated. The split point that resulted in the largest difference in the indicator value between the two groups is selected as a potential threshold. A permutation test is then used to test whether the potential threshold is "statistically significant." The underlying mechanism of leading to the multiple comparison trap is the same as the first example – a statistical test is applied to the data selected to have the maximum difference in IV, among all possible groupings.

Although we did not find a clear description of the null and the alternative hypothesis of the test, the use of a permutation test implies a null hypothesis that the response variable distribution is not a function of the gradient. The response variable used for calculating the IVs is taxon abundance. For a specific split point, the sample sizes of the two subsets (n_1 and n_2) are known. The permutation test tabulates (more specifically, samples from) all possible permutations of splitting the data into two subsets of sample sizes n_1 and n_2. For each permutation, an IV is calculated. The collection of these IVs is used to form an empirical distribution, and the p-value is calculated as the fraction of these IVs exceeding the observed IV. In other words, the distribution of these permutation-derived IVs is taken as the null hypothesis distribution. Consequently, the null hypothesis of the test is that taxon abundance is not a function of the gradient, which is equivalent to a constant expected value of taxon abundance along the gradient.

To simulate data from the null hypothesis model, we can, as in the previous example, either draw random samples from a specific null hypothesis distribution or randomly permute the existing abundance observations along the gradient. In the online supporting materials, we provide the computer code for the Monte Carlo simulations. When using simulated or synthetic data, we explored the relationship between the type I error probability and sample size. Using the Everglades data accompanying Baker and King [2010], we illustrate the consistency between the results of the synthetic data simulation and of the real data based simulation.

For both examples, we explicitly present the null and alternative hypotheses to show that rejecting the null in both cases does not imply a threshold response. R code and example data are available online at `https://github.com/songsqian/perils`.

2.2 Why Bootstrapping Doesn't Work

In many applications of the two methods, bootstrapping is used to develop a confidence interval of the derived threshold, which is then used to determine whether the threshold is statistically significant. The bootstrapping approach is, however, flawed for this particular situation. Bühlmann and Yu [2002] detailed statistical background of bootstrapping and provided mathematical proof that bootstrap estimated confidence interval is always too narrow in a split point problem. Accordingly, Banerjee and McKeague [2007], after empirically showed that the bootstrap estimated confidence interval is narrower than expected, developed a classical statistics method for calculating the confidence interval of a split point. Qian [2016] used a Monte Carlo simulation algorithm to study the behavior of a bootstrap estimated confidence interval and he showed that a bootstrap estimated 95% confidence interval often contains the true split point less than 70% of the time. With a narrower than expected confidence interval, we are more likely to conclude that the resulting threshold is "statistically significant." Intuitively, the failure of the bootstrapping method for a split point problem is because that the pool of available split points are more or less fixed in a data set. When using bootstrapping to estimate the standard deviations of another statistic, such as the mean, repeaded bootstrap sampling leads to bootstrap samples with different values of the statistic. When sampling for a split point, a bootstrap sample of the data will result in a subset of the same potential slit points as in the original data, hence the less than expected variability in the bootstrap estimated split point distirbution.

3 Results

3.1 Multiple Comparison Trap

3.1.1 The CART-based Method

We performed two sets of simulations. First, we simulate the distribution defined by the null hypothesis, no change of the response variable along the gradient, by drawing random numbers from a normal distribution paired with a set of gradient values (also randomly drawn). The deviance reduction method is applied to the resulting data and a two-sample t-test is used and the p-value is recorded. This process is repeated many times (10,000) and the fraction of the p-values smaller than 0.05 is an estimate of the test's type I error probability. These simulations show that the type I error probability of the test is a function of the sample size (Figure 1), and will exceed 0.2 when the sample size is larger than 40.

Using the data from Qian et al. [2003], we perform the simulation by randomly permuting the gradient (TP) data. For each random permutation, we apply the deviance reduction method to divide the data into two subsets and carry out a two-sample t-test to compare the means. After repeating the process many times (10,000), the fraction of tests with p-values below 0.05 is an estimate of the type I error probability. We estimated the type I error probability for BCD and percent of tolerant macroinvertebrate taxa for each sampling date. The estimated probabilities are similar to those obtained from the synthetic data simulation (Figure 1).

3.1.2 Permutation Test on IV

We simulate the permutation test null hypothesis by drawing abundance data from a Poisson distribution and randomly assign each abundance value to a gradient value. The resulting abundance distribution is not a function of the gradient. Under this null hypothesis, the expected abundance is constant along the gradient. For the synthetic data simulation, we draw random numbers from a Poisson distribution with a mean of 5. The resulting type I error probability is a function of sample size and consistently larger than 0.2 (Figure 2). Using the Everglades macroinvertebrate taxa abundance data from Baker and King [2010], the estimated type I error probabilities for individual taxa are consistently larger than 0.25 and follow the pattern of the synthetic data simulation but with a large variation. The variation of the type I error probability is inversely proportional to the number of zeroes in the data (Figure 3). This pattern suggests that taxa that are more tolerant to the disturbant (hence present along the entire gradient) are more likely to result in a false positive, that is, finding a threshold where there is not one. When these false positives are mixed with thresholds estimated for taxa that are sensitive to the disturbance, the resulting community threshold is indeed difficult to interpret.

3.2 Raven's Paradox

The null hypotheses for both examples describe a constant expected value of the response variable along the gradient of interest. For the CART-based method, the intended alternative hypothesis is clearly stated as a step-function model (the mean response does not change along the gradient before crossing the threshold and the mean jumps to a different stable value after crossing the threshold) [Qian et al., 2003]. This model is consistent with the CART model. Clearly, when the null hypothesis of a constant mean response is rejected, we cannot readily claim that the response pattern is a step function; the response pattern can be anything other than a constant one. Likewise, the null hypothesis of IV permutation test is a constant mean taxon abundance along the gradient. Rejecting the null hypothesis implies that the taxon abundance may vary along the gradient. But the significant test result yields no specific information on how the abundance will change. Concluding a threshold response in both examples is a fallacy of the Raven's Paradox.

4 Discussion

With the advent of inexpensive computing power, we are increasingly relying on computer-intensive methods. As a result, the multiple-comparison trap is increasingly relevant. Fortunately, with affordable computing power, we can easily implement Monte Carlo simulations to characterize the behavior of new methods. As a rule, we should always evaluate an NHT-based method's type I error probability before it can be presented for wide application.

Statistical null hypothesis testing (NHT) represented an intellectual leapfrog in the development of modern statistics [Stigler, 2016]. Using NHT allows us to formally summarize information in noisy data to support the hypothesis of interest. However, the NHT formulation of Neyman-Pearson emphasizes the evidence disproving the null hypothesis. Although accepting the alternative hypothesis is the logical consequence of rejecting the null, the alternative hypothesis is a composite hypothesis rather than a specific hypothesis we often have in mind. When we are interested in a specific alternative hypothesis (e.g., the step function model in the CART example), rejecting the null (of a constant response) is the same as observing a white shoe, which provides infinitesimally small support to the hypothesis that all ravens are black. We can easily avoid such a situation by clearly stating the null and alternative hypotheses, and the reasons why rejecting the null is evidence supporting the alternative hypothesis, that is, the hypothesis of interest.

Furthermore, the meaning of the term "statistically significant" is vague and often misinterpreted. For example, in both examples, when we say that the estimated threshold is statistically significant, a reader may interpret the statement to be that we have evidence to show the validity of (1) the existence of a threshold and (2) the estimated threshold value. The correct interpretation of the term "statistically significant" is that we have evidence against the null hypothesis. A significant result does not imply threshold response and is irrelevant to the estimated threshold value. It is, therefore, imperative that the meaning of the term "statistically significant" be clearly stated to help avoid the Raven's paradox.

In addition to the vagueness of what a statistically significant result is, the definition of an ecological threshold is also vague. By vague we mean that the definition does not allow the formulation of a specific quantitative model. Because these tests have a high type I error probability, they have a high probability of rejecting the null when the response is not a constant along the gradient. In other words, a significant result can indicate a threshold response, but also a linear response pattern, or any other non-constant patterns [Cuffney et al., 2011, Cuffney and Qian, 2013]. As in any statistical modeling problem, methods for detecting a threshold response should first establish the likely quantitative model forms [Qian, 2014a, Qian and Cuffney, 2012]. Using a formal model fitting and assessment process, we can then establish the validity of the proposed model as well as the model's inferential limit [McElreath, 2016]. The formulaic process of NHT is inadequate for establishing the statistical validity of a model. The difficulty of a modeling approach is the selection and justification of a proper model. New computer-intensive nonparametric methods are now available for us to carry out exploratory analysis to aid the development of a proper model. For example, [Wagenhoff et al., 2017] used boosted regression tree models to better understand ecosystem responses to agricultural stressors.

Statistics should be used to make a principled argument to support the hypothesis of interest. A principled argument starts with a clear statement of the goal. In our examples, we should start with a definition of a threshold response and derive a method to demonstrate its validity. As "a null hypothesis test is a ritualized exercise of devil's advocacy," it should not be used as the only piece of evidence; a persuasive case should be presented using the MAGIC (magnitude, articulation, generality, interestingness, and credibility) criteria of Abelson [1995].

5 Acknowledgement

We thank Ian Waite and Chad Wagner for reviewing an early version of the paper.

References

R.P. Abelson. *Statistics as Principled Argument.* Psychology Press, New York, 1995.

M.E. Baker and R.S. King. A new method for detecting and interpreting biodiversity and ecological community thresholds. *Methods in Ecology and Evolution*, 1(1):25–37, 2010.

M. Banerjee and I. W. McKeague. Confidence sets for split points in decision trees. *The Annals of Statistics*, 35(2):543–574, 2007.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

G.E.P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356): 791–799, 1976.

P. Bühlmann and B. Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

T.F. Cuffney and S.S. Qian. A critique of the use of indicator species scores for identifying thresholds in species responses. *Freshwater Science*, 32(2):471–488, 2013.

T.F. Cuffney, S.S. Qian, R.A. Brightbill, J.T. May, and I.R. Waite. Response to King and Baker: limitation on threshold detection and characterization of community thresholds. *Ecological Applications*, 21(7):2840–2845, 2011.

M. Dufrêne and P. Legendre. Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs*, 67(3), 1997.

A.M. Ellison, N.J. Gotelli, B.D. Inouye, and D.R. Strong. *P* values, hypothesis testing, and model selection: it's déjà vu all over again. *Ecology*, 95(3):609–610, 2014.

I.J. Good. *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesoda Press, Minneapolis, MN, 1983.

S. Holm. Simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

D. Kahneman and A. Tversky. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3:430–454, 1972.

R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman & Hall/CRC Press, 2016.

S.S. Qian. Ecological threshold and environmental management: A note on statistical methods for detecting thresholds. *Ecological Indicators*, 38:192–197, 2014a.

S.S. Qian. Statistics in ecology is for making a "principled" argument. *Landscape Ecology*, 29 (6):937–939, 2014b.

S.S. Qian. *Environmental and Ecological Statistics with R*. Chapman and Hall/CRC Press, 2nd edition, 2016.

S.S. Qian and T. F. Cuffney. To threshold or not to threshold? That's the question. *Ecological Indicators*, 15(1):1–9, 2012.

S.S. Qian, R.S. King, and C.J. Richardson. Two statistical methods for the detection of environmental thresholds. *Ecological Modelling*, 166:87–97, 2003.

S.M. Stigler. *The Seven Pillars of Statistical Wisdom*. Harvard University Press, Cambridge, MA., 2016.

J.W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5(2): 99–114, 1949.

A. Wagenhoff, A. Liess, A. Pastor, J.E. Clapcott, E.O. Goodwin, and R.G. Young. Thresholds in ecosystem structural and functional responses to agricultural stressors can inform limit setting in streams. *Freshwater Science*, 1:000–000 (DOI: 10.1080/690233), 2017.

6 Figures

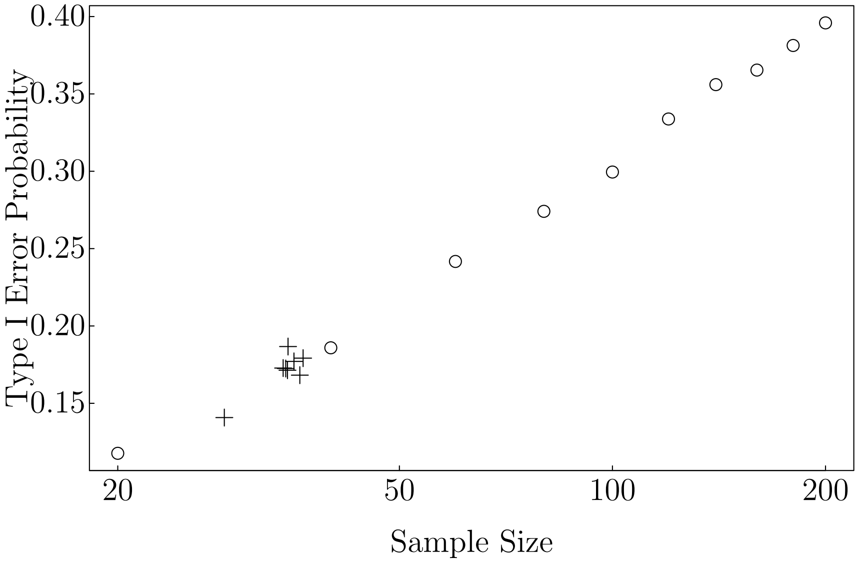

Figure 1: Simulation estimated type I error probability of the CART-based model is a function of sample size. Open circles are estimated from using fake data simulation and pluses are based on simulation using data from Qian et al. [2003] (with sample size jittered for better visual).

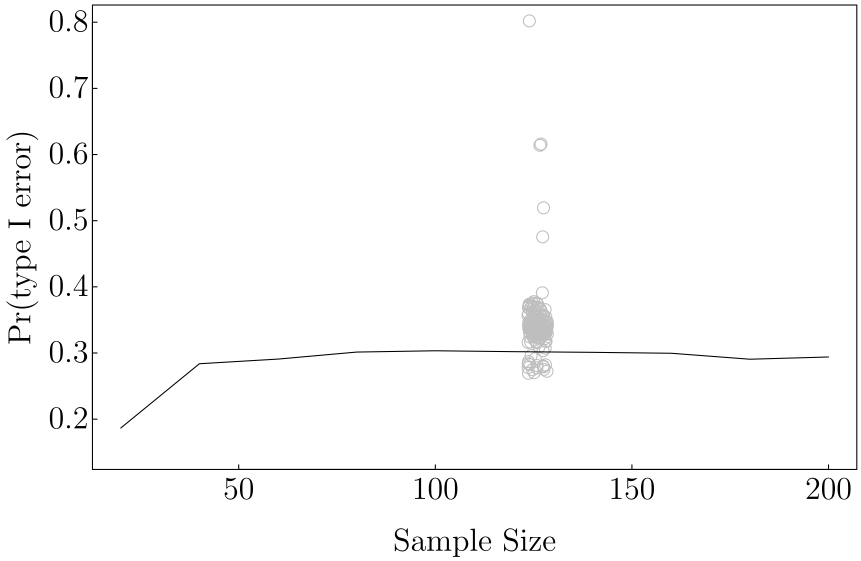

Figure 2: The type I error probability of TITAN's permutation test is a function of sample size. The line is estimated from using fake data simulation and shaded data points are based on simulation using the Everglades data from Baker and King [2010] (with sample size slightly jittered for easy visualization).

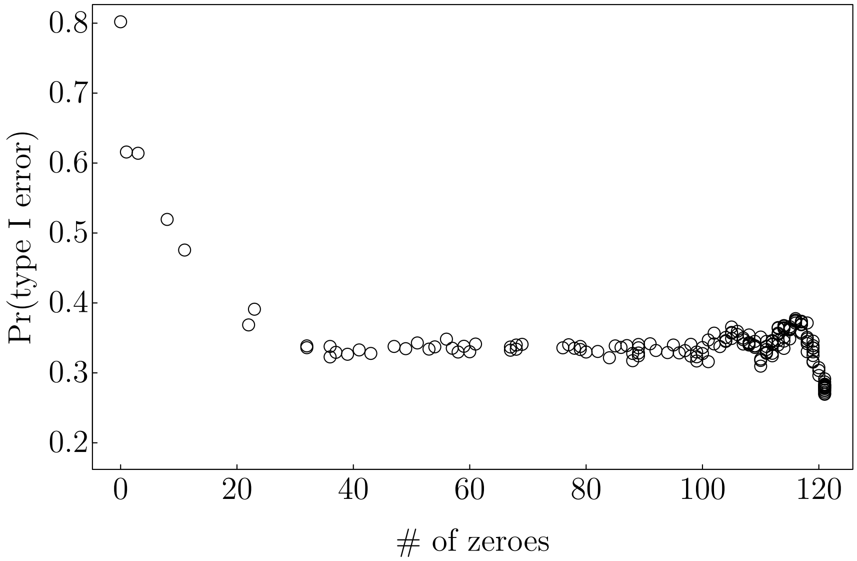

Figure 3: Estimated type I error probability of TITAN using the Everglades data from Baker and King [2010] is a function of the number of zeroes in the abundance data.