Conservative hypothesis tests and confidence intervals using importance sampling

BY MATTHEW T. HARRISON

Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912, U.S.A.
matthew_harrison@brown.edu

SUMMARY

Importance sampling is a common technique for Monte Carlo approximation, including that of p -values. Here it is shown that a simple correction of the usual importance sampling p -values provides valid p -values, meaning that a hypothesis test created by rejecting the null hypothesis when the p -value is at most α will also have a Type I error rate of at most α . This correction uses the importance weight of the original observation, which gives valuable diagnostic information under the null hypothesis. Using the corrected p -values can be crucial for multiple testing and also in problems where evaluating the accuracy of importance sampling approximations is difficult. Inverting the corrected p -values provides a useful way to create Monte Carlo confidence intervals that maintain the nominal significance level and use only a single Monte Carlo sample.

Some key words: Exact inference; Monte Carlo simulation; Multiple testing; p -value; Rasch model.

1. INTRODUCTION

Importance sampling is a common technique for Monte Carlo approximation. Besides its use in situations where efficient direct-sampling algorithms are unavailable, it can be used to accelerate the approximation of small p -values as needed for multiple hypothesis testing. Importance sampling can also be used for Monte Carlo approximation of confidence intervals using a single Monte Carlo sample by inverting a family of hypothesis tests. The practicality of these two uses of importance sampling, accelerated multiple testing and Monte Carlo approximation of confidence intervals, would seem to be limited by the fact that each places strong requirements on the importance sampling procedure. Multiple testing controls can be sensitive to small absolute errors but large relative errors in small p -values, which would seem to demand either excessively large Monte Carlo sample sizes or unrealistically accurate importance sampling proposal distributions in order to reduce absolute errors to tolerable levels. The confidence interval procedure uses a single proposal distribution to approximate probabilities under a large family of target distributions, which again would seem to demand large sample sizes in order to overcome the high variability that is frequently encountered when the proposal distribution is not tailored to a specific target distribution. Nevertheless, as shown here, simple corrections of the usual p -value approximations can be used to overcome these difficulties, making importance sampling a practical choice for accelerated multiple testing and for constructing Monte Carlo confidence intervals.

The p -value corrections that we introduce require negligible additional computation and still converge to the target p -value, but are valid p -values, meaning the probability of rejecting a null hypothesis does not exceed α for any specified level α . Hypothesis tests and confidence intervals constructed from the corrected p -value approximations are guaranteed to be conservative,

regardless of the Monte Carlo sample size, while still behaving like the target tests and intervals for sufficiently large Monte Carlo sample sizes. The combination of being both conservative and consistent turns out to be crucial in many applications where the importance sampling variability cannot be adequately controlled with practical amounts of computing, including multiple testing, confidence intervals and any hypothesis testing situation where the true variability of the importance sampling algorithm is either large or unknown.

Let X denote the observed data. Assume that the null hypothesis specifies a known probability distribution P for X . For a specified test statistic $t(X)$, the goal is to compute the p -value $p(X)$ defined by

$$p(x) = \text{pr}\{t(X) \geq t(x)\}, \quad (1)$$

where the probability is computed under the null hypothesis. Importance sampling can be used to approximate $p(X)$ if the p -value cannot be determined analytically. Let $Y = (Y_1, \dots, Y_n)$ be a random sample from the proposal distribution Q whose support includes the support of the target distribution P . Then $p(X)$ can be approximated with either of

$$p_1(X, Y) = \frac{\sum_{i=1}^n w(Y_i) \mathbb{1}\{t(Y_i) \geq t(X)\}}{n}, \quad p_2(X, Y) = \frac{\sum_{i=1}^n w(Y_i) \mathbb{1}\{t(Y_i) \geq t(X)\}}{\sum_{j=1}^n w(Y_j)},$$

where $\mathbb{1}$ is the indicator function, and the importance weights are defined to be the ratio of probability mass functions in the discrete case, i.e., $w(x) = P(x)/Q(x)$, and the ratio of densities in the continuous case. Each p_c is a consistent approximation of $p(X)$ as the Monte Carlo sample size n increases. The approximation p_1 is unbiased and is especially useful for accelerating the approximation of extremely small p -values. The approximation p_2 can be evaluated even if the importance weights are known only up to a constant of proportionality. The reader is referred to [Liu \(2001\)](#) for details and references concerning importance sampling and to [Lehmann & Romano \(2005\)](#) for hypothesis testing.

Here we propose the following simple corrections of p_c that use the importance weight of the original observation, namely

$$\begin{aligned} \tilde{p}_1(X, Y) &= \frac{w(X) + \sum_{i=1}^n w(Y_i) \mathbb{1}\{t(Y_i) \geq t(X)\}}{1 + n}, \\ \tilde{p}_2(X, Y) &= \frac{w(X) + \sum_{i=1}^n w(Y_i) \mathbb{1}\{t(Y_i) \geq t(X)\}}{w(X) + \sum_{j=1}^n w(Y_j)}. \end{aligned} \quad (2)$$

We show that the corrected p -value approximations, while clearly still consistent approximations of the target p -value $p(X)$, are also themselves valid, meaning

$$\text{pr}\{\tilde{p}_c(X, Y) \leq \alpha\} \leq \alpha \quad (0 \leq \alpha \leq 1; n = 0, 1, \dots; c = 1, 2),$$

under the null hypothesis, where the probability is with respect to the joint distribution of data and Monte Carlo sample.

For the special case of direct sampling, i.e., $Q = P$ and $w \equiv 1$, the validity of $\tilde{p}_c(X, Y) = (1 + n)^{-1}[1 + \sum_{i=1}^n \mathbb{1}\{t(Y_i) \geq t(X)\}]$ is well known and is often recommended over p_c for use as a p -value ([Davison & Hinkley, 1997](#), Ch. 4). This paper was motivated by trying to extend the special case of direct sampling to the more general case of importance sampling. For Markov chain Monte Carlo approaches, [Besag & Clifford \(1989\)](#) demonstrate how to generate valid p -value approximations using techniques that are unrelated to the ones here. It would

be interesting to know whether the two approaches could be combined so that the importance samples could also be generated by a Markov chain.

2. MAIN RESULTS

The main result is that each $\tilde{p}_c$ is a valid p -value. We generalize the introductory discussion in two directions. First, we allow arbitrary distributions, so the importance weights become the Radon–Nikodym derivative dP/dQ . In the discrete case, this simplifies to the ratio of probability mass functions and, in the continuous case, to the ratio of probability density functions. Secondly, we allow the choice of test statistic to depend on $(X, Y_1, \dots, Y_n)$ as long as the choice is invariant to permutations of $(X, Y_1, \dots, Y_n)$. We express this mathematically by writing the test statistic, t , as a function of two arguments: the first is the same as before, but the second argument takes the entire sequence $(X, Y_1, \dots, Y_n)$, although we require that t is invariant to permutations in the second argument. For example, we may want to transform the sequence $(X, Y_1, \dots, Y_n)$ in some way, either before or after applying a test statistic to the individual entries. As long as the transformation procedure is permutation invariant, such as centring and scaling, or converting to ranks, the main result still holds. Transformations are often desirable in multiple testing contexts for improving balance (Westfall & Young, 1993).

We begin with the precise notation and assumptions for the main result. Let P and Q denote probability distributions defined on the same measurable space $(S, \mathcal{S})$ with $P \ll Q$, meaning that sets with positive P probability also have positive Q probability, and let w denote a fixed, nonnegative version of the Radon–Nikodym derivative dP/dQ . The set $\mathcal{M}$ denotes all $(n+1)!$ permutations $\pi = (\pi_0, \dots, \pi_n)$ of $(0, \dots, n)$. For $\pi \in \mathcal{M}$ and $z = (z_0, \dots, z_n)$, we define $z^{(\pi)} = (z_{\pi_0}, \dots, z_{\pi_n})$. Assume that $t : S \times S^{n+1} \mapsto [-\infty, \infty]$ is a measurable function satisfying $t(a, z) = t(a, z^{(\pi)})$ for all $a \in S$, $z \in S^{n+1}$ and $\pi \in \mathcal{M}$. For $z \in S^{n+1}$, define

$$\tilde{p}_1(z) = \frac{\sum_{i=0}^n w(z_i) \mathbb{1}\{t(z_i, z) \geq t(z_0, z)\}}{n+1}, \quad \tilde{p}_2(z) = \frac{\sum_{i=0}^n w(z_i) \mathbb{1}\{t(z_i, z) \geq t(z_0, z)\}}{\sum_{j=0}^n w(z_j)}, \quad (3)$$

where we use the convention $0/0 = 0$. Let X have distribution P and $Y_0, Y_1, \dots, Y_n$ be independent and identically distributed with common distribution Q , independent of X . For notational convenience, define $Z = (Z_0, Z_1, \dots, Z_n)$ by $Z_0 = X, Z_1 = Y_1, \dots, Z_n = Y_n$, so that the corrected p -values defined in (2) can be written as $\tilde{p}_c(Z)$ from (3). Then, we obtain the following theorem.

THEOREM 1. *For each $c = 1, 2$, we have $\text{pr}\{\tilde{p}_c(Z) \leq \alpha\} \leq \alpha$ for all $\alpha \in [0, 1]$.*

Proof. See the Appendix. □

The theorem does not require any special relationships among P , Q , t or n . For example, in parametric settings, Q does not need to be in the same model class as the null and/or alternative hypotheses. Validity of the corrected p -values is ensured even for unusual cases such as $n = 0$ or importance weights with infinite variance. We discuss some practical considerations for choosing Q in the next section; presumably, P , t and n are dictated by the application.

The theorem continues to hold if each Y_k is chosen from a different Q , say Q_k , as long as the sequence $Q_0, Q_1, \dots, Q_n$ is itself independent and identically distributed from some distribution over proposal distributions. The k th importance weight is now a fixed version of the Radon–Nikodym derivative dP/dQ_k . This generalization can be useful in practice when each Y_k is generated hierarchically by first choosing an auxiliary variable V_k and then, given V_k , choosing

Y_k from some distribution Q^{V_k} that depends on V_k . If the marginal distribution Q of Y_k is not available, one can use importance weights based on the conditional distribution Q^{V_k} of Y_k given V_k . This is equivalent to using random proposal distributions as described above. The drawback of combining this approach with the p -value corrections is that the importance weight of the original observation is evaluated using dP/dQ_0 , where Q_0 is a randomly chosen proposal distribution. This further increases the Monte Carlo variation inherent in the p -value approximations. Generalizing the theorem to allow for random proposals requires no additional work. If ν is the joint distribution of each (Q_k, Y_k) and ν_Q is the marginal distribution of each Q_k , simply apply the theorem with $\nu_Q \times P$ in place of P , ν in place of Q , (Q_0, X) in place of X and (Q_k, Y_k) in place of Y_k . If one uses a regular conditional distribution to define the importance weight $\{d(\nu_Q \times P)/d\nu\}(Q_k, x)$, then it will simplify to a version of $(dQ_k/dP)(x)$.

The theorem concerns p -values for the simple hypothesis test that $X \sim P$, but valid p -values for simple tests can be used to generate valid p -values for composite tests by maximizing over all distributions in the null hypothesis. Consider a collection of distributions, $\{P_\theta : \theta \in \Theta\}$, corresponding test statistics, $\{t_\theta : \theta \in \Theta\}$, and corresponding importance weights, $w_\theta = dP_\theta/dQ$, where the proposal distribution, Q , satisfies $P_\theta \ll Q$ for all $\theta \in \Theta$. For each θ , define the corrected p -value approximations $\tilde{p}_c(\theta, z)$ as in (3) using t_θ and w_θ . The theorem shows that each $\tilde{p}_c(\theta_0, Z)$ is a valid p -value for the simple null hypothesis that $X \sim P_\theta$ for $\theta = \theta_0$. An immediate corollary is that $\sup_{\theta \in \Theta_0} \tilde{p}_c(\theta, Z)$ is a valid p -value for the composite null hypothesis that $X \sim P_\theta$ for some $\theta \in \Theta_0$, where $\Theta_0 \subset \Theta$. Only a single Monte Carlo sample is required.

Finally, we recall the well-known fact that any valid family of p -values can be inverted to give valid confidence intervals. For $\alpha \in [0, 1]$ and $c = 1, 2$, define the random set

$$\tilde{C}_c^\alpha(Z) = \{\theta \in \Theta : \tilde{p}_c(\theta, Z) > \alpha\}. \quad (4)$$

Then $\tilde{C}_c^\alpha$ is a $1 - \alpha$ confidence set for θ because $\text{pr}_\theta\{\tilde{C}_c^\alpha(Z) \ni \theta\} = 1 - \text{pr}_\theta\{\tilde{p}_c(\theta, Z) \leq \alpha\} \geq 1 - \alpha$, where pr_θ is calculated under the null hypothesis of P_θ . Only a single Monte Carlo sample is required. Section 4.2 illustrates this application. The idea of using importance sampling to construct confidence intervals from a single Monte Carlo sample was pointed out in [Green \(1992\)](#). See [Bolvik & Skovlund \(1996\)](#) and [Garthwaite & Buckland \(1992\)](#) for examples of other ways to create Monte Carlo confidence intervals.

3. PRACTICAL CONSIDERATIONS

Validity of the corrected p -values holds for any n , but validity alone is not a useful property. The p -value approximations must also be close to the target p -value in order to have the appropriate power characteristics. Convergence to the target p -value typically requires large n , especially for importance sampling algorithms with high variance. Furthermore, for the special case of direct sampling, i.e., $Q = P$, excessively small n can adversely affect power ([Davison & Hinkley, 1997](#), § 4.2.5). Presumably, a similar effect holds for importance sampling, although the effect is more nuanced, because the power characteristics of the p -value approximations can be strongly influenced by the choice of Q for small n , and the resulting hypothesis tests need not be uniformly less powerful than the target test.

It is clear that each of the corrected p -values defined in (2) has the same asymptotic behaviour as its uncorrected counterpart, so for sufficiently large n they are essentially equivalent. But, in practice, n will often be too small to ensure this equivalence. The concern for the investigator is that the corrected p -values may behave much worse than the uncorrected ones for practical choices of n . In particular, power suffers when the null hypothesis is false, the target p -value

p is small, the uncorrected approximation p_c is also small, but the corrected approximation $\tilde{p}_c$ is large. Careful choice of Q can lower the chances of this.

Each of the corrected p -values can be expressed as an interpolation between their respective uncorrected versions and another valid p -value:

$$\tilde{p}_1(z) = \frac{1}{n+1}w(x) + \left(1 - \frac{1}{n+1}\right)p_1(z), \quad (5)$$

$$\tilde{p}_2(z) = \frac{w(x)}{w(x) + \sum_{j=1}^n w(y_j)} + \left\{1 - \frac{w(x)}{w(x) + \sum_{j=1}^n w(y_j)}\right\}p_2(z), \quad (6)$$

where we write $z = (x, y_1, \dots, y_n)$. In both cases, power suffers when X comes from a distribution in the alternative hypothesis, and Q is such that $w(X)$ is typically large relative to $w(Y_1), \dots, w(Y_n)$. Since $w(Y_k)$ always has mean 1, problems might arise for alternative distributions that tend to make $w(X)$ much larger than 1. This problem can be avoided by choosing Q so that it gives more weight than does P to regions of the sample space that are more typical under alternative distributions than they are under P . The problem can also be avoided by choosing Q to be similar to P so that the weights are close to 1 throughout the sample space. Most proposal distributions are designed with one of these two goals in mind, so the corrected p -values should behave well for well-designed proposal distributions. In practice, however, proposal distributions can be quite bad, and it can be helpful to look more closely at how the proposal affects the power of the corrected p -values when they have not yet converged to the target p -value.

From (5) we see that $\tilde{p}_1$ is an interpolation between p_1 and $w(x)$, the latter being a valid p -value. Validity of $w(X)$ can be seen either by Theorem 1 for $c = 1, n = 0$ or by the following calculation

$$\text{pr}\{w(X) \leq \alpha\} = \sum_x \mathbb{1}\{P(x) \leq \alpha Q(x)\}P(x) \leq \sum_x \alpha Q(x) = \alpha \quad (7)$$

for discrete settings that easily generalizes to more abstract settings. Testing $w(X) = P(X)/Q(X) \leq \alpha$ is simply a likelihood ratio test of P versus Q , although using the critical value α will give a test of size smaller than α . The test can be strongly conservative, because the bound in (7) can be far from tight. So, $\tilde{p}_1$ is an interpolation between a conservative likelihood ratio test of P versus Q and an importance sampling Monte Carlo approximation of the target p -value. If the effect of $w(X)$ on $\tilde{p}_1$ has not disappeared, such as for small n , then $\tilde{p}_1$ is a p -value for the original null hypothesis, P , versus a new alternative hypothesis that is some amalgamation of the original alternative hypothesis and the proposal distribution Q . If Q is not in the alternative hypothesis, as it often will not be, this modified alternative hypothesis is important to keep in mind when interpreting any rejections.

The effect of Q on interpreting rejections is especially important in multiple testing situations. In these settings, importance sampling is often used to accelerate the approximation of tiny p -values via p_1 . If $p_1 \approx 0$, however, then $\tilde{p}_1 \approx w(x)/(n+1)$ and Q plays a critical role in determining which null hypotheses are rejected after correcting for multiple comparisons. The investigator can take advantage of this by ensuring that a rejection of P in favour of Q is sensible for the problems at hand, and by ensuring that events with small p -values will be heavily weighted by Q .

Turning to (6), we see that $\tilde{p}_2$ is an interpolation between 1 and p_2 . Since $p_2 \leq 1$, we always have $\tilde{p}_2 \geq p_2$, so $\tilde{p}_2$ leads to a more conservative test than p_2 . Its conservatism is controlled by the ratio $w(x)/\sum_j w(y_j)$. If the ratio is small, then the correction ensures validity with little loss of power. If it is large, there may be a substantial loss of power when using the correction.

The ratio will be approximately $1/n$ if P and Q are similar, which is often the design criterion when using p_2 .

An important caveat for the main results is that Q is not allowed to depend on the observed value of X . This caveat precludes one of the classical uses of importance sampling, namely using the observed value of $t(X)$ to design a Q that heavily weights the event $\{x : t(x) \geq t(X)\}$. In many cases, however, since the functional form of t is known, one can a priori design a Q that will heavily weight the event $\{x : t(x) \geq t(X)\}$ whenever the p -value would be small. For example, given a family of proposal distributions $\{Q_\ell\}$, each of which might be useful for a limited range of observed values of $t(X)$, one can use finite mixture distributions of the form $Q = \sum_{\ell=1}^L \lambda_\ell Q_\ell$ to obtain more robust performance. For similar reasons, mixture distributions are also useful for creating Monte Carlo confidence intervals in which the same proposal distribution is used for a family of target distributions. The example in § 4.2 has more details, and [Hesterberg \(1995\)](#) contains a more general discussion of the utility of mixture distributions for importance sampling. This caveat about Q does not preclude conditional inference, in which case P is the null conditional distribution of X given some appropriate $A(X)$, and all of the analysis takes place after conditioning on $A(X)$. The choice of Q and t can thus depend on $A(X)$, but not on additional details about X .

4. APPLICATIONS

4.1. Accelerating multiple permutation tests

Consider a collection of N datasets. The i th dataset $X^i = (V^i, L^i)$ contains a sample of values $V^i = (V_1^i, \dots, V_{m_i}^i)$ and corresponding labels $L^i = (L_1^i, \dots, L_{m_i}^i)$. The distributions over values and labels are unknown and perhaps unrelated across datasets. We are interested in identifying which datasets show dependence between the values and the labels. From a multiple hypothesis testing perspective, the i th null hypothesis is that V^i and L^i are independent, or more generally, that the values of L^i are exchangeable conditioned on the values of V^i . Given a test statistic $t(x) = t(v, \ell)$, a permutation test p -value for the i th null hypothesis is

$$p(X^i) = (m_i!)^{-1} \sum_{\pi} \mathbb{1}[t\{V^i, (L^i)^{(\pi)}\} \geq t(V^i, L^i)], \quad (8)$$

where the sum is over all permutations $\pi = (\pi_1, \dots, \pi_{m_i})$ of $(1, \dots, m_i)$, and where the notation $\ell^{(\pi)} = (\ell_{\pi_1}, \dots, \ell_{\pi_{m_i}})$ denotes a permuted version of the elements of ℓ using the permutation π . Inference is conditioned on the values and the labels, but not their pairings. The p -value in (8) agrees with (1) by taking the i th null distribution, P^i , to be uniform over permutations of the labels. The Bonferroni correction can be used to adjust for multiple tests ([Lehmann & Romano, 2005](#)). In particular, rejecting null hypothesis i whenever $p(X^i) \leq \alpha/N$ ensures that the probability of even a single false rejection is no more than α . Although computing p exactly is often prohibitive, Monte Carlo samples from P^i are readily available for approximating p , and importance sampling can be used to accelerate the approximation of tiny p -values. The Bonferroni correction is still sensible, provided the Monte Carlo approximate p -values are valid p -values, which is ensured by using the corrected p -value approximations.

Consider $N = 10^4$ independent datasets, each with $m_i = 100$ scalar values and corresponding binary labels. In each case, the 100 values are mutually independent and there are 40 labels of 1 and 60 labels of 0. In 9990 datasets, the values have a Cauchy distribution with median 0 and scale 1 irrespective of the labels, i.e., the null hypothesis is true. In 10 of the datasets, the value associated with label $\ell \in \{0, 1\}$ has a Cauchy distribution with median 2ℓ and scale 1, i.e., there

Table 1. Bonferroni-corrected performance versus Monte Carlo sample size for the simulation in § 4.1

n	No. of correct rejections out of 10				No. of incorrect rejections out of 9990			
	10^1	10^2	10^3	10^4	10^1	10^2	10^3	10^4
p_P	10	10	10	10	908	87	12	1
$\tilde{p}_P$	0	0	0	0	0	0	0	0
p_1	9	8	7	5	7579	1903	2	0
$\tilde{p}_1$	7	6	6	5	0	0	0	0
q_1	6	5	4	3	4545	585	0	0

is a relationship between values and labels, and the null hypothesis is false. This example was chosen because standard tests like the two-sample t -test and the Wilcoxon rank-sum test perform poorly. For the alternative hypothesis where label 1 tends to have larger values than label 0, a sensible test statistic is the difference in medians between values associated to label 1 and values associated to label 0, namely $t(x) = t(v, \ell) = \text{median}(v_j : \ell_j = 1) - \text{median}(v_j : \ell_j = 0)$.

Table 1 shows the $\alpha = 0.05$ Bonferroni-corrected performance of several different Monte Carlo approximations as n increases. The approximations p_P and $\tilde{p}_P$ refer to p_1 and $\tilde{p}_1$, respectively, for the special case of $Q = P$ and $w \equiv 1$, i.e., direct sampling from the uniform distribution over permutations. Using either p_P or $\tilde{p}_P$ would be the standard method of approximating p . The approximations p_1 and $\tilde{p}_1$ are based on a proposal distribution that prefers permutations that pair large values of t with label 1, as described in the Appendix. In each case, we reject when the approximate p -value does not exceed α/N . The final approximation q_1 refers to a Wald $1 - \alpha/(2N)$ upper confidence limit for p_1 using the same importance samples to approximate a standard error. For q_1 , we reject whenever $q_1 \leq \alpha/(2N)$, which, if the confidence limits were valid, would provide the same guarantees as the Bonferroni correction.

Only $\tilde{p}_1$ works well. The approximations that are not guaranteed to be valid p -values, namely p_P , p_1 and q_1 , require excessively large n before the false detection rate drops to the Bonferroni target of zero errors. The cases $n = 10^3$ and $n = 10^4$ are computationally burdensome, and the situation only worsens when N increases. Furthermore, in a real problem, the investigator has no way of determining that n is large enough. The confidence limit procedure q_1 similarly failed because the standard error approximations are inaccurate for small n . The valid p -values, namely $\tilde{p}_P$ and $\tilde{p}_1$, ensure that the Bonferroni correction works regardless of n , but $\tilde{p}_P \geq 1/(n+1)$, so it is not useful after a Bonferroni correction unless n is extremely large. The good performance of $\tilde{p}_1$ requires the combination of importance sampling, which allows tiny p -values to be approximated with small n , and the validity correction introduced in this paper, which allows multiple testing adjustments to work properly.

4.2. Exact inference for covariate effects in Rasch models

Let $X = (X_{ij} : i = 1, \dots, M; j = 1, \dots, N)$ be a binary $M \times N$ matrix. Consider the following logistic regression model for X . The X_{ij} are independent Bernoulli (p_{ij}), where

$$\log \frac{p_{ij}}{1 - p_{ij}} = \kappa + \alpha_i + \beta_j + \theta v_{ij}, \quad (9)$$

for unknown coefficients κ , $\alpha = (\alpha_1, \dots, \alpha_M)$, $\beta = (\beta_1, \dots, \beta_N)$ and θ , and known covariates $v = (v_{ij} : i = 1, \dots, M; j = 1, \dots, N)$. The special case $\theta = 0$ is called the Rasch model and is commonly used to model the response of M subjects to N binary questions (Rasch, 1961). In this example, we discuss inference about θ when κ , α , β are treated as nuisance parameters.

Consider first the problem of testing the null hypothesis of $\theta = \theta'$ versus the alternative hypothesis of $\theta \neq \theta'$. Conditioning on the row and column sums, say $r = (r_1, \dots, r_M)$ and $c = (c_1, \dots, c_N)$, removes the nuisance parameters (Cox, 1958; Mehta & Patel, 1995), and the original composite null hypothesis reduces to the simple, conditional, null hypothesis of $X \sim P_{\theta'}$, where $P_{\theta'}$ is the conditional distribution of X given the margins r, c for the model in (9) with $\theta = \theta'$, namely,

$$P_{\theta}(x) \propto \exp \left(\theta \sum_{ij} x_{ij} v_{ij} \right) \left\{ \prod_i \mathbb{1} \left(\sum_j x_{ij} = r_i \right) \right\} \left\{ \prod_j \mathbb{1} \left(\sum_i x_{ij} = c_j \right) \right\}. \quad (10)$$

A sensible test statistic is the minimal sufficient statistic for θ under the conditional distribution P_{θ} defined in (10), namely $t(x) = \sum_{ij} x_{ij} v_{ij}$. Since $t(X)$ has power for detecting $\theta > \theta'$ and $-t(X)$ has power for detecting $\theta < \theta'$, it is common to combine upper- and lower-tailed p -values into a single p -value, $p^{\pm}(\theta', X)$, defined by

$$\begin{aligned} p^+(\theta, x) &= \text{pr}_{\theta}\{t(X) \geq t(x)\}, & p^-(\theta, x) &= \text{pr}_{\theta}\{-t(X) \geq -t(x)\}, \\ p^{\pm}(\theta, x) &= \min[1, 2 \min\{p^+(\theta, x), p^-(\theta, x)\}], \end{aligned} \quad (11)$$

where pr_{θ} uses $X \sim P_{\theta}$. There are no practical algorithms for computing $p^{\pm}(\theta', X)$ nor for direct sampling from $P_{\theta'}$. We can, however, create a proposal distribution, say Q_{θ} , for any P_{θ} , as described in the Appendix. Although $Q_{\theta}(x)$ can be evaluated, $P_{\theta}(x)$ is known only up to a normalization constant, so we must use p_2 and $\tilde{p}_2$. To approximate $p^{\pm}$, we use two Monte Carlo p -value approximations, one for each of t and $-t$, and combine them as in (11). Validity of the underlying p -values ensures validity of the combined p -value.

Now consider the problem of creating confidence intervals for θ by inverting a family of Monte Carlo approximations of $p^{\pm}$ as in (4). Inference is still conditioned on the row and column sums, giving a unique conditional distribution P_{θ} for each θ . Unlike hypothesis testing, where the test of $\theta = \theta'$ can use a different proposal distribution for each θ' , the use of (4) requires a common proposal distribution for all θ . To create a common proposal distribution that might work well in practice, we use a mixture of proposals designed for specific θ . In particular, define $Q = L^{-1} \sum_{\ell=1}^L Q_{\theta_{\ell}}$, where $(\theta_1, \dots, \theta_L)$ are fixed. In the example below, we use $L = 601$ and $(\theta_1, \dots, \theta_L) = (-6.00, -5.98, -5.96, \dots, 6.00)$. For any θ , the importance weights are

$$w(\theta, x) = \frac{P_{\theta}(x)}{Q(x)} = c_{\theta} \frac{\exp\{\theta t(x)\}}{\sum_{\ell} Q_{\theta_{\ell}}(x)}$$

for any binary matrix x with the correct row and column sums, where c_{θ} is an unknown constant that is not needed for computing p_2 and $\tilde{p}_2$. We will use $p_2^+(\theta, z)$ and $p_2^-(\theta, z)$ to denote $p_2(z)$ when computed using the importance weights $w(\theta, \cdot)$ and the test statistics $+t$ and $-t$, respectively, where $z = (x, y_1, \dots, y_n)$. We will use $p_2^{\pm} = \min\{1, 2 \min(p_2^+, p_2^-)\}$ as in (11). Similarly, define $\tilde{p}_2^+$, $\tilde{p}_2^-$ and $\tilde{p}_2^{\pm}$ for $\tilde{p}_2$. Inverting these p -values gives the confidence sets $C^{\alpha}(z) = \{\theta \in \mathbb{R} : p_2^{\pm}(\theta, z) > \alpha\}$ and $\tilde{C}^{\alpha}(z) = \{\theta \in \mathbb{R} : \tilde{p}_2^{\pm}(\theta, z) > \alpha\}$. For fixed z , the approximate p -values are well-behaved functions of θ , so it is straightforward to numerically approximate the confidence sets, which will typically be intervals. From Theorem 1 and the discussion in § 2, we see that $\tilde{C}^{\alpha}$ maintains the nominal significance level, whereas C^{α} might not.

Table 2 describes the results of a simulation experiment that investigated the coverage properties of C^{α} and $\tilde{C}^{\alpha}$. In that experiment, we took $M = 200$, $N = 10$, and fixed the parameters κ , θ , α , β and the covariates v . We set $\theta = 2$ and generated 1000 datasets and importance

Table 2. Performance of 95% confidence intervals versus Monte Carlo sample size for the simulation in § 4.2

n	% coverage probability				Median length			
	10	50	100	500	10	50	100	500
C^α	27.8	77.7	85.7	93.3	0.34	0.96	1.08	1.16
$\tilde{C}^\alpha$	99.6	98.1	96.9	95.2	2.09	1.52	1.40	1.22

samples in order to approximate the true coverage probabilities and the median interval lengths of various confidence intervals. Confidence intervals based on the corrected p -values maintain a coverage probability of at least $1 - \alpha$ without becoming excessively long, while those based on uncorrected p -values can have much lower coverage probabilities. As the number of importance samples increases, the two confidence intervals begin to agree. They also roughly agree with the confidence intervals given by the approximate conditional inference methods reviewed in [Brazzale & Davison \(2008\)](#) and implemented in the cond R package ([Brazzale, 2005](#)), which had coverage probabilities for this simulation in the range of 94.5–94.7% and median lengths of 1.15–1.16, depending on the method of approximation. The conservatism of exact inference for discrete problems ([Brazzale & Davison, 2008](#)) is not an issue here, because the typical dataset in this simulation gives a conditional distribution of $t(X)$ with $\sim 10^{200}$ distinct values; for all practical purposes, the distribution is continuous.

Comparisons with other exact methods are not available, because existing enumeration software ([StataCorp, 2009](#); [Cytel, 2010](#)) are unable to enumerate the state space or generate direct samples with a few gigabytes of memory, and the corresponding Markov chain Monte Carlo methods ([Zamar et al., 2007](#)) often fail to adequately sample the state space, even after many hours of computing. Classical, unconditional, asymptotic confidence intervals also behave poorly, because of the large number of nuisance parameters. For this simulation, 95% Wald intervals had only 82.2% coverage probability with a median length of 1.24, emphasizing the utility of conditional inference.

5. DISCUSSION

The practical benefits of using $\tilde{p}_c$ over p_c are clear. Valid p -values are always crucial for multiple testing adjustments. But even for individual tests, $\tilde{p}_c$ protects against false rejections resulting from high and possibly undiagnosed variability in the importance sampling approximations. There is almost no computational penalty for using the corrections, and little or no loss of power for well-behaved importance sampling algorithms. These advantages extend to confidence intervals constructed by inverting $\tilde{p}_c$.

The corrections are designed to improve the interpretability of tests and intervals, but they are not designed to improve the accuracy of the p -value approximation. If one is interested in approximating the probability of the event in (1), but this probability will not be interpreted as a p -value, then p_c may be preferable to $\tilde{p}_c$. The corrections can play a diagnostic role, since large differences between $\tilde{p}_c$ and p_c indicate convergence failure for at least one of the approximations, but close agreement between p_c and $\tilde{p}_c$ need not indicate convergence. All four p -value approximations can be in close agreement but still far from p .

In many cases, it may be sensible to report both $\tilde{p}_c$ and p_c . Reporting $\tilde{p}_c$ allows for hypothesis tests that respect the nominal level and approximate the power characteristics of p . Reporting p_c , along with further approximations of the Monte Carlo standard error, provides information about the numerical value of p . The poor performance of q_1 in § 4.1 shows that p_c cannot be

meaningfully corrected by relying on approximate standard errors. To ensure the nominal level of tests and intervals, $\tilde{p}_c$ is required.

ACKNOWLEDGEMENT

This work was supported in part by the U.S. National Science Foundation and the U.S. National Institutes of Health, partly while the author was at the Department of Statistics at Carnegie Mellon University. The author thanks Stuart Geman, Jeffrey Miller, Lee Newberg, an associate editor and a referee for helpful comments.

APPENDIX

Proposal distributions

The simulation example in §4.1 uses conditional inference, so the target and proposal distributions for each dataset i can depend on the observed values $V = (V_1, \dots, V_m)$ and the fact that there are r one-labels and $m - r$ zero labels, but cannot depend on the observed pairing of labels and values. We suppress the dependence on i in the notation. Let $J = (J_1, \dots, J_m)$ be a permutation that makes V nonincreasing, i.e., $V_{J_1} \geq \dots \geq V_{J_m}$. Choose a random permutation $\Pi = (\Pi_1, \dots, \Pi_m)$ according to the distribution

$$\text{pr}(\Pi = \pi) = \frac{\exp\{\theta \sum_{i=1}^r \mathbb{1}(\pi_i \leq r)\}}{r!(m-r)! \sum_{k=0}^m \binom{r}{k} \binom{m-r}{r-k} \exp(\theta k)} \quad (\theta \in \mathbb{R}; r = 0, \dots, m), \quad \binom{a}{b} = \frac{a!}{b!(a-b)!},$$

where the binomial coefficients are defined to be zero if $a < 0$, $b < 0$ or $a < b$. Leaving V in the original observed order and permuting L so that $L_{J_{\Pi_1}} = \dots = L_{J_{\Pi_r}} = 1$ and $L_{J_{\Pi_{r+1}}} = \dots = L_{J_{\Pi_m}} = 0$ gives a random pairing of values with labels. The case $\theta = 0$ is the uniform distribution over permutations, i.e., the target null conditional distribution. We used $\theta = 3$ for the proposal distribution that assigns higher probability to those permutations that tend to match the label one with larger values.

The simulation example in §4.2 also uses conditional inference, so the target and proposal can depend on the observed margins, r and c , of the matrix X . For the target distribution P_θ defined in (10), we create the proposal Q_θ by modifying the importance sampling algorithm described in the author's 2009 unpublished preprint, arXiv:0906.1004v1, which is designed for the uniform distribution, i.e., $\theta = 0$. Specifically, we use equation (3.12a) of that preprint, and we include the additional weights $\{\exp(\theta v_{ij})\lambda_{ij}\}^{b_i}$ inside the product in equation (3.6) of that preprint when sampling from the j th column, where λ is a solution to the system of equations

$$\lambda_{ij} = (N - j) \left/ \left[M \sum_{\ell=j+1}^N \exp(\theta v_{i\ell}) \left\{ \sum_{k=1}^M \lambda_{kj} \exp(\theta v_{k\ell}) \right\}^{-1} \right] \right. \quad (i = 1, \dots, M; j = 1, \dots, N - 1).$$

Proofs

We prove Theorem 1 with a succession of lemmas. The notation and conventions follow §2 with the following additions. We use E to denote expectation. For any sequence $y = (y_0, \dots, y_n)$ and any $k \in \{0, \dots, n\}$, let $y^k = (y_k, y_1, \dots, y_{k-1}, y_0, y_{k+1}, \dots, y_n)$, which is the sequence obtained by swapping the 0th element and the k th element in the original sequence y . Let Π denote a random permutation chosen uniformly from $\mathcal{M}$ and chosen independently of Z . Let $U = Z^{(\Pi)}$. If $\pi \in \mathcal{M}$ is a fixed permutation, then Π and $\Pi^{(\pi)}$ have the same distribution, which means U and $U^{(\pi)}$ have the same distribution. In particular, U and U^k have the same distribution. Lemma A1 is the key algebraic inequality.

LEMMA A1. *For all $t_0, \dots, t_n \in [-\infty, \infty]$ and all $\alpha, w_0, \dots, w_n \in [0, \infty]$,*

$$\sum_{k=0}^n w_k \mathbb{1} \left\{ \sum_{i=0}^n w_i \mathbb{1}(t_i \geq t_k) \leq \alpha \right\} \leq \alpha.$$

Proof. Let H denote the left side of the desired inequality. We can assume that $H > 0$, since the statement is trivial otherwise. The pairs $(t_0, w_0), \dots, (t_n, w_n)$ can be reordered without affecting the value of H , so we can assume that $t_0 \geq \dots \geq t_n$. This implies that $\sum_{i=0}^n w_i \mathbb{1}(t_i \geq t_k)$ is increasing in k and that there exists a k^* defined as the largest k for which $\sum_{i=0}^n w_i \mathbb{1}(t_i \geq t_k) \leq \alpha$. So, $H = \sum_{k=0}^{k^*} w_k = \sum_{i=0}^{k^*} w_i \mathbb{1}(t_i \geq t_{k^*}) \leq \alpha$. $\square$

LEMMA A2. For any nonnegative, measurable function f on S^{n+1} ,

$$E\{f(Z)\} = E \left\{ (n+1)^{-1} \sum_{k=0}^n w(Y_k) f(Y^k) \right\}.$$

Proof. Recall that $Z_i = Y_i \sim Q$ for $i \neq 0$ and that $Z_0 = X \sim P$. A change of variables from X to Y_0 gives $E\{f(Z)\} = E\{w(Y_0)f(Y)\}$. Since the distribution of Y is invariant to permutations, we have $E\{w(Y_0)f(Y)\} = E\{w(Y_k)f(Y^k)\}$ for each $k = 0, \dots, n$, which means that $E\{f(Z)\} = E\{w(Y_0)f(Y)\} = (n+1)^{-1} \sum_{k=0}^n E\{w(Y_k)f(Y^k)\}$. Moving the sum inside the expectation completes the proof. $\square$

LEMMA A3. Let P_Z and P_U denote the distributions of Z and U , respectively, over S^{n+1} . Then $P_Z \ll P_U$ and

$$\frac{dP_Z}{dP_U}(u) = \frac{(n+1)w(u_0)}{\sum_{j=0}^n w(u_j)}$$

almost surely.

Proof. Let g be any nonnegative, measurable function on S^{n+1} . It is enough to show that

$$E\{g(Z)\} = E \left\{ \frac{(n+1)w(U_0)}{\sum_{j=0}^n w(U_j)} g(U) \right\}. \quad (\text{A1})$$

For any $\pi \in \mathcal{M}$, there is a unique inverse permutation $\pi^{-1} \in \mathcal{M}$ with $\pi_{\pi_i}^{-1} = \pi_{\pi_i^{-1}} = i$, for each $i = 0, \dots, n$. Comparing with the proof of Lemma A2, for any nonnegative, measurable function f , we have

$$E\{f(Z^{(\pi)})\} = E\{f(Y^{(\pi)})w(Y_0)\} = E\{f(Y)w(Y_{\pi_0^{-1}})\}, \quad (\text{A2})$$

where the first equality is a change of variables from Z to Y and the second equality follows from the fact that the distribution of Y is permutation invariant, so, in particular, Y and $Y^{(\pi^{-1})}$ have the same distribution. Using (A2) gives

$$\begin{aligned} E\{f(U)\} &= \frac{1}{(n+1)!} \sum_{\pi \in \mathcal{M}} E\{f(U) \mid \Pi = \pi\} = \frac{1}{(n+1)!} \sum_{\pi \in \mathcal{M}} E\{f(Z^{(\pi)})\} \\ &= \frac{1}{(n+1)!} \sum_{\pi \in \mathcal{M}} E\{f(Y)w(Y_{\pi_0^{-1}})\} = \frac{1}{(n+1)!} \sum_{j=0}^n \sum_{\substack{\pi \in \mathcal{M}: \\ \pi_0^{-1}=j}} E\{f(Y)w(Y_j)\} \\ &= \frac{1}{(n+1)!} \sum_{j=0}^n n! E\{f(Y)w(Y_j)\} = E \left\{ \frac{\sum_{j=0}^n w(Y_j)}{n+1} f(Y) \right\}. \end{aligned} \quad (\text{A3})$$

Applying (A3) to the function $f(u) = (n+1)g(u)w(u_0)/\sum_{j=0}^n w(u_j)$ gives

$$E \left\{ \frac{(n+1)w(U_0)}{\sum_{j=0}^n w(U_j)} g(U) \right\} = E \left\{ \frac{\sum_{j=0}^n w(Y_j)}{n+1} \frac{(n+1)w(Y_0)}{\sum_{j=0}^n w(Y_j)} g(Y) \right\} = E\{w(Y_0)g(Y)\} = E\{g(Z)\},$$

where the last equality is a change of variables as in (A2). This gives (A1) and completes the proof. $\square$

LEMMA A4. For any nonnegative, measurable function f on S^{n+1} ,

$$E\{f(Z)\} = E\left\{\sum_{k=0}^n \frac{w(U_k)}{\sum_{j=0}^n w(U_j)} f(U^k)\right\}.$$

Proof. Changing variables from Z to U and using Lemma A3 gives

$$E\{f(Z)\} = E\left\{\frac{(n+1)w(U_0)}{\sum_{j=0}^n w(U_j)} f(U)\right\}. \quad (\text{A4})$$

Since the distribution of U is invariant to permutations, we have

$$E\left\{\frac{(n+1)w(U_0)}{\sum_{j=0}^n w(U_j)} f(U)\right\} = E\left\{\frac{(n+1)w(U_k)}{\sum_{j=0}^n w(U_j)} f(U^k)\right\} \quad (k = 0, \dots, n). \quad (\text{A5})$$

Combining (A4) and (A5) and averaging over k gives

$$E\{f(Z)\} = E\left\{\frac{(n+1)w(U_0)}{\sum_{j=0}^n w(U_j)} f(U)\right\} = \frac{1}{n+1} \sum_{k=0}^n E\left\{\frac{(n+1)w(U_k)}{\sum_{j=0}^n w(U_j)} f(U^k)\right\}.$$

Moving the sum inside the expectation and cancelling the $(n+1)$ terms completes the proof. $\square$

Proof of Theorem 1 for $c = 1$. Applying Lemma A2 to the function $f(z) = \mathbb{1}\{\tilde{p}_1(z) \leq \alpha\}$ gives

$$\begin{aligned} \text{pr}\{\tilde{p}_1(Z) \leq \alpha\} &= E[\mathbb{1}\{\tilde{p}_1(Z) \leq \alpha\}] = E\left[(n+1)^{-1} \sum_{k=0}^n w(Y_k) \mathbb{1}\{\tilde{p}_1(Y^k) \leq \alpha\}\right] \\ &= E\left((n+1)^{-1} \sum_{k=0}^n w(Y_k) \mathbb{1}\left[(n+1)^{-1} \sum_{i=0}^n w(Y_i) \mathbb{1}\{t(Y_i, Y) \geq t(Y_k, Y)\} \leq \alpha\right]\right). \end{aligned}$$

The quantity inside the final expectation is always at most α , which follows from Lemma A1 by taking $t_\ell = t(Y_\ell, Y)$ and $w_\ell = w(Y_\ell)/(n+1)$ for each $\ell = 0, \dots, n$. $\square$

Proof of Theorem 1 for $c = 2$. Applying Lemma A4 to the function $f(z) = \mathbb{1}\{\tilde{p}_2(z) \leq \alpha\}$ gives

$$\begin{aligned} \text{pr}\{\tilde{p}_2(Z) \leq \alpha\} &= E[\mathbb{1}\{\tilde{p}_2(Z) \leq \alpha\}] = E\left[\sum_{k=0}^n \frac{w(U_k)}{\sum_{j=0}^n w(U_j)} \mathbb{1}\{\tilde{p}_2(U^k) \leq \alpha\}\right] \\ &= E\left(\sum_{k=0}^n \frac{w(U_k)}{\sum_{j=0}^n w(U_j)} \mathbb{1}\left[\sum_{i=0}^n \frac{w(U_i)}{\sum_{j=0}^n w(U_j)} \mathbb{1}\{t(U_i, U) \geq t(U_k, U)\} \leq \alpha\right]\right). \end{aligned}$$

The quantity inside the final expectation is always at most α , which follows from Lemma A1 by taking $t_\ell = t(U_\ell, U)$ and $w_\ell = w(U_\ell)/\{\sum_{j=0}^n w(U_j)\}$ for each $\ell = 0, \dots, n$. $\square$

REFERENCES

- BESAG, J. & CLIFFORD, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* **76**, 633–42.
- BOLVIKEN, E. & SKOVLUND, E. (1996). Confidence intervals from Monte Carlo tests. *J. Am. Statist. Assoc.* **91**, 1071–8.
- BRAZZALE, A. R. (2005). *hoa: an R package bundle for higher order likelihood inference*. *Rnews* **5**, 20–7. ISSN 609–3631, ftp://cran.r-project.org/doc/Rnews/Rnews_2005-1.pdf.
- BRAZZALE, A. R. & DAVISON, A. C. (2008). Accurate parametric inference for small samples. *Statist. Sci.* **23**, 465–84.
- COX, D. R. (1958). The regression analysis of binary sequences. *J. R. Statist. Soc. B* **20**, 215–42.
- CYTEL (2010). *LogXact 9*. Cambridge, MA: Cytel Inc.
- DAVISON, A. C. & HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge, UK: Cambridge University Press.

- GARTHWAITE, P. H. & BUCKLAND, S. T. (1992). Generating Monte Carlo confidence intervals by the Robbins–Monro process. *J. R. Statist. Soc. C* **41**, 159–71.
- GREEN, P. J. (1992). Discussion of the paper by Geyer and Thompson. *J. R. Statist. Soc. B* **54**, 683–4.
- HESTERBERG, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37**, 185–94.
- LEHMANN, E. L. & ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd edn. New York: Springer.
- LIU, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.
- MEHTA, C. R. & PATEL, N. R. (1995). Exact logistic regression: theory and examples. *Statist. Med.* **14**, 2143–60.
- RASCH, G. (1961). On general laws and the meaning of measurement in psychology. In *Proc. 4th Berkeley Symp. Math. Statist. Prob.: Prob. Theory*, vol. 4, Ed. J. Neyman. Berkeley, CA: University of California Press, pp. 321–34.
- STATA CORP (2009). *Stata Statistical Software: Release 11*. College Station, TX: StataCorp LP.
- WESTFALL, P. H. & YOUNG, S. S. (1993). *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*. New York: Wiley-Interscience.
- ZAMAR, D., MCNENEY, B. & GRAHAM, J. (2007). Elrm: software implementing exact-like inference for logistic regression models. *J. Statist. Software* **21**, 1–18.

[Received April 2011. Revised October 2011]