

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Karavarsamis, N;Guillera-Arroita, G;Huggins, RM;Morgan, BJT

Title:

The score test for the two-sample occupancy model

Date:

2020-03-01

Citation:

Karavarsamis, N., Guillera-Arroita, G., Huggins, R. M. & Morgan, B. J. T. (2020). The score test for the two-sample occupancy model. Australian and New Zealand Journal of Statistics, 62 (1), pp.95-115. <https://doi.org/10.1111/anzs.12288>.

Persistent Link:

<https://hdl.handle.net/11343/275595>

# Author Manuscript

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/anzs.12288](https://doi.org/10.1111/anzs.12288)

This article is protected by copyright. All rights reserved

# The score test for the two-sample occupancy model

N. Karavarsamis<sup>1,2\*</sup>, G. Guillera-Aroita<sup>3</sup>, R.M. Huggins<sup>1</sup>, B.J.T. Morgan<sup>4</sup>

*Department of Mathematics and Statistics, La Trobe University*

## Summary

The score test statistic from the observed information is easy to compute numerically. Its large sample distribution under the null hypothesis is well known and is equivalent to that of the score test based on the expected information, the likelihood-ratio test and the Wald test. However, several authors have noted that under the alternative hypothesis this no longer holds and in particular the score statistic from the observed information can take negative values. We extend the anthology on the score test to a problem of interest in ecology when studying species occurrence. This is the comparison of two zero-inflated binomial random variables from two independent samples under imperfect detection. An analysis of eigenvalues associated with the score test in this setting assists in understanding why using the observed information matrix in the score test can be problematic. We demonstrate through a combination of simulations and theoretical analysis that the power of the score test calculated under the observed information decreases as the populations being compared become more dissimilar. In particular, the score test based on the observed information is inconsistent. Finally, we propose a modified rule that rejects the null hypothesis when the score statistic is computed using the observed information is negative or is larger than the usual chi-square cut-off. In simulations in our setting this has power that is comparable to the Wald and likelihood ratio tests and consistency is largely restored. Our new test is easy to use and inference is possible. Supplementary material for this article is available online as per journal instructions.

**Key words:** eigenvalues; observed information; occupancy modelling; power of hypothesis test; zero-inflation

## 1. Introduction

The score statistic computed using the observed information matrix has several practical advantages. It only requires the computation of estimates under the null hypothesis and the use of the observed rather than the expected information does not require the computation

\*Author to whom correspondence should be addressed.

<sup>1</sup> School of Mathematics and Statistics, The University of Melbourne, Victoria 3010, Australia.

<sup>2</sup> Department of Mathematics and Statistics, La Trobe University, Victoria 3086, Australia.

Email: n.karavarsamis@latrobe.edu.au

<sup>3</sup> School of Biosciences, The University of Melbourne, Victoria 3010, Australia.

<sup>4</sup> School of Mathematics, Statistics & Actuarial Science, University of Kent, UK.

**Acknowledgment.** This work is funded by a Discovery Early Career Research Award from the Australian Research Council and a Merit Fellowship from the University of Kent.

of possibly complex expectations. A comparison of the observed and expected information matrices by Efron & Hinkley (1978) for the one-parameter case supported the use of the observed information over the expected Fisher information in the likelihood-ratio test. They claim that the inverse of the observed information matrix leads to estimates of variance which are closer to the data than those resulting from the expected Fisher information, as well as agreeing more closely with Bayesian and even with fiducial analyses. They also make the point that the observed information can be easier to compute, especially for complicated cases such as that given in Cox (1958). They give detailed evaluations of numerical results for moderate sample sizes, in addition to the general asymptotic theory. Recall that the observed information can be calculated via a numerical approximation to the Hessian. This is regularly done to estimate standard errors in complex statistical modelling Cowen et al. (2017). In addition, McCrea & Morgan (2011) evaluated score tests using the expected information, adopting a property of multinomial distributions, and used numerical approximations to first-order derivatives, rather than evaluate the expected information directly.

The asymptotic distribution of the score statistic under the null hypothesis is  $\chi^2$  and the null hypothesis is thus rejected when the score statistic is too large. (e.g. Freedman 2007, page 293). There are reservations on the use of the score statistic based on the observed information matrix. Freedman (2007) asks ‘How can the score test be inconsistent?’. It has been observed that this version of the score statistic can be negative. Examples of hypothesis tests that return negative score test statistics for the observed information matrix are given in Lawrance (1987); Yang & Abeyasinghe (2003); Godfrey & Orme (2001). Situations that returned negative eigenvalues of the observed information matrix are given in Catchpole & Morgan (1996); Storer, Wacholder & Breslow (1983) and Hosking (1984). Morgan, Palmer & Ridout (2007) examined the score statistic for the zero-inflated Poisson (ZIP) model and showed how it could be negative (Figure 1 of Morgan, Palmer & Ridout 2007) when testing for zero inflation. Motivated by this result, Freedman (2007) went on to further evaluate the use of the score statistic and in his Section 7 related inconsistency to negative eigenvalues of the observed information matrix. They gave one practical resolution that evaluated the observed information matrix at the full maximum likelihood estimates. However, this requires computation of the unrestricted maximum likelihood estimates which negates the computational advantage of the score test. Verbeke & Molenberghs (2007) summarise related issues with the score test.

We re-examine this problem in the context of comparing two binomial proportions under imperfect detection and subsequently propose a solution. This setting is very common in ecology, and occurs when comparing the proportion of sites occupied by a species in two regions or times, referred to as two ‘samples’ throughout (Guillera-Aroita & Lahoz-Monfort 2012). Detection of wildlife sightings is generally imperfect, and species may remain

completely undetected at sites they occupy. A common approach to account for imperfect detectability is to conduct several survey visits to sites. The statistical model to describe these data is based on a zero-inflated Binomial (ZIB) distribution (MacKenzie et al. 2002). The two-sample occupancy model involves the probabilities of occupancy ( $\psi_1$  and  $\psi_2$ ) and detection ( $p_1$  and  $p_2$ ) of the species in each of the two regions (or times). The null hypothesis of interest is that the two occupancy probabilities are the same:  $H_0 : \psi_1 = \psi_2$ . Thus under the null hypothesis there are three parameters and four under the alternative hypothesis.

In this setting, we examine the score test statistic from observed information, both with simulations and analytically. For this model we can compute the expected information matrix and in practice the score statistic based on this expectation would be of course preferred. However, our interest is more theoretical and is in explaining the anomalous behaviour of the score test statistic based on the observed information matrix. In particular we can examine the expected value of the observed information matrix computed under  $H_0$  when  $H_1$  is true and examine its eigenvalues. Using simulations we show that the commonly used rejection rule that rejects the null hypothesis if the value of the score statistic using the observed information matrix is greater than the usual  $\chi^2_1$  cut-off can have power that is much lower than that of the likelihood-ratio test (LRT), the Wald test and the score test based on the expected information matrix (Engle 1984; Bickel & Doksum 2001; Casella & Berger 2002; Lehmann & Romano 2005). Moreover, we also see that the power decreases as the distance between  $\psi_1$  and  $\psi_2$  increases. However, again in simulations we show that, if the rejection rule based on the score test statistic from observed information is modified to reject if its value is greater than the usual  $\chi^2_1$  or when it is negative, the power is restored. It outperforms the LRT even for highly unbalanced samples, and the Wald and score test from expected information for unbalance up to at least about 75%. A reduction in power is observed for all tests for highly unbalanced samples. Under the null hypothesis the score statistic is positive so the new test does not alter the size (level of significance) of the test. This is re-assuring as the observed score test may be obtained readily from numerical approximation and is computationally fast. Whereas the expected score test requires derivations of expectations which in general may be difficult or not available in closed form. This is not the case in our relatively simple setting, but this allows us to compute the expected score test statistic for our comparisons. Our conclusion is that, with this new rejection rule, the score test based on the observed information matrix is indeed useful and can still be used, making inference always possible even when the statistic is negative.

We organise our paper as follows. In Section 2, we define the score test. In Section 3, we conduct a simulation study to compare the likelihood-ratio test, the Wald test and the two versions of the score test. Our results confirm that the score test based on the observed information matrix can be used, making inference always possible even when the statistic is negative. This article is protected by copyright. All rights reserved.

power, with power decreasing as the alternative hypothesis moves away from the null. In Section 4 (and Appendix I) we examine the eigenvalues of the observed information matrix and its analytically computed expectation, under the alternative hypothesis. We see that, as the alternative hypothesis moves away from the null, the median over the simulations of the smallest eigenvalue of the observed information matrix becomes negative and so does the smallest eigenvalue of the expectation of this matrix. This means that the observed information matrix is indefinite and hence quadratic forms can be positive or negative. This gives rise to negative values of the score statistic. In Section 5, we introduce our modified rejection rule and evaluate its performance in simulations. The paper ends with discussion in Section 6.

## 2. The score test for the species occupancy model

Occupancy modelling is useful for large-scale monitoring programs where possible differences in occupancy between two samples (or two regions) may be of interest. For example, occupancy decline is of particular interest for species conservation and occupancy increase for invasive control programs (Guillera-Arroita & Lahoz-Monfort 2012; MacKenzie et al. 2017). As defined here, samples require multiple visits to a site where the population of the species under consideration is closed over the visits. Here the samples themselves vary spatially (Guillera-Arroita 2017).

In ecology, data from ‘occupancy models’ are used to estimate the probability  $\psi$  that a species is present at a site (or equivalently, the proportion of the landscape that it occupies), while accounting for imperfect detection (MacKenzie et al. 2002; Guillera-Arroita 2017). To allow for detectability, detection/non-detection data are usually collected over repeated visits to the sites. A second, nuisance, parameter, the detection probability  $p$ , is included in the model, and is the probability of detecting the species in a survey visit at a site where it is present.

In the one sample occupancy model, detections are described as independent Bernoulli trials. Let  $Y_i$  be the number of detections over  $K$  visits at site  $i$ ,  $i = 1, \dots, N$ . Then

$$\begin{aligned}\Pr(Y_i = 0) &= 1 - \psi + \psi(1 - p)^K \\ \Pr(Y_i = y_i) &= \psi p^{y_i} (1 - p)^{K - y_i}, \quad y_i = 1, 2, \dots, K \quad i = 1, 2, \dots, N.\end{aligned}\quad (1)$$

As the species is absent from some sites, the number of detections follows a zero-inflated binomial distribution (ZIB), with the level of zero-inflation set by  $1 - \psi$ . Let us now suppose that we want to compare species occupancy ( $\psi$ ) across two samples. Typically these two samples may represent two different geographical regions. We label the two regions 1 and 2

with associated occupancy probabilities  $\psi_1, \psi_2$  and detection probabilities  $p_1$  and  $p_2$ . First consider the likelihood for a single region  $j$ . The probability of detecting the species at least once at an occupied site from region  $j$  is  $\theta_j = 1 - (1 - p_j)^{K_j}$ . Let  $s_{d_j}$  denote the number of sites where any individuals were detected and  $d_j$  the total number of detections, for region  $j$ . Note that for the single region case  $d = \sum_{i=1}^N Y_i$ . Then, following Guillera-Arroita, Ridout & Morgan (2010), the likelihood component for region  $j$  is

$$L_j = \{\psi_j^{s_{d_j}} p_j^{d_j} (1 - p_j)^{K_j s_{d_j} - d_j}\} (1 - \psi_j \theta_j)^{N_j - s_{d_j}}, \quad j = 1, 2.$$

The joint likelihood is the product of the likelihoods of the two regions  $j = 1, 2$ , under the assumption of independence. The score equations for region  $j$  are given by

$$S_{j1} = s_{d_j}/\psi_j - (N_j - s_{d_j})\theta_j/(1 - \psi_j \theta_j) = 0,$$

and

$$S_{j2} = d_j/p_j - (s_{d_j} K_j - d_j)/(1 - p_j) - (N_j - s_{d_j})\psi_j K_j (1 - p_j)^{K_j - 1}/(1 - \psi_j \theta_j) = 0,$$

$j = 1, 2$  for occupancy and detection.

Let  $\boldsymbol{\theta} = (\psi_1, p_1, \psi_2, p_2)^\top$  be the vector of model parameters for the two-sample model. The likelihood for the two-sample model is the product,  $L(\boldsymbol{\theta}) = L_1 \times L_2$  of the two single-sample likelihoods assuming independence, and the resulting unconstrained score function is  $\mathbf{S}(\boldsymbol{\theta}) = \partial \log L(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = (S_{11}, S_{12}, S_{21}, S_{22})^\top$ . Also let  $\mathbf{J}(\boldsymbol{\theta}) = -\partial \mathbf{S}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^\top = -\mathbf{S}'(\boldsymbol{\theta})$  be the observed information matrix for samples of size  $N_1$  and  $N_2$ . The usual Neyman-Pearson likelihood theory gives the large-sample null distribution (i.e.  $\chi_1^2$ ) of the score statistic using both the expected (Fisher) and observed information matrices.

Our interest is in the behaviour of the score test based on the observed information matrix, so we define this formally. The definitions of the likelihood-ratio, Wald test and the score test based on the expected information matrix are well known and omitted.

We denote the true parameter values by  $\boldsymbol{\theta}_T = (\psi_1, p_1, \psi_2, p_2)^\top$ . Consider  $H_0 : \psi_1 = \psi_2 = \psi$ .

Let  $\boldsymbol{\theta}_S = (\psi, p_1, p_2)^\top$  and

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

This article is protected by copyright. All rights reserved.

a change of coordinates matrix. Then  $\mathbf{M}\boldsymbol{\theta}_S = (\psi, p_1, \psi, p_2)$  and  $\mathbf{S}_0(\boldsymbol{\theta}_S) = \mathbf{M}^\top \mathbf{S}(\mathbf{M}\boldsymbol{\theta}_S)$  is the score function under  $H_0$ . When the alternative hypothesis is true, the model under  $H_0$  is misspecified (White 1982). Therefore we suppose throughout that  $\boldsymbol{\theta}'_S$  (the restricted model parameters to be estimated) is implicitly defined to satisfy  $\mathbf{E}_{\boldsymbol{\theta}_T}(\mathbf{S}_0(\boldsymbol{\theta}'_S)) = \mathbf{0}$  and under  $H_0$  the maximum likelihood estimate  $\hat{\boldsymbol{\theta}}'_S$  is a solution of  $\mathbf{S}_0(\boldsymbol{\theta}_S) = \mathbf{0}$  (van der Vaart 2000; Davison 2003). Then the score statistic defined in terms of the observed information is

$$\mathbf{T}_O(\hat{\boldsymbol{\theta}}'_S) = \mathbf{S}(\mathbf{M}\hat{\boldsymbol{\theta}}'_S)^\top \mathbf{J}(\mathbf{M}\hat{\boldsymbol{\theta}}'_S)^{-1} \mathbf{S}(\mathbf{M}\hat{\boldsymbol{\theta}}'_S), \quad (2)$$

and replacing  $\mathbf{J}(\mathbf{M}\hat{\boldsymbol{\theta}}'_S)$  with  $\mathbf{E}_{\mathbf{M}\boldsymbol{\theta}_S}(\mathbf{M}\mathbf{J}(\boldsymbol{\theta}'_S))$  evaluated at  $\boldsymbol{\theta}'_S = \hat{\boldsymbol{\theta}}'_S$  gives the score test statistic from expected information  $\mathbf{T}_E(\hat{\boldsymbol{\theta}}'_S)$ . The asymptotic distribution of both versions of the score statistic under the null hypothesis is a chi-square distribution with one degree of freedom so that the usual tests are to reject  $H_0$  if  $\mathbf{T}_O(\hat{\boldsymbol{\theta}}'_S) > \chi_1^2(1 - \alpha)$  or  $\mathbf{T}_E(\hat{\boldsymbol{\theta}}'_S) > \chi_1^2(1 - \alpha)$ , where  $1 - \alpha$  indicates the upper tail probability of the  $\chi^2$  distribution.

### 3. Performance of competing tests

In this section, we conduct a simulation study to compare the powers of the Wald test, likelihood-ratio test and the two versions of the score test (based on the expected and observed information matrices) at particular combinations of sample sizes, number of visits and parameter values. We take  $p_1 = p_2 = 0.5$ ,  $K_1 = K_2 = 3$  visits and  $N_1 = N_2 = 50$  sites and refer throughout to this combination of parameter values as the ‘reference framework’. We consider  $\psi_1 = 0.8$  and  $\psi_2 = (1 - R)\psi_1$ , where  $0 < R < 1$  is the proportional decline in species occurrence, with varying  $R \in \{0, \dots, 0.9\}$  in steps of 0.025 to give different  $\psi_2$  values. Together with these combinations for  $\psi_1$  and  $\psi_2$ , these amount to 80 different parameter combinations for the reference framework. For each parameter combination, 50000 data sets were simulated. R code is available as Supplementary Material. Further simulation studies for a number of parameter combinations are presented in Appendix IV. Additional scenarios include values for  $N_1, N_2 \in \{20, 100, 200\}$  with combinations of survey occasions  $K_1, K_2 \in \{3, 15, 30\}$  and detection probabilities  $p_1, p_2 \in \{0.3, 0.5, 0.75\}$ .

Results, including power, are calculated from simulations where the numerical optimization did not fail and estimates produced were within the parameter space  $(0, 1)$  for  $\psi$  and  $p$ . Failure occurs when the optimization process does not converge to any value or the Hessian is not invertible, estimates are outside the boundaries, or when there are no detections. Non-existence of estimates and the violation of boundary conditions affects all methods similarly. Boundary conditions have been studied for the likelihood in the literature (e.g. White 1982; White 1984; Gaudin et al. 2004; Ridout & Morgan 2010). Also see

170 Karavarsamis et al. (2013) who found a failure rate of about 20%. This has not been formally  
 171 explored in the two sample case. The score tests based on the observed information cannot be  
 172 produced when the numerical Hessian cannot be computed or is singular. Thus it is possible  
 173 that the expected information exists but the observed information matrix does not. In this  
 174 case, the Wald test and score test based on the expected information can be computed but the  
 175 score test based on the observed information cannot. For example, for  $\psi_1 = 0.8$  and  $\psi_2$  for  
 176 each  $R(\leq 0.9)$  and for the reference framework, in our simulations we found that LRT had  
 177 few failed simulations, whereas the Wald, score (from observed information) and score (from  
 178 expected information) had approximately the same failure rate from this cause ( $\leq 3\%$ ).

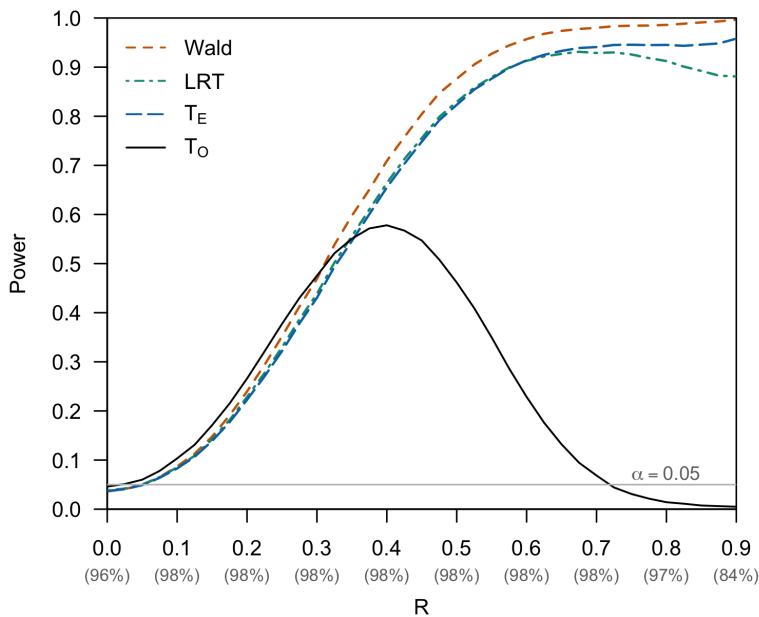


Figure 1. Power of LRT, Wald, score test for the expected ( $T_E$ ) and observed ( $T_O$ ) information matrix, with  $\psi_1 = 0.8$  under the reference framework for 50000 simulations for each value of  $R$ . Results are displayed for simulations where numerical optimization did not fail (percentages for  $T_O$  are listed in parenthesis under the  $x$ -axis). The significance level  $\alpha$  is marked at 0.05.

179 The results (Figure 1) show that the Wald test, the likelihood-ratio test (LRT) and  
 180 the score test based on the expected information matrix ( $T_E$ ) (evaluated at the maximum  
 181 likelihood estimates under  $H_0$ ) all have similar powers that increase as  $R$  increases. However  
 182 the score test from the observed information matrix ( $T_O$ ) (also evaluated at the maximum  
 183 likelihood estimates under  $H_0$ ) shows a decline in power after  $R \approx 0.4$ . Along

the lines of Figure 1 of Morgan, Palmer & Ridout (2007), if we evaluate the relationship between the score test statistic from the expected information and the observed information, the median of their ratio becomes negative at 0.5 as  $R$  increases (Figure 2(a)). At  $\psi_1 = \psi_2$ , the null hypothesis is true with  $R = 0$ . Then the score statistics are equal and their ratio is exactly equal to 1, as seen in Figure 2(a). When the medians of the simulated values of the score statistics are plotted separately (Figure 2(b)), we see that the median of the score test statistic from the observed information reaches a maximum at 0.4, then declines and later increases. It becomes negative at 0.5. That is, at  $R \approx 0.5$  half of the values of the score statistic from the observed information are positive and half are negative. This explains why the ratio of the expected to observed is a declining function and that it becomes negative at 0.5 (Figure 2(a)). At  $\approx 0.6$  the score statistic from the observed information ( $T_O$ ; long dashed line, Figure 2(b)) reaches another turning point and begins to increase however its median remains negative. Similar results and a decline at  $\approx 0.5$  are obtained when we take  $\psi_1 = 0.4$  (the power of  $T_O$ ; solid dashed line, Figure 9, Appendix III).

When we consider only those simulations where the score statistic from the observed information is positive ( $T_O^+$ ), we find there is good agreement between the expected ( $T_E$ ) and observed ( $T_O^+$ ) score test, i.e. both accept or reject the null hypothesis for a given dataset (Table 1(a)). The differences are predominantly where the test based on the score statistic from the observed information rejects the null hypothesis and that based on the score statistic from the expected information does not. This is made clear when  $T_O$  is plotted against  $T_E$  separately for each  $R$  (Figure 8, Appendix III). As  $R$  increases the number of simulations with non-negative test scores  $T_O^+$  decreases strongly and therefore power is reduced, e.g. when  $R = 0.8$  there are 1569 (3%) positive test values of the 50000 simulated datasets (Table 1(a)). We wish to adapt the rejection rule to give a greater power, which would be reflected in a higher number of positive test values when the alternative hypothesis is true.

#### 4. Eigenvalues

When we inspect the eigenvalues of the observed information matrix in our simulations, we find that, as  $R$  increases, the median of the smallest eigenvalue (the 4th) becomes negative at  $R \approx 0.5$  (Figure 3). To further investigate this result, we compute the expectation of the observed information matrix. Now, as  $\theta_T$  is the true value,  $\hat{\theta}'_S \xrightarrow{P} \theta'_S$  and  $E_{\theta_T}(J(M\theta'_S))$  may be readily computed. This requires computing  $\theta'_S$  for a given  $\theta_T$ . Then for  $j = 1, 2$ ,  $E(s_{d_j}) = N_j \psi_{jT} \theta_{jT}$ ,  $E(d_j | s_{d_j}) = K_j s_{d_j} p_{jT} / \theta_{jT}$  and  $E(d_j) = K_j N_j \psi_{jT} p_{jT}$ . In Figure 3, we plot the eigenvalues of  $E_{\theta_T}(J(M\theta'_S))$  for the case  $\psi_1 = 0.8$  under our reference framework. We find that the smallest eigenvalue has a value of zero at  $R \approx 0.5$ . Despite the eigenvalues of  $E_{\theta_T}(J(M\theta'_S))$  being all positive for  $R < 0.5$ , the estimator and the observed

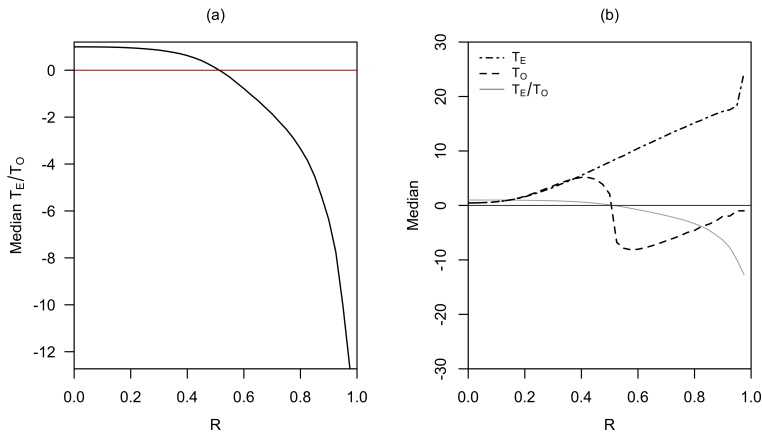


Figure 2. Score test statistics versus value of  $R$  for data simulated under the reference framework with  $\psi_1 = 0.8$ . In (a), median of the ratio between the score test statistics from the expected and observed information ( $T_E/T_O$ ). In (b), same ratio, together with the median of the score test statistics from expected information ( $T_E$ ) and from observed information ( $T_O$ ), plotted separately.

information matrix are random. Then for  $R < 0.5$  there will be some outcomes with negative eigenvalues leading to negative values of the score statistic from the observed information, as is apparent from Figure 2. That is, the observed information matrix is indefinite in this case.

The turning point at  $R = 0.4$  observed in Figure 2(b) reflects the turning point at 0.4 for the 2nd eigenvalue, as seen in Figure 3 for simulated and analytical results.

In Appendix I we examine the score statistic in more detail through the usual large-sample approximations. We give a general condition for a negative test value and apply this to the two sample occupancy model.

## 5. A modified rejection rule

We propose a new modified rejection rule based on rejecting according to being larger than the usual  $\chi^2$  critical value or when the score test statistic from the observed information is negative and examine its performance for our occupancy model. Our simulation results (Figure 4) clearly show that this new rejection rule,  $T_O^*$ , greatly improves the power of the score test from the observed information. It exceeds the power of the other tests, including the Wald and score test from expected information up to a high value of  $R$  ( $\approx 0.75$ ). Beyond  $R \approx 0.75$  our test outperforms the LRT although not the Wald or the score test from expected information. Figure 5 provides a visual display of the modified rejection rule.

Our rationale for the new improved rejection rule follows. As noted in Section 4 the usual This article is protected by copyright. All rights reserved.

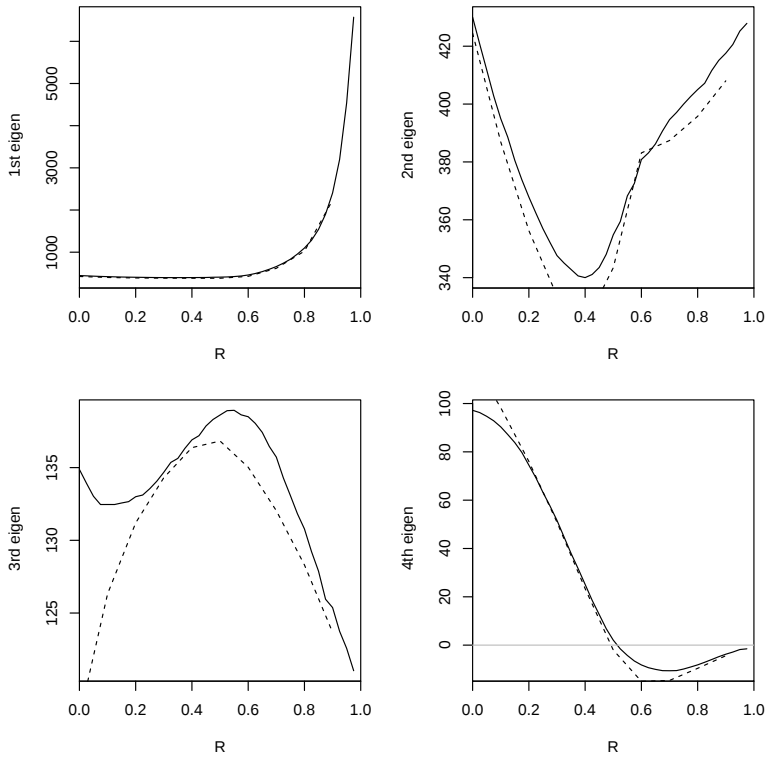


Figure 3. Eigenvalues for the observed information matrix, for different values of  $R$  under the reference framework with  $\psi_1 = 0.8$ . Solid lines are medians obtained from simulations (50000 at each value of  $R$ ). Dashed lines are eigenvalues of  $E_{\theta_T}(J(M\theta'_S))$ .

larger than the chosen  $\chi^2$  critical value. This is based on the null distribution where the observed information matrix is positive definite so that the score statistic from the observed information is positive.

However, we have seen that an indefinite information matrix is evidence that the null hypothesis is false. As demonstrated by the appearance of the negative 4th eigenvalue in Figure 3 as well from the negative values in Figures 6 and 7 in Appendix I. Because the observed information matrix may be indefinite under the alternative hypothesis, its inverse gives some positive and negative eigenvalues and the associated quadratic forms can be negative. Therefore, as these negative values only occur under the alternative hypothesis the power of the test based on the score statistic obtained using the observed information will be increased if we also reject the null hypothesis when the value of this score statistic is negative.

We find there is generally good agreement between the new modified test statistic  $T_O^*$  and  $T_O$ . This is supported by the fact that  $T_O^*$  is generally increased for greater  $R$  ( $> 0.4$ ) (Table 1(b)).

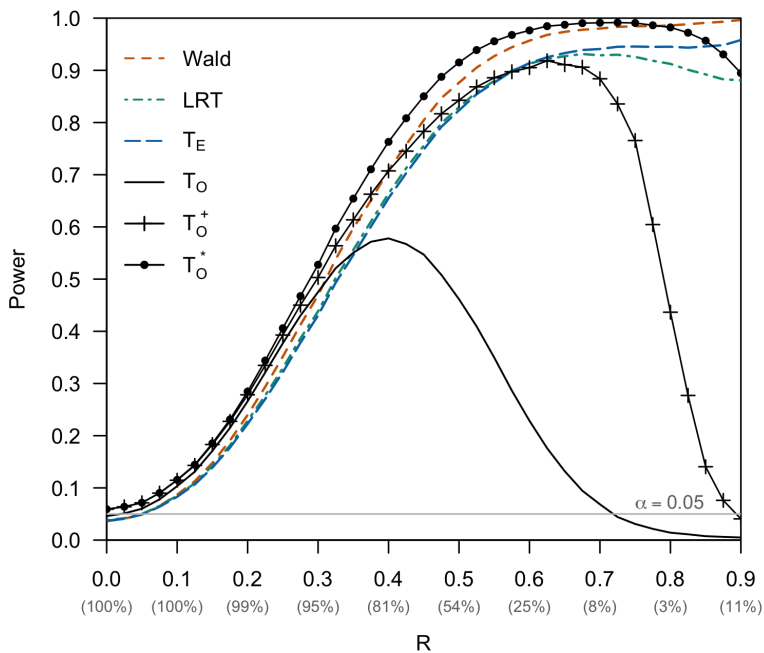


Figure 4. Power plot for  $\psi_1 = 0.8$  under the reference framework for the new modified score test based on the observed information matrix  $T_O^*$ , and for positive values of the score test based on observed information matrix  $T_O^+$ , compared to the other tests. Power is calculated from 50000 simulations for each  $R$  where numerical optimization did not fail. Percentages of simulations that returned a positive score test for observed information ( $T_O^+$ ) are listed in parenthesis under the  $x$ -axis.

Discrepancies are again where the new test based on the score statistic from the observed information  $T_O^*$  rejects the null hypothesis but that based on the score statistic from expected information  $T_E$  does not (Figure 8, Appendix III).

An analysis of the Type I error reveals our new test has the same level of significance as the positive score test from observed information, of approximately 5.9%. Furthermore, the Type I error from our new test is comparable to the Wald test (5.9%), and not much worse than the LRT and score test from expected information (4.9% for both).

Simulation results for  $\psi_1 = 0.4$  for the reference framework (Figure 9 and Table 2 in Appendix III) show similar results: the power is greatly improved when the new modified rule ( $T_O^*$ ) is used, compared to the original version of the score test from observed information ( $T_O$ ) as is consistency. We plot the power of the test statistics from simulation studies for this article is promoted by ongoing and future research in Appendix IV. We note these verify our new

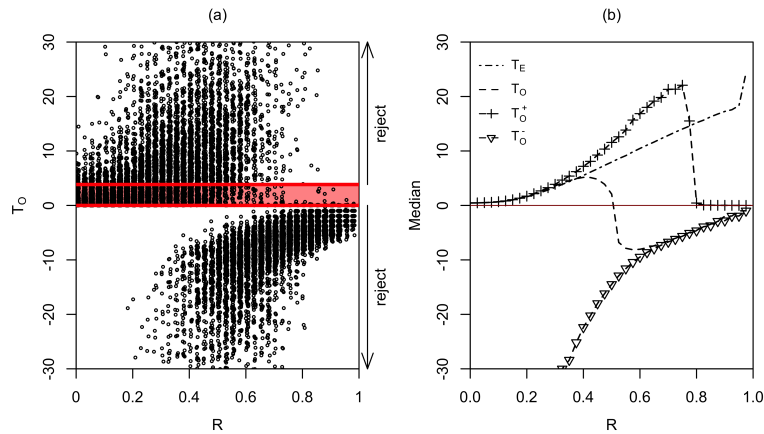


Figure 5. Visual display of the new modified rejection rule for  $\psi_1 = 0.8$  under the reference framework. In (a) score test statistic from the observed information  $T_O$  versus  $R$  (for clarity, only 500 simulations are shown for each  $R$ ). Horizontal lines at 0 and  $\chi^2_1$  for  $\alpha = 0.05$ , i.e. at 3.814, give bounds for the acceptance region. Power for each  $R$  is the proportion of simulations that lie outside the acceptance region. In (b), medians of positive ( $T_O^+$ ), negative ( $T_O^-$ ) and new modified rule ( $T_O^*$ ) values of the score statistic from the observed information, together with medians for all values ( $T_O$ ) and for the score from expected information ( $T_E$ ), obtained from 50000 simulations.

modified test results seen in the above case studies. Figures 11–16 demonstrate the tests are asymptotically equivalent, including our new modified test.

Table 1. Agreement (Agr.) for accepting/rejecting  $H_0$  between each of: in (a)  $T_O^+$  and  $T_E$ , and in (b)  $T_O^*$  and  $T_E$ . 50000 simulations are generated at each value of  $R$ , under the reference framework. The number of simulations with positive test statistics (sims) are listed in (a) and in (b) the number of positive and negative test values (sims). Discrepancy of ‘sims’ from 50000 simulations in (b) occurs when numerical optimization did not converge.

|     | $R$     | 0     | 0.1   | 0.2   | 0.3   | 0.4   | 0.5   | 0.6   | 0.7   | 0.8   | 0.9   |
|-----|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| (a) | Agr.    | 0.99  | 0.98  | 0.95  | 0.91  | 0.88  | 0.85  | 0.85  | 0.86  | 0.94  | 0.99  |
|     | $T_O^+$ |       |       |       |       |       |       |       |       |       |       |
|     | sims    | 49974 | 49948 | 49574 | 47542 | 40502 | 26986 | 12388 | 3805  | 1569  | 5317  |
| (b) | Agr.    | 0.99  | 0.98  | 0.95  | 0.91  | 0.90  | 0.91  | 0.94  | 0.95  | 0.95  | 0.96  |
|     | $T_O^*$ |       |       |       |       |       |       |       |       |       |       |
|     | sims    | 49981 | 49992 | 50000 | 49999 | 49998 | 49992 | 49982 | 49982 | 49929 | 48454 |

6. Discussion

Large-sample theory, as shown for example in Efron & Hinkley (1978), supports the use of the observed information over the expected in many settings. The score test based on the expected information matrix requires derivations of expectations which in general may be difficult to calculate or are not available in closed form. The score test based on the observed information matrix may be obtained readily from numerical approximation and is computationally fast. The two-sample occupancy model considered here is relevant to an

area of great importance in modern ecology (MacKenzie et al. 2017; Guillera-Arroita 2017). Our examination of the two sample occupancy model reveals problems consistent with those found for the zero-inflated Poisson (ZIP) model for a single parameter in Morgan, Palmer & Ridout (2007), as well as results of Freedman (2007) and comments made in Verbeke & Molenberghs (2007). In particular we note that in our application the tests are not equivalent for alternative hypotheses far from the null.

For the score test in our application the problem is that under the alternative hypothesis (i.e. when  $\theta_T \neq M\theta'_S$ ), we have seen that  $E_{\theta_T}(J(M\theta'_S))$  need not be positive definite. This was previously noted in the large-sample treatment, equations (11) and (12), of Freedman (2007). The consequence is some positive and some negative eigenvalues of the observed information matrix. As a result, the score test statistic from the observed information may be negative.

After an examination of these problems in this context we propose a new modified rejection rule. Our simulations reveal that for the two sample occupancy model the power is similar to that of the score test based on the expected information matrix and the significance level is unaffected, regardless of the sample size of the study. Type I error is unchanged. In addition consistency seems to be largely restored. We have not demonstrated theoretically that this test attains the same power as the score test based on the expected information matrix although we conjecture this is true both for the two-sample occupancy model and in general. Simulations in Appendix IV demonstrate our new modified test is asymptotically equivalent to the Wald, LRT and score test based on expected and observed information matrix. We find that there is good agreement between our new test  $T_O^*$  and the score test based on the expected information  $T_E$  when it comes to rejecting or accepting the null hypothesis.

## Appendix I

To examine the score statistic in more detail we adopt a classical approach. Note that  $\theta_S = M\theta_T$  and  $S_0(\theta_S) = M^\top S(M\theta_S)$ . We first outline general results when these assumptions hold. Then  $J_0(\theta_S) = S'_0(\theta_S) = M^\top S'(M\theta_S)M = M^\top J(M\theta_S)M$ . Define  $\mu = E_{\theta_T}(S(M\theta'_S))$  and  $\Sigma = \text{cov}_{\theta_T}(S'(M\theta'_S))$ . Then  $S(M\theta'_S) \sim N(\mu, \Sigma)$ .

301 Now, the usual Taylor series arguments yield  $0 = \mathbf{S}_0(\hat{\boldsymbol{\theta}}'_S) \approx \mathbf{S}_0(\boldsymbol{\theta}'_S) +$   
 302  $\mathbf{J}_0(\boldsymbol{\theta}'_S) (\hat{\boldsymbol{\theta}}'_S - \boldsymbol{\theta}'_S)$  or  $(\hat{\boldsymbol{\theta}}'_S - \boldsymbol{\theta}'_S) \approx -(\mathbf{J}_0(\boldsymbol{\theta}'_S))^{-1} \mathbf{S}_0(\boldsymbol{\theta}'_S)$ . Hence

$$\begin{aligned} \mathbf{S}(\mathbf{M}\hat{\boldsymbol{\theta}}'_S) &\approx \mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S) + \mathbf{J}(\boldsymbol{\theta}'_S)\mathbf{M}((\hat{\boldsymbol{\theta}}'_S - \boldsymbol{\theta}'_S)) \\ &\approx \mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S) - \mathbf{J}(\boldsymbol{\theta}'_S)\mathbf{M}(\mathbf{J}_0(\boldsymbol{\theta}'_S))^{-1} \mathbf{S}_0(\boldsymbol{\theta}'_S) \\ &= \mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S) - \mathbf{J}(\boldsymbol{\theta}'_S)\mathbf{M} \left( \mathbf{M}^\top \mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S)\mathbf{M} \right)^{-1} \mathbf{M}^\top \mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S) \\ &= \left( \mathbf{I} - \mathbf{J}(\boldsymbol{\theta}'_S)\mathbf{M} \left( \mathbf{M}^\top \mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S)\mathbf{M} \right)^{-1} \mathbf{M}^\top \right) \mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S), \end{aligned}$$

303 where  $\mathbf{I}$  is the identity matrix. Then, taking  $\hat{\boldsymbol{\theta}}'_S \approx \boldsymbol{\theta}'_S$ , (2) is approximately

$$\begin{aligned} &\mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S)^\top \left( \mathbf{I} - \mathbf{M} \left( \mathbf{M}^\top \mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S)\mathbf{M} \right)^{-1} \mathbf{M}^\top \mathbf{J}(\boldsymbol{\theta}'_S) \right) (\mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S))^{-1} \\ &\times \left( \mathbf{I} - \mathbf{J}(\boldsymbol{\theta}'_S)\mathbf{M} \left( \mathbf{M}^\top \mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S)\mathbf{M} \right)^{-1} \mathbf{M}^\top \right) \mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S) \\ &= \mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S)^\top \left( (\mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S))^{-1} - \mathbf{M} \left( \mathbf{M}^\top \mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S)\mathbf{M} \right)^{-1} \mathbf{M}^\top \right) \mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S) \\ &= \mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S)^\top \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}^{1/2} \left( (\mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S))^{-1} - \mathbf{M} \left( \mathbf{M}^\top \mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S)\mathbf{M} \right)^{-1} \mathbf{M}^\top \right) \\ &\times \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Sigma}^{-1/2} \mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S). \end{aligned}$$

304 Define  $\mathbf{P}$  so that

$$\mathbf{P}^\top \boldsymbol{\Sigma}^{1/2} \left( (\mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S))^{-1} - \mathbf{M} \left( \mathbf{M}^\top \mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S)\mathbf{M} \right)^{-1} \mathbf{M}^\top \right) \boldsymbol{\Sigma}^{1/2} \mathbf{P} = \boldsymbol{\Lambda},$$

305 for a diagonal matrix  $\boldsymbol{\Lambda}$  and  $\mathbf{P}^\top \mathbf{P} = \mathbf{P}\mathbf{P}^\top = \mathbf{I}$ . Then the diagonal of  $\boldsymbol{\Lambda}$  is the vector of  
 306 eigenvalues of  $\boldsymbol{\Sigma}^{1/2} \left( (\mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S))^{-1} - \mathbf{M} \left( \mathbf{M}^\top \mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S)\mathbf{M} \right)^{-1} \mathbf{M}^\top \right) \boldsymbol{\Sigma}^{1/2}$  or equiva-  
 307 lently  $\left( (\mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S))^{-1} - \mathbf{M} \left( \mathbf{M}^\top \mathbf{J}(\mathbf{M}\boldsymbol{\theta}'_S)\mathbf{M} \right)^{-1} \mathbf{M}^\top \right) \boldsymbol{\Sigma}$ . Thus (2) is approximately

$$\mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S)^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{P} \boldsymbol{\Lambda} \mathbf{P}^\top \boldsymbol{\Sigma}^{-1/2} \mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S). \quad (3)$$

308 Write  $\mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S) = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \mathbf{Z}$  where  $\mathbf{Z} \sim N_4(0, \mathbf{I})$ . That is,  $\mathbf{Z} =$   
 309  $\boldsymbol{\Sigma}^{-1/2} (\mathbf{S}(\mathbf{M}\boldsymbol{\theta}'_S) - \boldsymbol{\mu})$ . Let  $\mathbf{U} = \mathbf{P}^\top \mathbf{Z}$  and  $\mathbf{b} = \mathbf{P}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$ . Then (3) is

$$(\mathbf{b} + \mathbf{U})^\top \boldsymbol{\Lambda} (\mathbf{b} + \mathbf{U}) = \sum_{j=1}^4 \lambda_j (b_j + U_j)^2.$$

This article is protected by copyright. All rights reserved.

310 To examine this, we replace  $J(M\theta'_S)$  by  $E_{\theta_T}(J(M\theta'_S))$ , the expected information matrix.  
 311 Note that if the null hypothesis is true, then this is  $\Sigma$  and  $\mu = 0$ .  
 312 Now  $\Sigma^{1/2} \left( \Sigma^{-1} - M \left( M^\top \Sigma M \right)^{-1} M^\top \right) \Sigma^{1/2}$  is easily seen to be idempotent with  
 313 trace 1. Thus we should consider

$$\left( \left( E_{\theta_T}(J(M\theta'_S)) \right)^{-1} - M \left( M^\top (E_{\theta_T}(J(M\theta'_S))) M \right)^{-1} M^\top \right) \Sigma. \quad (4)$$

314 We now examine equation (4) for our two sample occupancy model under our reference  
 315 framework with  $\psi_1 = 0.8$  and take  $\psi_2 = \psi_1(1 - R)$  for  $0 \leq R < 1$  in steps of 0.01. The  
 316 calculations in Appendix II allow us to determine  $\Sigma$  and we have previously determined  
 317  $E_{\theta_T}(J(M\theta'_S))$ . We observe that, for each value of  $R$ , only the first eigenvalue is nonzero.  
 318 In Figure 6 we plot the inverse of this nonzero eigenvalue as a function of  $R$ . As in our earlier  
 319 examination in Section 4 of  $E_{\theta_T}(S(M\theta'_S))$ , we see that the eigenvalue becomes negative  
 320 at  $R \approx 0.5$ . This confirms that the negative values of the score statistic are not just due to  
 321 random variation.

322 Clearly, if there is only one nonzero eigenvalue and this is negative then the  
 323 matrix must be negative definite. However, the values of the score statistic under the  
 324 observed information matrix were found in our simulations to be positive and negative.  
 325 To examine this, we simulate data under our reference framework, with  $\psi_1 = 0.8$  and  
 326  $\psi_2 = 0.6$ . This gives  $\theta'_S = (0.673, 0.532, 0.336)^\top$ . For each set of data, we compute  
 327  $S(M\theta'_S)$  and  $J(M\theta'_S)$ . We take  $\Sigma$  to be the empirical covariance matrix of the  
 328  $S(M\theta'_S)$  computed for the simulated score functions and then compute the eigenvalues of  
 329  $\left( (J(M\theta'_S))^{-1} - M \left( M^\top J(M\theta'_S) M \right)^{-1} M^\top \right) \Sigma$ . The inverse of the first eigenvalue  
 330 is plotted in Figure 7. It is apparent that the eigenvalues for the observed information matrix  
 331 can be negative or positive. That is, random variation leads to the positive eigenvalues and  
 332 hence positive values of the score statistics.

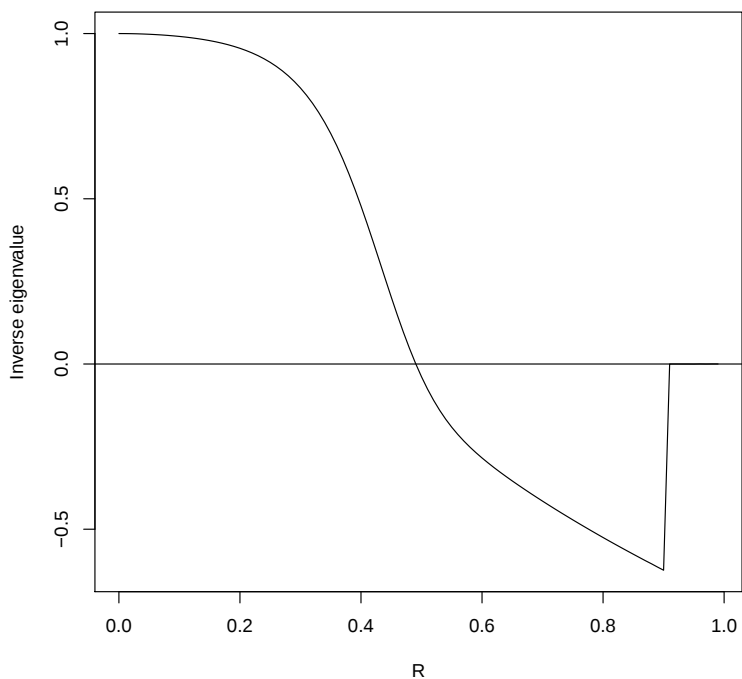


Figure 6. Inverse of the first eigenvalues, as a function of  $R$ , of the result found to be  $\left( \left( E_{\theta_T} (J(M\theta'_S)) \right)^{-1} - M \left( M^\top (E_{\theta_T} (J(M\theta'_S)) M) \right)^{-1} M^\top \right) \Sigma$ .

This article is protected by copyright. All rights reserved.

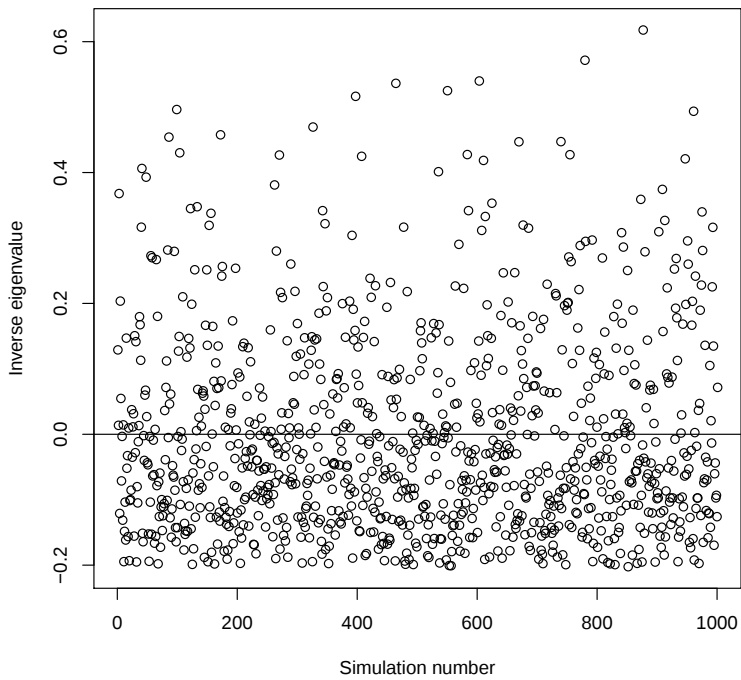


Figure 7. Inverse of the first eigenvalue, when  $R = 0.6$  for 1000 simulations, of  $\left( (J(M\theta'_S))^{-1} - M(M^\top J(M\theta'_S)M)^{-1}M^\top \right) \Sigma$ .

## Appendix II

To determine the eigenvalues we need to compute  $\Sigma$ . Recall that for a single region,

$$S_1 = \frac{s_d - \psi\theta N}{\psi(1 - \psi\theta)},$$

$$S_2 = \frac{d - s_d K p}{p(1 - p)} - \frac{(N - s_d)\psi K(1 - \theta)}{(1 - \psi\theta)(1 - p)},$$

$s_d \sim \text{Bin}(N, \psi_T \theta_T)$  and given  $s_d$ ,  $d$  is the sum of  $s_d$  independent positive binomial random variables. Then

$$\begin{aligned} E(s_d) &= N\psi_T\theta_T, \\ \text{var}(s_d) &= N\psi_T\theta_T(1 - \psi_T\theta_T), \\ E(d|s_d) &= s_d \frac{Kp_T}{\theta_T}, \\ \text{var}(d|s_d) &= s_d \left( \frac{K^2 p_T^2 - K p_T^2 + K p}{\theta_T} - \frac{K^2 p_T^2}{\theta_T^2} \right), \\ E(d) &= N\psi_T K p_T, \\ \text{var}(d) &= E(\text{var}(d|s_d)) + \text{var}(E(d|s_d)), \\ &= N\psi_T\theta_T \left( \frac{K^2 p_T^2 - K p_T^2 + K p}{\theta_T} - \frac{K^2 p_T^2}{\theta_T^2} \right) + N\psi_T\theta_T(1 - \psi_T\theta_T) \left( \frac{K p_T}{\theta_T} \right)^2, \\ \text{cov}(d, s_d) &= E(ds_d) - E(s_d)E(d), \\ &= E(s_d^2 K p_T / \theta_T) - N^2 \psi_T^2 \theta_T K p_T, \\ &= (N\psi_T\theta_T(1 - \psi_T\theta_T) + N^2 \psi_T^2 \theta_T^2) K p_T / \theta_T - N^2 \psi_T^2 \theta_T K p_T \\ &= N\psi_T(1 - \psi_T\theta_T) K p_T. \end{aligned}$$

Thus

$$\mu_1 = \frac{N(\psi_T\theta_T - \psi\theta)}{\psi(1 - \psi\theta)},$$

$$\mu_2 = \frac{N\psi_T K(p_T - p\theta_T)}{p(1 - p)} - \frac{N(1 - \psi_T\theta_T)\psi K(1 - \theta)}{(1 - \psi\theta)(1 - p)},$$

$$\Sigma_{11} = \text{var}(\mathbf{S}_1) = \frac{N\psi_T\theta_T(1-\psi_T\theta_T)}{\psi^2(1-\psi\theta)^2},$$

$$\begin{aligned}\Sigma_{22} &= \text{var}(\mathbf{S}_2) = \text{E}(\text{var}(S_2|s_d)) + \text{var}(\text{E}(S_2|s_d)) \\ &= \frac{\text{E}(\text{var}(d|s_d))}{p^2(1-p)^2} + \text{var}(s_d) \left( \frac{Kp_T - Kp\theta_T}{\theta_T p(1-p)} + \frac{\psi K(1-\theta)}{(1-\psi\theta)(1-p)} \right)^2,\end{aligned}$$

$$\begin{aligned}\text{E}(\mathbf{S}_1\mathbf{S}_2) &= \text{E}\{(s_d - \psi\theta N)s_d\} \frac{Kp_T/\theta_T - Kp}{\psi(1-\psi\theta)p(1-p)} \\ &\quad - \text{E}\{(s_d - \psi\theta N)(N - s_d)\} \frac{\psi K(1-\theta)}{\psi(1-\psi\theta)^2(1-p)} \\ &= \{\text{E}(s_d^2) - \psi\theta N\text{E}(s_d)\} \frac{Kp_T/\theta_T - Kp}{\psi(1-\psi\theta)p(1-p)} \\ &\quad - \{N(\text{E}(s_d) - \psi\theta N) - \text{E}(s_d^2) + \psi\theta N\text{E}(s_d)\} \frac{\psi K(1-\theta)}{\psi(1-\psi\theta)^2(1-p)},\end{aligned}$$

$$\Sigma_{12} = \text{E}(\mathbf{S}_1\mathbf{S}_2) - \boldsymbol{\mu}_1\boldsymbol{\mu}_2.$$

Appendix III

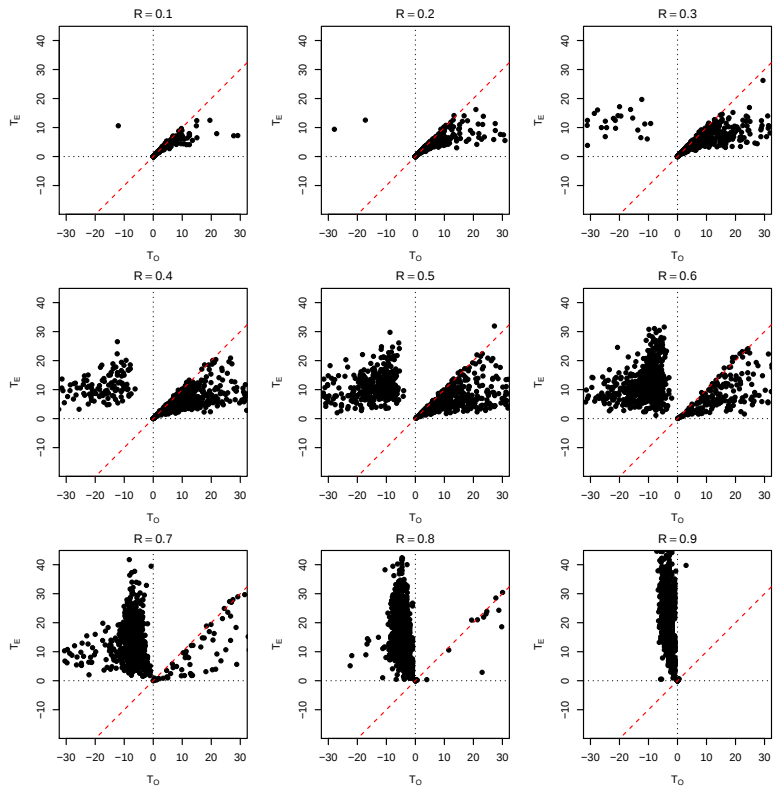


Figure 8. Observed ( $T_O$ ) versus expected ( $T_E$ ) score test statistic, for  $\psi_1 = 0.8$  under the reference framework. For clarity, 1000 of the 50000 simulations are displayed here, x-axis limits are set to  $\pm 30$  and y-axis limits set to  $(-10, 40)$ .

Table 2. Agreement (Agr.) for accepting/rejecting  $H_0$  between each of: in (a)  $T_O^+$  and  $T_E$ , and in (b)  $T_O^*$  and  $T_E$ . 50000 simulations are generated at each value of  $R$ , under the reference framework. The number of simulations with positive test statistics (sims) are listed in (a) and in (b) the number of positive and negative test values (sims). Discrepancy of ‘sims’ from 50000 simulations in (b) occurs when numerical optimization did not converge.

|    | $R$  | 0     | 0.1   | 0.2   | 0.3   | 0.4   | 0.5   | 0.6   | 0.7   | 0.8   | 0.9   |
|----|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| a) | Agr. | 0.97  | 0.97  | 0.95  | 0.92  | 0.88  | 0.85  | 0.82  | 0.84  | 0.92  | 0.99  |
|    | sims | 49721 | 49423 | 48648 | 46729 | 42786 | 36203 | 27412 | 18123 | 12380 | 11894 |
| b) | Agr. | 0.97  | 0.97  | 0.94  | 0.90  | 0.85  | 0.80  | 0.75  | 0.75  | 0.80  | 0.91  |
|    | sims | 50000 | 50000 | 49999 | 49997 | 49993 | 49976 | 49932 | 49593 | 48137 | 40369 |

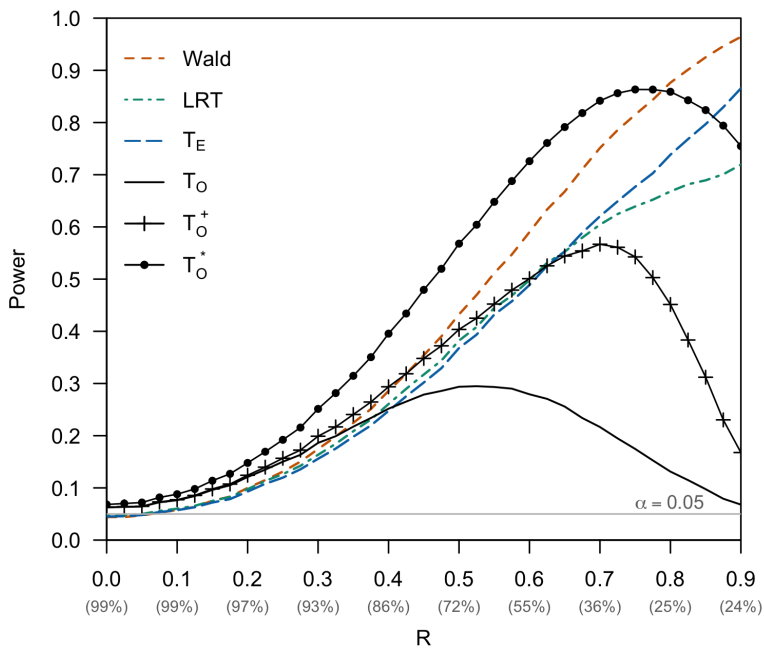


Figure 9. Power for the new modified test  $T_O^*$  and the score test under observed information matrix for positive values ( $T_O^+$ ), compared to the other tests. For each  $R$ , 50000 simulations are run with  $\psi_1 = 0.4$  under the reference framework ( $p_1 = p_2 = 0.5$ ,  $N_1 = N_2 = 50$ ,  $K_1 = K_2 = 3$ ). Percentage of simulations that did not fail and returned a positive score test under observed information ( $T_O^+$ ) are listed in parenthesis under the  $x$ -axis.

## Appendix IV

We provide simulation studies for combinations of occupancy ( $\psi_1 = \psi_2$ ), detection ( $p_1 = p_2$ ), sample size ( $N_1 = N_2$ ) and number of visits ( $K_1 = K_2$ ). We note that the tests are asymptotically equivalent, including our new modified test  $T_O^*$ , except for the score test under observed information  $T_O$  as evidenced in Figures 11, 12, 13, 14, 15 and 16.

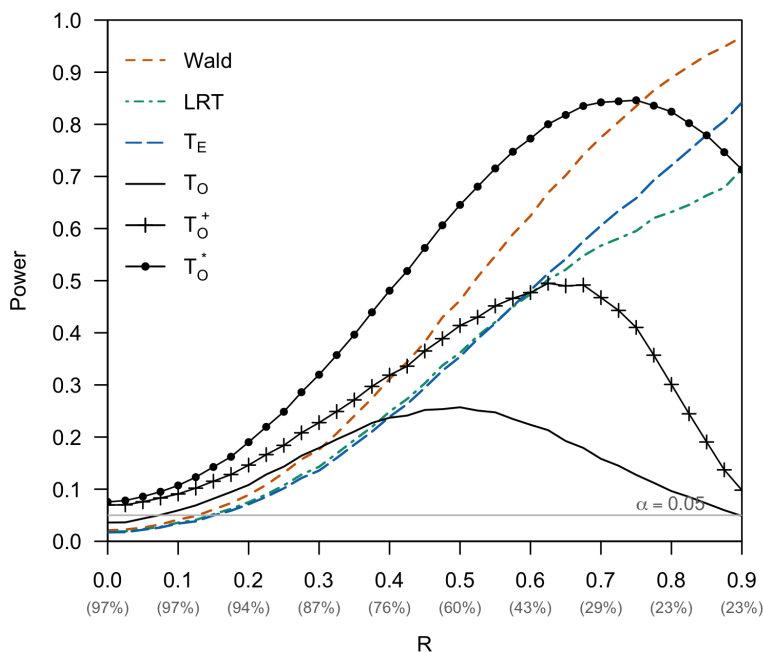


Figure 10. Power for the new modified test compared to other tests under  $\psi_1 = 0.8, p_1 = p_2 = 0.5, N_1 = N_2 = 20$  and  $K = 3$ , for 50000 simulations at each value of  $R$ . Percentage of simulations where numerical optimization did not fail and that returned a positive score test under the observed information ( $T_O^+$ ) are listed in parenthesis under the  $x$ -axis.

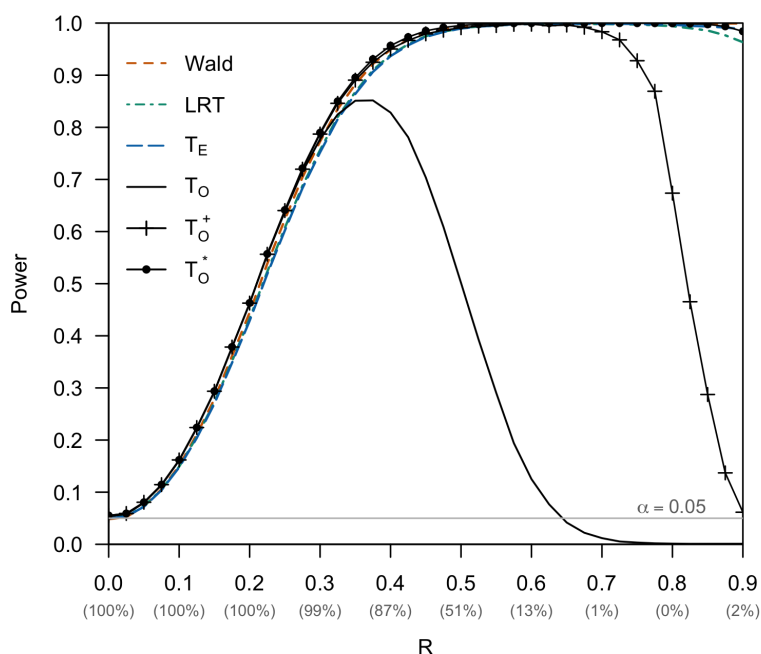


Figure 11. Power for the new modified test compared to other tests under  $\psi_1 = 0.8, p_1 = p_2 = 0.5, N_1 = N_2 = 100$  and  $K = 3$ , for 50000 simulations at each value of  $R$ . Percentage of simulations where numerical optimization did not fail and that returned a positive score test under the observed information ( $T_O^+$ ) are listed in parenthesis under the  $x$ -axis.

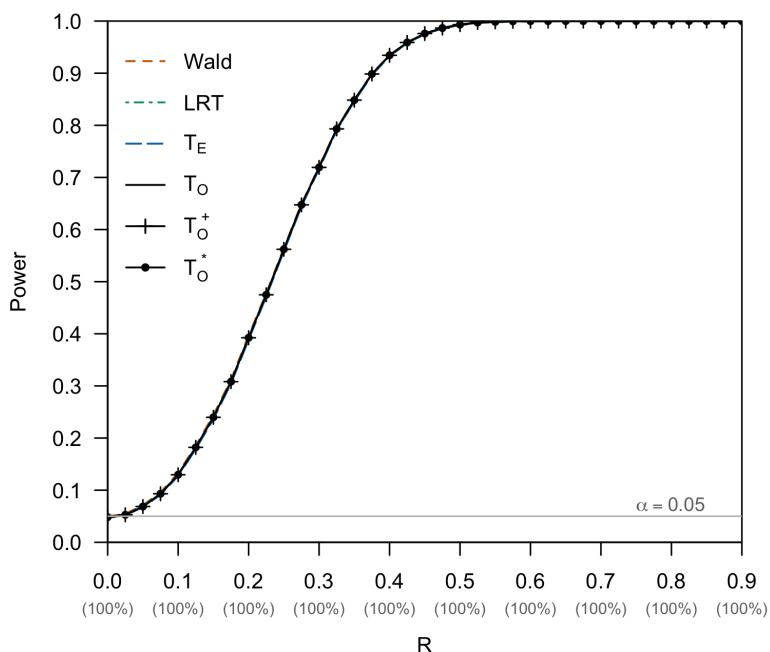


Figure 12. Power for the new modified test compared to other tests under  $\psi_1 = 0.4, p_1 = p_2 = 0.5, N_1 = N_2 = 200$  and  $K = 30$ , for 50000 simulations at each value of  $R$ . Percentage of simulations where numerical optimization did not fail and that returned a positive score test under the observed information ( $T_O^+$ ) are listed in parenthesis under the  $x$ -axis.

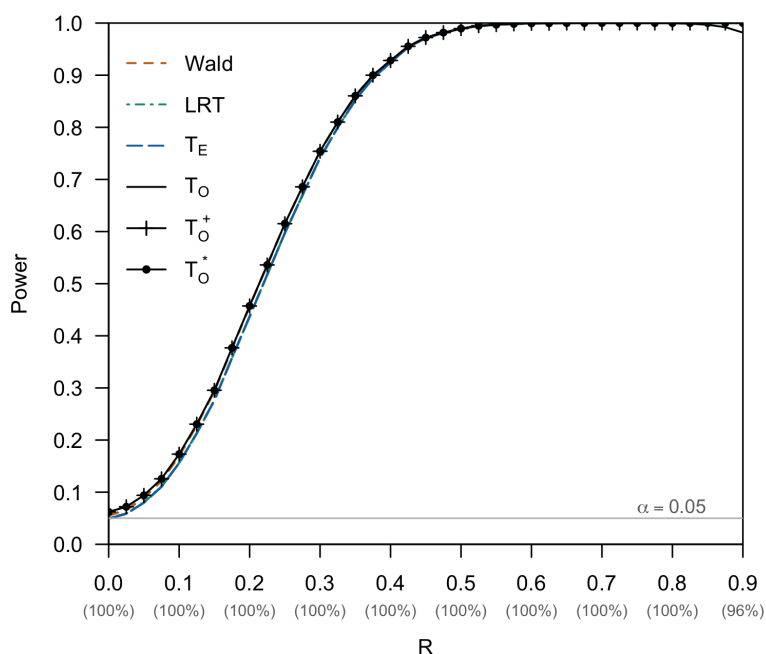


Figure 13. Power for the new modified test compared to other tests under  $\psi_1 = 0.8, p_1 = p_2 = 0.75, N_1 = N_2 = 50$  and  $K = 15$ , for 50000 simulations at each value of  $R$ . Percentage of simulations where numerical optimization did not fail and that returned a positive score test under the observed information ( $T_O^+$ ) are listed in parenthesis under the  $x$ -axis.

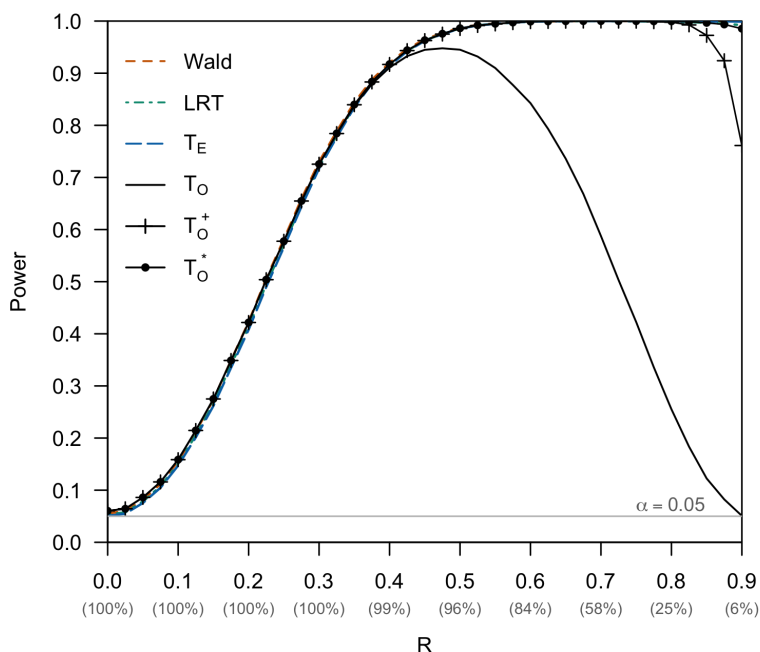


Figure 14. Power for the new modified test compared to other tests under  $\psi_1 = 0.8, p_1 = p_2 = 0.75, N_1 = N_2 = 50$  and  $K = 3$ , for 50000 simulations at each value of  $R$ . Percentage of simulations where numerical optimization did not fail and that returned a positive score test under the observed information ( $T_O^+$ ) are listed in parenthesis under the  $x$ -axis.

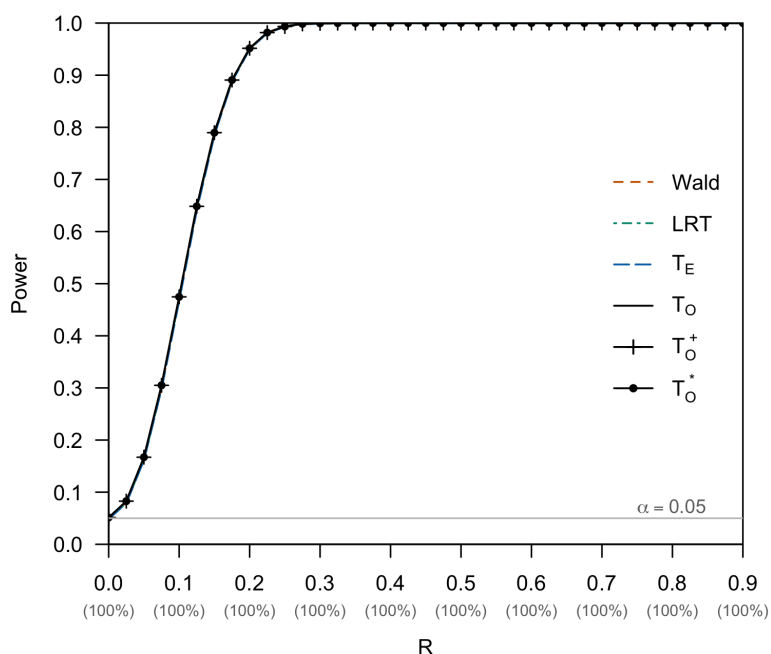


Figure 15. Power for the new modified test compared to other tests under  $\psi_1 = 0.8, p_1 = p_2 = 0.75, N_1 = N_2 = 50$  and  $K = 3$ , for 50000 simulations at each value of  $R$ . Percentage of simulations where numerical optimization did not fail and that returned a positive score test under the observed information ( $T_O^+$ ) are listed in parenthesis under the  $x$ -axis.

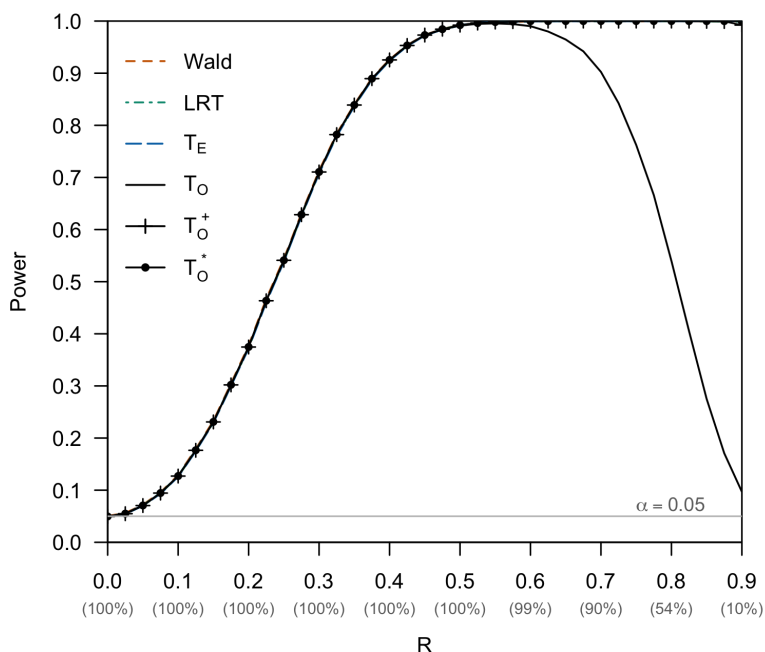


Figure 16. Power for the new modified test compared to other tests under  $\psi_1 = 0.4, p_1 = p_2 = 0.75, N_1 = N_2 = 200$  and  $K = 3$ , for 50000 simulations at each value of  $R$ . Percentage of simulations where numerical optimization did not fail and that returned a positive score test under the observed information ( $T_O^+$ ) are listed in parenthesis under the  $x$ -axis.

## References

- BICKEL, P.J. & DOKSUM, K.A. (2001). *Mathematical Statistics Basic Ideas and Selected Topics*. New York: Taylor Francis Inc.
- CASELLA, G. & BERGER, R. (2002). *Statistical Inference*. Pacific Grove, CA: Duxbury Press.
- CATCHPOLE, E. & MORGAN, B.J.T. (1996). Model selection in ring-recovery models using score tests. *Biometrics* **52**, 664–672.
- COWEN, L., BESBEAS, P., MORGAN, B.J.T. & SCHWARZ, C. (2017). Hidden markov models for extended batch data. *Biometrics* **73**, 1321–1331.
- COX, D.R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics* **29**, 357–372.
- DAVISON, A. (2003). *Statistical Models*. Cambridge: Cambridge University Press.
- EFRON, B. & HINKLEY, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**, 457–482.
- ENGLE, R. (1984). Chapter 13: Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handbook of Econometrics* **2**, 775–826.
- FREEDMAN, D.A. (2007). How can the score test be inconsistent? *The American Statistician* **61**, 291–295.
- GODFREY, L.G. & ORME, C.D. (2001). On improving the robustness and reliability of Rao's score test. *Journal of Statistical Planning and Inference* **97**, 153–176.
- GUILLERA-ARROITA, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography* **40**, 281–295.
- GUILLERA-ARROITA, G. & LAHOZ-MONFORT, J. (2012). Designing studies to detect differences in species occupancy: power analysis under imperfect detection. *Methods in Ecology and Evolution* **3**, 860–869.
- GUILLERA-ARROITA, G., RIDOUT, M.S. & MORGAN, B.J.T. (2010). Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution* **1**, 131–139.
- HOSKING, J.R.M. (1984). Testing whether the shape parameter is zero in the generalized extreme-value distribution. *Biometrika* **71**, 367–374.
- KARAVARSAMIS, N., ROBINSON, A.P., HEPWORTH, G., HAMILTON, A. & HEARD, G. (2013). Comparison of four bootstrap-based interval estimators of species occupancy and detection probabilities. *Australian and New Zealand Journal of Statistics* **55**, 235–252.
- LAWRANCE, A.J. (1987). The score statistic for regression transformation. *Biometrika* **74**, 275–279.
- LEHMANN, E.L. & ROMANO, J.P. (2005). *Testing Statistical Hypothesis*. New York: Springer.
- MACKENZIE, D.I., NICHOLS, J., ROYLE, J., POLLOCK, K., BAILEY, L. & HINES, J. (2017). *Occupancy Estimation and Modeling Inferring Patterns and Dynamics of Species Occurrence*. San Diego, CA.: Elsevier, 2nd edn.
- MACKENZIE, D.I., NICHOLS, J.D., LACHMAN, G.B., DROEGE, S., ROYLE, J. & LANGTIMM, C.A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**, 2248–2255.
- MCCREA, R. & MORGAN, B.J.T. (2011). Multi-site mark-recapture model selection using score tests. *Biometrics* **67**, 234–241.
- MORGAN, B.J.T., PALMER, K.J. & RIDOUT, M.S. (2007). Special section: score test oddities. *The American Statistician* **61**, 285–288.
- STORER, B.E., WACHOLDER, S. & BRESLOW, N. (1983). Maximum likelihood fitting of general risk models to stratified data. *Applied Statistics* **32**, 17–181.
- VAN DER VAART, A. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.
- VERBEKE, G. & MOLENBERGHS, G. (2007). What can go wrong with the score test? *The American Statistician* **61**, 289–290.
- WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

This article is protected by copyright. All rights reserved.

- 393 WINTLE, B., MCCARTHY, M., PARRIS, K. & BURGMAN, M. (2004). Precision and bias of methods for  
394 estimating point survey detection probabilities. *Ecological Applications* **14**, 703–712.
- 395 YANG, Z. & ABEYSINGHE, T. (2003). A score test for Box-Cox functional form. *Economics Letters* **79**,  
396 107–115.

Author Manuscript

This article is protected by copyright. All rights reserved.