How can the score test be consistent?

Natalie Karavarsamis[1]

Co-authors: Guillera-Arroita G.[2], Huggins R.[1], Morgan B.J.T.[3]

[1]School of Mathematics and Statistics, The University of Melbourne, Australia

[2] School of Biosciences, The University of Melbourne, Australia

[3]School of Mathematics, Statistics & Actuarial Science, University of Kent, UK

July 24, 2018

La Trobe – Kyushu joint seminar series

Comparing Occupancies

The Score Test for Species Occupancy Model

Eigenvalues

Results

New Test

Conclusions

Outline

Comparing Occupancies

The Score Test for Species Occupancy Model

Eigenvalues

Results

New Test

Conclusions

Comparing occupancies

A common question in ecology when studying occurrence of species is to compare two binomial proportions under imperfect detection as a way of comparing two occupancy samples or studies.

This leads to four parameters for estimation.

$H_0 : \psi_1 = \psi_2$

Comparing Occupancies

The Score Test for Species Occupancy Model

Eigenvalues

Results

New Test

Conclusions

Comparing occupancies

A common question in ecology when studying occurrence of species is to compare two binomial proportions under imperfect detection as a way of comparing two occupancy samples or studies.

This leads to four parameters for estimation.

$H_0 : \psi_1 = \psi_2$

Available tests include

- Wald
- Score
- LRT

Comparing Occupancies

The Score Test for Species Occupancy Model

Eigenvalues

Results

New Test

Conclusions

Comparing occupancies: Hypothesis tests

What we know

- Under H_0: asymptotic distribution under the null gives equivalence for Score Test, LRT and Wald
- Under H_1: these hypothesis tests are no longer equivalent
- Asymptotic theory no longer holds
- Negative score test values may be produced for the score test using the observed information and we examine this here (in paper).
- Observed information is easy to compute numerically
- Closed form expressions for expected information do not always exist, especially for more complex models
- We propose a new modified rule based on the observed score test.

Comparative Tests

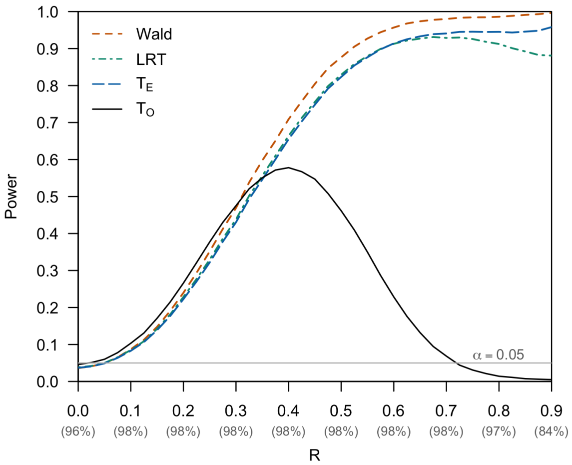


Figure 1: 50000 sims per R ($\psi_2 = \psi_1(1 - R)$) where numerical optimization did not fail.

Occupancy Model

In the standard model, detections are described as independent Bernoulli trials.
Y_i are the number of detections over K visits at site i, $i = 1, \ldots, N$. Then

$$Pr(Y_i = 0) = 1 - \psi + \psi(1-p)^K$$
$$Pr(Y_i = y_i) = \psi p^{y_i}(1-p)^{K-y_i},$$

$$y_i = 1, 2, \ldots, K \quad i = 1, 2, \ldots, N.$$

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 $1 - \psi$.

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

Note: no closed form expressions for the estimators (score equations)

Score Test: Two-sample model

We wish to compare occupancy for two independent studies (samples).

- The likelihood for the two-sample model is the product of the two single-sample likelihoods assuming independence,

  $\boldsymbol{S}(\boldsymbol{\theta}) = (S_{11}, S_{12}, S_{21}, S_{22})^T$ unconstrained score function

  $\boldsymbol{J}(\boldsymbol{\theta}) = \partial \boldsymbol{S}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T = \boldsymbol{S}'(\boldsymbol{\theta})$ observed information matrix

  $$\boldsymbol{T}_O(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{\theta})^T \boldsymbol{J}(\boldsymbol{\theta})^{-1} \boldsymbol{S}(\boldsymbol{\theta})$$

  and replace $\boldsymbol{J}(\boldsymbol{\theta})$ with $E(\boldsymbol{J}(\boldsymbol{\theta}))$ for expected score statistic $\boldsymbol{T}_E(\boldsymbol{\theta})$.
- Large-sample null distribution of the score statistic is χ^2_1 using both the expected and observed information matrices.

Summary

- At the unrestricted maximum, observed information will be usually positive definite. So if observed info. computed at unrestricted MLE then consistency (should be) Ok.
- We compute observed information at $\hat{\theta}_S$, the restricted maximum i.e. the parameter value maximising the log-likelihood over the null hypothesis.
- At the restricted max, the observed information can generate negative variance estimates - which makes inconsistency possible.
- Thus, score test can be inconsistent because at the MLE under the null hypothesis, the observed information matrix produces negative variance estimates.
- The test can also be inconsistent if the expected likelihood equation has spurious (multiple) roots.

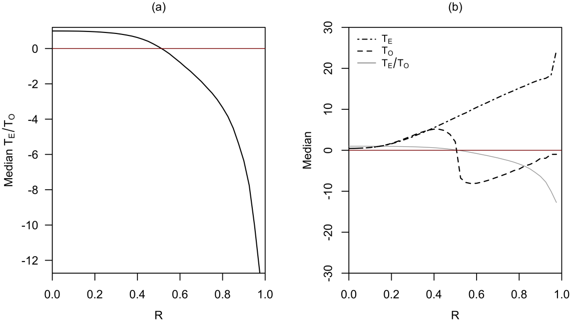


Figure 2: $\psi_1 = 0.8$ (a): median of T_E/T_O vs R. (b): same ratio, plus median of T_E and T_O.

- At $\psi_1 = \psi_2$ H_0 **is true with effect size equal to zero, i.e.** $R = 0$. Then $T_E = T_O$ and $T_E/T_O = 1$.
- At $R \approx 0.5$ half of the values of the observed score statistic are positive & half are negative.

The Score Test for Species Occupancy Model

Let

$\boldsymbol{\theta} = (\psi_1, p_1, \psi_2, p_2)^T$ be model parameters,

$\boldsymbol{\theta}_T = (\psi_{1T}, p_{1T}, \psi_{2T}, p_{2T})^T$ be true parameter values.

Consider

$H_0 : \psi_1 = \psi_2 = \psi$

then let

$\boldsymbol{\theta}' = (\psi, p_1, p_2)^T$ model parameters under H_0

$\boldsymbol{S}_0(\boldsymbol{\theta}')$ be the score function under H_0.

$\boldsymbol{\theta}'_S$ is restricted parameter subspace according to H_0, which satisfies $E_{\boldsymbol{\theta}_T}(\boldsymbol{S}_0(\boldsymbol{\theta}'_S)) = 0$, also let

$\widehat{\boldsymbol{\theta}}'_S$ be the MLE; a solution of $\boldsymbol{S}_0(\boldsymbol{\theta}') = 0$
i.e. maximises the log-likelihood subject to restriction S, the "subspace".

The Score Test

The score test statistic defined in terms of the observed information is

$$\boldsymbol{T}_O(\widehat{\boldsymbol{\theta}}'_S) = \boldsymbol{S}(\boldsymbol{M}\widehat{\boldsymbol{\theta}}'_S)^T \boldsymbol{J}(\boldsymbol{M}\widehat{\boldsymbol{\theta}}'_S)^{-1} \boldsymbol{S}(\boldsymbol{M}\widehat{\boldsymbol{\theta}}'_S)$$

- Replace $\boldsymbol{J}(\boldsymbol{M}\widehat{\boldsymbol{\theta}}'_S)$ with $E_{\boldsymbol{\theta}_T}(\boldsymbol{M}\boldsymbol{J}(\boldsymbol{\theta}'_S))$ evaluated at $\boldsymbol{\theta}'_S = \widehat{\boldsymbol{\theta}}'_S$ to give the expected score test statistic $\boldsymbol{T}_E(\widehat{\boldsymbol{\theta}}'_S)$.
- In our setting, this will have a chi-square distribution with one degree of freedom under H_0 asymptotically.
- As $\boldsymbol{\theta}_T$ is the true value $\widehat{\boldsymbol{\theta}}' \xrightarrow{P} \boldsymbol{\theta}'_S$, and $E_{\boldsymbol{\theta}_T}(\boldsymbol{J}(\boldsymbol{M}\boldsymbol{\theta}'_S))$ may be readily computed.
- This requires computing $\boldsymbol{\theta}'_S$ for a given $\boldsymbol{\theta}_T$.
- We examine these eigenvalues.

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

…detailed derivations in paper on Arxiv.org

Comparing Occupancies

The Score Test for Species Occupancy Model

Eigenvalues

Results

New Test

Conclusions

Eigenvalues

We plot the eigenvalues of $E_{\boldsymbol{\theta}_T}(\boldsymbol{J}(\boldsymbol{M}\boldsymbol{\theta}'_S))$ for the case $\psi_1 = 0.8$ (Figure 3)

- the smallest eigenvalue has a value of zero at $R \approx 0.5$.
- despite the eigenvalues of $E_{\boldsymbol{\theta}_T}(\boldsymbol{J}(\boldsymbol{M}\boldsymbol{\theta}'_S))$ being positive for $R < 0.5$, the estimator and the observed information matrix are random; there will be some outcomes with negative eigenvalues leading to negative values of the observed score statistic from fig. T_O/T_E and their medians (Figure 2). That is, the **observed information matrix is indefinite in this case**.
- The turning point at $R = 0.4$ (Figure 2(b)) reflects the turning point at 0.4 for the 2nd eigenvalue, as seen in Figure 3 for simulated and analytical results.

  (examined in detail in paper)

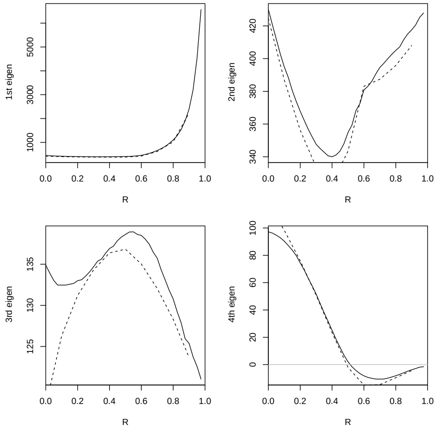


Figure 3: Eigenvalues for T_O vs R with $\psi_1 = 0.8$. Solid lines are medians obtained from simulations (50000 at each value of R). Dashed lines are eigenvalues of $E_{\boldsymbol{\theta}_T}(\boldsymbol{J}(\boldsymbol{M}\boldsymbol{\theta}'_S))$.

The Score Test

When the null hypothesis is true,

$$\boldsymbol{J}(\widehat{\boldsymbol{\theta}}'_S) \longrightarrow \boldsymbol{I}(\boldsymbol{\theta}_T)$$

When the null hypothesis is false, this is not so simple.

In our application the problem is that when

$$\boldsymbol{\theta}_T \neq \boldsymbol{M}\boldsymbol{\theta}'_S,$$

$E_{\boldsymbol{\theta}_T}(\boldsymbol{S}(\widehat{\boldsymbol{\theta}}'_S)) = f(\boldsymbol{\theta}_T, \boldsymbol{M}\boldsymbol{\theta}'_S)$ rather than $E_{\boldsymbol{\theta}_T}(\boldsymbol{S}(\widehat{\boldsymbol{\theta}}'_S)) = f(\boldsymbol{\theta}_T)$ then we have seen that $E_{\boldsymbol{\theta}_T}\left(\boldsymbol{J}(\boldsymbol{M}\boldsymbol{\theta}'_S)\right)$ need not be positive definite.

- This leads to an ambiguous score function producing some positive and some negative eigenvalues of the observed information matrix.
- As a result, the observed score test statistic may be negative.

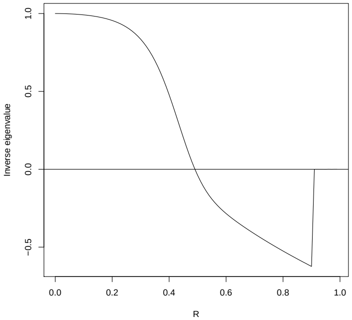
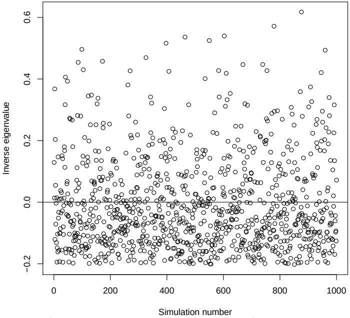


Figure 4: Inverse of the first eigenvalues of $\left(\left(E_{\boldsymbol{\theta}_T}\left(\boldsymbol{J}(\boldsymbol{M}\boldsymbol{\theta}_S')\right)\right)^{-1} - \boldsymbol{M}\left(\boldsymbol{M}^T\left(E_{\boldsymbol{\theta}_T}\left(\boldsymbol{J}(\boldsymbol{M}\boldsymbol{\theta}_S')\right)\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T\right)\boldsymbol{\Sigma}$ as a function of effect size R. As in our earlier examination of $E_{\boldsymbol{\theta}_T}\left(\boldsymbol{S}(\boldsymbol{M}\boldsymbol{\theta}_S')\right)$, we see that the eigenvalue becomes negative at $R \approx 0.5$.

This confirms that the negative values of the score statistic are not just due to random variation.



- Inverse of the first eigenvalue of $\left((\boldsymbol{J}(\boldsymbol{M\theta}'_S))^{-1} - \boldsymbol{M}\left(\boldsymbol{M}^T\boldsymbol{J}(\boldsymbol{M\theta}'_S)\boldsymbol{M}\right)^{-1}\boldsymbol{M}^T \right)\boldsymbol{\Sigma}$ when $R = 0.6$ (1000 sims).
- If there is only one nonzero eigenvalue and this is neg. then the matrix must be neg. definite.
- However, the values of the score statistic were observed in our simulations to be positive and negative.
- Eigenvalues for the obs. info. matrix can be neg. or pos. i.e. **random variation leads to the pos. eigenvalues and hence pos. values of the score statistics.** We may exploit this …

Positive and negative scores

- The differences between T_O vs T_E are predominantly where the test based on the observed score statistic rejects the null hypothesis and that based on the expected score statistic does not

- When we consider only those simulations where the observed score statistic 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. As R increases, the number of datasets with positive tests n decreases substantially. We wish to increase n.

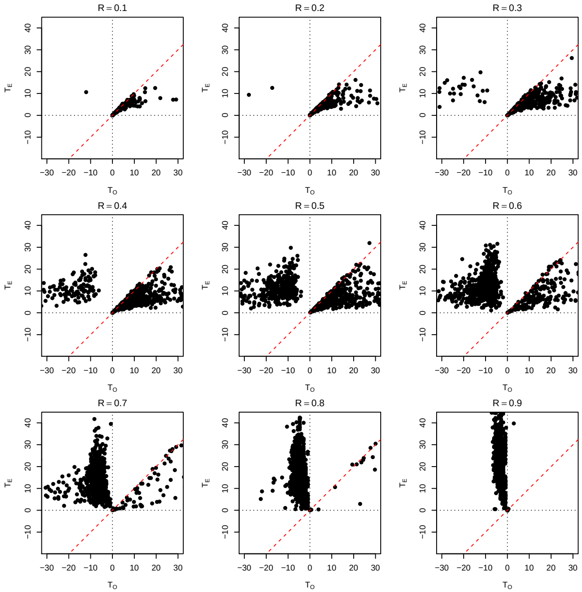


Figure 5: Observed (T_O) vs expected (T_E) score test statistic, for $\psi_1 = 0.8$. For clarity, 1000 of the 50000 sims and axes limited to ± 30.

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

- When we consider only those simulations where the observed score statistic 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.
- As R **increases, the number of datasets with positive tests** n **decreases substantially**, e.g. when $R = 0.8$ there are 1569 $(= n)$ positive tests of the 50000 simulated datasets (Table (a)). **We wish to increase** n**.**
- (note)The observed score statistic has a size that exceeds slightly the significance level, however this is not due to negative values of the statistic.

Naive test: Observed Score Test

Naive use of the observed score test results in

- a test of **low power**, with
- **power decreasing as the alternative hypothesis, H_1, moves away from the null**, as we saw in the power plot.

The New Test

We were able to improve the power of the hypothesis test for occupancy data even when the information matrix contains negative values.

Our modified rule has

- power that is mostly greater to any other test and
- largely restores consistency.

Comparing Occupancies

The Score Test for Species Occupancy Model

Eigenvalues

Results

New Test

Conclusions

The new test

The new test

Rejects the null hypothesis when the observed score statistic is larger than the usual chi-square cut-off or is negative.

Usual χ^2 rejection rule

$$T_O > \chi^2_{1,1-\alpha}$$

New rejection rule

$$T_O > \chi^2_{1,1-\alpha} \text{ or } \quad T_O < 0$$

New test is easy to use and inference is always possible.

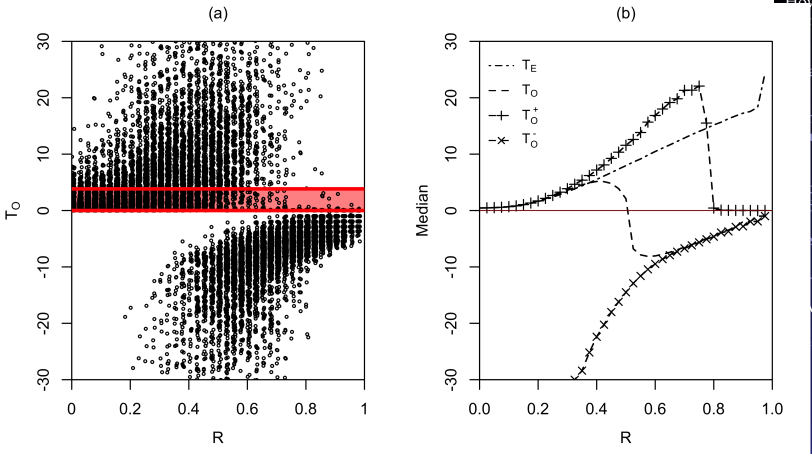


Figure 6: (a) Visual display of the new modified rejection rule for $\psi_1 = 0.8$ (for clarity, only 500 simulations are shown for each R). Power for each R is the proportion of simulations that lie outside the acceptance region (AR: shaded red).

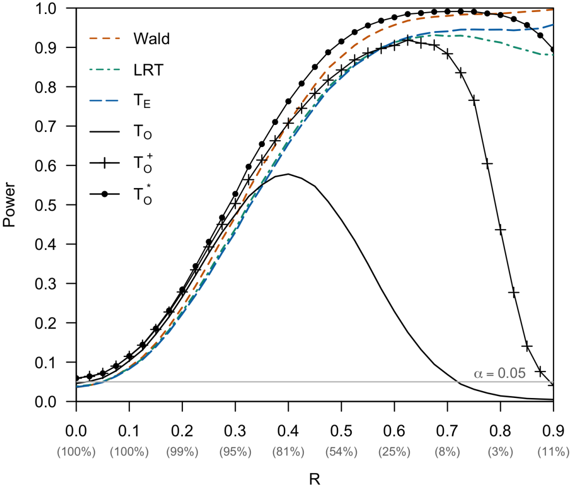


Figure 7: Power plot for $\psi = 0.8$.

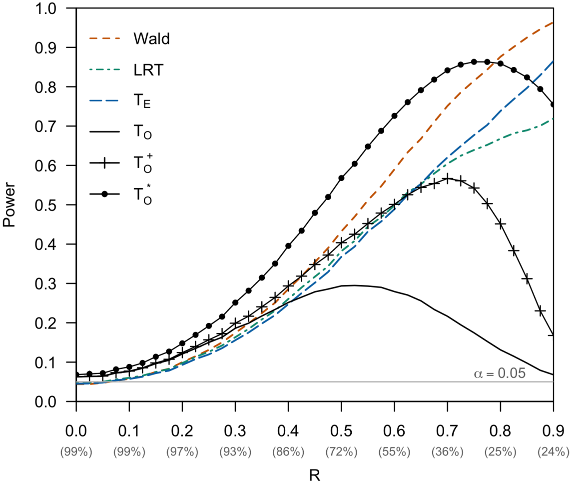


Figure 8: Power plot for scenario $\psi_1 = 0.4$

Background

Follows work done for zero-inflated Poisson by

- Freedman: How can the score test be inconsistent? (2007, *The American Statistician*, 61(4):291–295)
- Special section: Score Test oddities. Morgan BJT, Palmer KJ and Ridout MS (2007, *The American Statistician*, 61(4):291–295)

Summary

- At the unrestricted maximum, observed information will be usually positive definite.
- We compute observed information at $\hat{\theta}_S$, the parameter value maximising the log-likelihood over the null hypothesis, this is the restricted maximum.
- At a restricted max, the observed information can generate negative variance estimates - which makes inconsistency possible.

Problem

- The score test can be inconsistent because at the MLE under the null hypothesis, the observed information matrix produces negative variance estimates.
- The test can also be inconsistent if the expected likelihood equation has spurious (multiple) roots.

Problem

Freedman found

- expected model under the alternative is not always the same as under H_0 ie the asymptotics don't always work,
- this means an indefinite observed information matrix
- hence quadratic forms can be positive or negative
- this means there are negative eigenvalues
- that give positive or negative values in the observed info matrix
- that give negative score values...
- which means that the observed Score test can't be used...

Conclusions

Our test outperforms all other tests over the entire range of R.

Our new test

- is mostly the most powerful in our comparison to any other test, including the Wald test and particularly the expected score test
- is easy to use and inference is always possible
- consistency is largely restored
- does not require lengthy algebra for obtaining analytic expressions for the expected score and expected Wald
- our modified rule overcomes when large sample assumptions fail and avoids contradictory results.
- there are no complications from fitting a model under the alternative.
- in practice it is likely that an experiment may produce an indefinite information matrix
- with our new test a hypothesis test is possible.

Comparing Occupancies

The Score Test for Species Occupancy Model

Eigenvalues

Results

New Test

Conclusions