Exact tests of non-inferiority from independent binomial data based on second order test statistics

Chris Lloyd, Melbourne Business School
Exact tests of non-inferiority from independent binomial data based on second order test statistics

BY CHRIS J. LLOYD
Melbourne Business School, Carlton, 3053, AUSTRALIA
c.lloyd@mbs.edu

SUMMARY
Recent advances in likelihood asymptotics (Reid, 2003) lead to pivotal quantities that are closer to standard normal than standard pivotals and also respect some kind of conditionality. It is less clear the extent to which these methods work for discrete models. On the other hand, in the context of binomial trials conditional pivotals can lead to more efficient unconditional inferences, see Boschloo (1970) and Lloyd and Moldovan (2007). This suggests that second order pivotals that respect local conditionality might provide more powerful exact tests. For testing the rate ratio from independent binomial samples, we investigate 5 first order pivotals and the second order pivotal. Each of these is used to generate an exact test my maximising with respect to the nuisance parameter. We also consider the effect of pre-estimating the nuisance parameter.

Some Key Words: nuisance parameters; exact test; tests of independence; r-star

1 Introduction

In a clinical trial comparing a new treatment to a standard treatment, one may want to demonstrate that the new treatment is not practically inferior to the standard treatment. If \( p_1 \) is the probability of a positive end-point with the new treatment and \( p_0 \) the probability under the old, then we want to demonstrate that \( p_1 \) is not too much smaller than \( p_0 \). We consider so-called non-inferiority tests of the hypotheses

\[
\mathcal{H}_0 : \frac{p_1}{p_0} \leq 1 - \delta \quad \text{versus} \quad \mathcal{H}_1 : \frac{p_1}{p_0} > 1 - \delta,
\]

for some pre-chosen non-inferiority margin \( \delta \), often 0.1 or 0.2. Rejecting the null hypothesis means that the new treatment is not practically inferior to the old treatment, and may actually be better. Such tests were first considered by Chan (1998). Several
authors including Chan have looked at test involving \( p_1 - p_0 \) as the measure of relative efficacy rather than \( p_1/p_0 \).

Another context in which the same test arises is test of vaccine efficiency. In this case, let \( p_1 \) be the disease incidence rate amongst those vaccinated and \( p_0 \) the rate amongst those not vaccinated. The vaccine efficiency is often measured by \( e = 1 - p_1/p_0 \) which for an ideal vaccine would equal zero. In practice however, there will be a minimal level of efficiency \( \delta \) to be anticipated. Testing the null hypothesis \( e \geq \delta \) versus the alternative \( e < \delta \) is obviously identical to the hypotheses above.

The data comprise \( y_0 \) responses from \( n_0 \) independent individuals with treatment not applied, and \( y_1 \) responses from \( n_1 \) independent individuals with treatment applied. The parameters of interest are the probabilities of response \( p_0, p_1 \) with and without treatment.

## 2 First order test statistics

Denoting the interest parameter by \( \psi = \log(p_1/p_0) \), the non-inferiority hypotheses are

\[ H_0 : \psi \leq \psi_0 \quad \text{versus} \quad H_1 : \psi > \psi_0, \]

where \( \psi_0 = \log(1 - \delta) \approx -\delta \). We take the nuisance parameter to be \( \lambda = \log(p_0) \) and denote the full set of parameters by \( \theta = (\psi, \lambda) \). The log-likelihood is

\[
\ell(\psi, \lambda; y_0, y_1) = y_1(\psi + \lambda) + (n_1 - y_1) \log \left( 1 - e^{\psi + \lambda} \right) + y_0 \lambda + (n_0 - y_0) \log \left( 1 - e^\lambda \right) \tag{1}
\]

where \( t = y_0 + y_1 \) is the total number of successes. The score vector has \( \lambda \) component

\[
\frac{\partial \ell}{\partial \lambda} = \frac{y_1 - n_1 p_1}{1 - p_1} + \frac{y_0 - n_0 p_0}{1 - p_0} = \frac{y_1 - n_1 p_0 e^\psi}{1 - p_0 e^\psi} + \frac{y_0 - n_0 p_0}{1 - p_0}
\]

and so the MLE \( \hat{p}_{0\psi} \) of \( p_0 \) when \( \psi \) is fixed is obtained by solving a quadratic whose coefficients turn out to equal

\[
c_2 = (n_0 + n_1) e^\psi, c_1 = -(y_1 + n_0) - (y_0 + n_1) e^\psi, c_0 = t.
\]

This estimator, in a different form, was first given by Miettinen & Nurminen (1985). The restricted ML estimator of \( p_1 \) is \( \hat{p}_{1\psi} = e^\psi \hat{p}_{0\psi} \) and of \( \lambda \) is \( \hat{\lambda}_\psi = \log(\hat{p}_{0\psi}) \).
There are three basic test statistics that we will consider, each with an asymptotic standard normal distribution. The likelihood ratio statistic is

\[ r(\psi) = \text{sign}(\hat{\psi} - \psi) \left[ 2 \left\{ \ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}) \right\} \right] \]

Alternatively, start with the ML estimator of \( \psi \) whose asymptotic variance is

\[ \sigma^2(p_0, p_1) = \frac{1 - p_0}{n_0 p_0} + \frac{1 - p_1}{n_1 p_1}. \]

Replacing \( p_j \) by empirical estimators \( \hat{p}_j \) gives the estimator by \( \hat{\sigma}^2 \) while replacing \( p_j \) with restricted estimators \( \hat{p}_{j\psi} \) gives the restricted estimator \( \hat{\sigma}^2_{\psi} \). Then we have two Wald-type statistics

\[ W_1(\psi) = \frac{\hat{\psi} - \psi}{\hat{\sigma}}, \quad W_2(\psi_0) = \frac{\hat{\psi} - \psi}{\hat{\sigma}_{\psi}} \]

Finally, express the alternative hypothesis in the form \( p_1 - p_0 (1 - \delta) > 0 \). The ML estimator of this parameter has variance

\[ \tau^2(p_0, p_1) = \frac{p_1(1 - p_1)}{n_1} + (1 - \delta)^2 \frac{p_0(1 - p_0)}{n_0} \]

with unrestricted estimator \( \hat{\tau}^2 \) and restricted estimator \( \hat{\tau}^2_{\psi} \) which gives rise to two more statistics

\[ C_1(\psi) = \frac{\hat{p}_1 - e^\psi \hat{p}_0}{\hat{\tau}}, \quad C_2(\psi) = \frac{\hat{p}_1 - e^\psi \hat{p}_0}{\hat{\tau}_{\psi}} \]

the second of which was considered by Chan (1998). P-values based on a standard normal approximation to any of these five statistics are said to be first order accurate and suffer errors of \( O(n^{-1/2}) \) for one-sided P-values and \( O(n^{-1}) \) for two sided P-values. Second order methods described below can give one-sided P-values with error \( O(n^{-1}) \).

3 Second order test statistics

So-called second order likelihood inference is based on a local decomposition of the model into a sufficient and ancillary component. An approximate formula for the conditional distribution is available (Bardorff-Nielsen & Cox, 1994) and the cumulative tail probability of this distribution is further approximated (Lugannin & Rice, 1980), see Reid (2003) for an extensive review. All the approximations are very accurate,
sometimes spectacularly so for continuous models. The theory is considerably simpler for models that are embedded in an exponential family, see Brazzale, Davison and Reid (2007). Suppose that the log-likelihood is of exponential family form

\[ \ell(\theta; y) = \varphi(\theta)^T v(y) - c(\varphi(\theta)) \]  

where \(\varphi(\theta)\) is the canonical parameter of dimension \(d\) and \(v(y)\) is the sufficient statistic. For our application, the canonical parameters are the log-odds parameters and \(d = 2\). The general formula requires derivation of an adjustment statistic, \(q(\psi)\), that depends on two quantities. The first is the observed information matrix in the \((\psi, \lambda)\) parametrisation which we denote \(j_\theta\). The second is the \(d \times d\) Jacobian matrix \(\varphi_\theta\) of the transform from \(\varphi\) to \(\theta\). Then the adjustment statistic is

\[ q(\psi) = \frac{|\hat{\varphi} - \hat{\varphi}_\psi|}{|\varphi_\theta(\hat{\theta})|} \times \frac{|j(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|} \]  

The notation \(\hat{\varphi} - \hat{\varphi}_\psi\) in the first matrix is shorthand for the column vector giving the deviation of the canonical parameter \(\varphi\) when estimated under the general and null models. When \(\psi\) is a component of the canonical statistic \(\varphi\) then the first term reduces to \(\hat{\psi} - \psi\). The extent to which \(q(\psi)\) and likelihood root \(r(\psi)\) differ is a measure of the extent to which the first order approximation to the log-likelihood is inadequate. Second order inference is instead based on the modified likelihood root

\[ r^*(\psi) = r(\psi) + r(\psi)^{-1} \log \{q(\psi)/r(\psi)\} \]  

and its approximate standard normal distribution. The approximate one-sided P-value is \(p^*(\psi) = \Phi(-r^*(\psi))\). Note that \(r^*(\psi)\) breaks down in the centre of the distribution where \(\hat{\theta}\) and \(\hat{\theta}_\psi\) are close so that \(r(\psi) \approx 0\). This is of little practical interest however since there is, in this case, no statistical evidence against the value \(\psi\).

We now derive an explicit expression for \(q(\psi)\) for the binomial model. For future work, and it turns out also for simplicity, we generalise to testing \(\psi = h(p_1) - h(p_0)\) for a general link function \(h\) and take the nuisance parameter as \(\lambda = h(p_0)\). The inverse transformation is

\[ p_1 = h^{-1}(\psi + \lambda), p_0 = h^{-1}(\lambda). \]
Let $\varphi$ denote the logit transform which is also the canonical link. With abuse of notation, we later denote the logit parameters logit($p_j$) by $\varphi_j$. The log-likelihood is

$$\ell(\psi, \lambda; y_0, y_1) = y_0 \varphi\{h^{-1}(\lambda)\} + y_1 \varphi\{h^{-1}(\psi + \lambda)\} + n_0 \log\left(1 - h^{-1}(\lambda)\right) + n_1 \log\left(1 - h^{-1}(\psi + \lambda)\right)$$

Since $\partial \varphi / \partial p = 1 / (p(1 - p))$, the score functions are given by

$$U_\psi(\psi, \lambda) := \frac{\partial \ell}{\partial \psi} = \frac{y_1 - n_1 p_1}{p_1(1 - p_1)h'(p_1)}$$

$$U_\lambda(\psi, \lambda) := \frac{\partial \ell}{\partial \lambda} = \frac{y_0 - n_1 p_0}{p_0(1 - p_0)h'(p_0)} + \frac{y_1 - n_1 p_1}{p_1(1 - p_1)h'(p_1)}$$

and we will denote henceforth $w(p) = p(1 - p)h'(p)$. The restricted MLE $\hat{p}_0\psi$ satisfies the equation

$$0 = \frac{y_0 - n_1 p_0}{w(p_0)} + \frac{y_1 - n_1 p_1}{w(p_1)} \Leftrightarrow 0 = (y_0 - n_1 p_0)w(p_1) + (y_1 - n_1 p_1)w(p_0)$$

where $p_1 = h^{-1}(\psi + h(p_0))$. Solution requires numerical methods in general, but in our case the equation reduces to a quadratic. When $h$ is the identity link it reduces to a cubuc. Munk, Skipka and Stratmann (2005) give some theory on restricted ML estimation under this formulation.

Taking the covariance of the score functions gives the information terms

$$j_{\psi\psi} = V_1/w_1^2, j_{\psi\lambda} = V_1/w_1^2, j_{\lambda\lambda} = V_1/w_1^2 + V_0/w_0^2$$

where for $j = 0, 1$, we denote $V_j = n_j p_j(1 - p_j)$ and $w_j = w(p_j)$. Hence the non-orthogonality term is

$$\frac{|j(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta})|} = \frac{\hat{V}_1\hat{V}_0/(\hat{w}_1^2 \hat{w}_0^2)}{V_{0\psi}/\hat{w}_0^2 + V_{1\psi}/\hat{w}_1^2}$$

It remains to find the first term in (3) which depends on the transformation

$$(\psi, \lambda) \rightarrow (\varphi_1, \varphi_0) = (\beta(\psi + \lambda), \beta(\lambda)).$$

where $\beta(v) = \varphi(h^{-1}(v))$. The derivative of this function is

$$\beta'(v) = \frac{\varphi'\{h^{-1}(v)\}}{h'(h^{-1}(v))} = \frac{1}{p(1 - p)h'(p)} = \frac{1}{w(p)}$$
where \( p = h^{-1}(v) \). Therefore the Jacobian matrix is

\[
\varphi_\theta(\theta) = \begin{bmatrix}
\beta'(\psi + \lambda) & \beta'(\psi + \lambda)
\end{bmatrix} = \begin{bmatrix}
w_1^{-1} & w_1^{-1}
0 & w_0^{-1}
\end{bmatrix}
\]

with determinant \( 1/(w_1w_0) \) and so the first term in (3) is

\[
|\varphi_\theta(\hat{\theta})|^{-1} |\hat{\varphi} - \hat{\varphi}_\psi - \lambda(\hat{\theta}_\psi)| = \hat{w}_1 \hat{w}_0 \{ \hat{w}_0^{-1}(\hat{\varphi}_1 - \hat{\varphi}_1\psi) - \hat{w}_1^{-1}(\hat{\varphi}_0 - \hat{\varphi}_0\psi) \}
\]

Multiplying this by the square root of (6) gives

\[
q(\psi) = \frac{\{ \hat{w}_0^{-1}(\hat{\varphi}_1 - \hat{\varphi}_1\psi) - \hat{w}_1^{-1}(\hat{\varphi}_0 - \hat{\varphi}_0\psi) \} \sqrt{\hat{V}_1\hat{V}_0}}{\sqrt{\hat{V}_0\hat{w}_0^2 + \hat{V}_1\hat{w}_1^2}}
\]

and a slightly cleaner form is

\[
q(\psi) = \frac{\{ \hat{w}_1(\hat{\varphi}_1 - \hat{\varphi}_1\psi) - \hat{w}_0(\hat{\varphi}_0 - \hat{\varphi}_0\psi) \} \sqrt{\hat{V}_1\hat{V}_0}}{\sqrt{\hat{V}_0\hat{w}_1^2 + \hat{V}_1\hat{w}_1^2}}
\]

For inference on the logged rate ratio, \( h(p) = \log p \) and so \( w(p) = (1 - p) \). The formula is parametrisation invariant though Davison, Fraser and Reid (2006) seem to only claim invariance to linear transformations. There is an apparent problem that \( q(\psi) = 0 \) whenever \( \hat{V}_0\hat{V}_1 = 0 \) which occurs when \((y_0, y_1)\) is on the boundary of the sample space. We already noted that \( q(\psi) = 0 \) at the centre of the distribution as well. These anomalies are mentioned but not addressed in Davison, Fraser and Reid (2006), and the R-functions in bundle \textit{hoa} do not return an answer for boundary data sets. For exact tests, it is essential that the basic inference method performs adequately across the entire sample space. We will deal with these important details later in the paper.

4 Exact P-values

A much cleaner theory of 'exact' P-values can be given than of 'exact' tests, a term that can cause some confusion for discrete models. The defining property of a P-value \( P(Y) \) when nuisance parameters are present is

\[
\Pr(P(Y) \leq P(y); \lambda, \psi_0) \leq P(y) \quad \forall y \in \mathcal{Y}, \forall \lambda
\]

P-values with this property are called \textit{guaranteed} since they are guaranteed to give a valid, though possible conservative, test for any nominal size. A P-value is \textit{exact} if, for
each $y \in \mathcal{Y}$, the supremum of the left hand side of (8) over $\lambda$ equals $P(y)$, in other words the bound is tight for all $y$. This does not imply that the implied tests will have size, or even supremum size, exactly equal to nominal, which is typically unachievable.

Denote the left hand side of (8) by $\Pi(P, y, \lambda)$, which gives the exact significance of the observed $P(y)$ for a given $\omega$. Considered as a function of $\lambda$, I call this the significance profile, or simply profile, of $P(y)$. The profile not only describes the true significance of the observed $P$-value $P(y)$. It can also be used, in various ways, to construct an exact $P$-value, as defined above, starting from any approximate $P$-value $P(Y)$. It is shown in Lloyd (2008) that the maximised $P$-value $P^*(Y)$ where

$$P^*(y) := \sup_{\lambda} \{\Pi(P, y, \lambda)\} \tag{9}$$

is exact, is minimal amongst guaranteed $P$-values that order the sample space in the same order as $P(Y)$ and that any exact $P$-value must be expressible as the supremum of the profile of some $P(Y)$. In principle then, all $P$-values should be maximised. An additional theoretical use of this M-step is in evaluating how close a given $P$-value is to exact in a manner independent of nominal test size. We simply measure the closeness of $P(Y)$ to $P^*(Y)$. We will utilise this idea later.

Note that $P^*(Y)$ depends on the data $Y$ only through the ordering that $P(Y)$ induces on the sample space. So if we are going to ultimately use $P^*(Y)$ then the absolute accuracy of $P(Y)$ is not important, only the way it orders each sample point as relatively more or less hostile to the null. I call $P(Y)$ the generating statistic for $P^*(Y)$, noting that it is only defined up to an increasing monotone transformation. The challenge with maximised $P$-values is to find a good generating statistic. In our case we have $P$-value based on five different first order statistics and one based on the second order test statistic as candidates.

We illustrate the ideas on an example due to Berger and Boos (1994) where $(y_0, y_1) = (48, 14)$ and $(n_0, n_1) = (283, 47)$ so $(\hat{p}_0, \hat{p}_1) = (0.170, 0.298)$ and $\hat{\psi} = 0.563$ with standard error $\hat{\sigma} = 0.260$. Consider testing $\delta = 0.1$ which corresponds to $\psi_0 = \log 0.9 = -0.1053$. The restricted ML estimate of $p_0$ is $\hat{p}_{0\psi} = 0.190$ and $\hat{\sigma}_\psi = 0.344$. The observed values of our various test statistics are $r = 2.316$, $W_1 = 2.574$, $W_2 = 1.945$, $C_1 = 2.084$, $C_2 = 2.469$ and $r^* = 2.331$. The differences in these statistics do not by
themselves mean that the first order results are inaccurate. Indeed, the closeness of \( r \) and \( r^* \) suggests that the first order statistic \( r \) may be accurate here, if one believes the folklore that has arisen in the higher order asymptotics literature.

For discrete data, it is not sufficient that \( r \) and \( r^* \) be close to have confidence in the first order methods. Figure 1 plots profiles for four candidate generating P-values with the nuisance parameter \( \lambda \) taken to be \( p_0 \). The solid horizontal line is the observed value of the generating P-value. For \( W_2, C_2 \) and \( r \) the profiles contain a strong local maximum compared to which the quoted P-value strongly exaggerates the evidence against the null. These spikes are due to the differing accuracy of normal approximation across the parameter space. The profile for \( r^* \) is much better behaved.

One approach to this problem, suggested by Berger and Boos (1994), is to quote the P-value

\[
P_\gamma(y) := \sup \{ \Pi(P, y, \lambda) : \lambda \in C_\gamma(y) \}
\]

where \( C_\gamma(y) \) is a \((1-\gamma)\) confidence set for \( \lambda \) assuming the null hypothesis. They proved that \( P_\gamma(Y) \) is guaranteed. It is shown in Lloyd (2008) that they are not exact but for small \( \gamma \) they are almost exact and perform well. A theoretical problem is how to choose \( C_\gamma \), especially in higher dimensions.

The second approach to problematic profiles is to modify the generating statistic. A simple way to do this is to replace \( P(Y) \) with the estimated P-value

\[
\hat{P}(y) := \Pi(\hat{P}, y, \hat{\lambda}_\psi)
\]

where \( \hat{\lambda}_\psi \) is the ML estimate of \( \omega \) under the null. Estimated P-values have been around for a long time, see for instance Beran (1988) who showed that substituting an estimate prior to bootstrap leads to smaller asymptotic error rates than ordinary bootstrap. Kang and Chen (2000) have applied estimation to the statistic \( C_2(\psi_0) \) and compared it with the maximised P-value also based on \( C_2 \), and recommended caution in selecting which test to use.

The estimation step has two distinct effects. First, it seems to produce a P-value which is much closer to exact, but slightly liberal. It can be seen then as a computationally cheap but imperfect version of maximisation. Second, it has been demonstrated in Lloyd (2008) that the profile \( \Pi(\hat{P}, y, \lambda) \) of \( \hat{P}(Y) \) tends to be much flatter than the
original generator $P(Y)$. This is presumably due to it being based on an estimate of an exact significance value rather than a normal approximation whose accuracy is inconsistent across the sample space. An exact P-value based on the generating statistic $\hat{P}(Y)$ is $\hat{P}^*(y) := \sup_{\lambda} \{ \Pi(\hat{P}, y, \lambda) \}$ and is called the E+M P-value. For this reason, I will refer to the E-step as pre-estimation, since in principle it is applied prior to the maximisation or M-step. Both the M and EM P-values are exact and are determined by an initial approximate P-value, which we will call the generator.

The dotted lines in Figure 1 show the profiles $\Pi(\hat{P}, y, \lambda)$ for the pre-estimated P-values generated from the four basic P-values $W_2, C_2, r$ and $r^*$. Each of these profiles are quite flat and is very close to the observed P-value $\hat{P}(y)$, which can be read off from the original profile where it intersects the vertical line at $\lambda_\psi = 0.190$. A summary of the results is in Table 1. There is considerable disagreement between the three first order statistics. Maximisation reduces the significance of the first order statistics but has little effect on $r^*$. All four pre-estimated P-values agree quite closely and maximisation has very little effect on these, but reduces the significance of each slightly.

Table 1: Observed Z-values for data of Berger & Boos (1994) Each P-value is for testing alternative $\psi > -0.105$ and has been converted to the standard normal scale, for ease of interpretation.

<table>
<thead>
<tr>
<th>Generator</th>
<th>Base</th>
<th>M</th>
<th>E</th>
<th>E+M</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_2$</td>
<td>1.945</td>
<td>1.632</td>
<td>2.321</td>
<td>2.317</td>
</tr>
<tr>
<td>$C_2$</td>
<td>2.469</td>
<td>1.598</td>
<td>2.305</td>
<td>2.297</td>
</tr>
<tr>
<td>$r$</td>
<td>2.316</td>
<td>2.051</td>
<td>2.324</td>
<td>2.310</td>
</tr>
<tr>
<td>$r^*$</td>
<td>2.331</td>
<td>2.250</td>
<td>2.325</td>
<td>2.310</td>
</tr>
</tbody>
</table>

The last example suggests that pre-estimation is desirable for two reasons. First it gives a P-value which is almost exact. If true, this would allow the computationally problematic M-step to be dropped in practice. In general the computational difficulty of the M-step depends on the dimension of the nuisance parameter. Second, the EM P-value tends to be smaller (Z-value is larger) than the M P-value. If true generally, pre-estimated P-values will lead to a more powerful exact test that the original asymptotic P-value.

The results also suggested that when the generating P-value is $r^*$, pre-estimation
may be unnecessary. If true generally, this would allow the E-step to be dropped. In general the computational difficulty of the M-step depends on the cardinality of the sample space $N$, and is typically $O(N \log N)$ but monte-carlo methods can probably be developed when this become infeasible. The results of a comprehensive numerical investigation will be reported in later sections.

5 Further illustrations

Before giving the general results, we will display some plots to give some intuition for the causes of these results. We look first look at sample size $(n_0, n_1) = (50, 200)$ and have chosen the sample $(y_0, y_1) = (39, 157)$ for illustration. The upper panels of Figure 2 show profiles based on $r^*(\psi_0)$ at left and the ordinary likelihood root $r(\psi_0)$, with $\psi_0 = -0.105$ corresponding to a rate ratio of 0.9. The dotted curve is the profile based on the pre-estimated P-value and in each case the profile is very flat. The profile for the likelihood root displays a spike which is not present for the modified likelihood root. This results in the $r$-based P-value increasing from 0.0719 to 0.1386 with maximisation whereas the $r^*$-based P-value increases from 0.0682 to 0.0796. For this data set then, pre-estimation results in a smaller exact P-value based on $r$. The improvement is 0.1386-0.0796=0.0590. Starting with $r^*$ however, the P-value only improves from 0.0790 to 0.0796.

The next two panels show this particular data set in the context of all statistically relevant data sets. By statistically relevant we mean those data sets for which the maximised P-value, based on either $r$ or $r^*$, falls between 0.01 and 0.15. The vertical axis shows the improvement (i.e. reduction) in the maximised P-value from pre-estimation. It is apparent that pre-estimation has a large effect on P-values based on $r$ but little effect on those based on $r^*$. The bottom panels give histograms of this improvement. The mean improvement for P-values based on $r$ is 0.042 while for those based on $r^*$ is 0.009. So it appears from this analysis that the E-step could be omitted for the generator $r^*$.

Figure 3 repeats this analysis but for $(n_0, n_1) = (200, 50)$ and the particular data set $(y_0, y_1) = (45, 16)$. A similar conclusion is reached and the mean improvement from
the pre-estimation step for P-values based on \( r \) is 0.020 and for those based on \( r^* \) is 0.009.

6 Numerical study of exactness

There are six basic approximate statistics, five based on first order theory and one based on second order theory. Each of these can have the E-step applied giving a further six approximate statistics. Bearing in mind the computational burden of the M-step, we are interested in the deviation of these approximate P-values from exactness.

Let us define the distance between two P-values \( P_1(Y) \) and \( P_2(Y) \) by

\[
\delta_k(P_1, P_2; \theta) = E^{1/k}\{(P_1(Y) - P_2(Y))^k|R, \theta}\}
\]

where \( R \) is a subset of the sample space considered relevant. For instance, sample points where both P-values are between 0.9 and 1.0 are of no interest. For our application we will take \( R = \{ \hat{\psi} \geq \psi_0 \} \) which corresponds to there being some evidence against the null and any reasonable P-values being less than 0.5.

To measure how close a given P-value \( P(Y) \) is to exact we measure how stochastically close \( P(Y) \) is to \( P^*(Y) \), since the latter P-value is the ‘correct’ P-value based on the ordering induced by \( P \). Since \( \delta_k \) depends on the parameter values, we substitute the null value for \( \psi \) and take Lesbegue mean value with respect to the nuisance parameter \( p_0 \). This gives the exactness measure

\[
\delta_k(P^*, P) = \int_0^1 \delta_k(P^*, P; \psi_0, p_0) dp_0.
\]

For \( k = 1 \) this measures bias, positive values indicating that \( P \) is liberal. For \( k = 2 \) it measures root mean square deviation from exactness in either direction.

The upper section of Table 2 shows values of \( \delta_1(P^*, P) \) for all six approximate P-values described. The figures are all positive indicating liberalism and suppress a tendency for \( P^*(Y) - P(Y) \) to be smaller/larger when both P-values are smaller/larger. Approximate P-values based on \( r^* \) are least liberal though P-values based on the rather crude statistic \( W_1 \) are almost as good. The bottom section gives bias for the pre-estimated P-values. In most cases the bias is very small except for \( W_1 \). Indeed, \( W_1 \) is unique in there appearing to be little effect of pre-estimation on bias.
Table 2: **Liberalism of approximate P-values** Values of $\delta_1(P^*, P)$ for generating statistics given in column headings. Upper section is for basic generating statistic, lower section of table is for pre-estimated P-values.

<table>
<thead>
<tr>
<th>$(n_0, n_1)$</th>
<th>$W_1$</th>
<th>$W_2$</th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$r$</th>
<th>$r^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(20, 50)</td>
<td>0.059</td>
<td>0.195</td>
<td>0.122</td>
<td>0.064</td>
<td>0.097</td>
<td>0.048</td>
</tr>
<tr>
<td>(50, 20)</td>
<td>0.112</td>
<td>0.065</td>
<td>0.109</td>
<td>0.064</td>
<td>0.078</td>
<td>0.062</td>
</tr>
<tr>
<td>(50, 75)</td>
<td>0.050</td>
<td>0.125</td>
<td>0.070</td>
<td>0.048</td>
<td>0.065</td>
<td>0.042</td>
</tr>
<tr>
<td>(75, 50)</td>
<td>0.084</td>
<td>0.081</td>
<td>0.082</td>
<td>0.063</td>
<td>0.068</td>
<td>0.041</td>
</tr>
<tr>
<td>(60, 60)</td>
<td>0.066</td>
<td>0.098</td>
<td>0.069</td>
<td>0.056</td>
<td>0.067</td>
<td>0.044</td>
</tr>
<tr>
<td>(50, 200)</td>
<td>0.028</td>
<td>0.271</td>
<td>0.190</td>
<td>0.073</td>
<td>0.127</td>
<td>0.048</td>
</tr>
<tr>
<td>(200, 50)</td>
<td>0.069</td>
<td>0.036</td>
<td>0.066</td>
<td>0.042</td>
<td>0.045</td>
<td>0.038</td>
</tr>
<tr>
<td>(20, 50)</td>
<td>0.030</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
<td>0.008</td>
</tr>
<tr>
<td>(50, 20)</td>
<td>0.008</td>
<td>0.002</td>
<td>0.003</td>
<td>0.002</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>(50, 75)</td>
<td>0.026</td>
<td>0.004</td>
<td>0.003</td>
<td>0.004</td>
<td>0.003</td>
<td>0.004</td>
</tr>
<tr>
<td>(75, 50)</td>
<td>0.018</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.004</td>
</tr>
<tr>
<td>(60, 60)</td>
<td>0.030</td>
<td>0.011</td>
<td>0.011</td>
<td>0.011</td>
<td>0.011</td>
<td>0.011</td>
</tr>
<tr>
<td>(50, 200)</td>
<td>0.045</td>
<td>0.015</td>
<td>0.015</td>
<td>0.015</td>
<td>0.015</td>
<td>0.015</td>
</tr>
<tr>
<td>(200, 50)</td>
<td>0.004</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.001</td>
<td>0.004</td>
</tr>
</tbody>
</table>

7 **Numerical study of mean size**

If two P-values agree in their ordering and one is smaller than the other then the power of the smaller one uniformly dominates the power of the larger one. If two exact P-values are both exact then we would prefer the one that is stochastically smaller. While this does not completely replace a power study, it gives a comparison that is independent of nominal size. The first issue is the extent to which pre-estimation results in a better i.e. smaller exact P-value. We measure this by $\delta_1(P^*, \hat{P}^*)$ and positive values indicate that the pre-estimated P-values leads to smaller i.e. better exact P-values. We secondly look at which of the exact P-values is smallest by measuring it mean size. Write $\delta_1(P_1, P_2) = \mu(P_1) - \mu(P_2)$ where

$$\mu(P^*) = \int_0^1 E(P^*|R, \psi_0, p_0)dp_0.$$  

Throughout this section, the conditioning set $R$ taken to be the subset of the sample space for which either $\hat{P}$ or $\hat{P}^*$, both based on the LR statistic, is less that 0.5. Table 3 list values of $\delta_1(P^*, \hat{P}^*)$ for the six basic pivotals in this study. In virtually every case, pre-estimation reduces the mean size of the exact P-value, in most cases by a non-
trivial amount. As anticipated from earlier results, the advantage of pre-estimation is consistent less for P-values based on $r^*$. This effect is clearest for larger sample sizes. In other words, if one is to go to the trouble of calculating and exact P-value by maximisation, the preliminary E-step which is computationally intensive for larger sample sizes, is less critical, though still advantageous.

Table 3: **Advantage of pre-estimation for six generating P-values.** Values of $\delta_1(P^*, \hat{P}^*)$ for generating statistics given in column headings. In almost every case, the exact P-values based on the pre-estimated P-value is systematically smaller than the exact P-value based on the original P-value.

<table>
<thead>
<tr>
<th>$(n_0, n_1)$</th>
<th>$W_1$</th>
<th>$W_2$</th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$r$</th>
<th>$r^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(20, 50)</td>
<td>0.034</td>
<td>0.150</td>
<td>0.082</td>
<td>0.030</td>
<td>0.059</td>
<td>0.023</td>
</tr>
<tr>
<td>(50, 20)</td>
<td>0.074</td>
<td>0.044</td>
<td>0.079</td>
<td>0.043</td>
<td>0.047</td>
<td>0.046</td>
</tr>
<tr>
<td>(50, 75)</td>
<td>0.022</td>
<td>0.104</td>
<td>0.052</td>
<td>0.031</td>
<td>0.048</td>
<td>0.029</td>
</tr>
<tr>
<td>(75, 50)</td>
<td>0.038</td>
<td>0.053</td>
<td>0.051</td>
<td>0.032</td>
<td>0.037</td>
<td>0.028</td>
</tr>
<tr>
<td>(60, 60)</td>
<td>0.024</td>
<td>0.070</td>
<td>0.038</td>
<td>0.025</td>
<td>0.036</td>
<td>0.024</td>
</tr>
<tr>
<td>(50, 200)</td>
<td>-0.001</td>
<td>0.273</td>
<td>0.187</td>
<td>0.056</td>
<td>0.116</td>
<td>0.027</td>
</tr>
<tr>
<td>(200, 50)</td>
<td>0.054</td>
<td>0.040</td>
<td>0.062</td>
<td>0.037</td>
<td>0.039</td>
<td>0.029</td>
</tr>
</tbody>
</table>

Table 4 lists mean sizes for 12 exact P-values, with the conditioning set taken to be the subset of the sample space for which either $P$ or $\hat{P}$, both based on the LR statistic, is less that 0.5. The results suggest strongly that if pre-estimation is not to be employed then P-values based on $r^*$ are preferred. The results also indicate that pre-estimation leads to exact P-values which have almost identical performance, apart from $W_1$. It would seem that the E-step results in P-values that order the sample space in the same way, almost regardless of the original generating P-value. This is confirmed by Table 5 which displays the values $\delta_2(\hat{P}_1^*, \hat{P}_2^*)$ for different basic generating P-values ($P_1, P_2$). For instance, the figure 0.04446 in the first row indicates that the rms difference between EM P-values based on $W_1$ and those absed on $W_2$ is 0.0446. It is clear from this table that, apart from $W_1$, it makes not practical difference which of the other 5 P-values are used to generate the EM P-value. The same conclusion is true for the other sample sizes (results not presented).
Table 4: **Mean size of alternative exact P-values.** Values of $\mu(P^*)$ for generating statistics given in column headings. The conditioning set $R = \{P^*(Y) \leq 0.5\} \cup \{\hat{P}^*(Y) \leq 0.5\}$. Lower values are preferred. The upper section of the table is for the basic P-values. The lower section of the table is for the pre-estimated P-value.

<table>
<thead>
<tr>
<th>$(n_0, n_1)$</th>
<th>$W_1$</th>
<th>$W_2$</th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$r$</th>
<th>$r^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(20,50)</td>
<td>0.321</td>
<td>0.416</td>
<td>0.348</td>
<td>0.295</td>
<td>0.325</td>
<td>0.288</td>
</tr>
<tr>
<td>(50,20)</td>
<td>0.334</td>
<td>0.298</td>
<td>0.334</td>
<td>0.297</td>
<td>0.301</td>
<td>0.300</td>
</tr>
<tr>
<td>(50,75)</td>
<td>0.301</td>
<td>0.361</td>
<td>0.308</td>
<td>0.288</td>
<td>0.304</td>
<td>0.286</td>
</tr>
<tr>
<td>(75,50)</td>
<td>0.308</td>
<td>0.307</td>
<td>0.304</td>
<td>0.285</td>
<td>0.290</td>
<td>0.284</td>
</tr>
<tr>
<td>(60,60)</td>
<td>0.293</td>
<td>0.318</td>
<td>0.286</td>
<td>0.273</td>
<td>0.284</td>
<td>0.272</td>
</tr>
<tr>
<td>(50,200)</td>
<td>0.262</td>
<td>0.503</td>
<td>0.417</td>
<td>0.286</td>
<td>0.346</td>
<td>0.257</td>
</tr>
<tr>
<td>(200,50)</td>
<td>0.309</td>
<td>0.293</td>
<td>0.315</td>
<td>0.290</td>
<td>0.292</td>
<td>0.284</td>
</tr>
<tr>
<td>(20,50)</td>
<td>0.287</td>
<td>0.266</td>
<td>0.266</td>
<td>0.265</td>
<td>0.266</td>
<td>0.265</td>
</tr>
<tr>
<td>(50,20)</td>
<td>0.260</td>
<td>0.254</td>
<td>0.255</td>
<td>0.254</td>
<td>0.254</td>
<td>0.254</td>
</tr>
<tr>
<td>(50,75)</td>
<td>0.279</td>
<td>0.257</td>
<td>0.256</td>
<td>0.257</td>
<td>0.256</td>
<td>0.257</td>
</tr>
<tr>
<td>(75,50)</td>
<td>0.270</td>
<td>0.254</td>
<td>0.253</td>
<td>0.253</td>
<td>0.253</td>
<td>0.253</td>
</tr>
<tr>
<td>(60,60)</td>
<td>0.269</td>
<td>0.248</td>
<td>0.248</td>
<td>0.248</td>
<td>0.248</td>
<td>0.248</td>
</tr>
<tr>
<td>(50,200)</td>
<td>0.263</td>
<td>0.230</td>
<td>0.230</td>
<td>0.230</td>
<td>0.230</td>
<td>0.230</td>
</tr>
<tr>
<td>(200,50)</td>
<td>0.255</td>
<td>0.253</td>
<td>0.253</td>
<td>0.253</td>
<td>0.253</td>
<td>0.255</td>
</tr>
</tbody>
</table>

References.


Table 5: Closeness of 6 EM P-values. Sample size \((n_0, n_1) = (50, 75)\). Tabulated values are \(\delta_2(\hat{P}_1^*, \hat{P}_2^*)\) for each pair \(\langle P_1, P_2 \rangle\) of generating P-values. Apart from \(W_1\), there is hardly any practical difference between the alternative EM P-values.

<table>
<thead>
<tr>
<th></th>
<th>(W_1)</th>
<th>(W_2)</th>
<th>(C_1)</th>
<th>(C_2)</th>
<th>(r)</th>
<th>(r^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(W_1)</td>
<td>0.0000</td>
<td>0.0446</td>
<td>0.0444</td>
<td>0.0445</td>
<td>0.0445</td>
<td>0.0439</td>
</tr>
<tr>
<td>(W_2)</td>
<td>0.0446</td>
<td>0.0000</td>
<td>0.0037</td>
<td>0.0022</td>
<td>0.0031</td>
<td>0.0077</td>
</tr>
<tr>
<td>(C_1)</td>
<td>0.0444</td>
<td>0.0037</td>
<td>0.0000</td>
<td>0.0033</td>
<td>0.0023</td>
<td>0.0075</td>
</tr>
<tr>
<td>(C_2)</td>
<td>0.0445</td>
<td>0.0022</td>
<td>0.0033</td>
<td>0.0000</td>
<td>0.0030</td>
<td>0.0075</td>
</tr>
<tr>
<td>(r)</td>
<td>0.0445</td>
<td>0.0031</td>
<td>0.0023</td>
<td>0.0030</td>
<td>0.0000</td>
<td>0.0076</td>
</tr>
<tr>
<td>(r^*)</td>
<td>0.0439</td>
<td>0.0077</td>
<td>0.0075</td>
<td>0.0075</td>
<td>0.0076</td>
<td>0.0000</td>
</tr>
</tbody>
</table>


Figure 1: Profiles for four generating P-values. Data is \((y_0, y_1) = (48, 14), (n_0, n_1) = (283, 47)\) and \(\delta = 0.1\). Each panel shows the profile \(\Pi(P, y, \lambda)\) versus \(\lambda\). The horizontal line is \(P(y)\) and the vertical lines is \(\hat{\lambda}_\psi = 0.190\). The dotted curve is the profile function \(\Pi(P, y, \lambda)\) of the estimated P-value \(\hat{P}(y)\).
Figure 2: Modified and ordinary likelihood roots compared. Data is \((y_0, y_1) = (39, 157), (n_0, n_1) = (50, 200)\) and \(\delta = 0.1\). Upper panel shows the profiles for ordinary and pre-estimated P-values. Central panels show reduction from pre-estimation. Lower panels show distribution of reduction from pre-estimation.
Figure 3: Modified and ordinary likelihood roots compared. Data is \((y_0, y_1) = (45, 16), (n_0, n_1) = (200, 50)\) and \(\delta = 0.1\). Upper panel shows the profiles for ordinary and pre-estimated P-values. Central panels show reduction from pre-estimation. Lower panels show distribution of reduction from pre-estimation.