Melbourne Business School

From the SelectedWorks of Chris J. Lloyd

Summer 2008

On approximate conditioning and higher order asymptotics for 2x2 tables

Chris Lloyd

Available at: https://works.bepress.com/chris_lloyd/9/
On approximate conditioning and higher order asymptotics for $2 \times 2$ tables

BY CHRIS J. LLOYD

Draft: December 7, 2007

SUMMARY

For testing canonical parameters in a continuous exponential family, P-values based on higher order asymptotic formulas such as $p^*$ approximate the exact conditional P-value with great accuracy. For discrete models, the conditional distribution can be extremely discrete or even degenerate which raises the questions (a) should one try to approximate the conditional P-value, (b) what does $p^*$ approximate? Pierce and Peters (1999) have argued that $p^*$ approximates an approximately conditional P-value and that this approximately conditional P-value is an inferentially sensible quantity worth approximating. Their arguments and numerical results are oriented towards problems where the conditioning variable has 3 or 4 dimensions. We investigate the performance and logic of approximately conditional P-values for the case of $2 \times 2$ tables, as well as the extent to which $p^*$ functions as an approximation to these P-values. We conclude that approximately conditional P-values have rather erratic properties and suffer from a logical flaw. We also find that the mid-P value approximates them as well or better than $p^*$, but that neither approximations work well when the observed data is near the boundary of the sample space.

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

1 Introduction

Recent advances in likelihood theory have seen so-called higher order approximations to P-values for scalar parameters. An overview of the theory may be found in Barndorff-Nielsen and Cox (1994) and Reid (2003), and many details of practical implementation are in Brazzale, Davison and Reid (2007). For canonical parameters in continuous exponential families, the methods reduce to a double saddlepoint approximation to the tail probability of the estimator conditional on the sufficient statistics for the nuisance

---

1Melbourne Business School, Carlton, 3053, AUSTRALIA. email: c.lloyd@mbs.edu
parameters. The approximations are extremely accurate and conditioning for continuous models is well accepted. While there are various versions of these P-values, we will refer to them generically as \( p^* \).

For discrete models, the situation is less clear. Firstly, there is controversy over whether a conditional P-value is appropriate. Secondly, conditional P-values can become degenerate in which case it is not clear what \( p^* \) is approximating. Thirdly, it is not clear if a continuity correction should be applied. Pierce and Peters (1992) show that a continuity corrected version of \( p^* \) can give accurate approximations to conditional P-values, for both logistic regression and inference on a common log-odds ratio from several \( 2 \times 2 \) tables. Explicit error rates or not given.

Davison, Fraser and Reid (2006) argue that \( p^* \) approximates P-values from the conditional distribution with error \( O(n^{-1}) \) and verify this numerically for several examples including a single \( 2 \times 2 \) table. When a continuity is not applied, \( p^* \) approximates the mid P-value from this same conditional distribution. This follows an interpretation of Davison and Wang (2002) based on embedding the discrete model in a continuous model. The conditional distributions involved in their example are not close to degenerate and the data is not near the boundary of the sample space. Pierce and Peters (1999) argue that \( p^* \) without continuity correction is an approximation to a kind of ‘approximate conditional’ P-value which does not suffer from the degeneracy problems of fully conditional P-values. One is given the impression from reading this literature that the issues of conditionality are automatically resolved by using \( p^* \).

The arguments are developed in detail for canonical exponential families where there is an exact conditional distribution to approximate. However, the arguments are heuristic and their ideas, if correct, should apply more widely to non-canonical parameters where the distribution being approximated is based on an approximate local conditionality resolution.

There is a quite different strand of recent research into exact unconditional tests and confidence intervals where the nuisance parameter is eliminated by maximisation and other methods (Lloyd, 2008). One of the themes of this research is the search for good approximate statistics which generate the exact methods (Kabaila and Lloyd, 2002). A key requirement is that the generating statistics should perform well, more importantly equally well, across the entire sample space. It is already known that using conditional P-values and confidence limits to generate exact unconditional tests and confidence limits can give higher efficiency than using approximate unconditional generators such as those based on the likelihood root or Wald statistic, see Bochloo.
(1970) in the testing context and Lloyd and Moldovan (2007) in the confidence limits context. This research raises the possibility of using higher order asymptotic statistics as generators of exact inference. The question then arises as to what $p^*$ approximates, how sensitive it is to nuisance parameters and how well it performs in extreme cases.

The immediate purpose of this paper is first to examine the extent to which $p^*$ approximates the approximately conditional inference better than the conceptually simple mid-P and secondly to consider whether approximately conditional P-values respect the logic of conditionality. We resolve both issues in the negative. A broader purpose is to understand how $p^*$ performs in the kind of low dimensional binomial models that arise in clinical trials and to anticipate their performance as generators of exact procedures.

2 Conditional inference

There is a long history of controversy concerning conditional inference. There are a range of quite different arguments for conditioning. One argument is that conditioning on a sufficient statistic $T$ for the nuisance parameters gives a model depending only on the interest parameters which is available for inference. This is a convenience argument as nuisance parameters can be eliminated in other ways (Basu, 1977), for instance by maximisation or several levels of estimation. A second argument is that in full exponential families uniformly most powerful unbiased tests are based on conditional tail probabilities for continuous models and on randomised versions for discrete models (Lehmann, 1986). This suggests that the conditional likelihood contains, in some sense, all the relevant information about the interest parameter. These first two arguments are practical in the sense that if there exist other inferential methods that are as convenient and powerful one would not be compelled to perform conditional inference.

The remaining arguments for conditioning are logical and involve a claim that unconditional inferences are, or can be, epistemologically ‘wrong’. The first of these is that correct inferences should recognise the different levels of precision that may arise in an experiment (Cox, 1958) and that this can be captured by conditioning on an appropriate statistic $T$. Such a situation arises when the sample size is chosen by a random number generator in which case the sample size should be treated as fixed even though it is random. For this purpose, the conditioning variable $T$ should be an index of precision and may not even be sufficient for the nuisance parameters or distributed free of the interest parameter. A quite different argument against unconditional inference is
that when $T$ contains no information about the parameter of interest, observed values of $t$ should not be counted for or against hypotheses about the parameter. I develop this new argument in section 6 and show that it applies also to approximate conditioning.

The main argument against conditioning is that, in discrete examples, a great deal of statistical power can be sacrificed in eliminating the nuisance parameter (Berkson, 1978). It is a mathematical fact that P-values, conditional or not, derived from discrete data are conservative because unconditionally they tend to be larger than uniform. Conditional P-values supposedly suffer more from this than unconditional P-values since they are more discrete. However, this argument assumes what it is trying to prove since unconditional P-values are just as discrete as conditional ones when viewed conditionally. A more convincing argument against conditional inference is that when the conditional distribution becomes extremely discrete the discreteness dominates the inference, and can easily lead to no inference at all. For instance, in a simple logistic regression, inference on the intercept parameter is typically degenerate if the covariates are real valued. Unless one believes that frequentist inference on the intercept of a logistic regression is impossible, this is a powerful argument against dogmatic elimination of nuisance parameters by conditioning.

A possible way forward is to condition on a cruder version of the conditioning variable, sometimes called approximate conditioning (Cox, 1984). The hope is that the approximately conditional inference respects the logic of conditioning, namely the precision indexing property, while producing a model which need not suffer from excessive discreteness. The approximately conditional model will depend on the nuisance parameter but, it is argued, only slightly. Two obvious obstacles remain, namely how approximate should the conditioning be and how does one calculate the approximately conditional P-value?

The plan of the remainder of the paper is as follows. Section 3 gives notation and formulas for exact conditional and higher order inference on the odds ratio of a $2 \times 2$ table. Section 4 explains approximate conditioning for this same model. The results of a numerical investigation into the relationship between $p^*$, the mid P-value and approximately conditional P-values are reported in Section 5. We raise some logical problems with approximate conditioning in Section 6. Section 7 concludes the article with some discussion and general observations.
3 Inference on two binomials

The main issues can be explicated with the example of testing for independence in a $2 \times 2$ table. Suppose that 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. Letting $\varphi_0, \varphi_1$ denote the corresponding log-odds. We take the interest parameter to be $\psi = \varphi_1 - \varphi_0$ and consider testing

$$H_0 : \psi \leq \psi_0, \quad H_1 : \psi > \psi_0,$$

for some pre-chosen $\psi_0$. Taking the nuisance parameter to be $\lambda = \varphi_0$, the log-likelihood

$$\ell(\psi, \lambda; y) = \psi y_1 + \lambda t - n_1 \log(1 + e^{\psi + \lambda}) - n_0 \log(1 + e^\lambda)$$

where $t$ is the observed value of $T = Y_0 + Y_1$.

The distribution of $Y_1$ given $T = t$ is free of $\lambda$ and supported on the integers from $y_{\min} = \max(0, t - n_0)$ to $y_{\max} = \min(t, n_1)$ inclusive with probability function

$$p_c(y_1; t, \psi) := \Pr(Y_1 = y_1 | T = t; \psi) = \binom{n_1}{y_1} \binom{n_0}{t - y_1} e^{y_1 \psi} / \kappa(\psi)$$

where $\kappa(\psi)$ is a normalising constant. This gives two standard P-values for testing $\psi > \psi_0$ namely

$$P_{\text{tail}}(y_1; t, \psi_0) = \sum_{y=y_1}^{y_{\max}} p_c(y, \psi_0), \quad P_{\text{mid}}(y_1; t, \psi_0) = P_{\text{tail}}(y_1; t, \psi_0) - 0.5p_c(y_1; t, \psi_0).$$

The mid P-value is not a P-value in the strict sense of having a quasi-uniform distribution conditional on $T = t$ but to the extent that it is computed from the conditional distribution it respects conditionality.

First order approximate inference is based on quadratic approximations to the log-likelihood (1). The likelihood root statistics is

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

where $\hat{\lambda}$ is the restricted maximum likelihood estimate of $\lambda$ which solves $\partial \ell(\psi, \lambda)/\partial \lambda$ which reduces to a quadratic in this case. Second order inference is achieved by referring a modified likelihood root statistics $r^*(\psi)$ to the standard normal distribution. This modified root is given by

$$r^*(\psi) = r(\psi) + r(\psi)^{-1} \log \{ q(\psi) / r(\psi) \}$$
where \( q(\psi) \) is in general complicated but in this case reduces to

\[
q(\psi) = (\hat{\psi} - \psi) \sqrt{\frac{\hat{V}_0 \hat{V}_1}{V_{0\psi} + V_{1\psi}}},
\]

where \( V_j = n_j \pi_j (1 - \pi_j) \) and subscript-\( \psi \) indicates restricted ML estimation. The second order P-value is \( p^*(\psi) = \Phi(-r^*(\psi)) \). Note that \( r^*(\psi) \) breaks down in the centre of the distribution where \( \hat{\psi} \) and \( \psi \) 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 \( \psi \).

Davison, Fraser and Reid (2006) argue that \( p^* \) approximates \( p_{mid} \) with relative error \( O(n^{-1}) \). Pierce and Peters (1999) suggest that \( p^* \) approximates an approximately conditional inference, and that it does so better than \( p_{mid} \).

### 4 Approximate conditioning

As described in Section 2, the main difficulty with using the conditional distribution (2) is that the support can become so small that the discreteness has more effect on the inference that the information in the distribution. Discreteness also results in the unconditional distribution of \( P_{tail}(Y_1; T, \psi_0) \) being stochastically larger than uniform.

The idea of approximate conditioning is to restrict attention to \( T \) being within a neighbourhood \( \mathcal{N}_r(t) \) of \( t \) where \( r \) represents some kind of distance. The hope is that the resulting inference respects conditionality while producing a less discrete distribution. Let \( \pi(Y) \) be a possibly approximate P-value that is used to order the elements of the sample space in terms of their hostility to the null and let \( \pi_{obs} \) be the observed value. Then for any \( r \geq 0 \) an approximately conditional P-value is

\[
P_r(y, \psi_0, \lambda) = \Pr \{ \pi(Y) \leq \pi_{obs} | T \in \mathcal{N}_r(t); \lambda \} = \sum_{\tau \in \mathcal{N}_r(t)} \Pr \{ \pi(Y) \leq \pi_{obs} | T = \tau \} \Pr \{ T = \tau | T \in \mathcal{N}_r(t); \lambda \}
\]

with all probabilities calculated under the null. The ordering function \( \pi \) only affects the results through its ordering of the sample space. Peters and Pierce suggest using \( \pi = p^* \) though it turns out in our examples that identical results are obtained from \( \pi = P_{tail} \).

Unless \( r = 0 \) when the P-value is fully conditional, \( P_r(y, \psi_0, \lambda) \) depends on \( \lambda \) through the distribution of \( T \) given \( \mathcal{N}_r(t) \). Pierce and Peters suggest this dependence will be slight, on the basis of which they recommend replacing \( \lambda \) with the null estimate \( \hat{\lambda}_0 \), giving the computable P-value \( P_r(y, \psi_0, \hat{\lambda}_0) \).
From (5) we see that $P_r(y, \psi_0, \lambda)$ is the mean value of $\pi(T) := \Pr(\pi(Y) \leq \pi_{\text{obs}} | T)$ with respect to the conditional distribution of $T$ given $N_r(t)$. It is not at all obvious that the distribution of $T$ given $N_r(t)$ will depend only slightly on $\lambda$. So dependence of $P_r(y, \psi_0, \lambda)$ depends on how $\pi(T)$ depends on $T$. It is easy to show that $\pi(T)$ equals $P_{\text{tail}}$ for the observed value of $T$ and is smaller than this for all other values of $T$. The more discrete the conditional distributions the smaller these alternative values will be. So it would appear that in those examples where the dimension of $T$ is higher than 1 and where the conditional distribution becomes closer to degenerate, that $P_r(y, \psi_0, \lambda)$ may depend significantly on $\lambda$ even when $r$ is small.

There are two suggestions in Pierce and Peters (1999). Firstly, they suggest that $P_r(y, \psi, \hat{\lambda}_0)$ is a sensible inferential quantity with good performance characteristics. Secondly, they argue that $p^*(\psi_0)$ is an approximation to $P_r(y, \psi_0, \hat{\lambda}_0)$ for some appropriate but unspecified margin of conditioning $r$. They clearly have in mind models where the dimension of the nuisance parameter is greater than 1. In such circumstances, conditional distributions can be both difficult to compute and extremely discrete, so an alternative procedure is well worth exploring. In the investigation to follow, we are looking at the a very low dimensional problem.

5 Numerical investigation

For inference on two binomials, the conditioning statistic $T$ is one dimensional so it is natural to take $N_r(t) = \{t - r \leq T \leq t + r\}$. We will investigate the example of Pierce and Peters (1999) where $(n_0, n_1) = (15, 85)$ and $(y_0, y_1) = (4, 45)$ and we want to test for the alternative $\psi > 0$. It is easy to calculate $p^* = 0.0828$, $P_{\text{tail}} = 0.1312$ and $P_{\text{mid}} = 0.0882$ (Pierce and Peters quote the erroneous value 0.098). The left panel of Figure 1 displays $P_r(y; \lambda)$ as a function of $\lambda$ for $r = 1$.

The idea of approximately conditional inference is that for small values of $r$, the P-value is almost conditional and so dependence on the nuisance parameter should be slight. It does not appear that the dependence on $\lambda$ is slight nor does it even appear to be less for smaller values of $r$. Moreover, the actual curves seem to vary rather erratically with $r$. The vertical lines give the estimated value and 90% limits for $\lambda$. The dark horizontal line is $p^*$ and the dashed horizontal line is $p_{\text{mid}}$. In the region of the estimate $\hat{\lambda}$, $p^*$ does seem to be an adequate approximate to $P_r(y, \psi_0, \hat{\lambda}_0)$ though $P_{\text{mid}}$ seems rather better. The right panel plots $P_r(r, \psi_0, \hat{\lambda}_0)$ against $r$ for $r = 0, ..., 10$. For this data set $P_{\text{mid}}$ is close to $P_3$ or $P_4$ while $p^*$ is smaller even than the unconditional
P-value, which is practically identical to $P_{10}$.

As a second example, we take $(n_0, n_1) = (19, 7)$ and $(y_0, y_1) = (1, 5)$ from Brazzale, Davison and Reid (2007, p24). In fact, we will allow $y_1$ to vary from 2 to their observed value 5. The panels of Figure 2 correspond to these four different values of $y_1$ and $\psi_0 = 0$. Each panel displays $P_r(y, \psi_0, \lambda)$ as a function of $\lambda$ for $r = 1, ..., 7$. There is considerable dependence on $\lambda$, which is perhaps to be expected since the sample sizes are smaller. It is not clear whether $P_{\text{mid}}$ or $p^*$ is a better approximation to $P_r$.

To investigate the issue of how close $P_r(y, \psi_0, \hat{\lambda}_0)$ is to $p^*$ or $P_{\text{mid}}$ more generally, we calculated all possible values of these statistics for three values of $(n_0, n_1)$. The results are summarised in Table 1. The numbers reported in the table are the average of the absolute value of the proportional difference between two P-values. The average is restricted to data points of 'statistical interest' i.e. where $P_{\text{tail}} \in (.005, 0.2)$. A similar pattern is obtained for an unrestricted average but the mean errors give a pessimistic assessment of the practical accuracy.

It is apparent that $p_{\text{mid}}$ approximates $P_r$ better than does $p^*$. The last line shows how well $p^*$ approximates $p_{\text{mid}}$. There does not appear to be any case for interpreting $p^*$ as an approximation to $P_r$ rather than the simpler interpretation that it is an approximation to $p_{\text{mid}}$. The same conclusion is reached when the difference between
Table 1: **Approximations to** $P_r$. Each figure measures the mean absolute percentage difference between two P-values, restricted to the statistically interesting part of the sample space. The column headings are sample sizes $(n_0, n_1)$.

<table>
<thead>
<tr>
<th></th>
<th>(10, 15)</th>
<th>(20, 35)</th>
<th>(5, 95)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>5.6</td>
<td>26.7</td>
<td>7.7</td>
</tr>
<tr>
<td>$P_2$</td>
<td>15.0</td>
<td>32.9</td>
<td>7.1</td>
</tr>
<tr>
<td>$P_3$</td>
<td>18.9</td>
<td>35.3</td>
<td>9.6</td>
</tr>
<tr>
<td>$P_4$</td>
<td>20.7</td>
<td>33.8</td>
<td>11.3</td>
</tr>
<tr>
<td>$P_5$</td>
<td>21.2</td>
<td>33.7</td>
<td>11.9</td>
</tr>
<tr>
<td>$P_6$</td>
<td>21.4</td>
<td>33.3</td>
<td>12.5</td>
</tr>
<tr>
<td>$p_{mid}$</td>
<td>27.1</td>
<td>17.4</td>
<td>23.8</td>
</tr>
</tbody>
</table>

two P-values is measured by their difference rather than proportional difference.

6 **Spurious deflation**

Continuing with the same example, we look at the case $(y_0, y_1) = (0, 7)$ which is the most hostile data set to the null, see Figure 3. For this data set, $p^*$ grossly over-estimates the approximately conditional P-values. The reason for this is worth investigating since it uncovers the fact that approximately conditional P-values do no respect the logic of conditionality. This is not an argument against using $p^*$ as it is still a good approximation to $p_{mid}$ which does respect conditionality.

In general, let $y_{min}$ be the data set most hostile to the null, giving the minimum value $p_{min}$ of $\pi(y)$. Let $T$ be any candidate conditioning variable and let $T(y_{min}) = t_{min}$. Conditional on $T = t \neq t_{min}$, $\pi(Y)$ is always greater than $p_{min}$ and so $\Pr(\pi(Y) \leq p_{min}|T = t)$ is zero. Thus the only non-zero term in expression (5) is the $\tau = t_{min}$ term. For the $\tau = t_{min}$ term, $\Pr(\pi(Y) \leq p_{min}|T = t) = \Pr(Y = y|T = t)$ and supposing the ordering function $\pi$ is such that $y_{min}$ is on the boundary of the sample space this must equal $P_{tail}$. Hence for the most hostile data set (5) simplifies to

$$P_r(y; \lambda) = P_{tail}(y) \times \Pr(T = t|T \in \mathcal{N}_r(t); \lambda).$$

The approximately conditional P-value equals the conditional P-value $P_{tail}$ deflated by a factor measuring the conditional probability of observing $T = t$. Note that for the most extreme data set, $P_{mid}$ equals $P_{tail}$ multiplied by the factor 0.5.
For our case, $T = Y_0 + Y_1$ is the total number of positive responses and $\Pr(T = t | T \in N_r(t); \lambda)$ is readily calculable from the binomial distribution of $T$ and takes a maximum value very near to $\hat{\lambda}$. The conditioning event $N_r(t)$ contains roughly $2r + 1$ times as many points as the set $\{T = t\}$ and the conditional probability is a strictly decreasing function of $r$. It rather quickly converges to the unconditional probability that $T = t$.

The pertinent question then is why should the event that there were $t_{\min}$ positive responses in total be counted against the null hypothesis that $\psi = 0$? While many authors have argued that $T$ contains subtle information about $\psi$ in conjunction with $Y_1$, no-one has argued that it contains direct information by itself. Except for the case of fully conditional inference when $\Pr(T = t | T \in N_r(t)) = 1$, the approximately conditional P-values directly counts the observed value of $t$ against the null. I will call this phenomenon spurious deflation.

The problem is most easily understood for the most extreme data set but is not restricted to this case. Much the same phenomenon is occurring for the data set $(y_0, y_1) = (1, 7)$ in Figure 4, which is the second most hostile data set. It is interesting that $p^*$ is even larger than $P_{\text{tail}}$ in both this and the previous case. However, it should be mentioned that $p^*$ is calculated differently for these boundary cases. A continuity correction $(y_0 + \frac{1}{2}, y_1 - \frac{1}{2})$ is first applied to the data which makes $p^*$ an approximation to $P_{\text{tail}}$. This is then deflated by a factor 0.5 to give an approximation to $P_{\text{mid}}$. This is Pierce and Peter's recommendation.

Some explanation is in the right section of Table 2. The conditional P-value places $(y_0, y_1) = (1, 7)$ as the most extreme amongst the 8 outcomes with $t = 8$. Even relaxing this conditioning to the set $T \in \{7, 8, 9\}$ sees the observed counted amongst the two most extreme of 24 outcomes. Wider conditioning quickly sees these same two outcomes compared against an ever increasing reference set and smaller resulting P-values. For the less extreme outcome $(1, 6)$ this effect of simply adding points to the reference set becomes apparent after $r = 3$. This analysis suggests that the closer a sample point is to the most extreme, the more any approximately conditional P-value will be spuriously deflated. However, if $r$ is small and the sample point is not close to extreme then there does not seem to be a tendency for spurious deflation.

It should be conceded that Pierce and Peters have in mind problems where the nuisance parameter has dimension greater than 1, so that even for $r = 1$ the reference set around $T = t$ becomes much larger so that small values of $r$ are sufficient to enjoy the gains of a larger support. However, it is far from clear that spurious deflation is
Table 2: **The logic of approximately conditional P-values.** For different conditional ranges \( r \), the table lists the number of points in the set \( \{ t - r \leq T \leq t + r \} \) and the number of these that are more hostile to the null than the observed. The sample sizes are \((n_0, n_1) = (19, 7)\)

<table>
<thead>
<tr>
<th>( r )</th>
<th>((y_0, y_1) = (1, 6))</th>
<th>((y_0, y_1) = (1, 7))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2/8</td>
<td>1/8</td>
</tr>
<tr>
<td>1</td>
<td>4/23</td>
<td>2/24</td>
</tr>
<tr>
<td>2</td>
<td>5/37</td>
<td>2/39</td>
</tr>
<tr>
<td>3</td>
<td>6/50</td>
<td>2/53</td>
</tr>
<tr>
<td>4</td>
<td>6/62</td>
<td>2/66</td>
</tr>
<tr>
<td>5</td>
<td>6/73</td>
<td>2/78</td>
</tr>
<tr>
<td>6</td>
<td>6/83</td>
<td>2/89</td>
</tr>
<tr>
<td>7</td>
<td>6/92</td>
<td>2/99</td>
</tr>
</tbody>
</table>

less of a problem in higher dimensions. On the contrary, when \( T \) has dimension greater than 1, the probability of it taking any particular value will be smaller and so spurious deflation will be more of a problem.

### 7 Discussion

The fact that both \( p^* \) and \( P_{\text{mid}} \) tend to be larger than the approximately unconditional P-value for extreme data sets can be considered a good property. It supports the conjecture that these statistics might provide a good ordering of the sample space with respect to hostility to the null.

The dependence of \( P_r(y, \psi_0, \lambda) \) on \( \lambda \) in the above examples is not negligible and so replacing \( \lambda \) with an estimate does not really give an approximately conditional P-value. The other standard method of dealing with nuisance parameter is to maximise it. It is apparent from the plots and unreported numerical work that \( p^* \) approximates \( \sup_\lambda P_r(y, \psi_0, \lambda) \) more poorly than it does \( P_r(y, \psi_0, \hat{\lambda}) \).

We have identified spurious deflation as a logical problem with approximate conditional P-values. There is another logical difficulty which I call incoherence. The conditioning sets \( \mathcal{N}_r(t) \) do not partition the event space, rather they overlap. For instance with \( r = 1 \), a given sample point \( y \) may be embedded in one of three neighbourhoods. Not only is this hard to conceptualise, but it leads to the approximate conditional P-values \( P_r(y, \psi_0, \lambda) \) possibly violating the ordering \( \pi(y) \) (except for the
fully conditional and fully unconditional case). The incoherence property applies also to the usual interpretation of $p^*$, which is derived from a local conditionality resolution of the model, but where the approximate ancillary direction changes with each data set. The collection of conditioning sets again do not partition the sample space.

References.


Figure 2: **Approximately conditional P-values.** Each panel plots $P_r(y, \psi_0, \lambda)$ against $\lambda$ for $r = 1, \ldots, 7$. Vertical lines give statistical estimate and range for $\lambda$. Horizontal lines are $p^*$ (dark) and $P_{\text{mid}}$ (dashed).
Figure 3: **Extreme data set** \((y_0, y_1) = (0, 7)\). Left panel plots \(P_r(0, 7, \psi_0, \lambda)\) against \(\lambda\) for \(r = 1, \ldots, 7\). Right panel plots \(P_r(0, 7, \psi_0, \hat{\lambda})\) versus \(r\).

Figure 4: **Almost extreme data set** \((y_0, y_1) = (1, 7)\). Left panel plots \(P_r(1, 7, \psi_0, \lambda)\) against \(\lambda\) for \(r = 1, \ldots, 7\). Right panel plots \(P_r(1, 7, \psi_0, \hat{\lambda})\) versus \(r\).