



Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspiP-values based on approximate conditioning and p^*

Chris J. Lloyd

University of Melbourne, Carlton 3053, Australia

ARTICLE INFO

Article history:

Received 13 November 2008

Received in revised form

22 May 2009

Accepted 21 October 2009

Available online 29 October 2009

Keywords:

Nuisance parameters

Exact test

Tests of independence

 p^*

ABSTRACT

P-values based on higher order asymptotic formulas such as p^* are now readily available for practitioners. However, it is not always clear what these P-values mean for discrete models. For a canonical parameter, p^* should approximate a tail probability of the conditional distribution. Yet when this conditional distribution becomes degenerate, p^* still gives a non-degenerate answer. So there is the need for a more general interpretation of p^* . Pierce and Peters (1999) have argued that p^* approximates an approximately conditional P-value and, implicitly, that this is an inferentially sensible quantity worth approximating. We investigate these twin claims for the simple case of 2×2 tables. We find that approximately conditional P-values have rather erratic properties and that they are not especially well approximated by p^* , particularly when the observed data are near the boundary of the sample space. We also argue that approximately conditional P-values suffer from two, previously unrecognised, logical flaws. The consequences of these conclusions are discussed.

© 2009 Elsevier B.V. All rights reserved.

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-Neilsen and Cox (1994) and Reid (2003), and many details of practical implementation are in Brazzale et al. (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 parameters. For non-canonical parameters, the tail probability is conditional on an approximate sufficient statistic. The approximations are extremely accurate and conditioning for continuous models is well accepted. While there are various versions of these P-values, we 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 a non-degenerate p^* is approximating. Thirdly, it is unclear 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 2×2 tables. Explicit error rates are not given.

Davison et al. (2006) argue that p^* approximates the mid-P-value from the conditional distribution with error $O(n^{-1})$ and verify this numerically for several examples including a single 2×2 table. The conditional distributions they investigate are not close to degenerate, nor is the data near the boundary of the sample space. What does p^* mean when the conditional distribution approaches degeneracy? Pierce and Peters (1999) argue that it approximates what they call an 'approximately conditional' P-value. After reading this literature, one is left with the impression that issues of conditionality are automatically resolved by using p^* .

E-mail address: c.lloyd@mbs.edu

The purpose of this paper is to critically examine these ideas in the simple context of a 2×2 table. We assess numerically the extent to which p^* approximates the approximately conditional inference and uncover very little evidence for the assertion. We also look at some problematic logical properties of approximately conditional P-values which undermine them as a basis for justifying p^* . We note that p^* itself shares one of these logical defects.

The plan of the paper is as follows. Section 2 presents a brief overview of conditioning and the controversies surrounding this issue. We identify some of the key reasons for conditioning with a view to assessing whether approximately conditional inferences have these key features approximately. In Section 3, explicit formulas are given for the conditional P-value and p^* for the case of comparing two binomials. Section 4 explains approximate conditional P-values as well as some recent modifications of the idea. In Section 5 we numerically assess the extent to which approximate conditioning removes dependence on the nuisance parameter as well as how well it is approximated by p^* . The remainder of the paper concerns some previously unrecognised logical flaws in approximately conditional inference. Section 6 introduces the idea of spurious deflation, an undesirable inferential feature which conditioning avoids, and illustrates how approximately conditional P-values fails to avoid the problem, even approximately. In Section 7, we discuss a logical problem in interpreting both approximately conditional P-values and p^* . This problem is not shared by fully conditional or unconditional inference. It is the half-way nature of approximate conditioning that leads to the problem. The consequences of these conclusions are discussed in the final section.

2. Conditional inference

There is a long history of controversy concerning conditional inference and it is not our intention to resolve the issue here. There are, however, 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 then available for inference. This is a convenience argument as nuisance parameters can be accounted for in other ways (Basu, 1977). For instance if $P(Y, \theta_0, \psi)$ is a valid P-value for testing θ when the nuisance parameter ψ is known and if C_γ is a $(1 - \gamma)$ confidence region for ψ then

$$\sup\{P(Y, \theta_0, \psi); \psi \in C_\gamma\} + \gamma \tag{1}$$

is a valid P-value, as shown by Berger and Boos (1994). The case $\gamma = 0$ corresponds to full maximisation which is the usual definition of a P-value in the presence of nuisance parameters (Bickel and Doksum, 1977). By a P-value being valid, we mean that it is stochastically no smaller than a uniform variable under the null.

A second argument is that in full exponential families, conditional P-values generate uniformly most powerful unbiased tests (Lehmann, 1986). This suggests that the conditional likelihood contains, in some sense, all the relevant information about the interest parameter θ . 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 claims 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 $T = t$. That a different value of t might have occurred is deemed irrelevant. A second argument, which has not been as well articulated, is that when T contains no information about θ , observed values of t should not be counted for or against hypotheses about θ . I develop this idea in Section 6 and show that it argues against approximate conditioning.

For continuous models, conditional inferences are also correct unconditionally and there is little controversy about conditioning. However, for discrete examples, conditional tests are conservative when evaluated unconditionally. It is argued that conditioning sacrifices statistical power unnecessarily, all in the name of eliminating the nuisance parameter (Berkson, 1978). However, this argument assumes what it tries to prove since the evaluation is made from an unconditional viewpoint. 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 side-stepping the problems of excessive discreteness. The approximately conditional model will depend on the nuisance parameter but, it is argued, only slightly. This dependence can simply be ignored, by replacing ψ with an estimate, or accounted for by maximisation as in (1).

3. Inference on two binomials

The main ideas reviewed above can be explicated with the example of testing for independence in a 2×2 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 φ_0, φ_1 denote the corresponding log-odds, we

take the interest parameter to be $\theta = \varphi_1 - \varphi_0$ and test

$$\mathcal{H}_0 : \theta \leq \theta_0, \quad \mathcal{H}_1 : \theta > \theta_0$$

for some pre-chosen θ_0 . Taking the nuisance parameter to be $\psi = \varphi_0$, the log-likelihood

$$\ell(\theta, \psi; y) = \theta y_1 + \psi t - n_1 \log(1 + e^{\theta + \psi}) - n_0 \log(1 + e^\psi), \quad (2)$$

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

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

$$p_c(y; t, \theta) := \Pr(y = y | T = t; \theta) = \binom{n_1}{y} \binom{n_0}{t-y} e^{y\theta} \kappa(\theta), \quad (3)$$

where $\kappa(\theta)$ is a normalising constant. This gives two standard P-values for testing $\theta > \theta_0$ namely

$$P_{\text{tail}}(Y_1; t, \theta_0) = \sum_{y=y_1}^{y_{\max}} p_c(y; t, \theta_0), \quad P_{\text{mid}}(Y_1; t, \theta_0) = P_{\text{tail}}(Y_1; t, \theta_0) - 0.5p_c(Y_1; t, \theta_0).$$

The mid P-value is not conditionally valid but, to the extent that it is computed from the conditional distribution, it respects conditionality. Another possible P-value is the Liebermeister P-value which is the conditional P-value but applied to data that has been modified by adding a success to group 1 and a failure to group 0, as described in Seneta and Phipps (2001). This P-value has similar properties to mid-P and is not further considered.

First order approximate inference is based on quadratic approximations to the log-likelihood (2). The likelihood root statistic is

$$r(\theta) = \text{sign}(\hat{\theta} - \theta)[2\{\ell(\hat{\theta}, \hat{\psi}) - \ell(\theta, \hat{\psi}_\theta)\}],$$

where $\hat{\psi}_\theta$ is the restricted maximum likelihood estimate of ψ which solves $\partial \ell(\theta, \psi) / \partial \psi$ which reduces to a quadratic in this case. Second order inference is achieved by referring a modified likelihood root statistic

$$r^*(\theta) = r(\theta) + r(\theta)^{-1} \log\{q(\theta)/r(\theta)\} \quad (4)$$

to the standard normal distribution. In this case $q(\theta)$ reduces to

$$q(\theta) = (\hat{\theta} - \theta) \sqrt{\frac{\hat{V}_0 \hat{V}_1}{\hat{V}_{00} + \hat{V}_{10}}}, \quad (5)$$

where $V_j = n_j \pi_j (1 - \pi_j)$ and subscript- θ indicates restricted ML estimation. Details of the derivation of this formula are available from the author. The second order P-value is $p^*(\theta) = \Phi(-r^*(\theta))$. Note that $r^*(\theta)$ breaks down in the centre of the distribution where $\hat{\theta}$ and θ are close so that $q(\theta) \approx 0$ and $r(\theta) \approx 0$. This is of little practical interest, however, since there is, in this case, no statistical evidence against the value θ .

4. Approximate conditioning

As the support of the conditional distributions become smaller, discreteness effects start to dominate the inference. The conditional distribution (3) has support on $y_{\min}, \dots, y_{\max}$ which becomes smaller when t is close to 0 or $n_0 + n_1$. Even when t is moderate, discreteness results in the unconditional distribution of $P_{\text{tail}}(Y_1; T, \theta_0)$ being stochastically much larger than uniform and the implied test is conservative.

So-called approximate conditioning restricts 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 π_{obs} be the observed value. Then for any $r \geq 0$ an approximately conditional P-value is

$$P_r(y, \theta_0, \psi) = \Pr\{\pi(Y) \leq \pi_{\text{obs}} | T \in \mathcal{N}_r(t); \psi\} = \sum_{\tau \in \mathcal{N}_r(t)} \Pr\{\pi(Y) \leq \pi_{\text{obs}} | T = \tau\} \Pr\{T = \tau | T \in \mathcal{N}_r(t); \psi\} \quad (6)$$

with all probabilities calculated under the null. The ordering function π only affects the results through its ordering of the sample space. Peters and Pierce suggest using $\pi = p^*$.

When $r = 0$ and T is sufficient for ψ , the P-value is fully conditional, conditionally and unconditionally valid and does not depend on ψ . When $r > 1$, neither of these properties need hold. Firstly, P_r may be invalid. This follows from the fact that for sufficiently loose conditioning (i.e. for larger values of r) approximately conditional P-values will be close to unconditional and unconditional P-values are often liberal. Certainly no proof of the validity of P_r has ever been offered. Secondly, $P_r(y, \theta_0, \psi)$ depends on ψ through the distribution of T given $\mathcal{N}_r(t)$. Pierce and Peters suggest that dependence on ψ will be slight, on the basis of which they recommend replacing ψ with the null estimate $\hat{\psi}_0$, giving the computable P-value $P_r(y, \theta_0, \hat{\psi}_0)$. Yang and Kolassa (2005) suggest instead using the Berger-Boos device (1) with $\gamma = 0.001$ i.e. they

recommend

$$\sup\{P_r(y, \theta_0, \psi); \psi \in C_{0.001}\} + 0.001.$$

However, the Berger–Boos device only applies to valid P-values. Since P_r has not been demonstrated to be valid, neither is the suggested P-value of Yang and Kolassa necessarily valid though in practice this defect is probably very slight. It is not our intention to investigate their approach here.

How much does $P_r(y, \theta_0, \psi)$ depend on ψ ? From (6) we see that it is the mean value of $\alpha(T) := \Pr(\pi(Y) \leq \pi_{obs} | T)$ with respect to the conditional distribution of T given $\mathcal{N}_r(t)$. It is not at all obvious that the distribution of T given $\mathcal{N}_r(t)$ will depend only slightly on ψ . So dependence of $P_r(y, \theta_0, \psi)$ depends on how $\alpha(T)$ depends on T . It is easy to show that $\alpha(T)$ equals P_{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. It follows that where the dimension of T is higher than 1 (unlike our example) and where the conditional distribution becomes consequently more discrete that $P_r(y, \theta_0, \psi)$ may depend significantly on ψ even when r is small.

There are several suggestions in Pierce and Peters (1999). Firstly, they suggest that the distribution of $P_r(y, \theta, \psi)$ depends little on the nuisance parameter. Secondly, they argue that $p^*(\theta_0)$ is an approximation to $P_r(y, \theta_0, \hat{\psi}_0)$ for some appropriate but implicitly determined margin of conditioning r . In the next section we investigate these claims numerically.

5. Numerical investigation

For inference on two binomials, the conditioning statistic T is one dimensional so it is natural to take $\mathcal{N}_r(t) = \{t - r \leq T \leq t + r\}$. We will investigate the example of Pierce and Peters (1999) where $(n_0, n_1) = (15, 85)$ and $(y_0, y_1) = (5, 45)$ and we want to test for the alternative $\theta > \theta_0 = 0$. It is easy to calculate $p^* = 0.0828$, $P_{tail} = 0.1312$ and $P_{mid} = 0.0882$ (Pierce and Peters quote the erroneous value 0.098). The left panel of Fig. 1 displays $P_r(y; \psi)$ as a function of ψ for $r = 1, \dots, 7$. The vertical lines give the estimated value and 90% limits for ψ .

For small values of r , the approximately conditional P-value is almost conditional and so dependence on the nuisance parameter is expected to be slight. While what constitutes ‘slight’ is a subjective judgment, dependence on ψ is non-negligible. Nor does dependence seem to be less for smaller values of r . Moreover, the actual curves seem to vary rather erratically with r . The dark horizontal line is p^* and the dashed horizontal line is P_{mid} . In the region of the estimate $\hat{\psi}_0$, p^* does seem to be an adequate approximate to $P_r(y; \hat{\psi}_0)$ though the much simpler P_{mid} seems rather better. The right panel plots $P_r(y; \hat{\psi}_0)$ against r for $r = 0, \dots, 10$. For this data set P_{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 et al. (2007, p. 24). We allow y_1 to vary from 2 to their observed value 5. The panels of Fig. 2 correspond to these four different values of y_1 and $\theta_0 = 0$. There is considerable dependence on ψ , which is perhaps to be expected since the sample sizes are small. It is unclear if P_{mid} or p^* is a better approximation to P_r .

To investigate the issue of how close $P_r(y; \hat{\psi}_0)$ is to p^* or P_{mid} more generally, we calculate 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 absolute proportional difference between two P-values. The average is restricted to data points of ‘statistical interest’ i.e. where $P_{tail} \in (0.005, 0.2)$. A similar pattern is obtained for an unrestricted average.

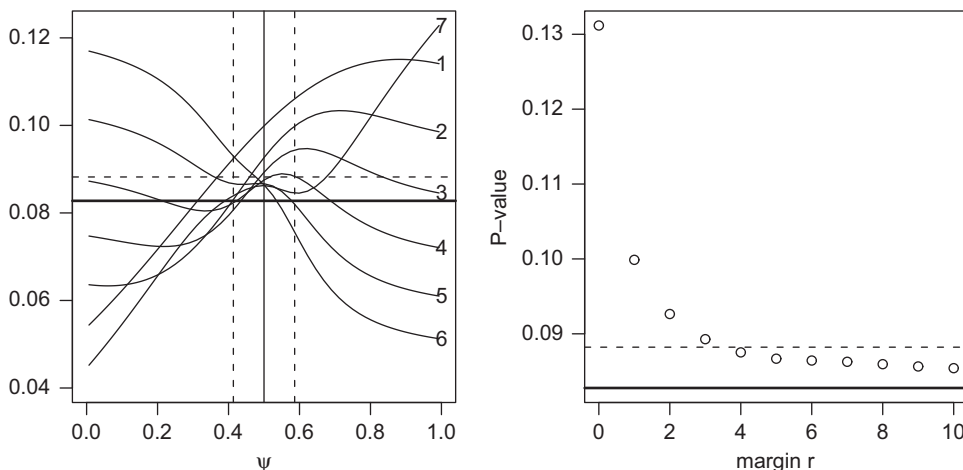


Fig. 1. Approximately conditional P-values. Left panel: $P_r(y; \lambda)$ against λ for $r = 1, \dots, 7$ for data of Pierce and Peters (1999). Vertical lines give statistical estimate and range for λ . Horizontal lines are p^* (dark) and P_{mid} (dashed). Right panel: Dependence of $P_r(y; \hat{\lambda}_0)$ on margin r .

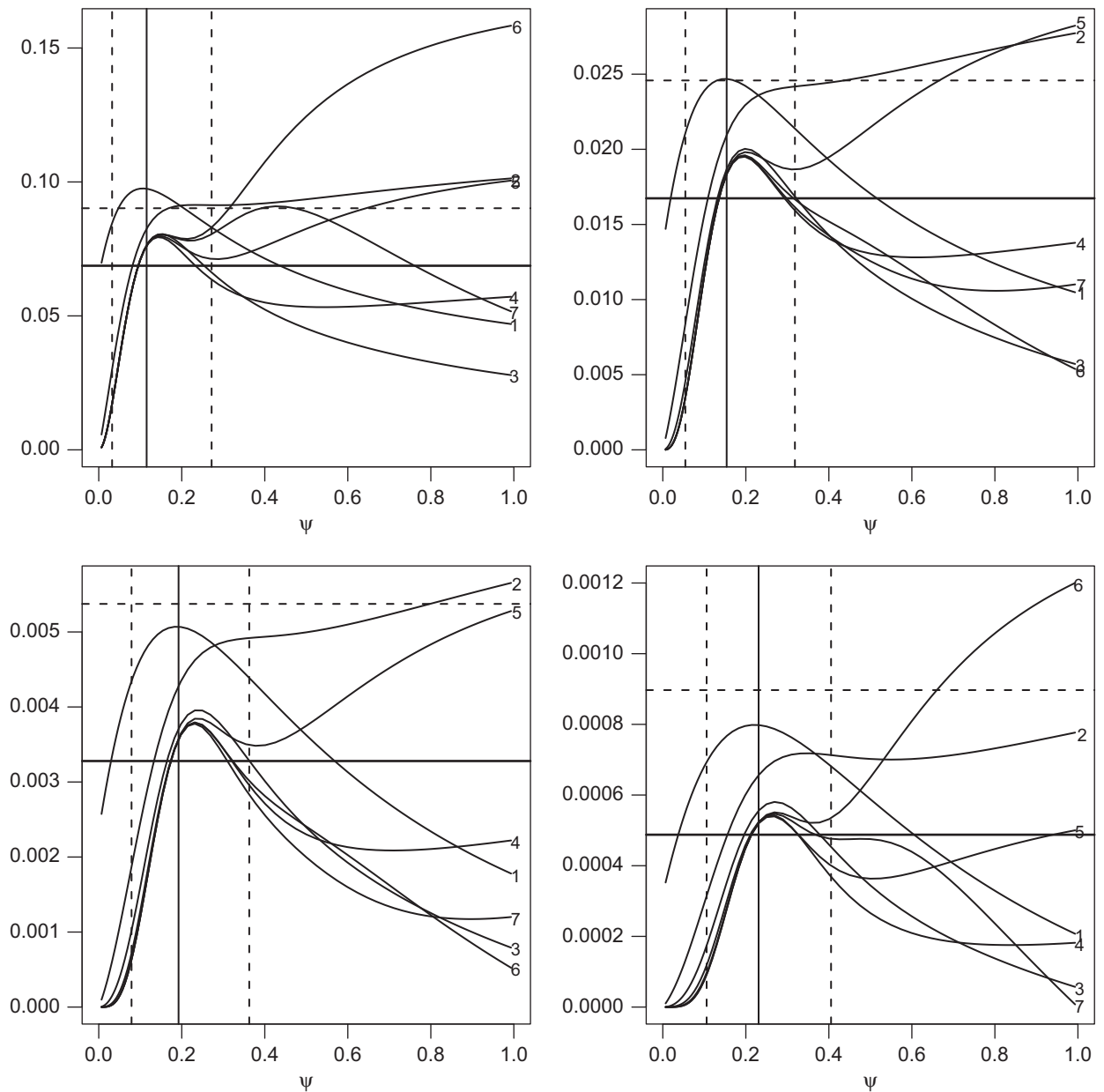


Fig. 2. Approximately conditional P-values. Each panel plots $P_r(y, \psi, \lambda)$ against λ for $r = 1, \dots, 7$. Vertical lines give statistical estimate and range for λ . Horizontal lines are p^* (dark) and P_{mid} (dashed).

It is apparent that P_{mid} approximates P_r better than does p^* . The last line shows how well p^* approximates P_{mid} . For 2×2 tables, 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_{mid} . It is acknowledged that [Pierce and Peters \(1999\)](#) have in mind models where the dimension of the conditioning variable is higher than 1, so that the support of the approximately conditional distribution is larger. Yet it is not clear why this should lead to a better approximation in any generality.

6. Spurious deflation

For the remainder of this paper, we point out that approximate conditioning fails to respect two logical properties, neither of which have been explicitly recognised in the literature. These defects undermine approximate conditioning as a foundational justification for p^* .

Consider Fisher's famous tea-tasting example where there are $y_1 = 4$ successes out of $n_1 = 4$ and $y_0 = 0$ successes out of $n_0 = 4$. For testing the alternative $\varphi_1 > \varphi_0$ the conditional P-value is $\Pr(Y_1 = 4 | T = 4) = 0.014$ which is computed from the

Table 1
Approximations to P_r .

	(10,15)		(20,35)		(5,95)	
	P_{mid}	p^*	P_{mid}	p^*	P_{mid}	p^*
P_1	5.6	26.7	7.7	21.5	18.7	29.1
P_2	15.0	32.9	7.1	19.8	8.0	23.9
P_3	18.9	35.3	9.6	18.8	4.6	23.3
P_4	20.7	33.8	11.3	16.5	7.8	23.4
P_5	21.2	33.7	11.9	15.8	11.9	23.9
P_6	21.4	33.3	12.5	15.2	15.1	24.5
P_{mid}		27.1		17.4		23.8

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) .

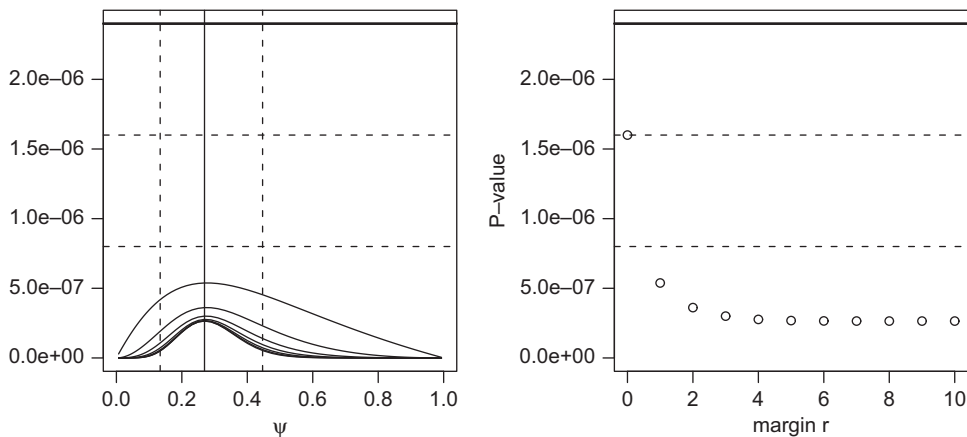


Fig. 3. Extreme data set $(y_0, y_1) = (0, 7)$. Left panel plots $P_r(0, 7; \psi_0, \lambda)$ against λ for $r = 1, \dots, 7$. Right panel plots $P_r(0, 7; \psi_0, \lambda)$ versus r . Horizontal lines are p^* (dark) and P_{mid} and P_{tail} (dashed).

hypergeometric distribution. Assuming that this outcome $(y_0, y_1) = (0, 4)$ is ranked as most hostile to the null, the unconditional P-value is just the unconditional probability of this outcome which can be written as

$$\Pr(Y_1 = 4|T = 4) \times \Pr(T = 4).$$

Thus the unconditional P-value of 0.014 is deflated by a factor $\Pr(T = 4)$. Typically the maximum value of this factor (which here is 0.273) is used. Quite explicitly then, the event that $T = 4$ is counted directly against the null hypothesis. The pertinent question then is why should the event that there were 4 positive responses in total be counted *against* the null hypothesis that the two success rates are equal? While many authors have argued that T contains subtle information about θ in conjunction with Y_1 , no-one has argued that it contains direct information by itself. I call this phenomenon *spurious deflation*. A P-value is spuriously deflated if it can be decomposed into a product of factors, at least one of which is logically irrelevant to the hypotheses under test. It is pertinent to note here that the mid-P-value equals $\Pr(Y_1 = 4|T = 4) \times 0.5$ but the deflation factor 0.5 is not a function of t and is applied on the basis that it adjusts for discreteness.

Do approximately conditional P-values avoid spurious deflation, at least approximately? To see that they do not, we continue with the earlier example with $(n_0, n_1) = (19, 7)$, but look at the data sets $(y_0, y_1) = (1, 7)$ and $(y_0, y_1) = (0, 7)$. Fig. 3 describes the latter data set, which is the most hostile data set to the null (as was the Fisher tea tasting example). The dashed horizontal lines are P_{tail} and P_{mid} and the dark line is p^* . The approximately conditional P-values are grossly smaller than all three of these. The fact that this is directly due to spurious deflation is worth elucidating since it uncovers the fact that approximately conditional P-values directly violate an important aspect of conditionality.

In general, let y_{max} be the data set most hostile to the null, giving the minimum value $\pi_{min} = \pi(y_{max})$ of $\pi(y)$. Let T be any candidate conditioning variable and define $t_{max} = T(y_{max})$. Refer to expression (6). Conditional on $T = \tau \neq t_{max}$, $\pi(Y)$ is always greater than π_{min} and so $\Pr(\pi(Y) \leq \pi_{min} | T = \tau)$ is zero. Thus, the only non-zero term in expression (6) is the $\tau = t_{max}$ term. For this term, $\Pr(\pi(Y) \leq \pi_{min} | T = t) = \Pr(Y = y_{max} | T = t_{max})$ which is just $P_{tail}(y_{max})$. Hence for the most hostile data set (6) becomes

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

The approximately conditional P-value equals the conditional P-value P_{tail} deflated by a factor measuring the conditional probability of observing $T = t$. Apart from the case $r = 0$ (when this factor equals 1), the approximately conditional P-value counts the event $T = t$ directly against the null hypothesis. The pertinent question then is why should the event that there were t_{max} positive responses in total be counted against the null hypothesis that $\theta = 0$? Except for the case of fully conditional inference, when $\Pr(T = t|T \in \mathcal{N}_0(t)) = 1$, the approximately conditional P-value directly counts the observed value of t against the null.

For our case, $T = Y_0 + Y_1$ is binomial and it is easy to verify numerically that the deflation factor $\Pr(T = t|T \in \mathcal{N}_r(t); \psi)$ takes a maximum value when ψ is very near to $\hat{\psi} = t/(n_0 + n_1)$ regardless of r and that this rather quickly converges to the unconditional probability that $T = t$ as r increases. While spurious deflation is most easily demonstrated for the most extreme data set, it is not restricted to this case. Much the same phenomenon is occurring for the data set $(y_0, y_1) = (1, 7)$ in Fig. 4, which is the second most hostile data set. Some explanation for how this happens 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 $t = 8$. For the approximately conditional P-value with $r = 1$, the conditioning set is $T \in \{7, 8, 9\}$ and the observed value is 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 is noted again that Pierce and Peters have in mind problems where the nuisance parameter has dimension greater than 1, so that even for small values of r the conditioning set has larger support. However, it is far from clear that spurious deflation is 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 is potentially more of a problem.

The conclusion seems unavoidable that approximately conditional P-values, like unconditional P-values, are subject to spurious deflation. This should not be counted against the use of p^* , which in these examples is pleasingly larger than the approximately conditional P-values. It does, however, undermine the interpretation that p^* is an approximation to an approximately conditional P-value.

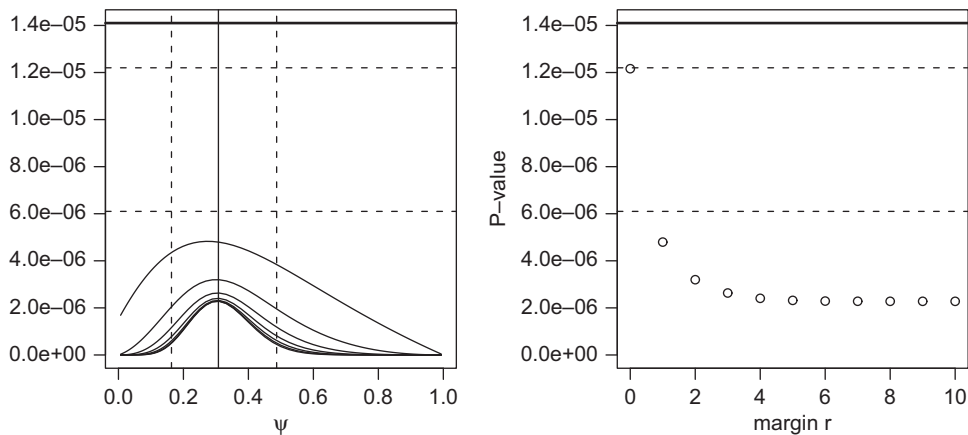


Fig. 4. Almost extreme data set $(y_0, y_1) = (1, 7)$. Left panel plots $P_r(1, 7; \psi_0, \lambda)$ against λ for $r = 1, \dots, 7$. Right panel plots $P_r(1, 7; \psi_0, \hat{\lambda})$ versus r . Horizontal lines are p^* (dark) and P_{mid} and P_{tail} (dashed).

Table 2
Tail sets and reference sets for approximately conditional P-values.

r	$(y_0, y_1) = (1, 6)$	$(y_0, y_1) = (1, 7)$
0	2/8	1/8
1	4/23	2/24
2	5/37	2/39
3	6/50	2/53
4	6/62	2/66
5	6/73	2/78
6	6/83	2/89
7	6/92	2/99

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 $(n_0, n_1) = (19, 7)$.

7. Incoherent conditionality

The final contribution of this paper is to identify a logical inconsistency in both p^* and P_r . The idea of a fully conditional P-value is that we restrict attention to the set $\{T = t\}$ in assessing the observed value of Y . In hypothetical future repetitions of the experiment, only those values of Y consistent with the conditioning set are countenanced. In our example where $\{T = 7\}$ the possible outcomes are $(0, 7), (1, 6), \dots, (6, 1), (7, 0)$ and the P-value for $(1, 6)$ is the sum of $\Pr(1, 6|T = 7)$ and $\Pr(0, 7|T = 7)$. The unobserved value $(0, 7)$ is assigned probability $\Pr(0, 7|T = 7)$ and in principle this can be done in advance of observing the data. A critical observation is that if we had observed $(0, 7)$ rather than $(1, 6)$ then it would still be the case that $T = 7$ and we would assign this same conditional probability to $(0, 7)$.

With approximately conditional P-values, the conditioning sets do not partition the sample space which leads to logically inconsistent treatment of unobserved sample points. To illustrate, consider the same outcome $(1, 6)$ but this time with the conditioning set $\{6 \leq T \leq 8\}$. There are 23 points in this set. The P-value for $(1, 6)$ adds up the conditional probability of four of these 23 data points, namely the observed $(1, 6)$ as well as the unobserved $(0, 6), (0, 7)$ and $(1, 7)$. In particular, the outcome $(0, 6)$ is assigned probability $\Pr(0, 6|6 \leq T \leq 8)$. However, if we had observed the outcome $(0, 6)$ then $T = 6$ and we would have instead conditioned on the set $\{5 \leq T \leq 7\}$. Therefore, the point $(0, 6)$ would have been assigned probability $\Pr(0, 6|5 \leq T \leq 7)$! There is an essential incoherence in how we model the unobserved sample points. Because if we observed them, we would assign a different probability to what we assert when they are unobserved. This is hard to accept, since the main purpose of a probabilistic theory of inference is to describe both what happens, and what could have happened, in a logically coherent manner.

This same problem applies to p^* for non-canonical models. In this case, p^* approximates a tail probability from a distribution conditional on an approximate sufficient statistic, since no exact sufficient statistic exists. This approximate sufficient statistic defines a subset of the sample space. However, the approximate sufficient statistic changes with different data sets. Consequently, the collection of the conditioning subsets does not partition the sample space and so the same logical problem arises.

8. Discussion

With recent advances in computational support for second order inference (Brazzale et al., 2007), these methods are likely to see increasing use. While they may perform adequately in many practical situations, it remains unclear what they approximate for discrete models. This paper has critically examined some explanations for p^* , especially that of Pierce and Peters (1999), with a view to bringing some of the difficulties to a wider audience. None of this should be taken as necessarily arguing against using p^* or its variants.

While there are certainly situations where approximately conditional P-values P_r are close to p^* , there are examples where they are not. Moreover, approximately conditional P-values themselves are problematic. Their dependence on ψ is not always negligible and they can also vary erratically with r . In addition, two logical weaknesses have been identified. The first is spurious deflation which is an undesirable property of unconditional P-values. Approximately conditional P-values can have this defect, even when r is small. The second relates to how the potential conditioning sets partition the sample space. This is a problem for any procedure where the conditioning variable is determined by the data.

There is a completely separate strand of recent research into exact unconditional inference, where the nuisance parameter is first estimated and then residual dependence eliminated by (partial) maximisation (Lloyd, 2008). One of the themes of this research is the importance of good approximate pivotals to generate the exact inference. In principle, P-values based on p^* should be closer to pivotal. At the same time, they involve some kind of conditioning which is known to lead to more powerful unconditional inferences, see Boschloo (1970) in the testing context and Lloyd and Moldovan (2007) in the confidence limits context. A broader purpose of this paper has been to understand how p^* methods perform in simple low dimensional models and to anticipate their performance as generators of exact procedures. Some preliminary results may be found in Lloyd (2010).

References

- Basu, D., 1977. On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.* 72, 355–366.
- Barndorff-Neilsen, O., Cox, D.R., 1994. *Inference and Asymptotics*. Chapman and Hall, London.
- Berger, R.L., Boos, D.D., 1994. P values maximised over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* 89, 1012–1016.
- Berkson, J., 1978. In dispraise of the exact text. *J. Statist. Plann. Inference* 2, 27–42.
- Bickel, P.J., Doksum, K.A., 1977. *Mathematical Statistics*. Holden-Day, Oakland.
- Boschloo, R.D., 1970. Raised conditional level of significance for the 2×2 table when testing the equalities of probabilities. *Statist. Neerlandica* 24, 1–35.
- Brazzale, A., Davison, A.C., Reid, N., 2007. *Applied Asymptotics: Case Studies in Small-sample Statistics*. Cambridge University Press, New York.
- Cox, D.R., 1958. Some problems connected with statistical inference. *Ann. Math. Statist.* 29, 357–372.
- Cox, D.R., 1984. Discussion of paper by F. Yates. *J. Roy. Statist. Soc. A* 147, 451.
- Davison, A.C., Fraser, D.A.S., Reid, N., 2006. Improved likelihood inference for discrete data. *J. Roy. Statist. Soc. B* 68, 495–508.
- Lehmann, E.L., 1986. *Testing Statistical Hypotheses*, second ed. Wiley, New York.
- Lloyd, C.J., 2008. Exact P-values for discrete models obtained by estimation and maximisation. *Austral. J. Statist.* 50, 329–346.
- Lloyd, C.J., 2010. Exact tests of non-inferiority based on pre-estimation and second order pivotals. *J. Statist. Comput. Simulation* 15, to appear.
- Lloyd, C.J., Moldovan, M., 2007. Unconditional efficient upper limits for the odds ratio based on conditional likelihood. *Statist. Med.* 26, 5136–5146.

- Pierce, D.A., Peters, D., 1992. Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *J. Roy. Statist. Soc. B* 54, 701–737.
- Pierce, D.A., Peters, D., 1999. Improving on exact tests by approximate conditioning. *Biometrika* 86, 267–277.
- Reid, N., 2003. Asymptotics and the theory of inference. *Ann. Statist.* 31, 1695–1731.
- Seneta, E., Phipps, M.C., 2001. On the comparison of two observed frequencies. *Biometrical J.* 43, 23–43.
- Yang, B., Kolassa, J.E., 2005. A refinement to approximate conditional inference. *Statist. Probab. Lett.* 72, 103–112.