*Biometrika* (1999), **86,** 2, *pp.* 265–277

*Printed in Great Britain*

# Improving on exact tests by approximate conditioning

BY DONALD A. PIERCE

*Radiation Effects Research Foundation, 5-2 Hijiyama Park, Hiroshima 732, Japan*

pierce@rerf.or.jp

AND DAWN PETERS

*Department of Statistics, Oregon State University, Corvallis, Oregon 97331, U.S.A.*

peters@stat.orst.edu

SUMMARY

The often conservative nature, for discrete data, of so-called exact tests seems usually the result of unnecessarily precise conditioning. We consider avoiding this by conditioning only approximately on the sufficient statistics for nuisance parameters. Modest relaxation of conditioning results in small loss in terms of the rationale for conditional inference, but can greatly reduce the difficulties caused by discreteness. Exact calculation of $p$-values based on approximate conditioning is possible, but unattractive both in terms of the amount of calculation involved and in requiring explicit specification of the extent to which conditioning is to be relaxed. It is shown that there is a highly accurate, easily computed and very natural asymptotic approximation that avoids these difficulties.

*Some key words*: Asymptotic methods; Conditional inference; Continuity correction; Discrete data; Exponential family; Fisher exact test; Logistic regression.

## 1. INTRODUCTION

For inference regarding a single canonical parameter in multiparameter discrete exponential families, we consider conditioning only approximately on the sufficient statistics for the nuisance parameters. The aim is to avoid the often poor operating characteristics of so-called exact tests, caused in most cases by unnecessary over-conditioning, while effectively maintaining the principles of conditional inference. Exact calculation of $p$-values based on approximate conditioning is possible, but unattractive in several respects. However, asymptotic considerations apply very naturally, and provide approximations to this with usually negligible error. Ultimately the prescription is very simple: apply the best of higher-order asymptotic methods as for continuous data, incorporating the approximate conditioning by making no continuity correction.

We will argue that the approximately-conditional $p$-value given by this prescription is a very good approximation to an exact $p$-value based on approximate conditioning, where the relaxation of conditioning is enough to eliminate effectively the discreteness problems, but not enough to vitiate the aims of conditioning.

The 'best of asymptotic methods' referred to in the prescription and described in § 3 would provide, with continuity correction, approximations to the standard exact tests with usually negligible error. The issues involve not so much the accuracy of asymptotic

approximations, but rather what should be approximated. We believe that our approximately-conditional $p$-value approximates very well an inferential quantity which is in virtually every respect preferable to the exactly-conditional $p$-value. The primary issues to be discussed involve the following:

(i) whether or not the gain from approximate conditioning, if the so-defined $p$-values were computed exactly, overcomes what is lost in the relaxation of exact conditioning;
(ii) the accuracy of higher-order asymptotic methods for computing $p$-values, either exactly or approximately conditional;
(iii) whether or not the approximately conditional $p$-value of the above prescription can achieve the goal of corresponding to suitable but not precisely-specified relaxation of the usual conditioning.

Regarding (i), there has been much discussion in the literature of difficulties introduced by exact conditioning; see for example Berkson (1978), Yates (1984), Upton (1982), Little (1989), Haviland (1990) and Agresti (1992). It is apparent that there is a remarkably persistent impasse. Some argue that conditioning is essential, not so much to obtain $p$-values exactly free of the nuisance parameter as for Fisherian reasons pertaining to relevant reference sets. Others argue that the resulting poor operating characteristics are so unacceptable as to override the more subtle Fisherian considerations, and that the resolution lies in approximate unconditional inference. All but the last of the references above focus on the Fisher exact test for comparing binomial samples, and since a surfeit of discussion on this example has left such opposing views we think it will help to consider the issue more generally and with different examples.

Few would claim there is no difficulty at all in an extremely limited conditional sample space. Extreme discreteness leads to poor unconditional operating characteristics for fixed-level tests, with actual test sizes typically about half the nominal level. This is caused by the variation, with the conditioning event, of the set of achievable $p$-values. Even conditionally, however, confidence intervals obtained by inversion of exact tests are conservative to about the same extent, because of the variation of the set of achievable $p$-values with hypothesised values for the interest parameter, rather than with the conditioning event.

It is a common view that if the aim is to compute and properly interpret $p$-values, rather than using fixed-level tests, then the difficulties indicated above are less serious. We think this view is wrong. Although under the hypothesis $\mathrm{pr}(p\text{-value} \leqslant p)$ is exactly $p$ for the achievable values of the $p$-value, one surely must be concerned about the sensitivity, or power, of the procedure. If, not atypically, the achievable $p$-values were $0{\cdot}001, 0{\cdot}30, 0{\cdot}45, \ldots,$ then for modest departures from the hypothesis one would be unlikely to obtain a $p$-value small enough to attract attention. This might be tolerable if it were an inherent limitation of a valid inferential procedure, but this is not the case.

Cox, in his discussion of Yates (1984), suggested approximate conditioning to resolve these difficulties, but with the caveat that it may not be advisable at a practical level. We think that the development here may serve to ameliorate his concerns. At any rate, it will become clear that approximate conditioning effectively eliminates the poor operating characteristics just referred to, and the issues to be investigated are the somewhat more subtle ones itemised in (i)–(iii) above.

Following some general development, we will begin our examples with a rather typical application of logistic regression, where it is seen that the conditional reference set is degenerate because the covariates carrying nuisance parameters are represented quite

precisely. One form of approximate conditioning, but as will be seen not a good one, is to represent the covariates less precisely. It is seen that the form investigated here is far more effective. Some other examples, including the Fisher exact test, are then considered.

## 2. Approximate conditioning

We consider discrete exponential families with the joint distribution of the dataset $x$ having density that can be expressed as

$$f_X(x; \psi, \nu) = h(x) \exp\{\psi s(x) + \nu t(x) - K(\psi, \nu)\},$$

where $\psi$ is a one-dimensional interest parameter and $\nu$ is a typically multidimensional nuisance parameter. We are interested in testing a hypothesis regarding $\psi$ based on data $x_{\text{obs}}$. In standard conditional inference the distribution of $s(X)$ given $t(X) = t(x_{\text{obs}})$, which depends only on $\psi$, is the basis for inference. We will generally, but not religiously, denote random variables by the capital of the letter denoting their values.

The settings of interest here, which are very common, are those where the unconditional distributions of suitable test statistics $g(S, T)$ are not very discrete, but the conditional distributions are highly discrete or degenerate. By 'not very discrete' we mean that the atoms of the distribution have small probabilities. A 'suitable test statistic' might be the likelihood ratio statistic $r$ defined in § 3, but for purposes of the point under consideration it could be any reasonable test statistic, including even the conditional $p$-value. The statistic $S$, which is often about equally discrete unconditionally and conditionally, is not a suitable test statistic in the sense intended, since the evidence against $\psi$ provided by $S = s$ depends on the value of $T$.

In the continuous setting, the simplest rationale for conditioning is the theory of similar tests, where $p$-values which do not depend on $\nu$ must agree with those developed conditionally; see for example Cox & Hinkley (1974, § 5.2). A perhaps stronger but less clear motivation is the Fisherian notion that the conditional reference set is more relevant to interpretation of the data at hand.

We investigate the possible theoretical advantage and practical feasibility of conditioning not on $t(X) = t(x_{\text{obs}})$ but on $t(X) \in \mathcal{N}$, where $\mathcal{N}$ is a relatively small neighbourhood of $t(x_{\text{obs}})$. In terms of the Fisherian rationale just mentioned, the essence of conditioning will clearly be maintained. With regard to similarity, one can expect that $p$-values and confidence limits based on such approximate conditioning will depend only to a small extent on $\nu$. The choice of $\mathcal{N}$, including arguments that it need not be explicitly specified, is discussed in later sections.

For observed data $(s_{\text{obs}}, t_{\text{obs}})$ the $\mathcal{N}$-conditional $p$-value is defined as

$$P_{\mathcal{N}}(s_{\text{obs}}, t_{\text{obs}}; \nu) = \text{pr}\{(S, T) \succeq (s_{\text{obs}}, t_{\text{obs}}) \mid T \in \mathcal{N}; \psi, \nu\}, \tag{2·1}$$

where $\succeq$ denotes an ordering of data according to evidence against the hypothesis. When we condition precisely on $t(X) = t$ the ordering $\succeq$ is provided by the value of $s$, but when we condition on $t(X) \in \mathcal{N}$ this must be extended to an ordering $\succeq$ of the sample points $(s, t)$ for $t \in \mathcal{N}$. We will assume that $\succeq$ agrees with the orderings given by the value of $s$ for each $t \in \mathcal{N}$, but then these must be linked together in some way for comparing datasets with different $t$-values. There are several natural approaches to this, which are virtually equivalent for practical purposes. Once a neighbourhood $\mathcal{N}$ and an ordering $\succeq$ are specified, one can in principle compute exactly an approximately-conditional $p$-value, albeit one that depends to some extent on $\nu$.

Write $P_t(s, t)$ for the exactly-conditional $p$-value based on data $(s, t)$. The following is more useful when the exactly-conditional reference sets, although perhaps quite limited, are not essentially degenerate. Since we assume that $\succeq$ agrees with the usual ordering conditionally on $T = t$, there is for each $t \in \mathscr{N}$ a critical value $\mathrm{CV}(t; s_{\mathrm{obs}}, t_{\mathrm{obs}})$ such that $(s, t) \succeq (s_{\mathrm{obs}}, t_{\mathrm{obs}})$ if and only if $P_t(s, t) \leqslant \mathrm{CV}(t; s_{\mathrm{obs}}, t_{\mathrm{obs}})$. This function $\mathrm{CV}(.)$ links together the orderings of data for different $t$'s in terms of exactly-conditional $p$-values $P_t(s, t)$. We then have that

$$\begin{aligned} P_{\mathscr{N}}(s_{\mathrm{obs}}, t_{\mathrm{obs}}; v) &= \sum_{t \in \mathscr{N}} \mathrm{pr}\{(S, T) \succeq (s_{\mathrm{obs}}, t_{\mathrm{obs}}) | T = t\} \, \mathrm{pr}(t | T \in \mathscr{N}; \psi, v) \\ &= \sum_{t \in \mathscr{N}} \mathrm{pr}\{P_t(S, t) \leqslant \mathrm{CV}(t; s_{\mathrm{obs}}, t_{\mathrm{obs}}) | T = t\} \, \mathrm{pr}(t | T \in \mathscr{N}; \psi, v) \\ &= \sum_{t \in \mathscr{N}} \max_s \{P_t(s, t) : P_t(s, t) \leqslant \mathrm{CV}(t; s_{\mathrm{obs}}, t_{\mathrm{obs}})\} \, \mathrm{pr}(t | T \in \mathscr{N}; \psi, v). \quad (2{\cdot}2) \end{aligned}$$

If $\mathscr{N}$ is not too large then the dependence on $v$ will be small, and substituting for $v$ the maximum likelihood estimator $\hat{v}_\psi$ from the observed data, under the hypothesised $\psi$, should suffice. Note that the dependence of $P_{\mathscr{N}}(s_{\mathrm{obs}}, t_{\mathrm{obs}}; v)$ on $v$ is only through the distribution $\mathrm{pr}(t | T \in \mathscr{N}; \psi, v)$. The terms in the last line of (2·2) which are being averaged with respect to this distribution are a collection of $p$-values independent of $v$, indexed by $t \in \mathscr{N}$. We will see that the discreteness of the distribution of $P_{\mathscr{N}}(s_{\mathrm{obs}}, t_{\mathrm{obs}}; v)$ is systematically and substantially reduced with increasing values of $\mathrm{card}(\mathscr{N})$, thus mitigating the difficulties arising from exact conditioning. The price of this is some dependence of the quantity on $v$, but we will show in examples that this dependence is usually very minor.

When the conditional reference sets are not too close to being degenerate, it is natural to consider taking $\mathrm{CV}(t; s_{\mathrm{obs}}, t_{\mathrm{obs}}) = P_{t_{\mathrm{obs}}}(s_{\mathrm{obs}}, t_{\mathrm{obs}})$; that is, defining the extended ordering in terms of exactly-conditional $p$-values. For our purposes, in order to avoid the degeneracy considerations, it will be better to rely on some other ordering, such as that given by the conditional likelihood ratio, but it is instructive to see where the ordering in terms of exactly-conditional $p$-values leads in the continuous setting. In this case the above argument shows that $P_{\mathscr{N}}(s_{\mathrm{obs}}, t_{\mathrm{obs}}; v) \equiv P_{t_{\mathrm{obs}}}(s_{\mathrm{obs}}, t_{\mathrm{obs}})$ for any choice of $\mathscr{N}$, including the entire range of $T$. This is because the maximising value of $P_t(s, t)$ in the last line of (2·2) is for every $t$ equal to $P_{t_{\mathrm{obs}}}(s_{\mathrm{obs}}, t_{\mathrm{obs}})$.

Thus, in the continuous case, if conditional reasoning is used to define $\succeq$ then subsequent conditioning becomes unnecessary. An ultimately unconditional solution is in fact standard for commonly-used continuous models, yielding, for example, Student's $t$-test and the variance ratio $F$-test. However, the result is usually arrived at differently; see e.g. Lehmann (1986, Ch. 5).

For the discrete case some writers (Boschloo, 1970; Berger, 1994) have considered unconditional inference based on the ordering $\succeq$ given by the exactly-conditional $p$-values. However, this is quite impractical by direct calculation. Most suggestions for unconditional inference have been in terms of rather crude asymptotic methods, and with little explicit attention to the choice of the ordering. What follows here has some bearing on notions of unconditional inference, as discussed in § 5, but it seems best for several reasons to focus on approximate conditioning.

## 3. ASYMPTOTIC CONSIDERATIONS

In this section we show that recently-developed higher-order asymptotic theory conforms quite naturally to the needs here. There are now excellent asymptotic approxi-

mations to the exactly-conditional $p$-value, which for discrete data involve continuity corrections to a normal approximation. The proposal here involves these same approximations, but without the continuity correction.

We begin with the directed likelihood ratio statistic for the hypothesis that the interest parameter has value $\psi$,

$$r_\psi = \text{sign}(\hat{\psi} - \psi)[2\{l(\hat{\psi}, \hat{v}) - l(\psi, \hat{v}_\psi)\}]^{\frac{1}{2}},$$

where $l(.)$ is the loglikelihood function corresponding to the observed data, $\hat{\psi}$ and $\hat{v}$ are the unrestricted maximum likelihood estimators, and $\hat{v}_\psi$ is the constrained maximum likelihood estimator when $\psi$ is taken as known. This $r_\psi$ is the square root of the usual chi-squared likelihood ratio statistic, having to first-order approximation a standard normal distribution.

There is for the present setting an easily-calculated modification of $r_\psi$, commonly denoted by $r_\psi^*$, which provides approximations to exact conditional $p$-values with ordinarily negligible error; see Pierce & Peters (1992), who were specialising to exponential families general results of Barndorff-Nielsen (1986, 1991). For discrete or continuous exponential families arising commonly in practice, it appears that the only situations where there is appreciable error in the use of $r_\psi^*$ are where $\dim(v)$ is very large in relation to the extent of the data. Primarily, this occurs when the coordinates of $v$ have the nature of incidental parameters; that is, roughly speaking, where the data comprise many strata, each providing very limited information regarding $\psi$, with a coordinate of $v$ associated with each stratum.

This modification of $r_\psi$ takes the form of an approximately standard normal statistic

$$r_\psi^* = r_\psi + \text{NP}_\psi + \text{INF}_\psi,$$

where $\text{NP}_\psi$ is an adjustment corresponding to elimination of the nuisance parameter $v$ by the conditioning in question, and $\text{INF}_\psi$ is an adjustment only appreciable when the adjusted information for $\psi$ is small. Formulae for these adjustments in the exponential family setting of § 2 are given in the Appendix, in terms of quantities readily available from the two maximum likelihood fits required to calculate $r_\psi$. For testing $\psi$ against the composite alternative of larger values the approximation to the exactly-conditional $p$-value, for continuously-distributed data, is then

$$\text{pr}\{S \geqslant s_{\text{obs}} \mid T = t_{\text{obs}}; \psi\} \doteq \bar{\Phi}(r_\psi^*), \tag{3·1}$$

where $\bar{\Phi}(.)$ is the complement of the standard normal distribution function. For a discrete setting where conditionally $s$ takes equal steps, taken henceforth to be 1, a continuity correction is made by evaluating $r_\psi^*$ at $(s_{\text{obs}} - 0{\cdot}5, t_{\text{obs}})$. There is an additional small correction that can be made, given in equation (14) of Pierce & Peters (1992), which is used in the following but has a relatively small effect. It is useful in the following arguments to make use of $r_\psi^*$ both with and without continuity correction, so the former will be distinguished as $r_\psi^{*c}$.

Similarly to $r_\psi(s, t)$, the statistic $r_\psi^*(s, t)$ is monotone in $s$ for given $t$, and so the event $S \geqslant s_{\text{obs}}$ in (3·1) can be replaced by $r_\psi^*(S, T) \geqslant r_\psi^*(s_{\text{obs}}, t_{\text{obs}})$. The ordering on $\{(s, t) : t \in \mathcal{N}\}$ based on $r_\psi^*$ agrees well, according to (3·1), with that based on exactly-conditional $p$-values for fixed $t$, and also serves to connect these orderings for different $t$ as discussed in § 2. With the aim of an ordering $\succeq$ which, in the discrete case, agrees with that provided by the exactly-conditional $p$-values when the conditional sample spaces are not close to

degenerate and is useful even when they are, we henceforth take this ordering as that provided by $r_\psi^*$, noting that that provided by $r_\psi^{*c}$ is essentially the same.

In the continuous case it follows from the above that $r_\psi^*(S, T)$ is stochastically independent of $T$ to the same degree of approximation as applies in (3·1). In the discrete case this breaks down to some extent since the achievable values of $r_\psi^*(s, t)$ vary with $t$. However, the main point of the paper is that, when we condition on $T \in \mathcal{N}$ rather than on $T = t$, the distribution of $r_\psi^*(S, T)$ becomes less discrete but is otherwise little changed.

For theoretical considerations it is useful to transform the data at the outset from $(s, t)$ to $\{r_\psi^{*c}(s, t), t\}$ so that the ordering corresponds to a single coordinate. We want to consider the exact and asymptotic bivariate distributions of $\{r_\psi^{*c}(S, T), T\}$, and in particular of the conditional distributions $r_\psi^{*c}(S, T)|T = t$ and $r_\psi^{*c}(S, T)|T \in \mathcal{N}$. The argument will be clearest if we consider situations where exactly-conditional $p$-values are uncomfortably discrete, but not so degenerate as to be useless. We will use the diagram in Fig. 1 to illustrate typical behaviour, corresponding to the first example in the following section, where the achievable $p$-values for $t_{\text{obs}}$ are 0·0045, 0·1059, 0·4475, . . . . In this example $t$ is three-dimensional and $\mathcal{N}$ is specified by allowing each coordinate of $t$ to vary by $\pm$ one 'step', as explained later.

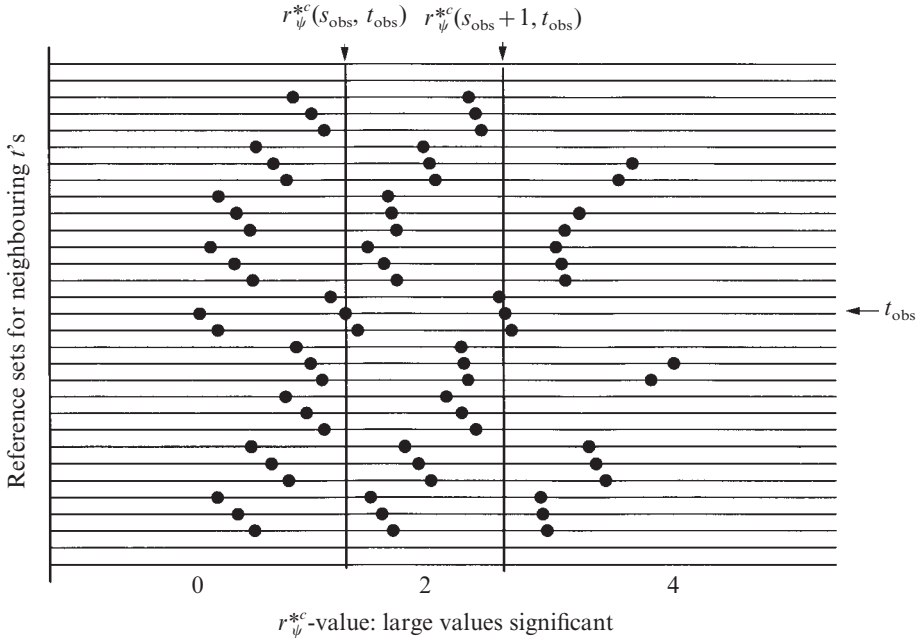


Fig. 1. Typical pattern of achievable values of $r_\psi^{*c}(s, t)$ for each $t \in \mathcal{N}$, showing that the distribution of $r_\psi^{*c}(s, t)|T \in \mathcal{N}$ is far less discrete than the exactly-conditional distribution, with steps of roughly $c/\text{card}(\mathcal{N})$.

Figure 1 illustrates the typical pattern of achievable values of the random variable $\{r_\psi^{*c}(S, T), T\}$ for $T \in \mathcal{N}$, where the marked centre-line corresponds to $t_{\text{obs}}$. The important thing to notice is that in the region of interest the values of $r_\psi^{*c}(s, t)$ lie on what can be referred to as a 'staggered lattice'; that is: (i) their spacings are approximately the same value $c$ for each $t \in \mathcal{N}$, and (ii) for $t \in \mathcal{N}$ the offsets of these lattices from that of $t_{\text{obs}}$ are approximately uniform modulo$(c)$. Consequently, as considered more carefully below, it

is evident that the distribution of $r_\psi^{*c}(S, T)|T \in \mathscr{N}$ will be far less discrete than that of $r_\psi^{*c}(S, T)|T = t$. As a result of (i), for each $t \in \mathscr{N}$ the smallest achievable $r_\psi^{*c}(s, t)$-value greater than $r_\psi^{*c}(s_{\text{obs}}, t_{\text{obs}})$ will usually fall between $r_\psi^{*c}(s_{\text{obs}}, t_{\text{obs}})$ and $r_\psi^{*c}(s_{\text{obs}} + 1, t_{\text{obs}})$. Together (i) and (ii) mean that, if the steps of $r_\psi^{*c}(s, t)$ for fixed $t$ are of approximate size $c$, then marginally for $t \in \mathscr{N}$ the steps are of approximate size $c/k$, where $k = \text{card}(\mathscr{N})$.

We present two forms of argument for the main result that an approximation to $P_{\mathscr{N}}(s_{\text{obs}}, t_{\text{obs}}; v)$, as defined by (2·2) for a suitable choice of $\mathscr{N}$, is provided by our approximately-conditional $p$-value $\bar{\Phi}\{r_\psi^*(s_{\text{obs}}, t_{\text{obs}})\}$, where $r_\psi^*$ is computed with no continuity correction.

First, consider the exact discrete joint distribution of $\{r_\psi^*(S, T), T\}$. Conditionally on $T = t$ this is essentially a discretised normal distribution, in the sense that $\text{pr}(R_\psi^* = r_\psi^* | T = t; \psi)$ is well approximated by the integral from $r_\psi^* - c/2$ to $r_\psi^* + c/2$ of the standard normal density. If we made the idealisation that the distribution $\text{pr}(t | T \in \mathscr{N})$ were uniform and the relations (i) and (ii) above were precisely true, then $r_\psi^*(S, T)|T \in \mathscr{N}$ would have a discretised normal distribution with $\text{pr}(R_\psi^* = r_\psi^* | T \in \mathscr{N}; \psi)$ well approximated by the integral from $r_\psi^* - c/(2k)$ to $r_\psi^* + c/(2k)$ of the standard normal distribution, where $k = \text{card}(\mathscr{N})$. Thus for the ordering $\succeq$ based on $r_\psi^*$ the $\mathscr{N}$-approximate $p$-value is well approximated by $\bar{\Phi}\{r_\psi^*(s_{\text{obs}}, t_{\text{obs}}) - c/(2k)\}$, which leads to our approximately-conditional $p$-value when $k$ is moderately large, say about 10 or more.

It is also useful to relate these considerations directly to the expression for $P_{\mathscr{N}}(s_{\text{obs}}, t_{\text{obs}}; v)$ provided by the last line of (2·2), where now we are taking

$$\text{CV}(t; s_{\text{obs}}, t_{\text{obs}}) = \bar{\Phi}\{r_\psi^{*c}(s_{\text{obs}}, t_{\text{obs}})\} \doteq P_{t_{\text{obs}}}(s_{\text{obs}}, t_{\text{obs}}).$$

Note that the $p$-values

$$p_t = \max_s [P_t(s, t) : P_t(s, t) \leqslant \bar{\Phi}\{r_\psi^{*c}(s_{\text{obs}}, t_{\text{obs}})\}]$$

being averaged there are to the approximation of (3·1) the $\bar{\Phi}$-transforms of those $r_\psi^{*c}$-values in Fig. 1 which are next largest in relation to $r_\psi^{*c}(s_{\text{obs}}, t_{\text{obs}})$. Thus it follows from observation (i) above that the values $\{p_t : t \in \mathscr{T}\}$ usually fall within the interval from $P_{t_{\text{obs}}}(s_{\text{obs}}, t_{\text{obs}})$ to $P_{t_{\text{obs}}}(s_{\text{obs}} + 1, t_{\text{obs}})$. This suggests two reasonable approaches to approximating (2·2) without explicit considerations of $\mathscr{N}$ and the probabilities $\text{pr}(t | T \in \mathscr{N}; \psi, v)$: one using the average of $P_{t_{\text{obs}}}(s_{\text{obs}}, t_{\text{obs}})$ and $P_{t_{\text{obs}}}(s_{\text{obs}} + 1, t_{\text{obs}})$, which is commonly referred to as the mid-$p$; and the other, which is usually a better approximation as well as simpler to compute, being the proposal of this paper.

The mid-$p$ can be an acceptable approximation to $P_{\mathscr{N}}(s_{\text{obs}}, t_{\text{obs}}; v)$, but the exactly-conditional $p$-values required for this are difficult to compute and not available when the conditional sample space is degenerate. The former difficulty can be avoided by making use of (3·1) to approximate $P_{t_{\text{obs}}}(s_{\text{obs}}, t_{\text{obs}})$ and $P_{t_{\text{obs}}}(s_{\text{obs}} + 1, t_{\text{obs}})$. This provides an excellent approximation to the mid-$p$, but usually the mid-$p$ is a poorer approximation to $P_{\mathscr{N}}(s_{\text{obs}}, t_{\text{obs}}; v)$ than the proposal of this paper. This is because the values $p_t$ defined above are less uniformly distributed on $P_{t_{\text{obs}}}(s_{\text{obs}}, t_{\text{obs}})$ to $P_{t_{\text{obs}}}(s_{\text{obs}} + 1, t_{\text{obs}})$ than are the corresponding values of $r_\psi^{*c}$ illustrated in Fig. 1. Since $r_\psi^{*c}(s_{\text{obs}}, t_{\text{obs}})$ is approximately midway between $r_\psi^{*c}(s_{\text{obs}}, t_{\text{obs}})$ and $r_\psi^{*c}(s_{\text{obs}} + 1, t_{\text{obs}})$, our approximately-conditional $p$-value $\bar{\Phi}\{r_\psi^*(s_{\text{obs}}, t_{\text{obs}})\}$ is near the median of the values of $p_t$ and is a better approximation to their weighted mean than is their mid-range, namely the mid-$p$.

In carrying out our proposal there will occasionally be the difficulty that when $s_{\text{obs}}$ is at one of the extremes of the range of the conditional sample space the maximum likelihood

estimator of $\psi$ may be infinite. Then it is not possible to compute $r_\psi^*$, but usually $r_\psi^{*c}$ presents no such difficulty. Our suggestion for such cases, where any reasonable $p$-value will be quite small, is to use the approximation to the mid-$p$ given by $\bar{\Phi}\{r_\psi^{*c}(s_{\text{obs}}, t_{\text{obs}})\}/2$.

## 4. EXAMPLES

We consider three examples, with the following design. In the first, $\dim(t) = 3$ and the exactly-conditional sample space is degenerate. We investigate an alternative type of approximate conditioning which would commonly be used to provide for an exact test, finding that it works far more poorly than the proposal here. In the second, $\dim(t) = 3$ but the exactly-conditional sample space is not degenerate, and here many would consider the exact test an ideal procedure. In the third, $\dim(t) = 1$ and the proposal here is rather stressed, since in order to make $\text{card}(\mathcal{N})$ substantial one must include a fairly wide range of $t$-values. In the first two examples we make some reasonable, if rather casual, choices of $\mathcal{N}$ in order to compute $P_{\mathcal{N}}(s_{\text{obs}}, t_{\text{obs}}; \hat{v}_\psi)$ exactly from (2·2). The aim is only to show that the asymptotic approximation is quite accurate; other reasonable choices of $\mathcal{N}$ lead to similar results in this regard. In the third example it is feasible to provide results for various selections of $\mathcal{N}$.

*Example* 1: *Logistic regression.* The artificial data of Table 1 are to represent an experiment such as a clinical trial; with Bernoulli response, a treatment with two levels whose effect is the parameter of interest, and nuisance parameters carried by two covariables which are initially represented with substantial precision. Conditioning pertains to the three-dimensional statistic $t$ whose coordinates are the overall number of successes and the sums of the two covariables over individuals whose response is success.

Table 1: *Example* 1. *Logistic regression data*

| Treatment 1 | | | Treatment 2 | | |
|---|---|---|---|---|---|
| 0 | 61·6 | 49·6 | 1 | 35·5 | 53·8 |
| 0 | 56·3 | 42·0 | 1 | 43·0 | 61·3 |
| 0 | 50·8 | 42·3 | 1 | 62·5 | 57·3 |
| 1 | 53·5 | 58·6 | 0 | 43·6 | 26·2 |
| 0 | 43·0 | 49·4 | 1 | 55·8 | 47·3 |
| 0 | 67·0 | 55·1 | 1 | 46·4 | 46·8 |
| 1 | 50·6 | 54·0 | 1 | 48·6 | 53·2 |
| 1 | 68·0 | 57·6 | 0 | 36·5 | 51·1 |
| 0 | 52·6 | 54·0 | 1 | 37·3 | 50·0 |
| 1 | 58·7 | 36·6 | 1 | 59·8 | 66·1 |

Not surprisingly, the conditional sample space is degenerate. Until recently this would be given little notice, since an asymptotic method would be used without explicit attention to conditioning, but with recent computational developments and software one might want to pursue a more exact test. A natural approach to circumventing the degeneracy would be to group the data by representing the covariables less precisely. Using the program LogXact (1992) we have prepared Table 2, which shows the conditional distributions of $S$ when the covariates are rounded to 2 digits, then to 1 digit and then grouped into three broad categories. The statistic $s$ is, for Table 2, represented as one-half of the difference between treatments of the number of successes.

The exactly-conditional $p$-values for the observed data, which has $s = 2$, are thus 1·0 for

Table 2: *Example* 1. *Exact conditional distribution of s for various groupings of the data by rounding the two covariables. The results corresponding to the observed value of s, namely* 2, *are in italic type*

| Rounding | $s$ | Frequency | Cum. prob. |
|---|---|---|---|
| 3 digits | *2* | *1* | *1* |
| 2 digits | *2* | *1* | *0·1* |
| | 1 | 3 | 0·4 |
| | 0 | 4 | 0·8 |
| | −1 | 2 | 1·0 |
| 1 digit | 3 | 4 | 0·005 |
| | *2* | *90* | *0·106* |
| | 1 | 304 | 0·448 |
| | 0 | 346 | 0·838 |
| | −1 | 132 | 0·987 |
| | −2 | 12 | 1·0 |
| 3 groups | 3 | 5 | 0·003 |
| | *2* | *128* | *0·087* |
| | 1 | 458 | 0·387 |
| | 0 | 592 | 0·775 |
| | −1 | 301 | 0·972 |
| | −2 | 42 | 1·0 |

Cum. prob., cumulative probability.

the ungrouped data, and 0·1, 0·106 and 0·087 for the three degrees of rounding considered. The set of achievable $p$-values remains quite limited even when each covariable is grouped into three categories. Smoothing of the distribution of $p$-values by this approach is remarkably ineffective. The approximately-conditional $p$-value proposed here, computed in terms of the original ungrouped data, is $\bar{\Phi}\{r_\psi^*(s_{\text{obs}}, t_{\text{obs}})\} = 0{\cdot}030$.

Our approach is most clearly evaluated by considering the data in the form where the covariables have been rounded to 1 digit. Some further comment on this grouping of the data is given later. In choosing a suitable $\mathcal{N}$ some account must be taken of the high correlation between the first coordinate of $t$, that is $\sum y_i$, and the other two coordinates. This will be done by centring the two covariates by subtracting 50, their approximate mean. Then $\mathcal{N}$ may be taken as those $t$'s where all coordinates are within one 'step' of those of $t_{\text{obs}}$, where a step means the smallest change achievable by varying the binary response vector; these step sizes are 1, 10 and 10 respectively.

For this choice the achievable values of $r_\psi^*(s, t)$ for each $t \in \mathcal{N}$ were shown in Fig. 1. The basic points are that it is essentially the value of $r_\psi^*$ rather than $s$ which represents the evidence in the enlarged reference set $t \in \mathcal{N}$, and that the distribution of $r_\psi^*$ is not very discrete in this reference set. As indicated in § 2 one can in principle compute the desired $p$-value $P_{\mathcal{N}}(s_{\text{obs}}, t_{\text{obs}}; v)$, using the ordering provided by $r_\psi^*$, without recourse to asymptotics, most conveniently by using (2·2). When this is evaluated at $\hat{v}_0$ the result is 0·040, and when $v$ is varied by $\pm 1$ standard deviations in each of the three principal coordinate directions of the sampling distribution of $\hat{v}_0$, $P_{\mathcal{N}}(s_{\text{obs}}, t_{\text{obs}}; v)$ varies by about $\pm 0{\cdot}002$ from its value at $\hat{v}_\psi$. The approximately-conditional $p$-value proposed here, when computed

from these grouped data, is $\bar{\Phi}\{r_\psi^*(s_{\text{obs}}, t_{\text{obs}})\} = 0{\cdot}043$. The mid-$p$, that is the average of $P_{t_{\text{obs}}}(s_{\text{obs}}, t_{\text{obs}})$ and $P_{t_{\text{obs}}}(s_{\text{obs}} + 1, t_{\text{obs}})$, is 0·055.

Although what has been given should testify to the basic accuracy of the $r_\psi^*$ asymptotics, we further note that the approximation to $P_{t_{\text{obs}}}(s_{\text{obs}}, t_{\text{obs}}) = 0{\cdot}1059$ based on $r_\psi^*$ with continuity correction is 0·1052. This is typical of results for other possible values of $(s_{\text{obs}}, t_{\text{obs}})$.

Finally, it was noted above that, in terms of the original precisely-represented covariables, $\bar{\Phi}\{r_\psi^*(s_{\text{obs}}, t_{\text{obs}})\} = 0{\cdot}030$, surprisingly different from the value 0·043 obtained from rounding the covariables to 1 digit. We generated a few more similar datasets, and such differences in one direction or the other are not uncommon. Since the exactly-conditional sample spaces are degenerate, it is more attractive to compute $P_{\mathcal{N}}(s_{\text{obs}}, t_{\text{obs}}; \hat{v}_\psi)$ by using direct simulation, based on the definition (2·1), than by using (2·2). With $10^6$ trials satisfying the condition $T \in \mathcal{N}$ the result obtained was 0·034. Thus the asymptotic approximation is very good, but we see again that rounding the covariables is not a good way to deal with the difficulties. The reason that we present the analysis above under such rounding, rather than simply giving this last result, is to make clearer the principles involved in approximate conditioning.

*Example* 2: *Comparing two pairs of binomial samples.* For the artificial data of Table 3, we take the objective as testing equality, over Treatment A, of the odds ratios for Treatment B. The exactly-conditional $p$-value is 0·096 and the mid-$p$ is 0·062. An exact $\mathcal{N}$-conditional $p$-value described below is 0·056, and the proposed approximately-conditional $p$-value is $\bar{\Phi}\{r_\psi^*(s_{\text{obs}}, t_{\text{obs}})\} = 0{\cdot}057$.

Table 3: *Example* 2. *Cross-classified binomial samples*

| Treatment A | 1 | | 2 | |
|---|---|---|---|---|
| Treatment B | 1 | 2 | 1 | 2 |
| Response | 60 | 45 | 40 | 30 |
| Sample size | 100 | 50 | 100 | 50 |

The sufficient statistic $t$ for the nuisance parameter can be represented in terms of covariables $(1, 1, 1, 1)$, $(1, 1, -1, -1)$ and $(1, -1, 1, -1)$, whose coordinates correspond to columns of the display. We will consider conditioning on $T \in \mathcal{N}$, where again each of the three coordinates of $t$ is within one 'step' of those of $t_{\text{obs}}$, and thus $\text{card}(\mathcal{N}) = 27$. The coordinates of $T$ as defined are close enough to being orthogonal for this rectangular $\mathcal{N}$ to be reasonable. The $\mathcal{N}$-conditional $p$-value from (2·2), with $v$ evaluated at $\hat{v}_0$, is 0·056. When $v$ is varied around $\hat{v}_0$ by $\pm 1$ standard error in the principal coordinate directions of the sampling distribution of $\hat{v}_0$, the results from (2·2) vary by about $\pm 0{\cdot}0001$.

It is noted that the $r_\psi^*$ approximation, with continuity correction, to the exactly-conditional $p$-value is correct to at least 3 digits, both for the observed data and for all the $(s, t)$ pairs arising in carrying out (2·2) for this $\mathcal{N}$.

*Example* 3: *Fisher's exact test.* We consider testing independence in a two-way table with entries $\{5, 10, 45, 40\}$. The exactly-conditional $p$-value is 0·131 and the mid-$p$ is 0·098. The approximately-conditional $p$-value proposed here is $\bar{\Phi}\{r_\psi^*(s_{\text{obs}}, t_{\text{obs}})\} = 0{\cdot}083$.

Approximately-conditional exact results involve the sampling rule, in this case Poisson or binomial. That is, the definition of $\mathcal{N}$ and hence the exact value of $P_{\mathcal{N}}(s_{\text{obs}}, t_{\text{obs}}; v)$ depend on the sampling rule, but our proposal $\bar{\Phi}\{r_\psi^*(s_{\text{obs}}, t_{\text{obs}})\}$ does not. The underlying

reason is that for the Poisson setting the profile and conditional likelihoods corresponding to eliminating the parameter for one marginal total are identical, as is well known.

First we consider Poisson sampling, with no marginal total fixed. If $\mathcal{N}$ is taken to allow each of the four marginal totals to vary by $\pm 1$, for which $\text{card}(\mathcal{N}) = 27$, then from (2·2) we find that $P_{\mathcal{N}}(s_{\text{obs}}, t_{\text{obs}}; \hat{v}_0) = 0{\cdot}084$. If $v$ is varied by $\pm 1$ standard error in each of the principal coordinate directions of the sampling distribution of $\hat{v}_0$, the values of $P_{\mathcal{N}}(s_{\text{obs}}, t_{\text{obs}}; v)$ vary by $\pm 0{\cdot}002$ from the value at $\hat{v}_0$.

For binomial sampling we will think of this as two samples of size $n = 50$. Here $t$ is one-dimensional, and the choice of $\mathcal{N}$ is more problematic since if $\text{card}(\mathcal{N})$ is to be substantial one must consider whether or not the points $t \in \mathcal{N}$ are 'suitably close' to $t_{\text{obs}}$ for the relaxation of conditioning to be innocuous. The calculations shown in Table 4 are useful in thinking about this. In successive rows $\mathcal{N}$ consists of data having total number of successes within $15 \pm 3$, $15 \pm 5$ and $15 \pm 7$.

Table 4: *Example* 3. *Exact and asymptotic results for some choices of* $\mathcal{N}$

| $k = \text{card}(\mathcal{N})$ | $P_{\mathcal{N}}(s_{\text{obs}}, t_{\text{obs}}; \hat{v}_0)$ | $r_\psi^*$ approximation with $c/(2k)$ cont. corr. | $v$-variation |
|---|---|---|---|
| 7 | 0·090 | 0·089 | 0·088–0·089 |
| 11 | 0·087 | 0·087 | 0·083–0·087 |
| 15 | 0·086 | 0·086 | 0·078–0·087 |

Cont. corr., continuity correction.

The second column provides exact results of (2·2). These values are somewhat larger than the proposed value of 0·083, but the differences are caused by discreteness of the $\mathcal{N}$-conditional $p$-values rather than the basic asymptotic approximations. This is seen from the third column which makes the continuity correction specific to $\text{card}(\mathcal{N})$; here $c$, the step size of $r_\psi^*(s, t)$ as $s$ varies, is approximately 0·55 for all $t \in \mathcal{N}$. As $\text{card}(\mathcal{N})$ becomes larger results converge towards the proposed value, but the dependence of $P_{\mathcal{N}}(s_{\text{obs}}, t_{\text{obs}}; v)$ on $v$ increases somewhat. The $v$-variation column gives the range of the $\mathcal{N}$-conditional $p$-value as $v$ varies $\pm 1$ standard error from $\hat{v}_0$.

Since $\mathcal{N}$ is simply a range and $v$ is one-dimensional, this example provides an opportunity to explore easily how the size of $\mathcal{N}$ affects the dependence of $P_{\mathcal{N}}(s_{\text{obs}}, t_{\text{obs}}; v)$ on $v$. Several authors, including Boschloo (1970), have discussed taking $\mathcal{N}$ as the entire range of $t$, that is, using an unconditional test.

Of course it would be too much to expect that everyone would agree on what to do for this example. Since $\text{card}(\mathcal{N})$ is rather small for what might be considered $t$'s 'near' $t_{\text{obs}}$, this application is indeed a rather extreme one from the viewpoint of this paper. However, we think that many will agree with us that $\bar{\Phi}\{r_\psi^*(s_{\text{obs}}, t_{\text{obs}})\}$ is a very good answer even for this problem.

## 5. Discussion

Although further work would be useful, we feel that the development and examples here indicate a quite positive assessment of the primary issues (i)–(iii) we raised in § 1.

It is our view that in applications a specific choice of $\mathcal{N}$ should not be considered, and that it would generally be a bad idea to calculate some $P_{\mathcal{N}}(s_{\text{obs}}, t_{\text{obs}}; \hat{v}_0)$ exactly. Moreover, we believe that our proposal, $\bar{\Phi}\{r_\psi^*(s_{\text{obs}}, t_{\text{obs}})\}$ without continuity correction, is usually an

excellent approximation to $P_{\mathscr{N}}(s_{\rm obs}, t_{\rm obs}; v)$ for an ideal but unspecified choice of $\mathscr{N}$, and for the range of $v$-values of interest.

We noted following (2·2) that in continuous settings, if exactly-conditional $p$-values are used to order data according to evidence against the hypothesis, then no further conditioning is required and indeed most significance tests for continuous data are in the final analysis expressed unconditionally. This is an important point in regard to the rationale for conditioning to eliminate nuisance parameters, which we think is rather poorly understood at present. In particular, there may be criteria for ordering data which are not based on exactly conditional $p$-values, but which lead to essentially the same unconditional tests. In this regard we should note that once one has decided to employ approximate conditioning then the approximately conditional $p$-values are not very discrete and the same rationale as for continuous data suggests that the conditioning could then be eliminated altogether. However, on grounds of avoiding the need for resolution of these rather deep issues, we prefer to continue the reliance on approximate conditioning for purposes here.

It is quite unfortunate that the term 'exact test' confers more than is justified, and also pre-empts a choice of adjective for tests which may be considered as superior. Although further research will be useful, we believe that the approximately-conditional $p$-value concept leads in the right direction. Since it is easily computed, we recommend that inferences based on this should normally include presentation of confidence intervals. These will consist of an interval of $\psi$-values such that the statistic $|r^*_\psi(s_{\rm obs}, t_{\rm obs})|$ is less than a Gaussian quantile. Even better is to present a graph, plotting the approximately-conditional $p$-value against the value of $\psi$, which portrays the entire family of confidence intervals. This is in many respects preferable to plotting the profile likelihood in that the NP adjustment in $r^*_\psi$ renders this more like the conditional likelihood, and the INF adjustment improves on the chi-squared approximation for going from likelihood to $p$-values.

## ACKNOWLEDGEMENT

This research was sponsored in part by the U.S. National Science Foundation. We thank Anthony Davison for suggesting to us, long ago, that asymptotic $p$-values might be more relevant than those from exact tests in this setting.

## APPENDIX

*Formulae for higher-order adjustments of* § 3

We provide here formulae for computing the adjustments $\text{NP}_\psi$ and $\text{INF}_\psi$ required for $r^*_\psi$. Derivation and discussion of these is given in Pierce & Peters (1992). It is critical in the following that $\psi$ and $v$ be taken as canonical parameters as specified in § 2. Write $I_{\psi\psi}(\psi, v)$, $I_{vv}(\psi, v)$, $I_{\psi v}(\psi, v)$ for partitions of the Fisher information evaluated at the function arguments, and

$$I_{\psi\psi\,|\,v}(\psi, v) = I_{\psi\psi}(\psi, v) - I_{\psi v}(\psi, v) I_{vv}(\psi, v)^{-1} I_{v\psi}(\psi, v)$$

for the adjusted information for $\psi$. Denote by $\hat\psi$ and $\hat v$ the unrestricted maximum likelihood estimators, and by $\hat v_\psi$ the constrained maximum likelihood estimator for known $\psi$. Write $w_\psi$ for the version of the Wald statistic given by $w_\psi = (\hat\psi - \psi) I_{\psi\psi\,|\,v}(\hat\psi, \hat v)^{\frac{1}{2}}$. If we let

$$\rho = \frac{|I_{vv}(\hat\psi, \hat v)|^{\frac{1}{2}}}{|I_{vv}(\psi, \hat v_\psi)|^{\frac{1}{2}}}$$

then

$$\text{NP}_\psi = \log(\rho)/r_\psi, \quad \text{INF}_\psi = \log(w_\psi/r_\psi)/r_\psi.$$

If $r_\psi = 0$, the adjustments may be computed by perturbing $\psi$ slightly from $\hat{\psi}$; that is, as functions of $\psi$ they have removable discontinuities at $\psi = \hat{\psi}$. We noted at the end of § 3 that when no continuity correction is made adjustments cannot be computed if the maximum likelihood estimate of $\psi$ is infinite, and suggested there a means of dealing with this.

## References


Agresti, A. (1992). A survey of exact inference for contingency tables (with Discussion). *Statist. Sci.* **7**, 131–77.

Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed likelihood ratio statistic. *Biometrika* **73**, 307–22.

Barndorff-Nielsen, O. E. (1991). Modified signed log likelihood ratio. *Biometrika* **78**, 557–63.

Berger, R. L. (1994). Letter to Editor. *Am. Statistician* **48**, 175.

Berkson, J. (1978). In dispraise of the exact test. *J. Statist. Plan. Infer.* **2**, 27–42.

Boschloo, R. D. (1970). Raised conditional level of significance for the $2 \times 2$ table when testing for the equality of two probabilities. *Statist. Neer.* **24**, 1–35.

Cox, D. R. (1984). Discussion of paper by F. Yates. *J. R. Statist. Soc.* A **147**, 451.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Haviland, M. G. (1990). Yates' correction for continuity and the analysis of $2 \times 2$ contingency tables. *Statist. Med.* **9**, 363–7.

Lehmann, E. (1986). *Testing Statistical Hypotheses*, 2nd ed. New York: Wiley.

Little, R. J. A. (1989). Testing the equality of two independent binomial proportions. *Am. Statist.* **43**, 283–8.

LogXact (1992). *A Software Package for Exact and Asymptotic Logistic Regression, Ver. 1.0*. Cambridge MA: Cytel Software.

Pierce, D. A. & Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with Discussion). *J. R. Statist. Soc.* B **54**, 701–37.

Upton, G. J. G. (1982). A comparison of alternative tests for the $2 \times 2$ comparative trial. *J. R. Statist. Soc.* A **145**, 86–105.

Yates, F. (1984). Tests of significance for $2 \times 2$ contingency tables (with Discussion). *J. R. Statist. Soc.* A **147**, 426–63.


[*Received August* 1997. *Revised April* 1998]