

SOME ISSUES ARISING FROM THE ANALYSIS OF 2×2 CONTINGENCY TABLES

C.J. LLOYD

University of Melbourne

Summary

We review the 40 year old controversy over the correct analysis of 2×2 tables. It is argued that conditioning on all marginal totals is appropriate by examining the likelihood based on these margins and secondly by comparing conditional and unconditional evaluations of certain specific outcomes. Several foundational issues which naturally arise are also discussed.

1. Introduction

Statistical inference is not easily placed in a rigid theoretical framework. There are, however, several basic principles (sufficiency, conditionality and likelihood principles) which apply to inference carried out in any framework. These principles seek to identify those aspects of the data which should be allowed to affect our conclusions; in particular, we should ignore any features of the data irrelevant to the issue at hand and ignore, as far as possible, a minimum of features which are relevant.

The purpose of this paper is three fold. Firstly, we draw attention to a fundamental dilemma in statistical theory, the issue of conditioning. We then study a particular case, namely testing homogeneity in a 2×2 contingency table, where this dilemma has very important practical ramifications. An unconditional approach to this problem was proposed first by Barnard (1945) but was retracted by Barnard (1982). The approach has attracted support from McDonald *et al.* (1977), Berkson (1978a,b) and Kempthorne (1978) but was criticised by Yates (1984). Our second aim is to argue that such unconditional tests, and the arguments commonly used to defend them, miss an essential aspect of the conditioning argument — precision indexing. Attention will be drawn to the complications which discrete distributions bring to the theory. Thirdly, and more generally, it

is suggested that procedures based on conditional likelihoods ought also to have their precisions evaluated conditionally.

2. The Conditionality Dilemma

The inferential situation studied in this paper comprises a set of observed data together with a family of probabilistic models parameterised by a finite dimensional parameter θ , describing the repeated sampling properties of the data. Most statisticians will be familiar with the notion of removing any part of an experiment which is irrelevant to inference on θ (or a component of interest) — the Irrelevance Principle. This immediately implies the Sufficiency Principle and also implies its handmaiden, the Conditionality Principle at least as applied to statistics which are distributed exactly free of θ .

It is one of the theses of this paper that Conditionality is a more profound concept than Sufficiency. This can be best appreciated by recalling the example of Cox (1958) where one of two machines measures a common quantity μ with differing variances, the machine used being independent of μ . The Sufficiency Principle does not, but the Conditionality Principle does, lead us to treat the machine actually used as fixed in the analysis. To do anything else would be widely acknowledged as misleading or even dishonest. The property of conditioning, that it leads us to consider the experiment actually obtained, rather than average over all those experiments which might have been obtained, we will call precision indexing.

An immediate consequence of recognising this property as important is that θ -freeness in distribution is no longer the sole requirement of an effective conditioning variable; a variable may be slightly dependent on θ but be an excellent precision indexer as is easily demonstrated by slightly modifying Cox's example. The fact that the notions of irrelevance and precision indexing can compete represents a fascinating philosophical dilemma in statistical inference. It is a dilemma as yet unresolved and affects the whole of statistics both theoretically and practically. In particular it is critical to the analysis of one of the most simple of statistical situations — clinical trials as represented by the 2×2 table.

3. Conditional and Unconditional Tests

We assume that the reader is familiar with the Neyman-Pearson theory of hypothesis testing. In this section we consider the special case of testing a scalar parameter $H_0: \theta = \theta_0$ on the basis of a pair of discrete statistics (X, T) say.

If R is any unconditional rejection region in the (X, T) plane then we may write it as $R = \bigcup R_t$ where $R_t = \{x : (x, t) \in R\}$. These are critical regions of conditional tests given $T = t$ with conditional powers

$$\beta_t(\theta) = Pr(X \in R_t | T = t; \theta)$$

and conditional size $\alpha_t = \beta_t(\theta_0)$. The power of the unconditional test is

$$\beta(\theta) = E_T(\beta_T(\theta)) = \sum \beta_t(\theta) Pr(T = t)$$

with size $\bar{\alpha} = \sum \alpha_t Pr(T = t)$. Because X is discrete, it is typically impossible for a target size α to be achieved by any α_t or $\bar{\alpha}$. However, attaching a uniform variable Z to the experiment we can produce *randomised* versions of the conditional tests with rejection regions

$$R_\alpha = (R_t, Z < c(\alpha, t))$$

such that $\alpha_t = \alpha = \bar{\alpha}$ exactly. Under certain conditions, if each conditional test is UMP then so is the unconditional test (Lehmann 1959). The 2×2 table is a case in point where this result applies. Clearly, both the sufficiency and conditionality principles prohibit the use of Z in inference on θ and 'post-randomisation' is quite correctly not used in practice. The notion is, however, a useful one; just as real numbers are usefully embedded in the complex plane so are non-randomised tests usefully viewed in this wider setting.

A full solution to the testing problem requires us to define a rejection region for any desired size, $\{R_\alpha : \alpha \in [0, 1]\}$. In evaluating the result of a particular test a common summary statistic is the P -value. We will denote the power of the test to which this corresponds (i.e. the power of the smallest test rejecting H_0) by

$$\bar{\beta}(\theta) = Pr(\text{observed or worse}; \theta)$$

and call it the upper observed power. Similarly $\underline{\beta} = Pr(\text{worse than observed})$ we call the lower observed power. In discrete models these values are distinct and the envelope $(\underline{\beta}, \bar{\beta})$ is a useful summary statistic in the testing context, their difference, $\bar{\beta} - \underline{\beta}$, being essentially the likelihood function. It is useful to think of $1 - \bar{\beta}(\theta)$ as measuring how hostile and $\underline{\beta}(\theta)$ as how favourable the observed is to H_0 . In the sequel, we will use both conditional and unconditional versions of the power envelope.

4. Tests of Association in 2×2 Contingency Tables

Suppose that two cancer treatments $T1$ and $T2$ are to be compared by giving N_1 patients $T1$ and N_2 patients $T2$ and counting the number of survivors after some period of time. We say $T1$ is better than $T2$ if $p_1 = Pr(\text{survival} | T1) > p_2 = Pr(\text{survival} | T2)$. If X_i is the number of survivors under T_i then we have the contingency table

	survived	died	Total
T1	X_1	$N_1 - X_1$	N_1
T2	X_2	$N_2 - X_2$	N_2
Total	T	$N - T$	N

with $X_i \stackrel{d}{=} Bi(p_i, N_i)$ under the natural assumption of independence between patients. The 'canonical' parameters are $\theta = \log(p_1/(1 - p_1)) - \log(p_2/(1 - p_2))$ and $\phi = \log(p_2/(1 - p_2))$ and we wish to test $H_0: \theta = 0$ against $H_1: \theta > 0$. The joint density of the data X_1, X_2 (or equivalently X_1, T) is then

$$p(x_1, t) = \binom{N_1}{x_1} \binom{N_2}{t - x_1} (1 + e^{\theta + \phi})^{-N_1} (1 + e^{\phi})^{-N_2} \exp(\theta x_1 + \phi t).$$

Clearly, T is completely sufficient for ϕ and X_1 for θ , the conditional and marginal distributions being, with $\tau = \min(N_1, t)$, $v = \max(0, t - N_2)$

$$p(x_1 | t; \theta) = \frac{\binom{N_1}{x_1} \binom{N_2}{t - x_1} e^{\theta x_1}}{\sum_{i=v}^{\tau} \binom{N_1}{i} \binom{N_2}{t - i} e^{\theta i}} \quad x = v, \dots, \tau$$

$$p(t; \theta, \phi) = \frac{e^{\phi t} \sum_{i=v}^{\tau} \binom{N_1}{i} \binom{N_2}{t - i} e^{\theta i}}{(1 + e^{\theta + \phi})^{N_1} (1 + e^{\phi})^{N_2}} \quad t = 0, \dots, N.$$

Let $h(t, \theta)$ be the denominator of $p(x_1 | t, \theta)$. Then $E(X_1 | t) = h'/h$ and so the conditional efficient score is simply $X_1 - E(X_1 | t; \theta)$. Thus the conditional MLE of θ is the solution of a polynomial of degree τ and, although complicated, results of Godambe (1980) show that this is the optimal estimating equation for θ . Since $p(x_1 | t; \theta)$ is an exponential family in θ , the conditional UMP test of $\theta = 0$ against $\theta > 0$ at target size α is simply to reject H_0 if $X_1 \geq c(\alpha, t)$, post-randomised at the boundary. We denote this test by $\tilde{\Pi}_t$. The test $\tilde{\Pi}$ defined by 'reject H_0 if $T = t$ and H_0 is rejected by $\tilde{\Pi}_t$ ' is unconditionally UMP at size α . It is interesting to note the parallels with some classical notions of conditional inference. Just as conditioning on an ancillary is supposed not to affect the value of our

estimate, so here the decision to accept or reject H_0 is entirely unaffected. (The conditional and unconditional MLE's of θ however differ because T is not an exact ancillary). Rather, conditioning affects the estimated precision of estimation, in this case the quoted power of the test. The powers of $\tilde{\Pi}_t$ and $\tilde{\Pi}$ here are quite different although their powers at H_0 , or size, have been made equal to α by post-randomisation.

If post-randomisation is now removed then each conditional test (denoted Π_t) will have size $\alpha_t \leq \alpha$ and the unconditional test obtained by joining these (denoted Π) size $\bar{\alpha}$ also $\leq \alpha$. Since $\tilde{\Pi}$ is unconditionally UMP, it is to be expected that Π should not be too far from the optimum among non-randomised tests. Any test likely to compete with Π should have critical sub-regions R_t not differing largely from the randomised ones of $\tilde{\Pi}$. It often turns out that Π can apparently be improved upon by certain tests which allow some sub-regions to have mass $\alpha_t > \alpha$ and some $< \alpha$. There are two advantages claimed for these rival tests;

- (i) They can be constructed to have size closer to α than is $\bar{\alpha}$
- (ii) They have higher power.

Thus in the unconditional framework of evaluation, Π may not be the best test though it is rarely far from best. The essentially arbitrary nature of a target size makes (i) rather unconvincing. More pertinent is (ii). However, we should note that the Π P -values tend to be lower than its competitors'. This is because the minimal critical region it measures, $R = \bigcup R_t$, comprises sub-regions each with conditional mass less than the conditional mass of the observed R_t . The observed powers do not typically differ remarkably if the influence of the lower Π P -values is taken into account. In fact any differences between Π and other non-randomised approximations to $\tilde{\Pi}$ are essentially due to (usually minor) local distortions which discreteness brings to the problem.

The more interesting issue from a foundational point of view is whether we should evaluate these tests, i.e. Π and its competitors, conditionally or unconditionally. To answer this we must study what inference is possible from T alone as well as the extent to which T 'indexes the actual precision attained'.

5. Inference Based on the Last Marginal Total Alone

What inference can be made about θ on the basis of T alone? It should be first pointed out that, inasmuch as the observed value of T restricts the possible values of X_1 , some information about θ can be adduced. For

instance if $T = 2$ then $X_1 = 0, 1$ or 2 from which we could unconditionally estimate θ as $-\infty$, $\log(N_2 - 1) - \log(N_1 - 1)$ or ∞ . This property of T has been noted by Berkson (1978b). However this triple estimate is of clearly little inferential value; being a consequence entirely of the discrete distribution of X_1 it is of less use the larger the sample. What should be noted is that information can be gleaned about θ from T alone only via inferring values of X_1 .

A more common estimation procedure with well documented optimality properties is the method of maximum likelihood. It is relatively straightforward to show that equating the score statistics based on T

$$U_\theta = E(X_1; \theta, \phi) - E(X_1 | T; \theta) = n_1 p_1(\theta, \phi) - h'(t, \theta)/h(t, \theta)$$

$$U_\phi = T - E(T; \theta, \phi) = T - N_1 p_1(\theta, \phi) - N_2 p_2(\theta, \phi)$$

to zero gives the unique solution $\theta = 0$, $\phi = \log(T/(N - T))$. This is only a saddlepoint of the likelihood surface however (see Plackett (1977)) and the maximum is actually attained at $\theta = \pm\infty$. Thus the MLE based on T alone is non-unique and further, *it is the same for every observed value of T* .

No similar test based on T alone can exist. This is because the distribution of T is complete in ϕ when $\theta = 0$ (actually for every θ) and so there exists no quantity $P(T, \theta)$ whose expectation is free of ϕ . In tests based in the full (X, T) -plane, points are ranked as favourable to H_0 according to $\eta(x | t)p(t; \theta, \phi)$ where η is some conditional ranking. However, we have already seen $p(t; \theta, \phi)$ is of no use by itself for ranking points whereas, except for distortions caused by the discrete nature of X_1 which are readily removed by post-randomisation, the optimal choice of η is $p(x_1 | t)$ for ranking points either in the plane or in each stratum where T is constant. The tests in the plane defined by Berkson and others such as Upton (1982), who studies 17 different tests, are essentially *ad hoc* having no more theoretical support than many other proposed tests.

Sprott (1975) claims that 'on occasion the marginal tables can apparently contain information' and gives an example where, from a sequence of n tables with $N = 2$, reasonable inference about θ results. However, there is no question that multiple observations of T provide meaningful inference about θ . The problem is that by the sufficiency principle any set of n tables must be combined into a single table and the resulting value of T (the sum of the n t_i 's) is again problematic.

There are several attempts at formalising the notion of 'no information about θ in T ' the earliest being exact ancillarity or equivalently, zero Fisher

information. Certainly, as the present discussion demonstrates, statistics which are not exact ancillaries may still provide no meaningful inference about a parameter. It may be routinely checked that T satisfies the definitions of ancillarity implied by Godambe (1980), Barndorff-Nielsen (1973) and Cox (1958) but not Fraser (1956), Spratt (1975).

6. The Precision Indexing Properties of T

It seems then that T is, by itself, of very little use for inference on the log-odds ratio θ , regardless of what status it may have theoretically. What then are the merits of conditioning on it? We contend that unconditional tests must be avoided if we are not to count certain observed values of T against H_0 . For this purpose we will compare tests Π and Π_t with $N_1 = N_2 = \frac{1}{2}N$. Suppose that we observe data $(X_1, T) = (N_1, N_1)$, the most unfavourable outcome for H_0 in both a conditional or unconditional sample space. The conditional upper observed power is

$$\bar{\beta}_{N_1}(\theta) = Pr(X_1 = N_1 \mid T = N_1) = e^{\theta N_1} / \sum \binom{N_1}{i} e^{\theta i}.$$

The unconditional upper observed power function is

$$\bar{\beta}(\theta, \phi) = \bar{\beta}_{N_1}(\theta) Pr(T = N_1 : \theta, \phi).$$

The ratio of P -values is $Pr(T = N_1; \phi)$, which is a central binomial probability tending to zero as $N \rightarrow \infty$. Hence the difference in conditional and unconditional conclusions can be extreme. Essentially the reason for this is that Π counts the event $\{T = N_1\}$ against the hypothesis that $\theta = 0$ which is rather similar to counting the fact that a particular machine was used in Cox's machine example against a particular value of μ . True there are other unconditional tests proposed besides Π but evaluating them conditionally we can always show that a certain value/s of T is counted against H_0 . The only way of avoiding this is to condition on T .

Another point against unconditional inference concerns the treatment of the most H_0 -favourable outcomes $X_1 = X_2$, in particular $(X_1, X_2) = (0, 0)$. Taking $N_1 = N_2 = 4$ for example we would consider (2,2) more evidence for H_0 than (0,0). Thus suppose we rank these H_0 favourable points in the order (2,2), (1,1), or (3,3), (0,0) or (4,4). The upper unconditional observed power of outcome (0,0) is then

$$\bar{\beta}(\theta) = Pr((X_1, X_2) = (0, 0) \text{ or } (4, 4); \theta, \phi) + Pr(X_1 \neq X_2; \theta, \phi).$$

In other words, we have a nontrivial test, albeit with a large size, which rejects H_0 on the outcome (0,0). The alternative is to rank points with $X_1 = X_2$ equally but then no distinction can be made between the qualitatively different outcomes (0,0) and (2,2). On the other hand

$$\underline{\beta}_0(\theta) = 0, \quad \bar{\beta}_0(\theta) = 1.$$

The conditional envelope powers of the outcome (0,0) accurately reflect the lack of information about θ which no survivals in the table implies. The conditional envelope power of (2,2) on the other hand is

$$\underline{\beta}_2(\theta) = 1 - Pr(X_1 = 2 \mid T = 4; \theta), \quad \bar{\beta}_2(\theta) = 1$$

and $\underline{\beta}_2(0) = 93/128 \simeq 0.73$. Thus the largest size test accepting H_0 from (2,2) has size 0.73 whereas only the trivial test of size 0 accepts H_0 from the outcome (0,0). In other words, (2,2) is more favourable to H_0 than is (0,0). If it be agreed that (0,0) tells us nothing about H_0 , whereas (2,2) is evidence in favour of it then it is essential to condition on T . The problem with the unconditional treatment of (0,0) is more clearly illustrated in the randomised test $\tilde{\Pi}$, although we do not propose post-randomised tests be used. Nevertheless, it is relevant to compare the conditional and unconditional evaluation of $\tilde{\Pi}$, *if it were used*. For a test at size $\alpha = 0.05$, with data (0,0), we would observe a uniform random variable Z and reject H_0 if $z < 0.05$. The conditional power function of this test is constant at 0.05 whereas the unconditional power function indicates that we have performed a test with good power properties, in fact the most powerful test possible in the (X, T) plane! It should be obvious that the conditional power is the more honest evaluation of the precision of this statistical procedure.

Examination of the conditional likelihood function for θ from X_1 given T reveals a large variation in the sensitivity of the experiment between differing values of T . With $N_1 = N_2 = \frac{1}{2}N$, the Fisher information is

$$t(N-t)/(4(N-1))$$

which equals 0 when $t = 0, N$. Figures 1 and 2 illustrate the function of t as a precision index. With $N = 40$, the outcomes (12,8) and (18,16) both lead to the MLE $\hat{\theta} = \log(9/4)$ and conditional MLE's differing from this by less than 1%. Conditional Fisher information is respectively 2.56, 1.31. This is reflected in the conditional loglikelihoods which are almost parabolic with radii of curvature 0.624 and 0.874 respectively. A more

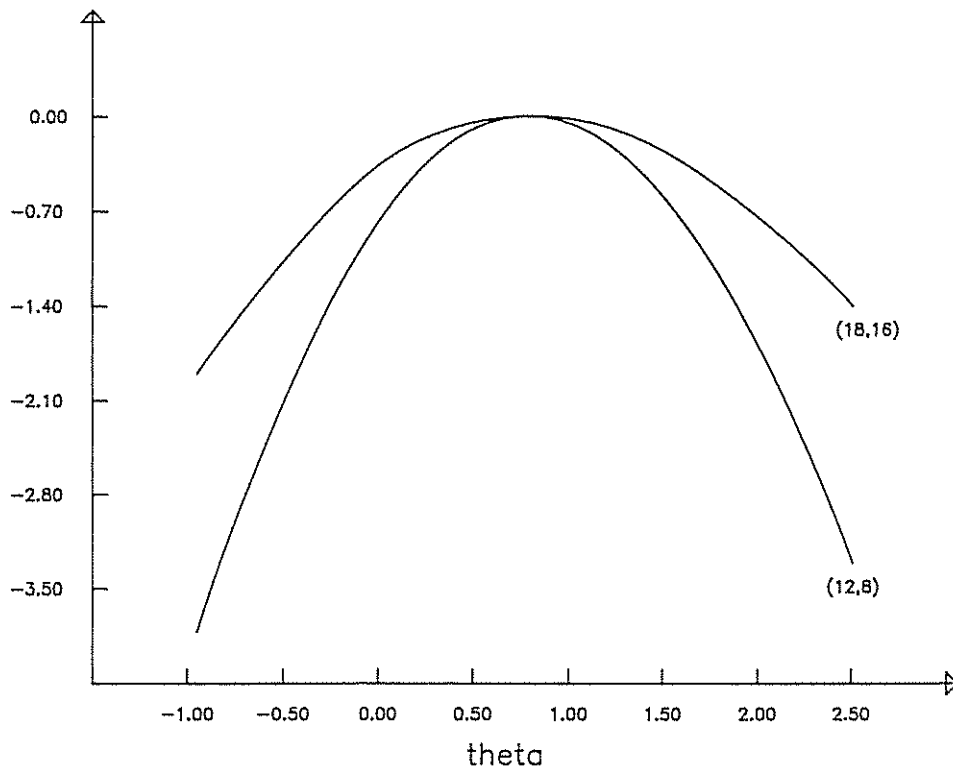


Fig. 1.—Conditional loglikelihoods compared for $N=40$, two cases where CMLE of θ is $0.8 \pm 1\%$.

graphic illustration is given in Figure 2 where $N = 10$ and conditional loglikelihoods of outcomes for which $\hat{\theta} = 0$ are compared. The curvature radii of these approximate parabolas are 1.224 for $t = 4$ or 6, 1.5 for $t = 2$ or 8 and ∞ for $t = 0$ or 10. The force of the precision indexing property is brought home when we consider that the same likelihoods would result from measuring 0 on one of three randomly chosen machines with approximately normal errors and respective variances 1.224, 1.5 and ∞ .

In conclusion, the experimental precision varies rather largely with different values of t , rather like a set of $N + 1$ machines with different variances. On the other hand the likelihood function based on T alone has been shown to contain very little information about θ . These are precisely the characteristics we require of a good conditioning variable — that it contain little information itself but effectively index the differing precisions the experiment can deliver.

Finally, we study the claim that Π_t is conservative compared to other unconditional tests. Firstly, since the former is a conditional test it is inappropriate to compare it with an unconditional test. The rather confusing analyses of Berkson (1978) and Upton (1982) on this point amount to observing that the size of Π_t exceeds the size of its unconditional analogue Π by a greater amount than do other tests. Thus, when we quote a con-

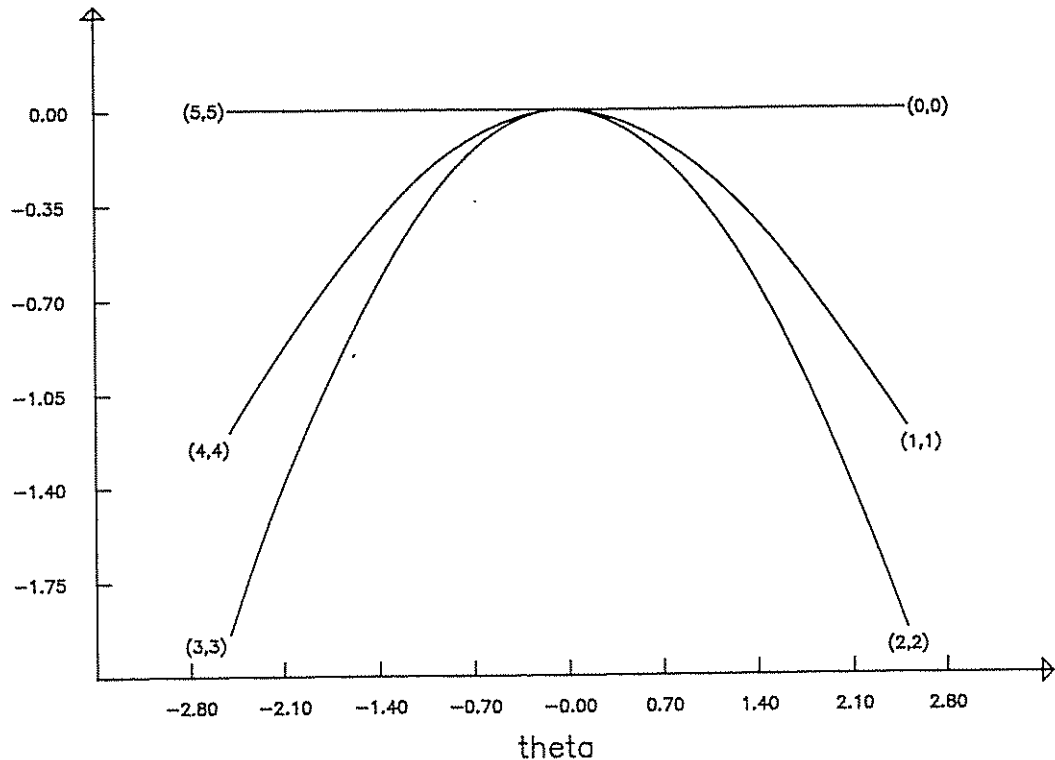


Fig. 2.—Conditional loglikelihoods compared for $N=10$, six cases where CMLE of θ is 0.

ditional P -value p_t we exceed the unconditional value \bar{p} by a rather large amount so that, at any level $\alpha * \epsilon(\bar{p}, p_t)$, H_0 will not be rejected by Π_t when unconditionally it would; this is the source of the label 'conservative' for Π_t . However, rather than discredit the test Π_t this merely draws attention to the fact that the effect of conditioning here is rather important. This should be counted for, not against, the conditional viewpoint and the lower P -value \bar{p} recognised as artificially low, just as quoting an average variance in Cox's machine example when the high variance machine is actually used.

In the conditional framework, the test Π_t lays unchallenged claim to optimality; it is UMP when randomised to the desired level. Further, the test corresponding to the observed power is UMP at the observed P -value. This is not the case with any of the unconditional non-randomised tests which is, of course, the reason why so many different ones have been proposed. Finally, since T is sufficient for ϕ , the conditional powers depend only on the parameter of interest, θ .

8. Discussion

It must be admitted that there is a fundamental difference between the 2×2 table and Cox's machine example, namely that the distribution of the conditioning variable depends on θ in the former case. However, the

discussion of section 2 makes it clear that this would not automatically disqualify T as a candidate conditioning variable. What is important is how much the conditioning variable depends on θ compared to how much the conditional distributions differ with the observed value of the conditioning variable. This seems a topic for further research. One approach may be the following. Cox (1971) proposed evaluating alternative exact ancillaries by the variance (with respect to T) of the conditional Fisher information, $I(X | T)$, since the larger this variance the greater the change in the experimental precision with changing values of the conditioning variable. When exact ancillaries are unavailable, as they are in the 2×2 table, then conditioning involves loss of information; the full information may be written as

$$I(X) = E(I(X | T)) + I(T)$$

under usual regularity conditions. Thus, in order to better estimate the precision as $I(X | t)$, rather than its average value, precision must be sacrificed. We should therefore compare the gain in quoting $I(X | t)$ with the loss $I(T)$. We tentatively propose that the Cox measure of merit of T be compared with the Fisher information in T as a general measure of conditioning merit and T conditioned on if this comparison exceeds some 'rule of thumb' value.

Finally, we raise a more general issue. There are several results relating to optimality of conditional tests and estimators. Lehmann (1959) derives an optimum power property, Godambe (1980) an optimal property of the conditional score function and Andersen (1970) asymptotic properties of conditional MLE's. It seems to the author that not enough consideration has been given in these studies to evaluating the proposed procedure conditionally rather than unconditionally. Certainly, unconditional properties of conditional tests are of interest in any general theory of conditioning. However, serious consideration ought to be given to a final quoted precision based on the conditional distribution. In this regard, it is interesting to wonder in what framework small sample partial likelihood estimators are to be treated.

References

- ANDERSEN, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *J. Roy. Statist. Soc. B* **32**, 283–301.
- BARNARD, G.A. (1945). A new test for 2×2 tables. *Nature*, **156**, 177.
- BARNARD, G.A. (1982). Conditionality versus similarity in the analysis of 2×2 tables. In *Essays in honour of C.R. Rao*, Amsterdam: North Holland.
- BARNDORFF-NIELSEN, O. (1973). On M -ancillarity. *Biometrika* **60**, 447–455.

- BERKSON, J. (1978a). In dispraise of the exact test. *J. Statist. Planning and Inference* **2**, 27-42.
- BERKSON, J. (1978b). Do the marginal totals of the 2×2 table contain relevant information concerning the table proportions? *J. Statist. Planning and Inference* **2**, 43-44.
- COX, D.R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357-372.
- COX, D.R. (1971). The choice between alternative ancillary statistics. *J. Roy. Statist. Soc. B* **33**, 251-255.
- FRASER, D.A.S. (1956). Sufficient statistics and nuisance parameters. *Ann. Math. Statist.* **27**, 838-842.
- GODAMBE, V.P. (1980). On sufficiency and ancillarity in the presence of nuisance parameters. *Biometrika* **17**, 155-162.
- KEMPTHORNE, O. (1978). In dispraise of the exact test: reactions. *J. Statist. Planning and Inference* **2**, 199-213.
- LEHMANN, E.L. (1959). *Testing Statistical Hypotheses*. New York: Wiley.
- MCDONALD, L.L., DAVIES, B.M. & MILLIKEN, G.A. (1977). A non-randomised unconditional test for comparing proportions in a 2×2 table. *Technometrics* **19**, 145-150.
- PLACKETT, R.L. (1977). The marginal totals of 2×2 tables. *Biometrika* **64**, 37-42.
- SPROTT, D.A. (1975). Marginal and conditional sufficiency. *Biometrika* **62**, 599-605.
- UPTON, G.J.G. (1982). A comparison of alternative tests for the 2×2 comparative trial. *J. Roy. Statist. Soc. A* **145**, 86-105.
- YATES, F. (1984). Tests of significance for 2×2 contingency tables. *J. Roy. Statist. Soc. A* **147**, 426-463.

Received January 1987; revised May 1987