J. R. Statist. Soc. A (1984),
147, *Part* 3, *pp.* 426–463

Tests of Significance for 2 × 2 Contingency Tables

By F. YATES

Rothamsted Experimental Station

[*Read before the* Royal Statistical Society *on Wednesday, March 21st, 1984, the President* Professor P. Armitage *in the Chair*]

SUMMARY

Fisher's exact test, and the approximation to it by the continuity-corrected χ^2 test, have repeatedly been attacked over the past 40 years, recently with the support of extensive computer exercises. The present paper argues, on commonsense grounds, supported by simple examples, that these attacks are misconceived, and are mainly due to uncritical acceptance of the Neyman–Pearson approach to tests of significance, the use of nominal levels, and refusal to accept the arguments for conditioning on the margins.

Two-sided tests have also added to the confusion; it is argued that the best definition of a two-sided probability is twice the observed one-tail probability.

Keywords: ANCILLARY STATISTICS; BINOMIAL PROBABILITIES; CONDITIONING; CONSERVATIVE TEST; CONTINGENCY TABLES; CONTINGENCY TEST; CONTINUITY CORRECTION; FISHER'S EXACT TEST; GOODNESS OF FIT; NEYMAN-PEARSON THEORY; ONE-TAIL PROBABILITIES; QUALITY CONTROL; SIGNIFICANCE TESTS; TWO-SIDED TESTS; YATES'S CORRECTION

1. INTRODUCTION

Tests of significance for evidence of association from data in 2 x 2 contingency tables have long been a matter of dispute. Ever since its introduction the legitimacy of Fisher's exact test has been under attack, mainly on the ground that it is too"conservative", i.e. that it gives fewer significant results than are justified by the evidence provided by the data, except for tables both margins of which are determined in advance.

These disputes are attributable to the fact that Neyman and Pearson, in the development of their theory of tests of significance, took it as axiomatic (or as Pearson preferred to call it, as a practical requirement) that the level of significance must be equal to the frequency with which the hypothesis is rejected in repeated sampling of any fixed population allowed by hypothesis. That this was indeed the basis of the earliest ideas on tests of significance is unquestionably true. These had their genesis in the concept of probability associated with games of chance, later extended, by the normal theory of errors, to errors in estimates based on continuous variables. The latter, of course, requires knowledge of the standard deviation or its estimation from the available data. It was Gosset's introduction of the *t*-test, which made due allowance for errors of estimation of σ from sparse data, that led to Fisher's recognition that conditioning on ancillary statistics (i.e. statistics that provide information on the accuracy of the estimated quantity but do not themselves contain any information on this quantity) is a fundamental and valuable extension of the theory of tests of significance.

The *t*-test was acceptable to the Neyman-Pearson school because it did not transgress the frequency requirements of repeated sampling. The marginal totals of a contingency table have a function similar to s^2 in that they provide no information, additional to that provided by the

Present address: Rothamsted Experimental Station, Harpenden, Herts AL5 2JQ.

0035–9238/84/147426 $2.00

body of the table, on lack of proportionality, but do provide information on the accuracy of estimates of association. However, like the Behrens-Fisher test, though for somewhat different reasons, the frequency requirements of repeated sampling are not satisfied.

The frequency property, when it holds, is undoubtedly an easy way of explaining what is meant by tests of significance; indeed Fisher lent support to this mode of explanation in much of his early writing. Any numerate person, for example, is aware of the probabilities associated with simple games of chance, such as tossing a coin or spinning a roulette wheel, and anyone who has any experience of errors of measurement can fully appreciate the implications of the normal theory of errors. Because the frequency requirement is not violated he is not likely to dispute the t-test, though he may feel some unease if the number of degrees of freedom for error is very small. He is much more disposed, however, to doubt the need for conditioning in tests of 2 x 2 contingency tables, particularly as he is often anxious to use the results of such tests to prove that some association is indicated, and is consequently the more ready to believe that the exact test is "conservative".

It is this mistaken belief that has prompted me to write this paper. The points at issue are illustrated by simple examples, mostly based on very small numbers, partly because of ease of presentation, but also because it is here that the contradictions between the two theories are most evident. As the size of a sample is increased the discrepancies between the different approaches are steadily reduced, though it should be remembered that these discrepancies are primarily dependent on the smallest expectation of any cell, not on the total number of observations in the table.

In addition to conditioning, the consequences of using nominal levels of significance, such as 5 and 1 per cent, are also discussed; this practice is a further defect of the Neyman-Pearson theory which undoubtedly adds to the general confusion. Two-sided tests are a further source of disagreement.

Fisher's exact test is closely related to the χ^2 test, which is itself a conditional test; indeed the continuity-corrected χ^2 gives close approximations to the exact test, except for tables with very small expectations. A brief history of the development of these tests, and some comments on some recent papers criticising the exact test, are included in the present paper.

A more mathematical matter that has entered into the disputes is the question of whether the marginal values of a table really contain no additional information on the existence of association, and therefore qualify as ancillary statistics. This is a more technical issue which is relegated to a short appendix.

2. NOTATION

In what follows the numbers in a 2 x 2 table are represented by the symbols of Table 1, where, in general, $m_1 \leqslant m_2$, $n_1 \leqslant n_2$ and $q_1 = 1 - p_1$, etc. With given margins, a, which is then the cell with the smallest expectation, can assume integral values of 0 to m_1 if $m_1 \leqslant n_1$, or 0 to n_1 if $m_1 > n_1$. The expectation e of a, when there is no association, equals $m_1 n_1/N$. In some contexts p_1 and p_2 can be regarded as estimates of binomial probabilities $\mathbf{p}_1$ and $\mathbf{p}_2$. If $\mathbf{p}_1 = \mathbf{p}_2 = \mathbf{p}$ say, a combined estimate of $\mathbf{p}$ is given by p.

TABLE 1
Notation

	B_1	B_2	*Total*	
A_1	a	b	n_1	$p_1 = a/n_1$
A_2	c	d	n_2	$p_2 = c/n_2$
Total	m_1	m_2	N	$p = m_1/N$

To save space numerical values for particular tables are given in the text in the form $(a, b;\ c, d)$.

Occasionally the form $\{m_1, m_2;\ n_1, n_2;\ N\}$ is used for marginal values. In the text, also, the word "table" is used in two senses, (i) a table with particular numerical values for a, b, c, d, (ii) the family of tables, $m_1 + 1$ or $n_1 + 1$ in number, which contain values of a, b, c, d conforming to the given numerical marginal totals. It will be obvious from the context which sense is implied.

If $m_1 \leqslant n_1$ and $p' = n_1/N = e/m_1$ the distribution of a will tend to the binomial distribution $(p' + q')^{m_1}$ as $N \to \infty$ with m_1 and e fixed; and similarly, with m_1 and n_1 interchanged, if $m_1 > n_1$.

3. EARLY HISTORY

In 1900 Karl Pearson introduced the χ^2 test for goodness of fit. The test has proved to be of great utility in many contexts, but unfortunately Pearson did not recognize that in addition to deducting one degree of freedom for the number in the sample an additional degree of freedom must be deducted for each additional parameter estimated from the data. In testing for association in contingency tables the expectations of the cell values are estimated from the marginal totals, and the number of degrees of freedom for an $r \times s$ table is therefore $(r-1)(s-1)$, not $rs - 1$. This error is particularly serious in 2×2 tables, for which χ^2 with one degree of freedom must be used, not three degrees of freedom as Pearson thought.

Udny Yule was also very concerned with contingency tables, and introduced a test for association in 2×2 tables in his textbook, *Introduction to the Theory of Statistics*, first published in 1911, using the large-sample estimate $\sqrt{(pq/n)}$ for the standard error of a proportion p. This gives an estimate of the standard error of the observed difference, $p_1 - p_2$, of the two probabilities, of

$$\sqrt{\left(\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)}.$$

He also noted that if there is no difference p_1 and p_2 can both be replaced by their combined estimate $p = m_1/N$. With this combined estimate Yule's test is equivalent to Pearson's χ^2 test with one degree of freedom. The most convenient formula for this is

$$\chi^2 = \frac{(ad - bc)^2 N}{m_1 m_2 n_1 n_2}. \tag{1}$$

Yule did not mention the χ^2 test in his textbook, but he evidently soon became aware of the discrepancy between his test and the χ^2 test with three degrees of freedom, as he drew attention to it in Greenwood and Yule (1915), and shortly afterwards constructed 350 2×2 tables and 100 4×4 tables by mechanical devices designed to give independent distributions, and compared the χ^2 distributions so obtained with those given by theory, but did not immediately publish his results.

The next event of importance was the publication by R. A. Fisher of his 1922 paper, in which he drew attention to Pearson's error. Although Yule was not fully satisfied with Fisher's proof he then simultaneously published the results of his sampling investigation, which, as was to be expected, confirmed Fisher's results. Pearson, as was his wont, did not immediately admit to any error, and a considerable controversy arose, but the correctness of Fisher's conclusions ultimately came to be generally accepted.

The χ^2 test is of course approximate and will not hold exactly when the expectations of the separate cells of a distribution or contingency table are small. In *Statistical Methods for Research Workers* (1925) Fisher advanced a rule of thumb that the expected number in any one cell should not be less than 5. This rule may in fact be adequate, indeed conservative, for tests involving more than one degree of freedom. It is more suspect for tests involving only a single degree of freedom. Such tests are special in that there are two separate tails, which should be kept distinct. A χ^2 test with 1 df is in fact equivalent, if the appropriate sign is attached to $\sqrt{\chi^2}$, to a test of a normal deviate with unit standard deviation.

If the exact distribution relevant to any particular problem is known the accuracy of the χ^2 test (or any other approximate method) can be investigated by comparing its performance with that given by the exact distribution over a range of typical examples. In 1933 I became interested in such an investigation. The exact form of a binomial distribution with given **p** was of course well known, but not that of a 2 x 2 table with given marginal totals. This was suggested to me by Fisher, and depends on the restriction that only sets of values conforming to both pairs of observed marginal totals are included in evaluating the probabilities, a restriction which is in fact also implicit in the χ^2 test, as the expectations of the cell values are calculated from the marginal totals. Although then unpublished, the exact form must have been known to Fisher for some years, as is indicated by a cryptic passage in an earlier paper (Fisher, 1926): "an exact discussion would show that [for tables with 35 entries] the average value of χ^2 should exceed unity by one part in 34".

The results of this investigation, reported in my 1934 paper, showed that the approximations given by χ^2 to both binomial and 2 x 2 exact probabilities, particularly when the parent distributions are approximately symmetrical, are greatly improved by deducting 1/2 from the observed deviations from expectations when calculating χ^2. This I termed the continuity correction. Formula (1) above then becomes

$$\chi_c^2 = \frac{(|ad-bc|-\frac{1}{2}N)^2 N}{m_1 m_2 n_1 n_2} .$$

If, however, the parent distribution, as for example a binomial distribution with **p** differing greatly from 0.5, is markedly asymmetrical, the one-tail probability given by χ_c, the square root of χ^2 corrected for continuity, will necessarily deviate somewhat from the true probability, because the normal distribution to which it is referred is symmetrical. I therefore produced a small table, covering 2 x 2 tables and binomial and Poisson distributions, of the χ_c values for the 2.5 per cent and 0.5 per cent significance levels, corresponding to the 1.96 and 2.58 values for the normal distribution. This was later included in *Statistical Tables* (Fisher and Yates, 1938). Fisher also added sections on the continuity correction and the exact test to the 5th edition (1934) of *Statistical Methods for Research Workers*, Sections 21.01, 21.02. He also (1935) stated his reason for believing that the exact test should always be used, whether or not the margins are determined in advance.

One might have thought that this would settle the matter, particularly as the χ^2 test had come to be recognized as the appropriate test when the expectations in all four cells of the table are reasonably large. However, in 1945 Barnard put forward a test which he claimed was more powerful than Fisher's exact test (Barnard, 1945). Taking the table to be generated by samples of n_1 and n_2 from two binomial distributions with probabilities $\mathbf{p}_1$ and $\mathbf{p}_2$, he argued that if $\mathbf{p}_1 = \mathbf{p}_2 = \mathbf{p}$ and $n_1 = n_2 = 3$, for example, the probability of getting the table (3,0; 0,3) is $\mathbf{p}^3\mathbf{q}^3$, which has the value 1/64 when **p** = 0.5, and is *less* than this for all other values of **p**, as opposed to a probability of 1/20 if both margins are regarded as fixed.

At first sight, Barnard's argument seems to make good sense. It is certainly true that if we take repeated pairs of samples from two binomials each with **p** = 0.5, 1 in 64 pairs on the average will give the table (3,0; 0,3). This, however, is equivalent to taking a sample of 6 from a single binomial, and dividing it at random into two triplets. From the binomial distribution $(1/2 + 1/2)^6$ the probability of getting values of 3,3 in the m_1, m_2 margin is 20/64. Thus the combined probability of getting the table (3,0; 0,3) is 20/64 x 1/20 = 1/64. The crucial question, therefore, is whether the factor 20/64 should be included in the calculation of the significance probability.

Barnard later (1947) elaborated his proposal into what he termed the *CSM* test. A somewhat similar proposal had also been made by E. B. Wilson (1941). Both proposals were soon abandoned, however, and Fisher was able to wiite in his discussion of the problem in *Statistical Methods and Scientific Inference* (1956):

> "Professor Barnard has since then frankly avowed [(1949)] that further reflection has led him to the same conclusion [that only samples conforming to the observed marginal totals rank for inclusion] as Yates and Fisher, as indeed Wilson with equal generosity had done earlier."

That this conclusion is still not accepted in many quarters, however, is very evident from numerous recent publications. The simple numerical examples in the following sections will, it is hoped, throw further light on the points at issue, and illustrate the ways in which tests of significance, correctly applied, can be of help in the interpretation of 2 x 2 data.

4. COMPARATIVE TRIALS

If, say, we wish to test whether inoculation with a new serum reduces the risk of contracting some infectious disease, a group of N individuals may be chosen for the test and n_1 of them selected *at random* for inoculation, leaving the remainder n_2 uninoculated. This determines the n_1, n_2 margin. Moreover if none of the N individuals were inoculated a given number m_1 (unknown to the experimenter) would be fated to contract the disease. If the inoculation has no effect this will not be changed by the experiment. Subject to this condition, therefore, the m_1, m_2 margin is also determined. The statistical problem, if there is an apparent beneficial effect of inoculation, is the evaluation of the probability that the observed or a greater apparent effect can be attributed to chance causes resulting from the random assignment of the inoculation treatment; and conversely if there is an apparent deleterious effect the evaluation of the probability of getting a negative effect of this or greater magnitude by chance.

Given the marginal values, and random selection for inoculation, the probability of the occurrence of any particular set of cell values $(a, b;\ c, d)$ when inoculation has no effect can be shown by combinatorial analysis to be

$$\frac{m_1!\, m_2!\, n_1!\, n_2!}{a!\, b!\, c!\, d!\, N!}.$$

This gives Fisher's exact distribution. If $m_1 \leqslant n_1$ there will be $m_1 + 1$ terms, with values of a from 0 to m_1; if $m_1 > n_1$ there will be $n_1 + 1$ terms. If a_0 is the observed value, summation of the probabilities from the lower or upper tail gives the probability of getting a value of $a \leqslant$ or $\geqslant a_0$. This provides an exact test of the significance of an apparent association. With given values of n_1 and n_2 there will be a set of such distributions, $N + 1$ in number, the relevant distribution being determined by the observed values of m_1, m_2. The figures in brackets in Table 2 show the 11 distributions obtained when $n_1 = n_2 = 5$.

It should be noted that Barnard did not, in his 1947 paper, discuss comparative trials, but uses the term, misleadingly in my opinion, for samples from two binomials.

5. SAMPLES FROM TWO BINOMIALS

In a comparative trial the individuals included are not necessarily chosen at random from a defined larger population—they may merely be selected as suitable experimental material. In many 2 x 2 tables, however, the data are in fact, or can be regarded as, samples from two defined large populations, in which case the two lines of the table constitute samples of n_1 and n_2 from two binomial distributions. If $\mathbf{p}_1$ and $\mathbf{p}_2$ are the binomial probabilities and there is no association, so that $\mathbf{p}_1 = \mathbf{p}_2 = \mathbf{p}$ say, a combined estimate $p = m_1/N$ from the m_1, m_2 margin provides a sufficient estimate of $\mathbf{p}$. Conditioning on this estimate, i.e. regarding the m_1, m_2 margin as "fixed", then gives Fisher's exact distribution. If, however, we do not impose this conditioning, and instead consider all combinations of the possible samples of n_1 and n_2 that can arise from the two binomial distributions, their associated probabilities, ranked in order of the values of $p_1 - p_2$, or in some other plausible manner, will provide a basis for an alternative test of whether $\mathbf{p}_1$ differs significantly from $\mathbf{p}_2$. This was the basis of Barnard's "more powerful" *CSM* test.

The following specific example may help to clarify thinking on this matter. Table 2 sets out

TABLE 2

Relative frequencies of the 36 2 × 2 tables generated by samples from two binomial distributions, $n_1 = n_2 = 5$, p = 1/2, *classified by values of the* m_1, m_2 *margin*

	m_1, m_2 margin												
$p_1 - p_2$	10, 0	9, 1	8, 2	7, 3	6, 4	5, 5	4, 6	3, 7	2, 8	1, 9	0, 10	*Total*	*Overall probability*
−1.0						1 (0.004)						1	0.001
−0.8					5 (0.024)		5 (0.024)					10	0.010
−0.6				10 (0.083)		25 (0.099)		10 (0.083)				45	0.044
−0.4			10 (0.222)		50 (0.238)		50 (0.238)		10 (0.222)			120	0.117
−0.2		5 (0.5)		50 (0.417)		100 (0.397)		50 (0.417)		5 (0.5)		210	0.205
0.0	1 (1.0)		25 (0.556)		100 (0.476)		100 (0.476)		25 (0.556)		1 (1.0)	252	0.246
+0.2		5 (0.5)		50 (0.417)		100 (0.397)		50 (0.417)		5 (0.5)		210	0.205
+0.4			10 (0.222)		50 (0.238)		50 (0.238)		10 (0.222)			120	0.117
+0.6				10 (0.083)		25 (0.099)		10 (0.083)				45	0.044
+0.8					5 (0.024)		5 (0.024)					10	0.010
+1.0						1 (0.004)						1	0.001
Total	1	10	45	120	210	252	210	120	45	10	1	1024	1.000

The first column contains the single table (5, 0; 5, 0), the second the two tables (4, 1; 5, 0), (5, 0; 4, 1), etc. The figures in parentheses are the elements of the hypergeometric distribution for given values of the m_1, m_2 margin.

the relative frequencies (probabilities × 1024) of the 36 possible outcomes with $n_1 = n_2 = 5$ and $\mathbf{p} = 1/2$. In the table the outcomes are classified by the differences between p_1 and p_2 and by the values of the marginal totals m_1, m_2.

The table enables the probability of obtaining samples with any given characteristics, with or without conditioning on the m_1, m_2 margin, to be calculated. For samples in which $p_1 - p_2 \geqslant 0.6$, for example, the probability without conditioning is $(10 + 25 + 10 + 5 + 5 + 1)/1024 = 0.055$, or directly by the marginal totals $(45 + 10 + 1)/1024$. These marginal totals are in fact 1024 times the relative frequencies, shown in the right-hand margin, of the binomial $(1/2 + 1/2)^{10}$. Thus if we were betting on $p_1 - p_2 \geqslant 0.6$ without knowledge of the m_1, m_2 margin the fair betting odds would be (1024-56) : 56 or 17:1 approximately.

If, however, we were informed, by the person taking the samples, of the marginal totals m_1, m_2 for each particular sample, the conditional probabilities (shown in brackets) would enable us to make much more discriminating bets. We would, if we were wise, only bet when marginal totals of 5, 5, 7, 3 or 3, 7 occurred. For 5, 5 the probability of a successful bet is $(25 + 1)/252 = 0.103$, and for 7, 3 and 3, 7 is $10/120 = 0.083$. For margins 6, 4 and 4, 6 the probability of success is only $5/210 = 0.024$ and for the remainder is zero.

Gambling of this type can easily be performed with two packs of cards. If, after shuffling, 5 cards are dealt from each pack, and if the numbers of red cards are at issue, this is equivalent, apart from the fact that the packs constitute finite populations, to independent samples of 5 from two binomial distributions each with $\mathbf{p} = 1/2$. If the sample cards are laid on the table face downwards all we can say about the probability that there are three, four, or five more red cards in the sample from pack *A* than in that from pack *B* is that its overall value is 0.055, and this should govern the betting. If, however, the 10 sample cards are shuffled and then displayed face upwards, the total numbers of reds and blacks are immediately apparent, although we still do not know to which pack each individual card belongs. If one of the opponents is aware of the value of this information, and takes cognisance of it, while the other pins his faith on the overall probability, the latter is clearly likely to find himself considerably out of pocket.

To obtain more precise probabilities account must be taken of the fact that the samples are from finite populations of 52 cards. In such samples the probabilities of getting 0, 1, . . ., 5 red cards are approximately (1, 6, 13, 13, 6, 1)/40, instead of the binomial probabilities of (1, 5, 10, 10, 5, 1)/32. Substitution of these new values in the diagonal margins of the square of values of Table 2, with corresponding adjustments to the values in the body of the square, gives an overall probability of 0.047 of an excess of 3 or more red cards in pack *A*, i.e. fair betting odds of 20:1 instead of 17:1; if the margins are revealed and bets are placed only for margins 5, 5, 7, 3, 3, 7 the average gain per bet at odds of 20:1 will be 70 per cent of the stake, compared with 68 per cent at odds of 17:1 for true binomial sampling.

The conditional probabilities differ somewhat from those given by the exact distribution for random sampling from two infinite populations with the same **p**. The last two values for the 5, 5 margin, for example, are 0.087 and 0.0024 instead of 0.099 and 0.0040. This may at first sight seem surprising, but a little consideration will show that generation of a table in this manner from two finite populations is not equivalent to the random allocation of treatments adopted for a comparative trial.

The frequencies in Table 2 relate to $\mathbf{p} = 1/2$. Those for $\mathbf{p} = 1/4$, for example, can be obtained by multiplying the values in the successive columns by 1, 3, 9, 27, ... This gives a total over the whole table of 4^{10} and an overall probability that $p_2 - p_1 \geqslant 0.6$ of 0.031 instead of 0.055. The conditional probabilities, based on knowledge of m_1, m_2, are, however, unaltered. The lower overall probability is due to differences in the frequency of occurrence of the different values of m_1, m_2; 3, 7 for example, will occur 81 times as frequently as 7, 3, and the five central m_1, m_2, which are the only ones in which $p_2 - p_1$ can possibly be $\geqslant 0.6$, will only occur in 55 per cent of all samples instead of 89 per cent.

From the above it will be seen that knowledge of the m_1, m_2 margin merely provides a measure of the sensitivity of the observed sample to departures from the null hypothesis. Although sometimes disputed (see the Appendix), it seems to me obvious, as it did to Fisher, that the margins of a 2 x 2 table, however generated, provide virtually no information on the existence of association. In samples from two binomials, for example, absence of association implies that $\mathbf{p}_1 = \mathbf{p}_2$: if $n_1 = n_2$, differences between p_1 and p_2 of a given magnitude but opposite signs occur with equal frequency, as is shown by Table 2; this does not hold if $n_1 \neq n_2$, but the mean value of $p_1 - p_2$ for given m_1, m_2 is still zero. m_1, m_2 are therefore ancillary statistics, in the Fisherian sense, and define a "recognizable subset" (*Statistical Methods and Scientific Inference*, pp. 32, 109). It is the probabilities of occurrence in the relevant subset that provide the correct basis for tests of significance. In other words, we must condition on the margins, whatever the origin of the table. Whether no, one or two margins are "fixed" in advance is irrelevant.

It is still sometimes represented (e.g. Kempthorne, 1979; Upton, 1982) that although conditioning on the margins is justified by the necessity for randomization in comparative trials such as the inoculation experiment described in Section 4, this does not apply to samples from two binomials, for which "more powerful" unconditional tests are available. This line of reasoning is fallacious. If, for example, the subjects for the inoculation trial of Section 4 had been obtained by selecting a random sample of 10 individuals from some larger population and then assigning these individuals at random to the inoculated–non-inoculated groups, we might alternatively have combined the two steps by taking samples of 5 individuals each from the population, which is notionally equivalent to dividing the population into two populations and taking a sample from each. The difference between a trial of this type and one on a haphazard collection of individuals is that (subject to the qualification that any extensive use of inoculation is likely to reduce the subsequent risk rate) any results that emerge relate to all the individuals in the parent population. Tests of significance are unaffected.

If the differences between two separate populations are being investigated the notional division above has a real existence, and the actual differences replace those produced by the imposed treatments. Statistically, therefore, the two situations are equivalent and the same tests of significance must be used.

With discontinuous data, subdivision of the possible outcomes into subsets does, of course, inevitably reduce the significance level of the more extreme outcomes, because only the probabilities of outcomes belonging to the relevant subset will enter into the calculation of the significance. It is this fact, I think, and the urge to find "more powerful" tests, regardless of their relevance, that gives rise to the fatal attraction of unconditional tests for discontinuous data. In continuous data conditional tests have long been accepted without question, at least provided that a significance level P is attained with frequency P in repeated sampling. In testing for significance of a linear regression, for example, the formula $V(b) = \sigma^2/S(x-\bar{x})^2$ is used if the variance σ^2 of y is known, or indeed if it is estimated from the observations, whether the values of x are preassigned or random, provided only that any unknown parameters in the distribution of x are unrelated to the parameters of interest.

6. CHANGES IN MARGINAL TOTALS DUE TO TREATMENT EFFECTS

Part of the reluctance to accept the fact that for the purpose of the test of significance in a comparative trial the m_1, m_2 margin must be taken as known, possibly stems from the knowledge that if the treatment does have an effect m_1 and m_2 will certainly be changed.

Consider a specific example. Suppose we are confident that an inoculation treatment provides sure protection against a certain disease and wish to demonstrate this by doing a trial on 10 subjects, 5 of which are to be inoculated, the other 5 not. The success of such a demonstration depends critically on the number of subjects "at risk", i.e. those who will contract the disease if uninoculated. Table 3 shows the distribution of significant and non-significant results for differing numbers at risk when the inoculation is completely successful. The table is constructed as follows. If all 10 subjects are at risk a table with the values (0, 5; 5, 0) will always be obtained, giving from Table 2 a significance level of 0.004. If only 8 are at risk then before inoculation m_1, m_2 will have the values 8, 2. The chances of obtaining initial cell values for inoculated, unin-

TABLE 3

Percentages of significant results in trials with 10 subjects when inoculation gives certain protection against infection

c, d	*Significance level (P)*	*Number at risk**								
		10	9	8	7	6	5	4	3	< 3
5, 0	0.004	100	50	22	8	2	0.4	–	–	–
4, 1	0.024	–	50	56	42	24	10	2	–	–
3, 2	0.083	–	–	22	42	48	40	24	8	–
2, 3 etc.	⩾0.222	–	–	–	8	26	40	74	92	100

* Those who will become infected if not inoculated.

oculated, of (3, 2; 5, 0), (4, 1; 4, 1), (5, 0; 3, 2) are consequently, from Table 2, 0.222, 0.556, 0.222. On completion of the experiment the first (inoculated) row of each of these tables will become 0, 5, giving significance levels, again from Table 2, of 0.004, 0.024, 0.083 respectively.

This example illustrates another point. If all 10 subjects are at risk, and one of those inoculated contracts the disease, the table (1, 4; 5, 0) will be obtained. This has the same significance level, 0.024, as (0, 5; 4, 1), the second of the two outcomes above when 9 subjects are at risk, but the interpretation is different: here the claim that the inoculation is always successful is definitely disproved, whereas the latter result merely indicates that at least one of the subjects was not at risk.

A statistician reporting on a 2 x 2 table, therefore, should not regard determination of a formal significance level as his sole duty. The two extreme outcomes in an inoculation trial, (0, 5; 0, 5) and (5, 0; 5, 0), for example, both give $P = 1.0$; the former merely indicates that all or most of the test subjects were not at risk, and that further trials should be made on more suitable material; the latter that inoculation is clearly not very effective.

The latter result does not, of course, imply that inoculation provides *no* protection. Upper limits to $\mathbf{p}_1$ at various significance levels (P) can be obtained from the limits of expectation for the $\mathbf{p}$ of a binomial distribution; for $a = 0$, $n_1 = 5$ and $P = 0.1$, 0.025, 0.005 the upper limits for $\mathbf{p}$ are 0.37, 0.52, 0.65 (*Statistical Tables*, Table VIII.1).

7. EFFECT OF INCREASING THE NUMBER OF CONTROLS

Many 2 x 2 comparative trials consist of the comparison of a new treatment against some standard or no treatment. In such cases additional controls can often be included at little extra cost. If so there will be a useful gain in the sensitivity of significance tests and in the accuracy of estimates of the difference between $\mathbf{p}_1$ and $\mathbf{p}_2$.

As an example consider the effect on tests of significance of doubling the number of uninoculated subjects in the test of the last section. We now have $n_1 = 5$, $n_2 = 10$, giving 66 possible outcomes of the test. The percentages of significant results with varying numbers of subjects at risk are shown in Table 4. These are calculated in the same manner as those of Table 3. To

TABLE 4

Effect of increasing the number of controls: percentages of significant results in trials with 5 inoculated and 10 uninoculated subjects when inoculation gives certain protection

c	*Significance level (P)*	*Number at risk*									
		> 12	12	11	10	9	8	7	6	5	< 5
10, 9, 8	< 0.01	100	74	41	17	5	1	–	–	–	–
7	0.019	–	26	44	40	24	9	2	–	–	–
6	0.042	–	–	15	35	42	33	16	4	–	–
5	0.084	–	–	–	8	25	39	39	25	8	–
< 5	⩾ 0.154	–	–	–	–	4	18	43	71	92	100

facilitate comparison between the two tables some significance levels have been grouped together in Table 4.

As is to be expected the addition of extra controls substantially increases the sensitivity of the test. If 80 per cent of the subjects are at risk, for example, then with additional controls 74 per cent of all tests attain a significance level < 0.01, and the remaining 26 per cent a level < 0.025, whereas without additional controls the corresponding figures are 22 per cent and 56 per cent, the remaining 22 per cent being non-significant.

8. FISHER'S TEA-TASTING EXPERIMENT

The example in the last section illustrates the insight that can be gained, when planning experiments on quantal data, by studying the performance of tests of significance under various circumstances, using postulated real effects. A further example which I found intriguing is an experiment described by Fisher in the *Design of Experiments*. Its object was to test a lady's claim that she could tell, when drinking tea, whether the milk had been poured before or after the tea. As designed by Fisher, the experiment provides a classic and somewhat rare example of the generation of a 2 × 2 table in which both margins are determined in advance.

In this experiment the lady was offered eight cups of tea, and was asked to decide which of these had the milk added first, and which last, having been informed there were in fact four of each kind. Suppose the lady has definite discriminating ability, but is sometimes in doubt as to the correct verdict. If there are no doubtful cases the outcome (4, 0; 0, 4) will result, giving a significance probability of 1/70. Uncertainty on one cup only will give the same result, as it will be assigned to the group with only three cups. The same will happen if there are two uncertainties both belonging to the same group. But if these belong one to each group, there will be a 50 per cent chance of wrong assignment. This will give the outcome (3, 1; 1, 3), and a probability of 17/70 of getting this or a better result. The unbracketed values in the "no rejects" columns of Table 5 summarize these results.

If the lady is not informed in advance that there are four cups of each kind the m_1, m_2 margin is not determined, and she will consequently no longer be able to make correct assignments with certainty for the 1, 0 and 2, 0 distributions; there is also an additional possible pair of outcomes in the 1, 1 case. These cases are shown in square brackets. They indicate the advantage to the subject of the information on the constitution of what is submitted for test.

If, however, the lady is permitted to declare her uncertainties, and these are omitted from the assessment of significance, we obtain the results shown in the last two columns of Table 5. These generate a nicely graded set of probabilities which give a fairer assessment of the subject's true power of discrimination, not marred by the potential chance failure in the 1, 1 case, but giving due credit to correct clear-cut judgements.

9. NOMINAL LEVELS OF SIGNIFICANCE

A contributory cause of confusion that affects discontinuous data is the use of conventional nominal levels of significance such as 5 and 1 per cent. This was partly engendered by the use of the nominal significance probability for the argument in tables of t and the normal distribution. This did tend to encourage practical workers to think that if an experiment gives a non-significant result at the chosen level not only is the existence of a real effect not established, but that there is in fact no effect. This mode of thought was further encouraged by the mathematical symbolism adopted by the Neyman–Pearson school: $H_0: \theta = 0, H_1: \theta \neq 0$, or the even more absurd, if θ can be negative, $H_0: \theta = 0, H_1: \theta > 0$.

In quantitative experiments the practice of ornamenting tables by one, two or three stars to denote 5, 1 and 0.1 per cent significance is a convenient way of drawing attention to the more outstanding effects, though this does not obviate the need to report standard errors. With discontinuous data, however, the use of nominal levels can be seriously misleading. The chance of getting 8 or more heads in 10 tosses of a coin, for example, is 0.055, and that for 9 or more heads is 0.011, as can easily be calculated from the binomial $(1/2 + 1/2)^{10}$. Here no conditioning (except

TABLE 5

Fisher's tea-tasting experiment. Configurations and significance probabilities (P) that will be obtained by subjects who either judge correctly or are in doubt about individual cups

Doubtful cases		*No rejects**				*Doubtful rejected*	
No.	*Distribution*	*Possible outcomes*		*P*		*Outcome*	*P*
0	–	(4, 0; 0, 4)		0.014		(4, 0; 0, 4)	0.014
1	1, 0	(4, 0; 0, 4)	[(3, 1; 0, 4)]	0.014	[0.071]	(3, 0; 0, 4)	0.029
2	1, 1	(4, 0; 0, 4), (3, 1; 1, 3)	[(3, 1; 0, 4), (4, 0; 1, 3)]	0.014, 0.243	[0.071, 0.071]	(3, 0; 0, 3)	0.05
2	2, 0	(4, 0; 0, 4)	[(3, 1; 0, 4), (2, 2; 0, 4)]	0.014	[0.071, 0.214]	(2, 0; 0, 4)	0.067

* Entries in square brackets are those for the additional outcomes that can occur when the subject is not informed that there are four cups of each type.

on the number in the sample) is involved, but provided the coin is unbiased only 1.1 per cent of all samples on average will be declared significant at a nominal level of 5 per cent. The actual significance probability attained should therefore always be given when reporting on discontinuous data.

Concentration on single-tail nominal levels of 2.5 and 0.5 per cent is a defect in my 1934 paper, which reflects the current thinking of that time. It may be noted, however, that although Fisher was himself in large part responsible for the widespread use of nominal levels, he always gave actual significance levels when discussing discontinuous examples.

10. QUALITY CONTROL

Tests based on 2 x 2 tables are sometimes required for quality control. An early example is provided by Pearson (1947). This was for testing the performance of batches of small armour-piercing anti-tank shot. To test any particular batch a random sample of 12 shot from the batch and a similar sample of standard shot were fired at a test plate. This procedure was adopted because of unavoidable variations between different test plates and the limited number of shot that could be fired at any one plate. The batch was rejected if its performance was significantly inferior at some chosen nominal level of significance to that of the standard.

Pearson actually recommended use of the χ^2 test without correction for continuity, on the grounds that this provided a reasonable approximation to Barnard's unconditional *CSM* test. (See comments on Table 8, Section 12, for discussion on this point.) What he overlooked was that extreme values of the m_1, m_2 margin in either direction will always give non-significant results, whether or not the batch is defective. Such extreme values will occur more frequently if both $\mathbf{p}_1$ and $\mathbf{p}_2$ are near 1, or both are near 0. Variation between the test plates affects $\mathbf{p}_1$ and $\mathbf{p}_2$ jointly, thereby increasing these frequencies. Such results should be labelled "No verdict".

What the practical man requires to know, therefore, is the percentages of batches with varying degrees of defect which are likely to be passed, rejected, or returned for further testing. The procedure for determining these percentages may be illustrated, without involving excessive arithmetic, for $n_1 = n_2 = 5$. The results are set out in Table 6. $\mathbf{p}_1$ and $\mathbf{p}_2$ are the assumed probabilities of penetration of a shot from the standard and batch samples respectively. Values of 2/3, 1/2 and 1/3 were taken for $\mathbf{p}_1$ and values giving odds ratios of 1 : 1, 4 : 1 and infinity (i.e. complete failure) for $\mathbf{p}_2$.

TABLE 6

Quality control: percentages of batches on which the test gives no verdict, and percentages of batches (other than those on which there is no verdict) which are rejected ($n_1 = n_2 = 5$)

		$\mathbf{p}_2$			(a) *No verdict* (%)			(b) *Rejection* (*% excluding* (a))		
Odds ratio	$\mathbf{p}_1$:	2/3	1/2	1/3	2/3	1/2	1/3	2/3	1/2	1/3
1 : 1		2/3	1/2	1/3	30	11	30	6.3	6.1	6.3
4 : 1		1/3	1/5	1/9	9	25	61	32.8	32.9	34.0
∞		0	0	0	21	50	79	100	100	100

$\mathbf{p}_1$ is the probability of penetration by the standard shot, $\mathbf{p}_2$ that for a shot from the batch. The assumed values of $\mathbf{p}_2$ are those given by the odds ratio, $\mathbf{p}_1 \mathbf{q}_2 / \mathbf{q}_1 \mathbf{p}_2$.

The results for $\mathbf{p}_1 = \mathbf{p}_2 = 1/2$ can be deduced directly from Table 2. As the table shows, the available reject levels are necessarily markedly different for alternating m_1, m_2 values. Those chosen were 0.083, 0.024, 0.103, 0.024, 0.083 for $m_1 = 7$ to 3. The relative frequencies of different m_1, m_2 are given by the bottom line of the table. Excluding outcomes for which $m_1 = 0$, 1, 2 or 8, 9, 10, therefore, the probability of obtaining a significant result when $\mathbf{p}_1 = \mathbf{p}_2 = \mathbf{p} = 1/2$ is

$$(120 \times 0.083 + 210 \times 0.024 + \ldots)/(120 + 210 + \ldots) = (10 + 5 + \ldots)/912 = 56/912 = 0.061.$$

This is merely a weighted mean of the significance levels forced upon us by the discontinuity of the data. The weights depend on the value of $\mathbf{p}$, but the weighted mean is little changed by changes in $\mathbf{p}$. The weights for $\mathbf{p} = 1/3$, for example, are given by the coefficients of the binomial $(1/3 + 2/3)^{10}$, and can be obtained by multiplying the values in the bottom line of Table 2 by 1, 2, 4, 8, . . ., 1024, and similarly, in reverse order, for $\mathbf{p} = 2/3$. This gives a probability of rejection of 0.063 for both $\mathbf{p} = 1/3$ and 2/3.

The proportion of excluded outcomes, for which a re-test will be required, is, however, much more dependent on the value of $\mathbf{p}$. For $\mathbf{p} = 1/2$ the proportion will be $2(1 + 10 + 45)/1024 = 0.109$, whereas for $\mathbf{p} = 1/3$ it will be $(1 + 20 + 180 + 11520 + 5120 + 1024)/3^{10} = 0.303$.

These values are exhibited, in percentage form, in the first line of Table 6. They tell us what may be expected if the batch being tested is equal to the standard. To see how effective the test is in detecting batches that are sub-standard we must ascertain what happens when $\mathbf{p}_2 < \mathbf{p}_1$. This can be done by constructing tables similar to Table 2 with new relative frequencies.

For $\mathbf{p}_1 = 1/2$, $\mathbf{p}_2 = 1/5$, (odds ratio 4:1), for example, the NW border line of the square is unchanged, but the NE border line must be replaced by the coefficients of the binomial $(1/5 + 4/5)^5$, i.e. by 1, 20, 160, 640, 1280, 1024. A new square of products is then formed, and the columns are summed to give the m_1, m_2 frequencies, which also serve as divisors for calculating the hypergeometric probabilities.

The results obtained from these further tables are shown in the second line of Table 6. The third line of Table 6 is easily obtained, as complete failure of a test batch can only give tables $(a, b;\ 0, 5)$ where a and b have the binomial distribution $(\mathbf{p}_1 + \mathbf{q}_1)^5$. Tables with $a = 3$, 4 or 5 will be significant.

It is obvious from Table 6, as indeed is to be expected, that samples as small as 5 give only very rough tests. Taking an odds ratio of 4:1 as representing a serious degree of defect, only one third of such batches will be rejected, and this at a cost of rejecting 6 per cent of the batches which are up to standard.

The amount of re-testing required for various $\mathbf{p}_1$ shows clearly that $\mathbf{p}_1 = 1/2$ is the value to aim at, if, as is to be expected, the majority of batches are up to standard. The differences in the values of this part of the table are a reflection of variations in the expected marginal totals with different values of $\mathbf{p}_1$ and $\mathbf{p}_2$. Note also that the actual value of $\mathbf{p}_1$ for any particular test is not under complete control: it depends on the test plate actually used. If it was, and if the other variables could be similarly controlled, there would be no need to include a standard sample in each test.

The above example is only intended as an illustration of method. The vital point that emerges is the importance of recognising that some samples do not give any worthwhile information on the point at issue, and that the proportion of such samples is substantially increased by variation in $\mathbf{p}_1$. These uninformative samples always give a non-significant P, but can be identified by their marginal values.

It would be interesting to see how the test performs with larger n_1 and n_2. The same procedure can be followed, but the arithmetic is tedious on a desk or pocket calculator. It would, however, be a simple matter to program a computer to do all or at least the more onerous parts of the calculations.

11. RECENT CRITICISMS OF THE EXACT TEST AND THE CONTINUITY CORRECTION

Failure to recognize the force of the arguments for conditioning outlined above, and evaluation of the performance of tests at nominal levels of significance, has resulted in numerous papers criticizing Fisher's exact test and the continuity correction, and many alternative tests have been devised. Upton (1982) examines no less than 22 tests for comparative trials, and gives references to 53 papers, 25 of them dated from 1970 onwards. This is by no means a complete bibliography of papers, even in English and American journals, on this subject.

Recently computers have been called into play to investigate more fully the performance of rival tests. Extensive tables have been published which, taken at their face value, may well deceive uncritical readers.

There is no need here to attempt any general review. A brief discussion of a paper by Berkson (1978) and four rejoinders to it by Barnard, Basu, Corsten and de Kroon, and Kempthorne (1979), together with Upton's paper, will serve to illustrate the present confusion of thought, and is particularly relevant because the papers by Berkson and Kempthorne are already being taken as guides to up-to-date thinking on the subject: Upton leans heavily on them in his preamble, and they were cited by Fienberg without adverse comment in a lecture series *R. A. Fisher: An Appreciation* (1980).

12. BERKSON'S "DISPRAISE"

Berkson considers three tests, which he denotes by T_N, T_C and T_E. T_N is what he terms the normal test. This he specifies in the same manner as did Yule, not specifically telling his readers (though he was clearly aware of it) that T_N is the same as the χ^2 test without the correction for continuity, the customary formula for which is much more convenient for computation. T_C is defined as the same test with "Yates' correction". T_E is the exact test. At the end of the paper he forcefully concludes that, "at least for a comparative trial with $n_1 = n_2$, T_N is preferable to T_E [and by implication to T_C] and T_E should not be used".

How did he reach this conclusion? "Following the ideas of the Neyman-Pearson theory of tests of significance" and adopting the two-binomial model, he determined the frequencies α_e with which significant verdicts would be given by T_N, T_C and T_E at nominal significance levels $\alpha = 0.05$ and 0.01 (single tail) in repeated sampling of tables with $n_1 = n_2$ when $\mathbf{p}_1 = \mathbf{p}_2$. The table giving his results covers values of $\mathbf{p} = 0.1$ (x 0.1) 0.9 and values of $n = 5, 10, 20, 50, 100, 200$ (the last for $\mathbf{p} = 0.5$ only). The production of this table, and an associated table of the power of the tests, involved Berkson and his associates in a considerable computer exercise.

Table 7 gives an extract of his table for T_N and T_E for $\alpha = 0.05$ and $\mathbf{p} = 0.5$ and 0.2, 0.8.

TABLE 7
Berkson's α_e for $\alpha = 0.05$

	$\mathbf{p} = 0.5$		$\mathbf{p} = 0.2,\ 0.8$	
n_1, n_2	T_N	T_E	T_N	T_E
5	0.0547	0.0107	0.0218	0.0023
10	0.0579	0.0211	0.0455	0.0150
20	0.0421	0.0213	0.0513	0.0226
50	0.0449	0.0287	0.0502	0.0288
100	0.0518	0.0384	0.0497	0.0350
200	0.0494	0.0400	–	–

(His values for T_C are, as is to be expected, for the most part the same as those for T_E.) The values for $n_1 = n_2 = 5$ when $\mathbf{p} = 0.5$ can be verified from Table 2. For T_E, only the tables in the last two lines attain significance at the 0.05 level. Their combined unconditional probability is $(5 + 5 + 1)/1024 = 0.0107$. For T_N the three tables in the next line also attain significance, giving a combined unconditional probability of $56/1024 = 0.0547$. Similar calculations with the frequencies in the successive columns of Table 2 multiplied by $1, 4, 4^2$, etc. and a divisor of 5^{10} give the values for $\mathbf{p} = 0.2, 0.8$.

The fact that in the full table the values of α_e given by T_N are close to the nominal α, except for small n, and $\mathbf{p}$ differing considerably from 0.5, convinced Berkson that T_E was extremely conservative and that T_N was the right test to use. The table is, however, irrelevant in any practical sense. The probabilities given by all three tests are in fact conditional, T_C and T_N because they are both based on χ^2 with 1 df. The justification for T_C is that it provides a close approximation to the exact conditional test T_E. Omission of the continuity correction from T_N results in much higher values for χ^2, particularly for tables with one or more small marginal values, and T_N consequently greatly exaggerates the conditional significance. It also exaggerates the unconditional significance, at least for $\mathbf{p} = 0.5$, as is shown in Table 8. For the table (4, 1; 1, 4), for example, $P_E = 0.103$, $P_C = 0.103$, $P_N = 0.029$, whereas the unconditional probability is 0.055.

TABLE 8

Values of P_E, P_C, P_N, and unconditional P for $n_1 = n_2 = 5$

					Unconditional P	
Table	$p_1 - p_2$	P_E	P_C	P_N	p = 0.5	p = 0.2, 0.8
(4, 1; 1, 4)	0.6	0.103	0.103	0.029	0.0547	0.0218
(3, 2; 0, 5) (5, 0; 2, 3)	0.6	0.083	0.084	0.019	0.0303	0.0192
(4, 1; 0, 5) (5, 0; 1, 4)	0.8	0.024	0.026	0.0049	0.0107	0.0023
(5, 0; 0, 5)	1.0	0.0040	0.0057	0.0008	0.0010	0.0001

Having established to his own satisfaction that T_N is the correct test, Berkson gives an example of the contrasting performance of T_N and T_E in a hypothetical clinical trial—he even specifies it as "double blind"—in which 30 out of 35 patients are cured by a new treatment and 24 out of 35 are cured by the currently used treatment. This gives $P_N = 0.0438$, $P_E = 0.0767$. Berkson suggests that the scientist concerned might reasonably consider T_E to be "destructive" rather than "conservative".

To be fair, Berkson, after referring to various authorities in support of his arguments in favour of T_N, does quote from Fisher's 1935 paper, and after some discussion concludes: "If the significance P is taken to represent, not the frequencies of errors of the first kind, but a measure of the subjective credibility of the null hypothesis, objectified by equating it to fair betting odds, then the exact test with randomizing is the correct test." He then, however, continues: "But what investigating scientist would decide which of two drugs is the more effective by tossing a coin? Perhaps this is a crucial case in reference to the question as to whether statistics is concerned with decision or inference." This misses the point. The coin tossing is only used to eliminate selection bias. Nor would a scientist expect decisions to be taken solely on the basis of a single trial of this size. The significance level of 1 in 13 given by P_E is by no means negligible evidence in favour of the new treatment.

For good measure, Berkson also questions Fisher's contention that the marginal values of a 2 x 2 table contain no information on lack of proportionality, citing various authorities, and even, in a supplementary paper (1978), advancing a most remarkable "proof" of his own!

13. REACTIONS TO BERKSON'S DISPRAISE

Berkson's attacks on the exact test elicited four replies. I need only comment briefly. They do little to clarify the real issues—indeed for the most part they only add further confusion. The most remarkable is that by Kempthorne. He starts by defining three "origins": I, a double dichotomy (only N determined by the observer); II, two binomials (n_1 and n_2 determined); III, a comparative trial (n_1 and n_2 determined, with random assignment between them). After lengthy discussion and much rhetorical abuse of Fisher's arguments in *Statistical Methods and Scientific Inference*, and regret at Barnard's disavowal of the *CSM* test, he concludes that

Berkson's study indicates that T_N (though perhaps not as good as the *CSM* test) is appropriate for Origin II data, and "does what is sought very nicely and easily". He bases this conclusion on the results of Berkson's computations, summarized in Table 7 above. If he had worked out the *CSM* values for $n_1 = n_2 = 5$ (not a difficult task, and included in my Table 8) he might have realized that the impression created by Berkson's table is misleading.

Kempthorne does recognize that for Origin III only the exact T_E test is appropriate, though on the somewhat weak ground that the individuals in the trial are *not* selected at random from a larger population. They can of course be so selected, and then assigned at random to the two treatments; if so, any results emerging from the trial will be relevant to the population from which the selection is made.

Even more surprisingly, "in the lack of additional knowledge" he opts for the exact test for Origin I data. Surely, if his arguments on Origin II were accepted, they would apply with equal or greater force to Origin I. Also it is scarcely a help to "readers with lack of time to read the whole of the material" to reproduce Berkson's definition of T_N in the last paragraph of his conclusion, instead of telling them that it is χ^2 without the correction for continuity. Was he unaware of this?

Basu's paper contains a much briefer but equally rambling discussion of the problem. His final advice is to "act like a Bayesian", adding: "Data interpretation is not a scientific method. There cannot be a mindless weighing of evidence. Can I be truly objective unless I am completely ignorant of the subject?!"

Corsten and de Kroon's short paper is much more sensible. Accepting, without argument, that conditioning is appropriate, they conclude that " comparison of T_N and T_E at the honest basis of conditioning on k [the m_1, m_2 margin in my notation] discredits T_N completely; the value of α_e is irrelevant in this context." They continue, however, with a section headed "Unconditional Testing" beginning: "Berkson's preference may still exist for those who reject conditional considerations at all in this problem." To cater for this preference they state: "It is customary advice to replace this [T_E] by . . . the (*unconditional*) test statistic T_E'." T_E' is in fact the same as that which Pearson used when correcting for continuity, and differs from T_C only in the substitution of the factor $N-1$ for N in the formula for χ_c^2. Their italics indicate that they think that use of a normal approximation in this way makes the test unconditional! I suspect that Berkson, and indeed Kempthorne, suffered from the same delusion.

Barnard's paper is mainly concerned with what he terms "test procedures", and is somewhat peripheral to Berkson's paper, but he very firmly recommends that only the exact test should be used by individual experimenters when reporting the results of their experiments.

14. UPTON'S PAPER

Upton's main object was to examine the performance of the many tests that have been proposed for 2 x 2 comparative trials. The definition he adopted for such trials was that of Barnard (1947), i.e. tables with one fixed margin. The main part of his paper is devoted to a description of 22 alternative tests, and a comparison of the performance of 17 of them in repeated unconditional sampling of two binomials with a common $\mathbf{p}$ and nominal $\alpha = 0.05$ (apparently two-tail), over the whole range, 0 to 1, of $\mathbf{p}$. In addition an attempt is made to assess "the overall accuracy" of the 17 tests, using several criteria. These criteria were evaluated on an assorted collection of no less than 20 tables, both for $\alpha = 0.05$ and $\alpha = 0.01$; only the results for $\alpha = 0.05$ are reported. All this must have involved an immense amount of work, only made possible by modern computing aids.

.In his summary Upton states: "Amongst other results it is shown that the exact test of Fisher, and the corresponding Yates correction to Pearson's χ^2 test, give tests which are both very conservative and inappropriate. The uncorrected χ^2 test performs well. On both empirical and theoretical grounds, the preferred test is the scaled version $(N-1)/N\,\chi^2$." Essentially his procedure is that adopted by Berkson; he could in fact have presented his results in tabular form, as in Table 7, and his summary echoes that of Berkson. My criticisms of Berkson therefore apply equally to

Upton. His suggested modification of the uncorrected χ^2 by the factor $(N-1)/N$ is trivial and lacks any sound theoretical basis. True, following Kempthorne, he does state in his recommendations at the end of the paper that "if the set of data being analysed cannot be regarded as a random sample from the population(s) of interest, as for example occurs in the self-selecting medical trial", only the exact test or the continuity-corrected χ^2 is appropriate, which is a slight advance on Berkson. Berkson did, however, confine his attention to one-tail tests, whereas Upton claims it is "more natural" to use two-tail tests, and adopts a prevalent but misleading procedure (described below) for deriving their associated probabilities. This has introduced further irregularities into his results.

Although there are numerous references to previous literature, it seems that Upton has made only a superficial study of many of the papers he cites, or has relied on comments on them by other authors. For example, his description of "Yates' correction to χ^2" begins: "Yates (1934) observed that, as N increases, the hypergeometric distribution is increasingly well approximated by the normal distribution." I made no such observation. Nor is the statement, in the form he gives it, true; if n_1 and p are held fixed, but n_2, and therefore N, tend to infinity, the hypergeometric distribution tends to a binomial with $n_1 + 1$ terms only. It was Pearson, in his 1947 paper, who applied a continuity correction to the hypergeometric normal approximation, $\{(N-1)/N\}\,\chi^2$. My continuity correction was applied directly to χ^2. Actually, though Pearson did not realize it, my correction performs on average better than his. Upton's further statement, based on a paragraph in Pearson's 1947 paper, that my correction "had been in common use since at least 1921", is incorrect, and results from a misinterpretation of Pearson's actual remarks.

The introductory sections suffer similarly. Had Upton studied Fisher's *Statistical Methods and Scientific Inference*, instead of accepting Kempthorne's "analysis" of it, he might have had doubts about the relevance of his investigation.

15. TWO-SIDED TESTS

With normally distributed continuous data the customary tabulation of t has encouraged statisticians to think in terms of two-sided tests. As the normal distribution and the associated t distributions are symmetrical this raises no problems, though it should be remembered that if an experiment shows a significant difference between two treatments at $P = 0.04$, say, and B has emerged as superior to A, this is equivalent to the statement that B is significantly better than A at the $P = 0.02$ level; also, if fiducial limits $\bar{x} \pm t_{.05}s_m$ are assigned to a mean m of which $\bar{x}$ is an estimate, the fiducial probability that m is below the lower limit is 0.025, not 0.05, and similarly it is 0.025 that m is above the upper limit.

If a continuous error distribution is symmetrical about the null value equal deviations in either direction will have equal one-tail probabilities; if the error distribution is not symmetrical these probabilities will be unequal. However any continuous distribution with a single maximum can be transformed into a normal distribution. Moreover in any one set of results the information on departures from the null hypothesis relates only to departures in the observed direction. Consequently the rule for determining the two-sided probability, if this is required, should be *to double the observed one-tail probability*. This is invariant under transformation, whereas basing two-sided probabilities on equal but opposite deviations is not.

Transformation of data to normal or approximately normal form is of course a well-known device for determining significance probabilities and fiducial limits. A classic example is provided by the correlation coefficient. As Fisher showed (see, for example, *Statistical Methods for Research Workers*, Section 35) the transformation

$$z = \tfrac{1}{2}\,\{\log_e(1+r) - \log_e(1-r)\}$$

gives a distribution which is closely approximated by a normal distribution with variance $1/(n'-3)$, where n' is the number of pairs of observations on which r is based. If, for example, we wish to assess the significance of the difference of an observed r from a theoretical expected value r_0 of $+0.75$ ($z_0 = +0.97$) the one-tail probability will be given directly by reference to a table of the standard normal integral with $x = (z - 0.97)/\sqrt{(n' - 3)}$.

With discontinuous distributions there is a further problem. In a 2 x 2 table with $n_1 = n_2$ there will be pairs of points representing the integral divisions on the two tails which are equidistant from the expected value, e. These will have equal hypergeometric probabilities, as is shown by Table 2. If $n_1 \neq n_2$, but $2e$ is integral, there will still be pairs of points equidistant from e, but also some points on the longer tail that are unpaired; the hypergeometric distribution will then be asymmetric, and the associated probabilities will be unequal. If $2e$ is not integral there will be no equidistant pairs. This last contingency may be termed mismatch.

Table 9 gives examples of the two-sided probabilities for the more extreme values of a, the observed value in the cell with the smallest expectation, obtained by (i) doubling the one-tail probability of the value actually observed, and by (ii) taking the sum of the one-tail probability of the observed value and the one-tail probability of the value on the opposite tail for which the deviation is equal to that of the observed value, or if there is mismatch ($2e$ not integral) the value with the next greater deviation.

TABLE 9

Two-sided probabilities for four tables with given margins: contrasts between (i) twice the exact one-tail probability; (ii) the sum of the exact probabilities of all values with deviations greater than or equal to that of the observed deviation, regardless of sign; (iii) the probability given by the continuity-corrected χ^2

	Table A		Table B			Table C			Table D		
a	(i) (ii)	(iii)	(i)	(ii)	(iii)	(i)	(ii)	(iii)	(i)	(ii)	(iii)
0	0.008	0.013	0.012	0.009	0.019	0.023	0.036	0.041	0.021	0.016	0.039
1	0.092	0.096	0.123	0.095	0.128	0.160	0.170	0.174	0.151	0.102	0.166
...	...	...	...	...	...	...	...	...	...	...	...
6	0.092	0.096	0.067	0.040	0.070	0.181	0.170	0.174	0.191	0.171	0.185
7	0.008	0.013	0.005	0.003	0.008	0.049	0.036	0.041	0.053	0.037	0.045
8	–	–	–	–	–	0.010	0.005	0.007	0.011	0.005	0.007
Exp'n (*e*)	3.5		3.325			3.5			3.544		
Table	a b	20	a	b	19	a	b	20	a	b	20
	c d	20	c	d	21	c	d	60	c	d	59
	7 33	40	7	33	40	14	66	80	14	65	79

In tables A and C $2e$ is integral, in tables B and D it is not. Table A is symmetric and the underlying distribution in table B is nearly so; tables C and D are markedly asymmetric. Consequently both methods give the same results in table A, but in table B all the probabilities given by method (ii) are one quarter to one third less than those of method (i). In table C the asymmetry of method (i) is obliterated by method (ii) except for unmatched extremes on the longer tail. This in itself seems unreasonable, as if $a = 0$ is observed, for example, its smaller probability should be regarded as giving stronger evidence for a departure from the null hypothesis than would the occurrence of $a = 7$.

Table D differs from table C only in the rejection of a single observation from cell d (which in table C must contain at least 46 observations). This increases the expectation of a slightly; it therefore seems reasonable to expect that the probabilities for $a = 0$ and 1 will be slightly decreased and those for $a = 6$, 7 and 8 will be slightly increased, as is indeed the case for method (i). For method (ii), however, the changes for $a = 6$, 7 and 8 are trivial, but the decreases for $a = 0$ and 1 are large. The practical worker, confronted with this fact, might well conclude that however "natural" method (ii) appears to be at first sight it stands condemned on common-sense grounds.

What were Fisher's views on this matter? So far as I know they were never expressed in print, but his reply to a letter by D. J. Finney, a copy of which came my way by chance after drafting the above argument, is, I think, of sufficient interest to reproduce here. I am most grateful to Professor J. H. Bennett of Adelaide University and to Professor Finney for permission to quote this correspondence.

Finney's query arose from Fisher's letter to *Science* (1941), in which Fisher gave the one-tailed test (treated, not treated) for Wilson's example (5, 1; 1, 5). Finney noticed that Wilson's original statement of the problem really required a two-sided test, and that although this presented no difficulty in Wilson's example the solution was not clear for an asymmetrical table, for example (5, 3; 1, 5). As he wrote (May 28th, 1946): "How is he to test the null hypothesis that *A* and *B* are equally harmful, while considering deviations from equality in either direction? Simply to double the total probability for (5, 3; 1, 5) and (6, 2; 0, 6) scarcely seems appropriate, as it does not correspond to any discrete subdivision of cases at the other tail such as (1, 7; 5, 1) and (0, 8; 6, 0). Nor does there appear to me any obvious reason for calculating the probabilities for the two most extreme configurations at the other tail (keeping marginal totals unaltered) and adding their total to the appropriate probability for the tail at which the observations occur.

"Am I missing something very simple here? I cannot remember having seen this problem discussed, and should be grateful for your views."

To this Fisher replied (May 31st): "My dear Finney, – Thanks for your letter. It is a good problem, but I believe I can defend the simple solution of doubling the total probability, not because it corresponds to any discrete subdivision of cases of the other tail, but because it corresponds with halving the probability, supposedly chosen in advance, with which the one observed is to be compared. That is to say, one may decide in advance that if the probability is less than one in forty in either direction then we shall consider if [that?], pending further investigation, the viruses are not pathologically equivalent.

"How does this strike you?"

16. USE OF χ_c^2 IN TWO-SIDED TESTS

A χ^2 test with 1 df is essentially a two-sided test. To obtain the one-tail probability, the probability obtained by reference to a χ^2 table must be halved, but its value is dependent on the deviation actually observed, regardless of whether the deviations on the two tails match or not. There will therefore be no underestimation of *P* in two-sided tests due to mismatch. As, however, equal deviations in opposite directions give equal χ^2 values, differences in significance due to asymmetry will be obliterated. This is apparent in table C of Table 9, where for example the values in column (iii) for $a = 0$ and 7 are both 0.041, whereas those in column (i) are 0.023 and 0.049 respectively. The effect of mismatch on column (ii) of table D, however, is eliminated in column (iii); the difference between the values 0.039 and 0.045 for $a = 0$ and 7 is solely due to the greater deviation from expectation for $a = 0$.

Table B, which differs trivially from table A, also illustrates the serious distortion of the column (ii) values due to mismatch: for $a = 6$, for example, the correct value 0.067 is reduced to 0.040, whereas the continuity-corrected χ^2 gives the value 0.070.

Comparisons of columns (i) and (iii) of Table 9 give an indication of the accuracy of the χ_c approximations in tables with small values of *e*. The largest discrepancies are of course those due to asymmetry, such as those in tables C and D. It is perhaps worth noting that linear interpolation in Table VIII of *Statistical Tables*, using χ_c as argument, gives considerably improved approximations. (The procedure is there illustrated in Example 5.) The approximations to twice the one-tail probabilities for table C, $a = 0, 7, 8$, for example, are 0.026, 0.049, 0.009, which agree well with 0.023, 0.049, 0.010.

The above approximations are based on the tabular values for the corresponding limiting contingency distributions, which will have margins {14, 42; 14, 42; 56}. Similar adjustments, using the tabular values for the corresponding binomial distribution $(3/4 + 1/4)^{14}$, give the approximations 0.016, 0.053, 0.011.

Now that computers are available it would be a relatively simple matter to provide a table for adjusting χ_c values which is more detailed and easier to use. If *corrections* to χ_c are tabulated, with χ_c as argument, and a table of the normal probability integral is available, almost exact significance probabilities could rapidly be obtained with even the most primitive pocket calculators.

17. RECENT ATTACKS ON χ_c^2

Adoption of what I consider to be an inappropriate definition of two-sided tests ((ii) above) has resulted in numerous warnings against the use of the continuity-corrected χ^2 for such tests. A remarkable investigation was made by Haber (1980). Using a specially written computer program, he compared the performance of five different tests on all 2 x 2 tables, some 150 000 in number, for which $N < 100$, $e \geqslant 1$, and for which what he termed "the exact exceedance probability" (i.e. definition (ii), here denoted by P_H) has a value between 0.001 and 0.1. The five tests considered were: the uncorrected χ^2, the continuity-corrected χ^2, two tests based on "Cochran's principle" (though Cochran himself did not support its use for two-sided tests), and a test proposed by Mantel. When $2e$ is an integer tests 3 and 5 are equivalent to the continuity-corrected χ^2, and test 4 is nearly so. (For a specification of these latter tests see Haber's paper.)

Haber tabulated his results in 60 groups, covering values of P_H in the ranges 0.001-0.01 and 0.01-0.1 and grouped according to values of N and e. Results for $\bar{R}, R_{\min}$ and $R_{\max}$ were reported, where for each test $R = P_A/P_H$, P_A being the probability given by the test. Table 10 shows a typical panel of his table, that for P_H (0.01-0.1), $3 \leqslant e < 5$, $40 \leqslant N < 60$. This includes contributions from 2924 tables.

TABLE 10
An example of Haber's results. R is the ratio of the test probability to Haber's "exact exceedance probability"

	Test	$\bar{R}$	$R_{\min}$	$R_{\max}$
1.	**Uncorrected χ^2**	**0.64**	**0.39**	**1.03**
2.	**Continuity-corrected χ^2**	**1.56**	**1.05**	**2.77**
3.	**Two tests based on**	**1.03**	**0.75**	**1.50**
4.	**"Cochran's principle"**	**1.00**	**0.74**	**1.39**
5.	**A test proposed by Mantel**	**1.13**	**0.80**	**1.56**

Taken at their face value, these results indicate that not only does the uncorrected χ^2 considerably underestimate the true significance probability, but also that the continuity-corrected χ^2 seriously overestimates it. The other three tests also exhibit wide variations, though the mean values of R are reasonably close to 1.

This, however, gives an entirely false picture, as Table 9 shows. For the continuity-corrected χ^2 the values of Haber's R are given by the ratios of the probabilities in columns (iii) and (ii). The major differences from unity are due to mismatch. For tables B and D the four R in the range 0.01-0.1 of P_H have values 1.35, 1.75, 2.43, 1.22, mean 1.69, whereas the two pairs in tables A and C have values 1.04 and 1.14. As the great majority of the 2924 tables in Table 10 (as in other parts of Haber's table) are subject to mismatch, the conformity of the four values for tables B and D with his reported results is not surprising.

If, however, the two-sided probability is defined as twice the one-tail probability of the observed value, the appropriate comparison for assessing the average accuracy of χ_c^2 is that between columns (i) and (iii), not (ii) and (iii). The averages for columns (i), (ii) and (iii) of the probabilities for which the column (i) value is in the range 0.01-0.1 are, for tables A and C, 0.053, 0.052, 0.056, and for tables B and D, 0.033, 0.021, 0.036. This shows, as is confirmed by Table VIII of *Statistical Tables*, that the average bias of $P(\chi_c^2)$, if definition (i) is adopted, is small, even for small e. This, of course, does not imply that errors in $P(\chi_c^2)$ or $P(\chi_c)$ are always negligible. They can be relatively large for tables with asymmetric distributions, as tables C and D show.

Had Haber segregated tables with mismatch in the presentation of his results, he would have produced a much more informative table, and one which is more relevant to practical requirements. In most planned comparative trials $n_1 = n_2$ or n_2 is a small multiple, λ, of n_1. If $n_1 = n_2$ there is never mismatch; if $\lambda = 2$ or 3, approximately one third or one half, respectively, of the resultant tables will be free of mismatch.

More fundamentally, if Haber had made comparisons of $P(\chi_c^2)$ with definition (i) as well as (ii) of the exact probability, and if he had subdivided his results according to the degree of asymmetry, the real causes of discrepancies in $P(\chi_c^2)$ would have been much more apparent. In a later paper (Haber, 1982) he does mention that a two-sided probability can be defined in several ways, but again adopts definition (ii) without discussion, and without even revealing what the alternatives are.

18. CONCLUSIONS

The following seem to be the most important conclusions that should be drawn from the above discussion:

1. In spite of the frequently expressed view that Fisher's exact test, based on conditioning on the marginal values, is too "conservative", it appears to be the only rational test, whether both, one, or neither of the margins are determined in advance. The marginal values determine the sensitivity of the test.
2. Unconditional tests, based on the two-binomial model, appeal because they are "more powerful" than the exact test, but stand condemned both by the general arguments for conditioning, and also because the random assignment of treatments in comparative trials leads to the exact test, in spite of only one margin being determined in advance.
3. The examples given in Table 9 confirm that the continuity-corrected χ^2 gives close approximations to the exact test, except when the underlying exact hypergeometric distribution is markedly skew. Condemnation of the continuity correction on the ground that it gives a test that is too conservative is merely the result of failure to recognize that the χ^2 test, like the exact test, is a conditional test.
4. Use of nominal levels of significance such as 5 and 1 per cent is a further source of confusion; the actual levels attained should always be given when analysing discontinuous data.
5. In general, one-tail probabilities should be used, but if a two-sided probability is required the best convention to adopt is to double the observed one-tail probability, as the χ^2 test does automatically. The common convention of taking the sum of the probabilities of all deviations greater than or equal to the observed deviation, regardless of sign, has no realistic justification.
6. In reporting on comparative trials and comparisons between different populations the responsibility of the statistician does not end with the evaluation of the significance probability. He should also comment on the actual p values, and their likely implications.
7. In planning quality control tests using 2 x 2 contrasts, the probabilities of acceptance, rejection and any required retesting should be calculated for various postulated levels of defect in a batch.

ACKNOWLEDGEMENTS

I should like to put on record my thanks to David Cox and George Barnard for helpful suggestions, most of which have been incorporated, to Donald Preece for help in various ways and to Dawn Johnson for her splendid work in the preparation of the final typescript.

REFERENCES

Barnard, G. A. (1945) A new test for 2 X 2 tables. *Nature*, **156**, 177.

——— (1947) Significance tests for 2 X 2 tables. *Biometrika*, **34**, 123–138.

——— (1949) Statistical inference. *J. R. Statist. Soc.* B, **11**, 115–139.

——— (1979) In contradiction to J. Berkson's dispraise: conditional tests can be more efficient. *J. Statist. Planning and Inference*, **3**, 181–187.

Basu, D. (1979) Discussion of Joseph Berkson's paper "In dispraise of the exact test". *J. Statist. Planning and Inference*, **3**, 189–192.

Berkson, J. (1978) In dispraise of the exact test. *J. Statist. Planning and Inference*, **2**, 27–42.

——— (1978) Do the marginal totals of the 2 × 2 table contain relevant information respecting the table proportions? *J. Statist. Planning and Inference*, **2**, 43–44.

Corsten, L. C. A. and de Kroon, J. P. M. (1979) Comment on J. Berkson's paper "In dispraise of the exact test". *J. Statist. Planning and Inference*, **3**, 193–197.

Cox, D. R. (1970) *The Analysis of Binary Data*. London: Methuen.

Fienberg, S. E. (1980) Fisher's contributions to the analysis of categorical data. In *R. A. Fisher: An Appreciation* (S. E. Fienberg and D. V. Hinkley, eds), pp. 75–84. Berlin: Springer-Verlag.

Fisher, R. A. (1922) On the interpretation of χ^2 from contingency tables, and the calculation of *P*. *J. R. Statist. Soc.*, **85**, 87–94.

——— (1925) *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd (5th edition, 1934).

——— (1926) Bayes' theorem and the fourfold table. *Eugen. Rev.*, **18**, 32–33.

——— (1935) The logic of inductive inference. *J. R. Statist. Soc.*, **98**, 39–54.

——— (1935) *The Design of Experiments*. Edinburgh: Oliver and Boyd.

——— (1941) The interpretation of experimental four-fold tables. *Science*, **94**, 210–211.

——— (1956) *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Fisher, R. A. and Yates, F. (1938) *Statistical Tables for Biological, Agricultural and Medical Research*. Edinburgh: Oliver and Boyd (6th edition 1963).

Greenwood, M. and Yule, G. U. (1915) The statistics of anti-cholera and anti-typhoid inoculations, and the interpretation of such statistics in general. *Proc. R. Soc. Med. (Epidemiology)*, **8**, 113–190.

Haber, M. (1980) A comparison of some continuity corrections for the χ^2 test on 2 × 2 tables. *J. Amer. Statist. Assoc.*, **75**, 510–515.

——— (1982) The continuity correction and statistical testing. *Int. Statist. Rev.*, **50**, 135–144.

Kempthorne, O. (1979) In dispraise of the exact test: reactions. *J. Statist. Planning and Inference*, **3**, 199–213.

Pearson, E. S. (1947) The choice of statistical tests illustrated on the interpretation of data classed in a 2 × 2 table. *Biometrika*, **34**, 139–167.

Pearson, K. (1900) On the criticism that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.* (5), **50**, 157–175.

Plackett, R. L. (1977) The marginal totals of a 2 × 2 table. *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. R. Statist. Soc.* A, **145**, 86–105.

Wilson, E. B. (1941) The controlled experiment and the four-fold table. *Science*, **93**, 557–560.

Yates, F. (1934) Contingency tables involving small numbers and the χ^2 test. *J. R. Statist. Soc. Suppl.*, **1**, 217–235.

———(1939) An apparent inconsistency arising from tests of significance based on fiducial distributions of unknown parameters. *Proc. Camb. Phil. Soc.*, **35**, 579–591.

Yule, G. U. (1911) *An Introduction to the Theory of Statistics*. London: Griffin.

APPENDIX

Justification for Regarding the Margins as Ancillary

Fisher introduced his argument for the exact test with the statement: "If it be admitted that these marginal frequencies by themselves supply no information on the point at issue, namely, as to the proportionality of the frequencies in the body of the table, we may recognize the information they supply as wholly ancillary;". The form of this statement is, I think, unfortunate. Certainly it has stimulated others to attempt to demonstrate that the margins do contain some information on proportionality. Had Fisher phrased his statement differently, by saying "If it be admitted that these marginal frequencies supply no information, additional to that contained in the body of the table, . . .", possibly mentioning that this follows from the fact that p_1 and p_2 are sufficient statistics for $\mathbf{p}_1$ and $\mathbf{p}_2$, his grounds for treating the margins as ancillary statistics would have been clearer.

That the margins of a 2 × 2 table by themselves do not, except in extreme cases and in repeated sampling, contain any information on proportionality, is certainly true. In the analogous case in which quantal observations are replaced by quantitative measurements, however, the situation is somewhat different. Such measurements can be arranged in tabular form, analogous to that of a 2 × 2 table, as in Table 11. Assuming that the observations are normally distributed with the same variance about means μ_1 and μ_2, a test of significance of $\bar{x}_1 - \bar{x}_2$ is provided by the ordinary *t*-test, and the fiducial distribution of $\mu_1 - \mu_2$ is similarly available. Suppose, however, that the measurements cannot be assigned to groups A_1 and A_2, as might conceivably happen if the

TABLE 11
The analogy between a 2×2 *table and* n_1 *and* n_2 *values of a quantitative variate x relating to two samples* A_1 *and* A_2

A_1	n_1 values of x_1	$\bar{x}_1$
A_2	n_2 values of x_2	$\bar{x}_2$
	N values of x	$\bar{x}$

identity of the objects had been concealed when being measured and the record of the code used was then lost. If $\mu_1 - \mu_2$ is large compared with its standard error the histogram of all the measurements will exhibit two separate distributions, thus permitting full reconstitution of the data. Note, however, that if $n_1 = n_2$ we cannot say which distribution appertains to A_1. If the distributions overlap, but with clear peaks, significance is still not in doubt, but an unbiased estimate of $\mu_1 - \mu_2$ and its standard error would require a curve-fitting exercise. If there are not two clear peaks there will be little information on $\mu_1 - \mu_2$ or on the significance of the difference. The fact that in certain cases there is quite definite information from the margin in no way invalidates the ancillarity argument on which the t-test and the associated fiducial distribution are based. This rests on the sufficiency of the estimates from the full data of the means and variances, and their independence. (See, for example, Yates, 1939.)

Turning now to 2×2 contingency tables, it is clear that separation of the marginal line into its A_1 and A_2 components is not in general possible. However, if $\mathbf{p}_1 = 1$ and $\mathbf{p}_2 = 0$ the table $(n_1, 0;\ 0, n_2)$ will always be obtained, giving $m_1 = n_1$ in the bottom margin, whereas if $\mathbf{p}_1 = \mathbf{p}_2 = \mathbf{p}$ then m_1 will have the binomial distribution $(\mathbf{p} + \mathbf{q})^N$ in successive samples. Thus if N is large and the difference between m_1 and n_1 is very small we might regard this as somewhat shaky evidence that $\mathbf{p}_1$ and $\mathbf{p}_2$ differ substantially. This, however, is an unprofitable speculation, as when n_1 and n_2 are large the values in the body of the table provide accurate information on $\mathbf{p}_1$ and $\mathbf{p}_2$. Note also that, as in the quantitative case, if $n_1 = n_2$ we cannot tell from the margin which of $\mathbf{p}_1$ and $\mathbf{p}_2$ is likely to be the greater.

This vestigial source of information from the margins may account for the results obtained by Plackett (1977), using the likelihood approach. Here I need only quote from his conclusions:

> "Fisher did not say that the marginal frequencies supply no information, but he argued as if this were the case. The following remarks seem to confirm the intuitive view that the likelihood function provides little information about λ.
>
> . . . .
>
> (d) The procedures of inference used here are known to be asymptotically best in many problems. Their application has been inconclusive."

Sprott (1975) also tackled this problem, and took as an example a set of matched pairs, one of each pair being selected at random for treatment A_1, the other being given treatment A_2. Any such pair must be one of four types: (a), $(1, 0;\ 0, 1)$; (b), $(1, 0;\ 1, 0)$; (c), $(0, 1;\ 0, 1)$; (d), $(0, 1;\ 1, 0)$. Only (a) and (d), which both have margins $\{1, 1;\ 1, 1;\ 2\}$, give any information on the difference between A_1 and A_2. The obvious procedure under most circumstances is therefore to reject (b) and (c) pairs and include only (a) and (d) pairs in the analysis, as Cox (1970), to whom Sprott refers, recommends.

As Sprott was concerned only with the margins he could not adopt this course. Instead he divided the pairs into two groups, (a) and (d), which he termed discordant, and (b) and (c), which he termed concordant. If, for any one pair, the binomial probabilities, in my notation, are $\mathbf{p}_1$ and $\mathbf{p}_2$, the probabilities of (a), (b), (c) and (d) are $\mathbf{p}_1\mathbf{q}_2$, $\mathbf{p}_1\mathbf{p}_2$, $\mathbf{q}_1\mathbf{q}_2$, $\mathbf{q}_1\mathbf{p}_2$ respectively. The null hypothesis that $\mathbf{p}_1 = \mathbf{p}_2 = \mathbf{p}$ then gives the probability of a discordant pair as $2\mathbf{pq}$, and of a concordant pair as $\mathbf{p}^2 + \mathbf{q}^2$. Since $2\mathbf{pq} \leqslant 1/2$ the probability that all n pairs are discordant is $\leqslant 1/2^n$, e.g. for 10 such pairs $P < 0.001$. From this he concluded that "it appears that some sort of information can on occasion be present in the marginal totals of a contingency table". All he is really saying is that if all pairs are discordant, one of the treatments is likely to be almost always a failure, and the other almost always a success. But even if $\mathbf{p}_1 = 0.1$ and $\mathbf{p}_2 = 0.9$, for example, with no variation between pairs, there is only a 1 in 7 chance that all of 10 pairs will be discordant. Nor can we tell from the margins which treatment is a success.

DISCUSSION OF DR YATES'S PAPER

Professor G. A. Barnard (Retired): As Dr Yates points out, arguments about 2 X 2 tables have now gone on for 70 years, so perhaps it would be too much to hope to forestall a centenary, though his paper should go far towards reducing the audience at any such celebration. We must be grateful to him for this, and for emphasizing that there is much more to the interpretation of data, even as simple as this, than simple significance testing, so-called.

Dr Yates has dealt so exhaustively with randomized comparative trials that any residual controversy must now be concerned with the two binomial case, and I shall confine my remarks to this. We need to remember that p_1 and p_2 serve to parameterize this case fully, so that we cannot hope to make a fully adequate summary of the message in the data if we confine ourselves to significance testing, and that in relation to one parameter only rather than two. I stress this point because the Neymannian and the Bayesian approaches to these problems share the feature that we appear to be able to *demand* inferences of a particular kind—for example about a parameter chosen by us as the "parameter of interest", irrespective of other so-called "nuisance parameters". Bayesians succeed in doing this by adding untestable assumptions to the data, while Neymannians introduce arbitrary "principles" such as "similarity" or "unbiasedness" of a test which sometimes, as with the t test, happen to produce sensible answers, but at other times—as with the 2 X 2 table—produce absurdities (Barnard, 1982a). A Fisherian approaches data in the hope that it may throw light on questions of interest, but recognising that a given data set may not allow us to provide unambiguous answers to all the questions we would wish to ask.

In the present case we want to say something about the "difference" between p_1 and p_2, without reference to any complementary parameter. We must first find two parameters, θ and ϕ, such that θ represents "difference", while ϕ represents the complementary parameter which is to be neglected. The two parameters must be range-independent, and θ should reverse its sign on interchange of p_1 and p_2. The simple difference $p_1 - p_2$ cannot be range-independent of any complementary parameter, and the simplest (and perhaps, essentially the only) parameters satisfying these conditions are $\theta = \frac{1}{2}\{\ln(p_1/q_1) - \ln(p_2/q_2)\}$ and $\phi = \frac{1}{2}\{\ln(p_1/q_1) + \ln(p_2/q_2)\}$, the semi-difference of log odds, and the semi-sum. We then enquire whether the 2 X 2 table data allow us to infer something about θ without reference to ϕ.

In parametric cases such as this, the kind of information provided by the data is discoverable from the likelihood function which, for the table (3, 0; 0, 3) is, writing λ for e^{θ} and ν for e^{ϕ}, $L(\lambda, \nu) = \nu^3\lambda^6/\{\lambda + \nu + \lambda\nu^2 + \nu\lambda^2\}^3$ which can be factorised into $L_1(\lambda)\, L_2(\lambda, \nu)$, where $L_1(\lambda) = \lambda^6/\{1 + 9\lambda^2 + 9\lambda^4 + \lambda^6\}$ while and $L_2(\lambda, \nu) = \nu^3\{1 + 9\lambda^2 + 9\lambda^4 + \lambda^6\}/\{\lambda + \nu + \lambda\nu^2 + \nu\lambda^2\}^3$. If the second factor involved ϕ only, we could immediately infer the possibility of making fully efficient inferences about θ without regard to ϕ. As it is, we recognize L_2 as the likelihood function provided by knowledge of the marginal totals $\{3, 3;\ 3, 3;\ 6\}$, while L_1 is the further likelihood provided by knowledge of the contents (3, 0; 0, 3) of the table, given the marginal totals. We can imagine ourselves being informed of the data, first by being told the marginal totals and then, knowing these, being told the contents of the table. If we are prepared to neglect such information about θ as is provided by the marginal totals, then inferences about θ, irrespective of ϕ, are possible on the basis of the conditional distribution.

Should we be prepared to neglect the information in the marginal totals? The null value of θ is 0, and a representative alternative might be taken as 1, corresponding to an odds ratio p_1q_2/p_2q_1 of about 7.4. The likelihood ratio for $\theta = 1$ against $\theta = 0$, from the conditional distribution, is $L_1(e^2)/L_1(e^0) = 8.3846$, while assuming the most likely value, 0, for ϕ, the likelihood ratio from the distribution of the marginal totals is $L_2(e^2, e^0)/L_2(e^0, e^0) = 1.1652$. Measuring the amount of information by the logarithm of the likelihood ratios, there is almost 14 times as much information in the conditional distribution as in that of the marginal totals. And, of course, to use the information in the marginal totals requires some guess concerning the value of