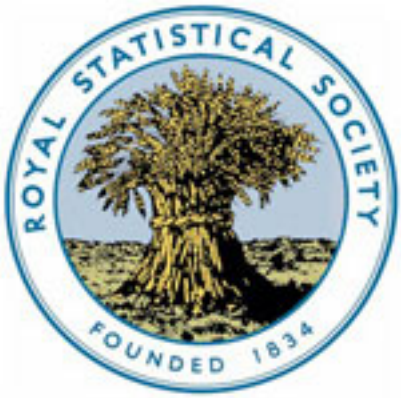



*J. R. Statist. Soc.* A (1992)
**155**, *Part* 3, *pp.* 395–402

# Fisher's Exact Test

By GRAHAM J. G. UPTON†

*University of Essex, Colchester, UK*

[Received May 1991. Revised August 1991]

SUMMARY

This paper reviews the problems that bedevil the selection of an appropriate test for the analysis of a $2 \times 2$ table. In contradiction to an earlier paper, the author now argues the case for the use of Fisher's exact test. It is noted that *all* test statistics for the $2 \times 2$ table have discrete distributions and it is suggested that it is irrational to prescribe an unattainable fixed significance level. The use of mid-*P* is suggested, if a formula is required for prescribing a variable tail probability. The problems of two-tail tests are discussed.

*Keywords*: GOODNESS OF FIT; MID-*P*; SIGNIFICANCE; TEST PROCEDURES; TWO-TAILED TESTS; $2 \times 2$ TABLE

## 1. INTRODUCTION

Barnard (1984) noted that 'arguments about $2 \times 2$ tables have now gone on for 70 years', while Cox (1984) described these arguments as representing 'a saga, a story with deep implications'. Barnard and Cox were respectively proposing and seconding the vote of thanks of the Royal Statistical Society for the paper by Yates (1984). Both speakers proffered the hope that Yates's paper would, in the words of Professor Cox, 'squash once and for all various misconceptions'. However, despite these wishes, that paper has by no means brought the discussions of this contentious topic to a halt. This is evidenced by the flurry of recent papers that include those by Haber (1986), Overall *et al.* (1987), Rice (1988), Lloyd (1988), D'Agostino *et al.* (1988), Barnard (1989), Little (1989), Camilli (1990), Richardson (1990), Storer and Kim (1990) and Cormack and Mantel (1991).

A decade ago I wrote a paper (Upton, 1982), frequently referenced subsequently, in which I made a case against the use of Fisher's exact test. The purpose of this paper is to announce my conversion, brought about by conversations with Professor Barnard and the stimulus of his recent papers (Barnard, 1989, 1990). The argument for the use of the conditional test statistic is set out briefly in the next section. The subsequent sections are concerned with the associated problems of significance levels and tail probabilities which have both helped in the past to confuse discussions of the problem.

## 2. PURPLE FLOWERS, TEST PROCEDURES AND THE EXACT TEST

I find that Fisher's example of the purple flowers, recounted by Barnard (1984) and set out at length in Table 3 of Camilli (1990), represents a clinching argument concerning the need for conditioning. A modified form of the same example is

†*Address for correspondence*: Department of Mathematics, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK.

 0035–9238/92/155395 $2.00

discussed by Little (1989), and a similar example is presented by Cormack and Mantel (1991).

The example is concerned with a number of plants (four in the original example) each of which may, or may not, flower. The question at issue is the proportion of plants that, on flowering, bear purple flowers. The essence of the argument is that, assuming that flowering is independent of flower colour, it cannot make sense for the test to depend on the number of plants that happen to flower.

Barnard (1979) distinguished between tests and test procedures, and it may help to set out a possible test procedure for this example. To make matters more realistic, assume that there are 12 (rather than four) plants, and that the hypotheses are $H_0$:$P$(purple) = 0.5 and $H_1$:$P$(purple) > 0.5. The proposed test procedure, which assumes that the nuisance parameter $\pi$, the probability of flowering, is independent of the flower colour, is set out in Table 1.

When $\pi = \frac{1}{2}$, the overall type I error of this procedure is 0.036, when $\pi = \frac{7}{8}$ it is 0.046 and for $\pi = 1$ it reaches a maximum at 0.073. However, it is not these overall probabilities that will concern the experimenter, but rather the *relevant* (conditional) probability given in the final column of Table 1. The word 'relevant' here applies to the actual number of plants that happen to flower. Once that number is known, the values in the other rows of the table are irrelevant.

The argument for the exact test exactly parallels that for the purple flowers. Suppose, for example, that we are interested in the equality ($H_0$), or otherwise ($H_1$), of the germination rates of seeds from two different sources and that six seeds have been obtained from each source. A possible test procedure is set out in Table 2.

If the common probability of germination is $\frac{1}{2}$, then this test procedure has a type I error of 0.039, whereas the chance of a situation arising in which no decision can be made is 0.146. This latter situation predominates if the true value of the nuisance parameter is near 0 or 1.

Common to these two examples is the precept that we would not consider the performance of a test procedure with respect to its use in *cases that have not occurred.* Indeed, if we were to do so, then it would be difficult to know where to stop—should we, perhaps, be aiming at a global significance level for all statisticians over all significance tests? This is patently absurd but is, I regret, no more than a logical development of my 1982 argument.

TABLE 1
*Proposed test procedure for the case of 12 seeds*

| *No. flowering* | *Condition for rejecting $H_0$* | *P(rejection)* |
|---|---|---|
| 12 | 9 or more plants have purple flowers | 0.073 |
| 11 | 9 or more plants have purple flowers | 0.033 |
| 10 | 8 or more plants have purple flowers | 0.055 |
| 9 | 8 or more plants have purple flowers | 0.020 |
| 8 | 7 or more plants have purple flowers | 0.035 |
| 7 | 6 or more plants have purple flowers | 0.062 |
| 6 | All 6 plants have purple flowers | 0.016 |
| 5 | All 5 plants have purple flowers | 0.031 |
| 4 | All 4 plants have purple flowers | 0.063 |
| <4 | No decision | |

TABLE 2
*Proposed test procedure for the case of six seeds from each of two sources*

| *No. germinating* | *Condition for rejecting* $H_0$ | *P(rejection)* |
|---|---|---|
| >8 | No decision | |
| 8 | 6 from one source and 2 from the other | 0.061 |
| 7 | 6 from one source and 1 from the other | 0.015 |
| 6 | 5 or 6 from one source, 1 or 0 from the other | 0.080 |
| 5 | 5 from one source and 0 from the other | 0.015 |
| 4 | 4 from one source and 0 from the other | 0.061 |
| <4 | No decision | |

## 3. SIGNIFICANCE TESTS AND THE 5% TRAP

Part of the blame for the continuing controversy over the selection of an appropriate test procedure can be laid at the door of Fisher himself, since it was Fisher who introduced tables with preselected tail probabilities for the $\chi^2$-table 'owing to copyright restrictions' (Fisher, 1958). Since the introduction of these, and similar tables, generations of statistics teachers have had to steer an uneasy line between the convenience of the tables and the inevitable brainwashing of their students that arises. Readers of this paper should need no reminding that an outcome associated with a tail probability (measured in some agreed way!) of 0.0501 should be treated in essentially the same way as an outcome that produces a tail probability of 0.0499. For all practical purposes the two outcomes are equally 'significant'.

Barnard (1990) refers to 5% and 1% as 'magic values' and gives an interesting historical account of the development of the increasing 'significance' of significance. Both Barnard (1990) and Camilli (1990) quote the remark by Fisher (1973) that

> 'no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and ideas'.

The experimenter must keep in mind that significance at the 5% level will only coincide with practical significance by chance! There are therefore legitimate conceptual reasons for departing from the 5% level, in addition to the pragmatic reason that it may be unattainable.

## 4. FLEXIBLE *P*-VALUES

The quote from Fisher in the previous section is, in effect, a directive to use 'flexible' *P*-values, varying from case to case. The determinants of this flexibility will be associated with the actions consequent on the decisions that we take. A decision that results in very costly consequences will not be taken lightly: the evidence will need to be very persuasive.

A second determinant will be the sample size on which the decision rests, since this will determine the power of the test procedure. Among others, Johnstone (1986) and Barnard (1989) advocate reducing the tail probability judged as critical, as the sample size increases. Both stress the need for flexibility in significance levels. McPherson

(1989) also advocates flexibility in a valuable extended critique of the use of *P*-values.

Detailed prescriptions for varying significance levels in the analysis of multi-dimensional contingency tables have been advocated by several researchers. Aitkin (1979, 1980) proposes a simultaneous test procedure for testing null hypotheses concerning groups of parameters which involves global significance levels of between 25% and 50% and results in individual significance tests at many intermediate levels. Raftery (1986) advocates the use of the Bayesian information criterion due to Schwarz (1978) for the selection of an appropriate model. This procedure results in a formula that directly relates significance levels to sample size. Support for this procedure is provided by the work of Koehler and Murphree (1988) and an example of its use is provided by Upton (1990).

## 5. THE EXACT TEST IS NOT CONSERVATIVE!

Some of the recent papers on $2 \times 2$ tables *begin* with the statement, taken as axiomatic, that the exact test (or the $\chi^2$-approximation due to Yates (1934)) is a conservative test (e.g. D'Agostino *et al.* (1988) and Storer and Kim (1990)). Others have conducted studies of a variety of contingency tables and have then come to that same conclusion (Overall *et al.*, 1987; Richardson, 1990). These latter papers are merely the most recent in a long line of studies which includes Upton (1982). It is now my view that all these papers are in error because they start with a preposterous premise, namely that a test procedure must be defined to have a constant significance level over all implementations.

As an example, consider the approach that I adopted in my 1982 paper. I asked the question 'Which of the competing tests, when employed at a nominal significance level $\alpha$, most nearly gives a true type I error equal to $\alpha$?'. This echoed the frequentist approach of Neyman and Pearson. However, my concern with this question was not based on theoretical considerations but was purely pragmatic. I took the view that, if users of a test believe that it has a type I error equal to $\alpha$, then their belief should not be too far from the truth. This led me to conclude that Fisher's exact test and the close approximation obtained using the Yates correction to the $\chi^2$-test ($Y$) were not appropriate because, in their usual implementation, their actual tail probabilities are usually much smaller than the nominal tail probability.

However, there are two key phrases near the beginning of the previous paragraph: 'nominal significance level' and 'if users of a test believe that it has a type I error equal to $\alpha$'. It is this idea of a pre-set significance level, which is rarely exactly attainable by the test statistic, that is the source of all the confusion. If we quote only the attainable significance levels consonant with the particular set of fixed margins for the data at hand then the exact test is *not* conservative.

Tocher (1950) showed that the exact test, augmented with an auxiliary randomization to achieve a desired significance level, was the uniformly most powerful unbiased test. This augmentation was described as 'repugnant' by Mantel and Greenhouse (1968). If we are content to work with *achievable* significance levels then the need for the auxiliary experiment disappears, but it is then difficult to make sense of the unbiasedness property. It is amusing to note that the much quoted case of the $2 \times 2$ table in which all four margins are equal to 3 and in which the cell frequencies are (3, 0, 0, 3) corresponds to a conditional single-tail probability of precisely 5%, without recourse to randomization.

## 6. DISCRETE DISTRIBUTIONS AND MID-*P*

Most of the problems that arise in considering tests for the $2 \times 2$ table arise because of the inherent discreteness of *any* test statistic that we may wish to use (unless augmented by the 'repugnant' randomization). The discreteness is most apparent for the exact test but is equally true for all other test statistics. For example, with the uncorrected '$\chi^2$-statistic', the true distribution is not $\chi^2$, but a closely approximating discrete distribution. The idea of a tail probability, so transparent when considering a continuous distribution, becomes, for a discrete distribution, more baffling the more that one ponders over it. For a continuous distribution we can happily calculate the value of $T(x)$, where $T(x) = P(X \geqslant x)$ and $x$ is the observed value of the random variable $X$. In this case, as required, $E\{T(X)\} = 0.5$. However, when $X$ is discrete, $E\{T(X)\} > 0.5$, implying that the Fisher tail areas 'are "biased" in an upward direction' (Barnard, 1989). To correct this problem Lancaster (1949) suggested the use of the *mid-P*-value $M(x)$, given by

$$M(x) = P(X > x) + 0.5\,P(X = x). \tag{1}$$

Further support for the use of mid-*P* is to be found in Stone (1969) and Anscombe (1981).

The test procedures specified in Tables 1 and 2 were based on nothing other than the requirement that the tail probabilities should be reasonably close to 0.05. No rigorous rule was used in their formulation. However, if this requirement is deemed to be sensible and a rule is required, then mid-*P* provides a convenient rule—one selects for rejection those cases corresponding to mid-*P* values less than 0.05 (Barnard, 1989). As it happens, the procedures suggested in Tables 1 and 2 satisfy this criterion. Barnard suggests that mid-*P* may be thought of as 'assessing the strength of evidence against the null hypothesis'. Barnard also observes that, since the exact test is free of nuisance parameters, it is simple to compute the power of a test procedure as a function of the odds ratio. An examination of the power function may influence a decision concerning the test procedure to be used.

## 7. YATES'S CORRECTION AND THE EXACT TEST

My support for Fisher's exact test might be construed as implying support for its close approximator, $Y$, the Yates-corrected version of the familiar $\chi^2$-test. However, I see a danger inherent in the use of $Y$ that does not occur with the use of Fisher's exact test. This danger is that the user of $Y$ may be seduced by the continuous nature of the $\chi^2$-distribution into forgetting that $Y$ is being used to *approximate* a sum of *discrete* probabilities. It is then only a short step to forgetting that the true type I error is not what it appears to be!

## 8. MID-*P* AND THE UNCORRECTED TEST

We begin by considering the general case in which $Z$ is a continuous unimodal random variable having the same mean and variance as a discrete variable $X$. Suppose that we are interested in the probability of exceeding a critical value $x$, lying in the upper tail of the distribution of $X$. Using the distribution of $Z$ to approximate that of $X$, and using the continuity correction, we have the familiar results

$$P(X > x) \simeq P(Z > x + \tfrac{1}{2}) = P(Z > x) - P(x < Z < x + \tfrac{1}{2}) \tag{2}$$

and

$$P(X=x) \simeq P(x-\tfrac{1}{2} < Z < x) + P(x < Z < x+\tfrac{1}{2}). \quad (3)$$

In many cases, a reasonable further approximation is that

$$P(x-\tfrac{1}{2} < Z < x) = P(x < Z < x+\tfrac{1}{2}), \quad (4)$$

in which case

$$P(X=x) \simeq 2P(x < Z < x+\tfrac{1}{2}). \quad (5)$$

Combining equations (1), (3) and (5), we have

$$M(x) \simeq P(Z > x). \quad (6)$$

These results treated a general case. In the specific instance of the $2 \times 2$ table, the equality in equation (4) must be replaced by an inequality, since the density function of $\chi_1^2$ is monotonically decreasing. For a two-tailed test the inequality works in an opposite fashion in each tail. The two discrepancies from inequality largely cancel themselves out, with the result that the use of mid-$P$ is tantamount to the use of the *uncorrected* $\chi^2$-test.

Fortunately these discussions are becoming increasingly academic owing to the increasing number of statistical packages that report exact tail probabilities. A notable example is STATXACT, reviewed by Sprent (1990), which also reports mid-$P$-values.

## 9. ONE-SIDED AND TWO-SIDED TESTS

Here we are faced with two contentious issues. First, can it make sense to perform a one-tailed test? Second, if we perform a two-tailed test, how do we reconcile the information from an observation in one of the tails with the probabilities (of events that did not occur, but might have occurred) in the other tail?

The argument against the existence of one-tailed tests is that, although the experimenter may be anticipating, or hoping for, an outcome in one tail of the distribution, he (or she) will surely not disregard an extreme result in the opposite tail. It may, for example, draw attention to some defect in the test procedure.

However, Fisher clearly often used one-tailed tests when analysing $2 \times 2$ tables. In 1957, writing to E. B. Wilson, Fisher stated that

> '[in the context of] the $2 \times 2$ table, when making an exact calculation I always use the single tail, and if I want to compare significance with cases where both tails are used, I simply *double the value obtained, without regard to the question of how lumpy the other tail may be* [my italics]. Usually, indeed, I think that the single tail is appropriate, though of course not always' (Bennett, 1990).

We can reconcile these views by requiring the experimenter to set out clearly his or her test procedure *before* conducting the experiment. This has many benefits. It avoids an unconscious bias on the part of the experimenter towards a choice of critical region that conveniently includes the observed outcome! It also draws attention to problems caused by small sample sizes: the hawk–owl experiment described by Rice (1988) might not have been conducted if it had been observed that the most extreme result gave a tail probability of 1/15—though *any* experimentation is better than none. Finally, although the experimenter may be expecting, or hoping for, a result in

one tail of the distribution, it forces him or her to consider the consequences of an outcome in the opposite tail.

The previous quotation from Fisher shows clearly that he considered each tail as a separate entity. In determining the form of the test procedure, we can presume that Fisher would have considered each tail separately. How then would Fisher have formulated a test procedure? It seems to me that, using either mid-$P$ or the tail probability, as appropriate, he would have made separate decisions for each tail concerning where, in that tail, he wished to draw the dividing line or lines between the possible decisions.

Note that there may be a range of outcomes for the original experiment for which the most appropriate conclusion may be that there is a need for the collection of further data before a final decision is reached. In this context the paper by Berger and Sellke (1987) on posterior probabilities is relevant, and we should perhaps take to heart the view of Good (1987) in the ensuing discussion that

> 'the conventional $P$ value of approximately .05 [should] be correctly interpreted: not as a good reason for rejecting $H_0$ but as a reason for obtaining more evidence provided that the original experiment was worth doing in the first place'.

## 10. SUMMARIZING REMARKS

Although the primary concern of this paper has been with the $2 \times 2$ table, the ramifications extend beyond this specialized situation. For example, the issues raised by unattainable significance levels apply to simple hypotheses concerning any discrete distribution.

In effect we have evolved a general prescription for the method of conducting a general significance test, based on the need to prescribe *before conducting the experiment* the nature of the conclusions to be drawn from the possible outcomes of the experiment. This might be thought to be standard practice—it is certainly an uncontentious conclusion. Nevertheless, implicit in the thoughtless use of a fixed but unattainable significance level in the context of $2 \times 2$ tables is an abuse of this practice.

## ACKNOWLEDGEMENT

I wish to thank Professor Barnard for the conversations that stimulated the production of this paper. To a considerable extent this paper plagiarizes his views expressed in the papers referenced earlier, and I am grateful for his comment to the effect that he is never averse to people repeating what he has been advocating! I am also grateful to him for bringing to my attention the letter from Fisher to Wilson, cited earlier. Nevertheless, despite the above, misrepresentations of facts or opinions are, of course, my own responsibility.

## REFERENCES

Aitkin, M. (1979) A simultaneous test procedure for contingency table models. *Appl. Statist.*, **28**, 233–242.

——— (1980) A note on the selection of log-linear models. *Biometrics*, **36**, 173–178.

Anscombe, F. J. (1981) *Computing in Statistical Science through APL*, pp. 288–289. New York: Springer.

Barnard, G. A. (1979) In contradiction to J. Berkson's dispraise: conditional tests can be more efficient. *J. Statist. Planng Inf.*, **3**, 181–187.

——— (1984) Discussion on Tests of significance for 2 × 2 contingency tables (by F. Yates). *J. R. Statist. Soc.* A, **147**, 449–450.

——— (1989) On alleged gains in power from lower *P* values. *Statist. Med.*, **8**, 1469–1477.

——— (1990) Must clinical trials be large?: the interpretation of *P* values and the combination of test results. *Statist. Med.*, **9**, 601–614.

Bennett, J. H. (1990) *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*, p. 239. Oxford: Oxford University Press.

Berger, J. O. and Sellke, T. (1987) Testing a point null hypothesis: the irreconcilability of *P* values and evidence: *J. Am. Statist. Ass.*, **82**, 112–122.

Camilli, G. (1990) The test of homogeneity for 2 × 2 contingency tables: a review of and some personal opinions on the controversy. *Psychol. Bull.*, **108**, 135–145.

Cormack, R. S. and Mantel, N. (1991) Fisher's exact test: the marginal totals as seen from two different angles. *Statistician*, **40**, 27–34.

Cox, D. R. (1984) Discussion on Tests of significance for 2 × 2 contingency tables (by F. Yates). *J. R. Statist. Soc.* A, **147**, 451.

D'Agostino, R. B., Chase, W. and Belanger, A. (1988) The appropriateness of some common procedures for testing equality of two independent binomial proportions. *Am. Statistn*, **42**, 198–202.

Fisher, R. A. (1958) *Statistical Methods for Research Workers*, ch. 4. Edinburgh: Oliver and Boyd.

——— (1973) *Statistical Methods and Scientific Inference*, p. 45. New York: Hafner.

Good, I. J. (1987) Comment. *J. Am. Statist. Ass.*, **82**, 125–128.

Haber, M. (1986) An exact unconditional test for the 2 × 2 comparative trial. *Psychol. Bull.*, **99**, 129–132.

Johnstone, D. J. (1986) Tests of significance in theory and practice. *Statistician*, **35**, 491–504.

Koehler, A. B. and Murphree, E. S. (1988) A comparison of the Akaike and Schwarz criteria for selecting model order. *Appl. Statist.*, **37**, 187–195.

Lancaster, H. O. (1949) Statistical control of counting experiments. *Biometrika*, **39**, 419–422.

Little, R. J. A. (1989) Testing the equality of two independent binomial proportions. *Am. Statistn*, **43**, 283–288.

Lloyd, C. J. (1988) Doubling the one-sided *P*-value in testing independence in 2 × 2 tables against a two-sided alternative. *Statist. Med.*, **5**, 629–635.

Mantel, N. and Greenhouse, S. W. (1968) What is the continuity correction? *Am. Statistn*, **22**, 27–30.

McPherson, G. (1989) The scientists' view of statistics—a neglected area. *J. R. Statist. Soc.* A, **152**, 221–240.

Overall, J. E., Rhoades, H. M. and Starbuck, R. R. (1987) Small-sample tests for homogeneity of response probabilities in 2 × 2 contingency tables. *Psychol. Bull.*, **102**, 307–314.

Raftery, A. E. (1986) A note on Bayes factors for log-linear contingency table models with vague prior information. *J. R. Statist. Soc.* B, **48**, 249–250.

Rice, W. R. (1988) A new probability model for determining exact *P* values for 2 × 2 contingency tables when comparing binomial proportions. *Biometrics*, **44**, 1–22.

Richardson, J. T. E. (1990) Variants of chi-square for 2 × 2 contingency tables. *Br. J. Math. Statist. Psychol.*, **43**, 309–326.

Schwarz, G. (1978) Estimating the dimensions of a model. *Ann. Statist.*, **6**, 461–464.

Sprent, P. (1990) Review of STATXACT. *Appl. Statist.*, **39**, 391–394.

Stone, M. (1969) The role of significance testing: some data with a message. *Biometrika*, **56**, 485–493.

Storer, B. E. and Kim, C. (1990) Exact properties of some exact test statistics for comparing two binomial proportions. *J. Am. Statist. Ass.*, **85**, 146–155.

Tocher, K. D. (1950) Extension of the Neyman–Pearson theory of tests to discontinuous variates. *Biometrika*, **37**, 130–144.

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

——— (1990) The exploratory analysis of survey data using log-linear models. *Statistician*, **40**, 169–182.

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

——— (1984) Tests of significance for 2 × 2 contingency tables (with discussion). *J. R. Statist. Soc.* A, **147**, 426–463.