The Choice of Statistical Tests Illustrated on the Interpretation of Data Classed in a 2 × 2 Table

Author(s): E. S. Pearson

Source: *Biometrika*, Vol. 34, No. 1/2 (Jan., 1947), pp. 139-167
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: http://www.jstor.org/stable/2332518
Accessed: 07-08-2016 20:18 UTC

THE CHOICE OF STATISTICAL TESTS ILLUSTRATED ON THE INTERPRETATION OF DATA CLASSED IN A 2 × 2 TABLE

By E. S. PEARSON

CONTENTS

(i) Introductory

1. The problem of testing the significance of a difference between two proportions is one which receives early attention in text-books on mathematical statistics, and it might be thought to be one of the questions whose final solution lies behind us. It is a problem whose simplicity makes it easy to examine the logical cogency of the methods put forward for its solution, but, on examination, it is evident that they have not yet been rounded off satisfactorily. The origin of the present paper lies partly in an investigation commenced in 1938 and discussed at the time in College lectures, and partly in recent correspondence in *Nature* in which G. A. Barnard (1945*a*, *b*) and R. A. Fisher (1945*a*) have taken part.* This correspondence has suggested that in a problem of such apparent simplicity, starting from different premises, it is possible to reach what may sometimes be very different numerical probability figures by which to judge significance.

2. Such a difference in levels of significance in the solution of an everyday problem is obviously puzzling to the users of statistical methods who are accustomed to accept the technique as an established procedure and have not the opportunity for a critical examination of the conditions under which probability theory is brought to bear as a guide to action. For the question here at issue is a fundamental one of why and how our judgement is influenced by the calculation of a probability, and the dilemma raised by the Barnard-Fisher correspondence can only be answered in terms of our views on the practical function of the theory. We may all agree that in practice we use probability figures derived from an analysis of numerical data to help us to make up our minds on the next step, whether in experimental research or executive action. But what form of presentation of the probability set-up is likely to result in the greater number of sound decisions is likely to be always a matter for differences of opinion.

3. All that I can do is to approach the problem of the 2 × 2 table from the viewpoint which appears most helpful to me. In the preceding paper Mr Barnard has elaborated the

* There was also an earlier discussion on the same subject between E. B. Wilson (1941, 1942) and R. A. Fisher (1941).

views expressed in his letters to *Nature*. Such discussion is, I believe, desirable, even though controversial issues are raised. For the value of the whole elaborate structure of the modern theory of mathematical statistics depends at least in part on the sense in which the individual statistician appreciates the meaning of the probability model he is using when drawing the practical conclusions from his analysis of data. I have used the words 'in part', for it is true that the analytical process of applying the statistical technique to experimental data may in itself be enormously illuminating even without paying any close regard to a final probability figure. Such is the case, for example, with the technique of analysis of variance, where the mere process of breaking up a total sum of squares into parts with which different sources of variability can be associated, brings with it a reward in clear thinking even without the application of a probability test.

4. There is a very wide variety in the types of situation in which probability theory is introduced to help in reaching a decision as to further action.

(A) At one extreme we have the case where repeated decisions must be made on results obtained from some routine procedure carried out under controlled conditions.

(B) At the other is the situation where statistical tools are applied to an isolated investigation of considerable importance in which many of the issues involved in the conclusion can hardly be assessed in numerical terms.

5. Two situations of this kind, in which the statistical technique involved is that of testing the significance of a difference between two proportions, may be illustrated from problems arising in the 'proof' of armour-piercing shot or shell.

6. *Example of type A*. In the proof of small anti-tank, armour-piercing shot it might be decided to set aside, as a standard, a batch of shot whose quality has been established by special trials; against this standard, later batches can be compared. The variable measured is the proportion of shot which fail to perforate a plate of specified thickness when fired with a given striking velocity. The use of standard shot is necessary for calibration purposes, because there are inevitable changes in toughness from one proof plate to another and only a limited number of shot can be fired at a single plate. Then the situation might be summed up as follows:*

Aim of proof. To ensure that as few batches as possible are passed into service which are less effective than the standard.

Method of proof. Twelve rounds of the standard and twelve of the batch under test to be fired, round for round, against a single test plate and a record kept of the number of failures in each group, say a and b.

Routine sentencing rule. This should lay down a ready means of determining, from a knowledge of a and b, whether to class the new batch as inferior to the standard or not.

Assumptions accepted in using rule. That the two samples of twelve shot have each been randomly selected from the much larger batches. That against the particular plate used, a proportion p_1 of the standard and p_2 of the new batch would fail to give satisfactory perforation at the specified striking velocity. That while p_1 and p_2 would be different for other plates, if $p_2 > p_1$ for one plate, it will be so for all other plates. The objective is to segregate batches of shot for which $p_2 > p_1$.

* It has been somewhat simplified for illustrative purposes, e.g. complete control of the striking velocity is not in practice possible.

7. *Example of type B.* Two types of heavy armour-piercing naval shell of the same calibre are under consideration; they may be of different design or made by different firms. Since the cost of producing and testing a single round of this kind runs into many hundreds of pounds, the investigation is a costly one, yet the issues involved are far reaching. Twelve shells of one kind and eight of the other have been fired; two of the former and five of the latter failed to perforate the plate. In what way can a statistical test contribute to the decision which must be taken on further action?

8. In dealing with Example A the guiding principle followed in seeking help from the theory of probability can be very simple. We can set as our object a rule which:

(i) will result in an increasing chance of detecting that $p_2 > p_1$, the larger the difference;

(ii) will leave only a small chance of segregating the new batch wrongly when, in fact, $p_2 \leqslant p_1$.

Diagrammatically the rule would consist in segregating the new batch when the point (a, b) falls within some such area as that shown shaded in Fig. 1. In this problem involving a routine procedure, it is the long-run frequency of different consequences of the proof sentencing which is of importance, and probability theory is introduced to provide a measure of expected frequency. This method of introducing the theory of probability into this proof problem is not necessarily the only one that could be adopted in fixing a routine procedure, but it is a simple one and, since simplicity has the merit of appealing to the user's understanding, it has great advantages.

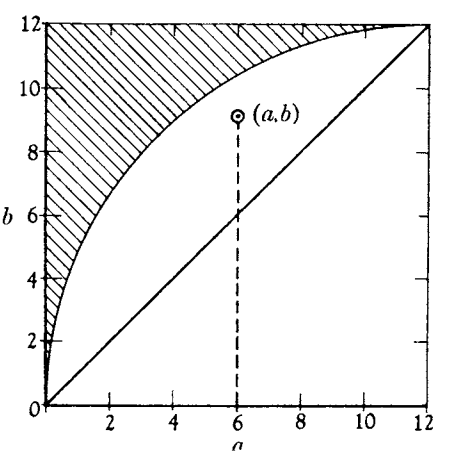

Fig. 1

9. When dealing with Example B a very considerable number of factors must be weighed in the balance, and the result of a statistical test of significance could never be the over-riding one. There will be other information as to the effect of changes in shell design, possibly from shell of different calibre; information as to the uniformity in quality of output of the firm or firms concerned; questions of cost and of general policy. He would be a bold man who would attempt to express these in numerical terms. Whereas when tackling problem A it is easy to convince the practical man of the value of a probability construct related to frequency of occurrence, in problem B the argument that 'if we were to repeatedly do so and so, such and such result would follow in the long run' is at once met by the commonsense answer that we never should carry out a precisely similar trial again.

10. Nevertheless, it is clear that the scientist with a knowledge of statistical method behind him can make his contribution to a round-table discussion, provided he has acquired a grasp of the practical issues. Starting from the basis that individual shell will never be identical in armour-piercing qualities, however good the control of production, he has to consider how much of the difference between (i) two failures out of twelve and (ii) five failures out of eight is likely to be due to this inevitable variability. There may be a number of ways of sizing up the position involving different assumptions or hypothetical constructs; he may follow one or several of these. The value of his advice is dependent almost

entirely on the soundness of his scientific judgement, and very little on whether his back-room calculations have been based on inverse or direct probability or on an appeal to fiducial argument.

11. How far, then, can one go in giving precision to a philosophy of statistical inference? It seems clear that in certain problems probability theory is of value because of its close relation to frequency of occurrence; such seems to be the case for my Example A. Tests can be built up to satisfy the practical requirements in this field. In other and, no doubt, more numerous cases there is no repetition of the same type of trial or experiment, but all the same we can and many of us do use the same test rules to guide our decision, following the analysis of an isolated set of numerical data. Why do we do this? What are the springs of decision? Is it because the formulation of the case in terms of hypothetical repetition helps to that clarity of view needed for sound judgement? Or is it because we are content that the application of a rule, now in this investigation, now in that, should result in a long-run frequency of errors in judgement which we control at a low figure? On this I should not care to dogmatize, realizing how difficult it is to analyse the reasons governing even one's own personal decisions.

12. That the frequency concept is not generally accepted in the interpretation of statistical tests is of course well known. With his characteristic forcefulness R. A. Fisher (1945*b*) has recently written: 'In recent times one often repeated exposition of the tests of significance, by J. Neyman, a writer not closely associated with the development of these tests, seems liable to lead mathematical readers astray, through laying down axiomatically, what is not agreed or generally true, 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. This intrusive axiom, which is foreign to the reasoning on which the tests of significance were in fact based seems to be a real bar to progress....'

13. But the subject of criticism seems to me less an intrusive mathematical axiom than a mathematical formulation of a practical requirement which statisticians of many schools of thought have deliberately advanced. Prof. Fisher's contributions to the development of tests of significance have been outstanding, but such tests, if under another name, were discovered before his day and are being derived far and wide to meet new needs. To claim what seems to amount to patent rights over their interpretation can hardly be his serious intention. Many of us, as statisticians, fall into the all too easy habit of making authoritative statements as to how probability theory should be used as a guide to judgement, but ultimately it is likely that the method of application which finds greatest favour will be that which through its simplicity and directness appeals most to the common scientific user's understanding. Hitherto the user has been accustomed to accept the function of probability theory laid down by the mathematicians; but it would be good if he could take a larger share in formulating himself what are the practical requirements that the theory should satisfy in application.

(ii) The choice of statistical tests

14. One approach to follow in determining tests to be applied to the 2×2 class of problem follows the lines that Neyman and I have adopted since 1928 in dealing with tests of statistical hypotheses. Let me first recapitulate in broad terms the steps in that approach when applied to a problem where the universe of possible observations can be represented by a

finite set of discrete points. A test of significance may be described as a method of analysis of statistical data which helps us to discriminate between alternative theories or hypotheses. In order to make use of the theory of probability in the sense here understood, a random process must either have been purposely introduced or be assumed to have been present in the collection of data; then the hypothesis very often concerns the values of parameters contained in the probability laws which, in the conceptual sphere, form the mathematical counterpart of the sampling distributions of experience.

15. We proceed by setting up a specific hypothesis to test, H_0 in Neyman's and my terminology, the null hypothesis in R. A. Fisher's. At the same time, in choosing the test, we take into account alternatives to H_0 which we believe possible or at any rate consider it most important to be on the look out for. Thus we wish the test to have maximum discriminating power within a certain class of hypotheses. Three steps in constructing the test may be defined:

Step 1. We must first specify the set of results which could follow on repeated application of the random process used in the collection of the data; this may be termed the experimental probability set.

Step 2. We then divide this set by a system of ordered boundaries or contours such that as we pass across one boundary and proceed to the next, we come to a class of results which makes us more and more inclined, on the information available, to reject the hypothesis tested in favour of alternatives which differ from it by increasing amounts.

Step 3. We then, if possible, associate with each contour level the chance that, if H_0 is true, a result will occur in random sampling lying beyond that level.

This rather crude statement of procedure will be developed in more detail in discussing the problems that arise in connexion with the 2×2 table.

16. *Notes on these points.* (*a*) *Step* 1. This involves the definition of what Neyman and I have termed the sample space, W. The application in three forms of the 2×2 problem is discussed in paragraphs 19, 27 and 46 below.

(*b*) *Step* 2. For a given hypothesis under test there may be a number of ways of deriving a system of contours, and only in certain cases can there be said to be complete agreement on which is the 'best'. Practical expediency will often carry weight in the choice. It is widely accepted that the choice cannot be made without paying regard to the admissible hypotheses alternative to H_0, whether this process is given formal precision or taken as a broad guide. In our first papers (Neyman & Pearson, 1928*a*, *b*) we suggested that the likelihood ratio criterion, λ, was a very useful one to employ in determining a family of contours which would be ordered in relation to our confidence in the hypothesis tested when set against the background of admissible alternatives. Thus Step 2 preceded Step 3. In later papers (Neyman & Pearson, 1933, 1936 and 1938) we started with a fixed value for the chance, ϵ, of Step 3 and determined the associated contour, taking account of what we termed the power of a test with regard to the alternative hypotheses. The family of Step 2 followed on giving decreasing values to ϵ. However, although the mathematical procedure may put Step 3 before 2, we cannot put this into operation before we have decided, under Step 2, on the guiding principle to be used in choosing the contour system. That is why I have numbered the steps in this order.

(*c*) *Step* 3. If this can be accomplished, we have what Neyman and I called control of the '1st kind of error'. In problems where, as below, we are concerned with discrete rather than

continuous probability distributions (e.g. for the binomial, the Poisson, the multinomial and the hypergeometric distributions), this objective cannot always be achieved, and it may be necessary to be satisfied with a knowledge of an upper limit of the chance of rejecting the hypothesis tested when it is true.

(iii) APPLICATION OF THIS APPROACH TO THE ANALYSIS OF DATA CLASSED IN A 2×2 TABLE

17. The frequencies of the data in the table may be defined in the following notation:

Table 1

	Col. 1	Col. 2	Total
Row 1	a	c	m
Row 2	b	d	n
Total	r	s	N

If we follow in turn the steps defined above to determine the method of interpretation of such data, the requirements of the appropriate tests are seen to follow very simply, although mathematical or computational difficulties arise in implementing them. On taking Step 1 we can separate out at once the three types of problem which Barnard has differentiated;* these I shall call Problems I, II and III. They are distinguished by the sample space having 1, 2 and 3 dimensions respectively. From the mathematical point of view it might seem more logical to take them in the reverse order, adding first one and then a second restriction to the 3-dimensioned case of Problem III. For a simple exposition, I think the reverse procedure of building up from I to III is preferable and this has been adopted in the following sections.

(iv) PROBLEM I

18. This may be described as the test of the significance of the difference between two treatments after these have been randomly assigned to a group of $N = m+n$ individuals (Barnard terms it the 2×2 independence trial). To use the terminology of a particular application, we may say that we are observing the presence or absence of 'reaction X'. The first treatment is applied to m and the second to n of the N individuals; as a result a/m and b/n show reaction X.

19. In this case the random process has been applied within the group of N individuals, and its repetition would simply involve other random reassignments of the two treatments among the N. No assumption is made as to how the N individuals were selected from some larger universe. The repetition may be hypothetical, in the sense that it often could not take place, e.g. if reaction X = death. Indeed, repetition under the same essential conditions is frequently impossible in practice. But this correspondence between the frequency of results upon hypothetical repetition and the probability distribution of the counterpart mathematical model forms an accepted part of the process of reasoning whereby (following

* Statisticians had, of course, all been more or less conscious of these differences, but, at any rate in my own case, it was discussion with Mr Barnard which made it easy to see the problem in its full clarity.

the present approach) we use probability theory as a basis for inference. The hypothesis tested is that while some individuals show reaction X and some do not, the result would be the same whichever treatment were applied *as far as these N individuals are concerned*. Thus, on the null hypothesis, there are $r = a+b$ individuals who will react and $s = c+d$ who will not, whatever the assignment of treatments.

20. The chance that a will react in m and $b = r-a$ in n is, therefore, if the hypothesis be true,

$$P_1\{a \mid N, r, m\} = \frac{m!\,n!\,r!\,s!}{a!\,b!\,c!\,d!\,N!}. \tag{1}$$

This expression is proportional to the coefficient of x^a in the hypergeometric series

$$F(\alpha, \beta, \gamma, x) = F(-r, -m, n-r+1, x). \tag{2}$$

Thus, taking $m \geqslant n$, a can assume values of

(i) $0, 1, \ldots, r$ if $r \leqslant n$,

(ii) $r-n, r-n+1, \ldots, r$ if $n < r \leqslant m$,

(iii) $r-n, r-n+1, \ldots, m$ if $r > m$.

For this probability distribution, it is known (K. Pearson (1899) and Kendall (1943, p. 127))

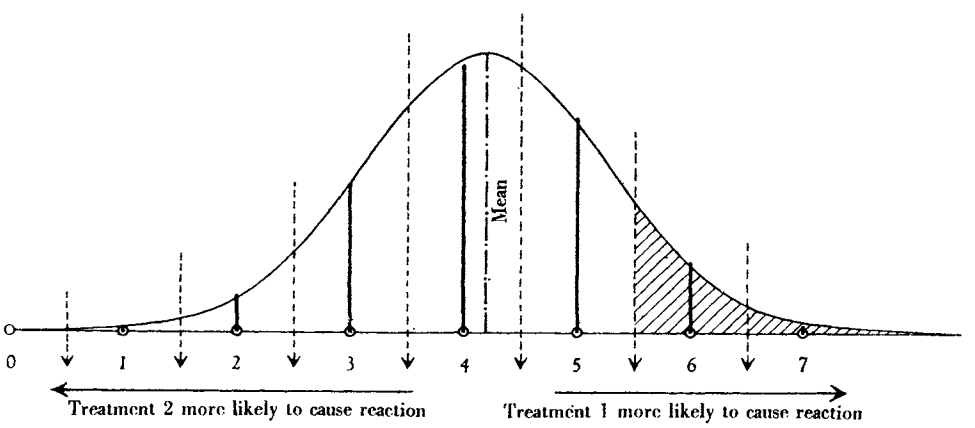

Fig. 2

that

$$\text{Mean } a = \frac{rm}{N}, \tag{3}$$

$$\text{Variance of } a = \sigma_a^2 = \frac{mnrs}{N^2(N-1)}. \tag{4}$$

21. For the particular case

$$N = 20, \quad r = 7, \quad m = 12, \quad n = 8,$$

the terms in the distribution of $P_1\{a \mid 20, 7, 12\}$ are shown as ordinates in Fig. 2 and given in the accompanying Table 2. The experimental probability set consists of the eight alternative values for a, viz. $0, 1, \ldots, 7$ with which the probabilities tabled are associated if H_0 is true. Further

$$\text{Mean } a = \bar{a} = 4{\cdot}2, \quad \sigma_a = 1{\cdot}0721. \tag{5}$$

22. Next consider step 2. The purpose of the investigation is to test the hypothesis that the difference between $a/12$ and $(r-a)/8$ has resulted simply from a random partition of 20 individuals, of whom r will show reaction X in whichever treatment group they are included. The experiment gives $r=7$. The contour levels fall between the 8 points of the set as shown in Fig. 2; the further a lies towards the right, the more inclined we shall be to accept the alternative hypothesis that $a/12>(r-a)/8$ because treatment 1 is more effective than treatment 2. The further a lies to the left, the more we shall incline towards the reverse alternative. To complete Step 3, we have only to calculate the sums of the tail terms of the hypergeometric series, as shown in Table 2 for the special case.

Table 2. *Problem I. Chances for special case* $N = 20$, $r = 7$, $m = 12$, *if* H_0 *is true*

a	Chance of a	Chance of a or less	
		True value	Normal approx.
0	0·0001	0·000	0·000
1	0·0043	0·004	0·006
2	0·0477	0·052	0·056
3	0·1987	0·251	0·257
4	0·3576	—	—
		Chance of a or more	
		True value	Normal approx.
5	0·2861	0·392	0·390
6	0·0954	0·106	0·113
7	0·0102	0·010	0·016

23. Having set up the machinery of the test, we come to the practical question. Beyond which contour levels must a fall before we infer that there is a treatment difference? Not, I think, in the example, if a were 3, 4 or 5; possibly if $a = 6$, more probably if $a = 2$ and almost certainly if $a = 0$, 1 or 7. Were we to fix as critical levels those between $a = 1$ and 2 on the one hand, and between $a = 6$ and 7 on the other, then we should be guided in our decision by the following knowledge: if there were no treatment difference, so that seven out of the twenty individuals would have shown reaction X whichever treatment were applied, then the chance under random assignment of treatments that $a < 2$ or > 6 is only 0·014 or 1 in 70. Had we taken the critical levels between 2 and 3 and between 6 and 7, the corresponding chance would be 0·062 or 1 in 16. This summing up in terms of probability helps towards the balanced decision on the next practical step to be taken, because it helps us to assess the extent of purely chance fluctuations that are possible. It may be assumed that in a matter of importance we should never be content with a single experiment applied to twenty individuals; but the result of applying the statistical test with its answer in terms of the chance of a mistaken conclusion if a certain rule of inference were followed, will help to determine

the lines of further experimental work and the degree of confidence with which we proceed provisionally to adopt a new technique.

24. An experiment falling under this head has the advantage that the random process introduced is under complete control. The analysis will give an answer in probability terms whether the N individuals have been randomly selected from a larger whole or not. But this answer is limited in the sense that it relates only to the N; if we wish to draw conclusions about a wider population or populations, then a random selection of the N or, separately, of both its parts m and n is needed. Thus we come to Problems II and III.

25. *Approximation to the hypergeometric terms.* When dealing with small numbers, the calculation of the tail terms of the series may not be laborious, but it soon becomes so when r is large. An obvious approximation is that obtained by using an integral under the normal curve with the mean and standard deviation of equations (3) and (4) to represent the sum of the hypergeometric terms. As usual when approximating to the sum of the terms for $x = a,\ a+1,\ a+2,\ \ldots$, etc., of a discrete probability distribution by the integral under a continuous curve, we take this integral from the point $x = a - \frac{1}{2}$. Thus Fig. 3 shows the normal curve

$$p(x) = \frac{1}{\sqrt{(2\pi)}\,\sigma_a} \exp\left[-\tfrac{1}{2}(x-\bar{a})^2/\sigma_a^2\right], \tag{6}$$

with $\bar{a}$ and σ_a as in equations (5), and the approximation to the sum of the hypergeometric terms for $a = 6$ and 7 is

$$\int_{5.5}^{\infty} p(x)\,dx,$$

represented by the area marked with cross-hatching. The approximations for different levels are shown in Table 2, and are seen in this case to be quite adequate for the purpose of the test. Further comparisons are made in the Appendix, and it appears that provided m and n are fairly nearly equal, as they are likely to be in most planned experiments of the Problem I type, the normal approximation is surprisingly good. Yates (1934) has suggested a method of further correction.

26. *The correction for continuity.* In the 2×2 table connexion, the improvement obtained by taking the normal integral (i) from $x = a - \frac{1}{2}$ if $a > \bar{a}$ or (ii) from $x = a + \frac{1}{2}$ if $a < \bar{a}$ (so that we are summing for the lower tail), was pointed out by Yates (1934) and has often been termed 'Yates's correction for continuity'. It is, however, the natural adjustment to make on the basis of the Euler-Maclaurin theorem, when approximating to a sum of ordinates by an integral and without wishing to detract from the value of Yates's suggestion in this particular problem, it should be pointed out that the adjustment was used by statisticians well before 1934, when employing a normal or skew curve to give the sum of terms of a binomial or hypergeometric series.*

(v) PROBLEM II

27. This may be described as the test of whether the proportion of individuals bearing a character A is the same in two different populations, from each of which a random sample has been drawn, i.e. the test of the hypothesis that

$$p_1(A) = p_2(A) = p, \tag{7}$$

* The method was in use in the Department of Applied Statistics when I joined the staff in 1921, and may have been current many years before that.

where p is some common but unspecified proportion. Barnard describes this as the case of the 2×2 comparative trial. Here m individuals have been drawn at random from the first population and n from the second, and it is found that a/m and b/n, respectively, bear the character A. The conditions are assumed to be such that if the random procedure of selection were repeated, the appropriate probability distributions for a and b would be given by the terms of binomial expansions. Table 3 shows the observed results.

Table 3

	No. with character A	No. without A	Total
1st sample	a	c	m
2nd sample	b	d	n
Total	r	s	N

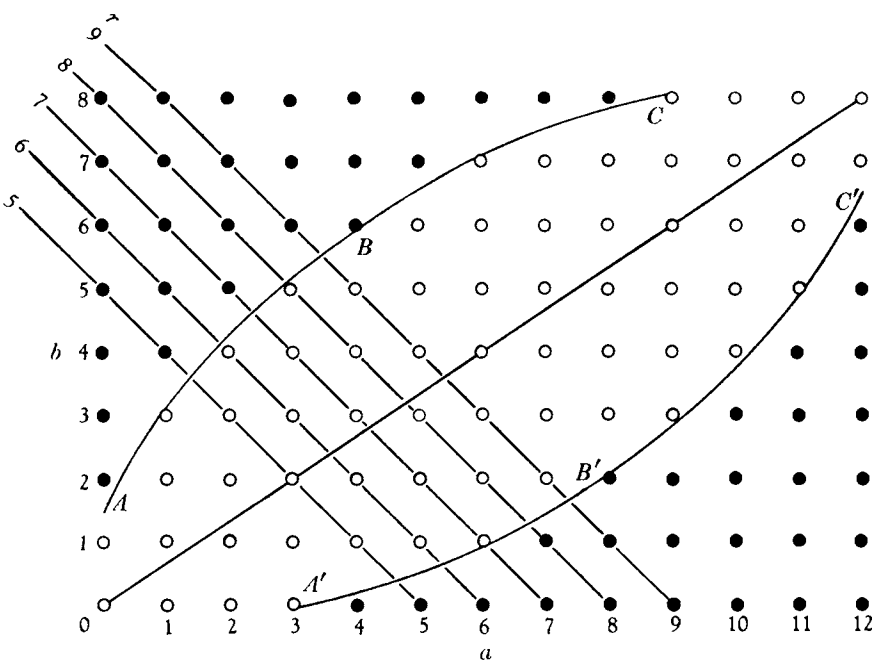

Fig. 3. The curves ABC and $A'B'C'$ represent the significance contours L_ϵ and L'_ϵ, respectively.

In this problem there have been two applications of a random selection process, not one as for Problem I, and the experimental probability set consists of the $(m+1)(n+1)$ alternative values of the doublet (a, b) $(0 \leqslant a \leqslant m, 0 \leqslant b \leqslant n)$ which can be represented in the lattice diagram shown in Fig. 3 for the special case $m = 12$, $n = 8$. It might, of course, be argued that in the hypothetical repetition of the selection process m and n need not remain constant, but this, I think, would introduce an unnecessary complication into the probability set-up.

28. The question before us is whether the result (a, b) is consistent with the hypothesis H_0 defined in equation (7) above, or whether it suggests that either $p_1 > p_2$ or that $p_1 < p_2$. A little reflexion shows that we have no reason to reject H_0 if the point (a, b) lies near the diagonal line on which $a/m = b/n$, but, broadly speaking, are more and more likely to do so the farther the point falls from this line in the direction of the corners $(0, n)$ and $(m, 0)$ of the lattice diagram. This statement requires amplification. In defining the significance contours we may consider the following question: If H_0 is not true, what departures from equality in p_1 and p_2 do we regard it of equal importance to detect? Should the power of the test be roughly the same for constant values, for example, of

$$(a) \quad p_1 - p_2, \qquad (b) \quad p_1/p_2 \qquad \text{or} \qquad (c) \quad \frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2}?$$

The procedure which I have adopted in the sections which follow is frankly one of expediency. I have not considered in detail how to choose a family of significance contours satisfying requirements formulated in advance, but have taken those suggested by the customary large-sample procedure which gives contours of the form ABC, $A'B'C'$ drawn in Fig. 3. These will, I believe, make the power of the test to detect a difference more nearly dependent on the ratio of the odds given by (c) than on either of the expressions (a) or (b). E. B. Wilson (1941) chooses the expression (a). This point, however, needs further investigation. It should be noted that a similar problem, in the case where the sampling distributions follow the Poisson law, was discussed very fully by Przyborowski & Wilenski (1939).

29. Besides involving a 2-dimensional instead of a 1-dimensional experimental probability set, Problem II differs from Problem I in that we need an answer which is independent of the unknown common probability p of the null hypothesis. In Problem I the part of p was played by the fraction r/N given by the data. We are concerned now with what Neyman and I (Neyman & Pearson, 1933) have termed a composite hypothesis, and were it possible would like the contour levels to bound regions which are 'similar to the sample space with regard to the parameter p' (loc. cit. p. 313) (i.e. are independent of p). The following considerations show the lines along which a first attack of the problem can proceed.

30. If H_0 is true and equation (7) holds, then the probability of the observed result may be written*

$$P_2\{a \mid p, m\} \times P_2\{b \mid p, n\} = \frac{m!}{a!\,c!} p^a (1-p)^c \times \frac{n!}{b!\,d!} p^b (1-p)^d \tag{8·1}$$

$$= \frac{N!}{r!\,s!} p^r (1-p)^s \times \frac{m!\,n!\,r!\,s!}{a!\,b!\,c!\,d!\,N!} \tag{8·2}$$

$$= P_2\{r \mid p, N\} \times P_1\{a \mid N, r, m\}. \tag{8·3}$$

Thus the probability of obtaining the doublet (a, b) in sampling from two populations with a common p may be regarded as the product of two terms:

(i) The probability that $a + b = r$ or that the point (a, b) in Fig. 3 falls on a diagonal line on which r = constant. This probability, $P_2\{r \mid p, N\}$, is the $(r+1)$th term in the expansion of the binomial

$$((1-p)+p)^N.$$

(ii) The relative probability, given r, of the observed partition into a and $b = r - a$; this is independent of p and is identical with the expression $P_1\{a \mid N, r, m\}$ of equation (1), i.e. is proportional to a term of the hypergeometric series (2).

* It will be seen that $P_1\{\ \}$ has been used to denote a hypergeometric probability and $P_2\{\ \}$ a binomial probability.

31. If, now, it were possible to draw a boundary line L_ϵ such as ABC shown in Fig. 3, cutting off at the end of each diagonal, r = constant, a group of points $(a, r-a)$ such that

$$\sum_a [P_1\{a \mid N, r, m\}] = \epsilon, \tag{9}$$

where ϵ is a fraction between 0 and 1 chosen at will, then the requirement of Step 3 would be satisfied. For in rejecting H_0 when (a, b) fall beyond this boundary,* the chance of doing so if H_0 were true would be

$$\sum_{r=0}^{N} [P_2\{r \mid p, N\} \times \epsilon] = \epsilon \times \sum_{r=0}^{N} [P_2\{r \mid p, N\}] = \epsilon, \tag{10}$$

i.e. would be independent of the unknown common p of the hypothesis tested. The test would then be analogous to 'Student's' test for the significance of the difference between two means, where we have a system of contour levels L_ϵ each associated with a chance ϵ, independent of the values of any unknown parameters which are irrelevant to the composite hypothesis tested.

32. Unfortunately, this objective cannot be achieved because we are not dealing with continuous probability distributions and $P_1\{a \mid N, r, m\}$ exists only at discrete, integral values of a. If we follow the present line of approach, all that is possible is to take contour or significance levels which cut off from an end of each diagonal, r = constant, a group of points for which

$$\sum_a [P_1\{a \mid N, r, m\}] = \beta_r \leqslant \epsilon. \tag{11}$$

Then, in rejecting H_0 when (a, b) falls beyond such a contour, we know that the chance of doing so, if H_0 is true, will be

$$\sum_{r=0}^{N} [P_2\{r \mid p, N\} \times \beta_r] \leqslant \epsilon. \tag{12}$$

It is clear that the amount by which the probability falls below ϵ will be a function of p, and that in taking Step 3 we are only associating with each significance level L_ϵ an upper limit, ϵ, to the probability of rejecting H_0 when it is true.

33. We have still, of course, to determine the most appropriate system of significance levels and to set out a ready means of finding an upper limit, ϵ, associated with the level on which an observed doublet (a, b) falls.† Mr Barnard has broken new ground in

(i) defining for this Problem II one systematic method of determining a family of levels L_ϵ based on certain clearly defined principles;

(ii) determining the true upper bound to the associated probability ϵ which, in the case of small samples at any rate, may be considerably below that which has hitherto been used.

Since, however, much tabling is needed before his theoretical advance can be followed by a practical working rule available for samples of any sizes, m and n, I think it is worth while describing the cruder handling of the lattice diagram which I had discussed in 1938–9

* There would be a similar series of boundaries, L'_ϵ, below the diagonal $a/m = b/n$, such as $A'B'C'$ of Fig. 3.

† The likelihood ratio λ might be used in determining the family of significance contours, as was suggested in connexion with the general χ^2 problem (Neyman & Pearson, 1928*b*, p. 283). In large samples λ would approximately equal $e^{-\frac{1}{2}u^2}$, where u is given by equation (22) below.

lectures. This involves, perhaps, not much more than a restatement of what may be termed the classical approach to Problem II (see paras. 43 and 44 below), but it does bring out the difference between Problems I and II, which I think important.

34. It may be well to emphasize here that this distinction between the handling of Problems I and II is not universally accepted. Fisher has set out his approach as follows in a paper read before the Royal Statistical Society (1935): 'To the many methods of treatment hitherto suggested for the 2×2 table the concept of ancillary information suggests this new one. Let us blot out the contents of the table, leaving only the marginal frequencies. 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; and therefore recognize that we are concerned only with the relative probabilities of occurrence of the different ways in which the table can be filled in, subject to these marginal frequencies.'

This view has also been supported by Yates (1934). As I understand it, Fisher would refer the observation (a, b) to a linear set (as in my Problem I), however the data have been collected; this attitude follows readily if we discard the requirement that the probability distribution used in the test must be related to the frequency distribution that would be generated by repeated application of the random sampling process employed in the experiment. It will be seen that with Fisher's approach there is a gain in simplicity in handling the analysis; it must remain a matter of opinion whether there is a loss in the relevance of the probability construct to the question at issue. It is, of course, only when handling small samples or in cases where (a, b) lies close to one of the corners $(0, 0)$ or (m, n) of the lattice that this need for choice between probability constructs is thrust upon us.

(vi) Solution of Problem II, using the normal approximation

35. If the samples are large, the calculation of hypergeometric terms becomes laborious and we turn naturally, as in so many other statistical problems, to the approximation using the normal curve. In fact, except when r or s are very small or m and n very different in magnitude, the normal curve with mean and standard deviation given by equations (3) and (4) provides a surprisingly good approximation to the relative probability distribution of a for fixed r, viz. $P_1\{a \mid N, r, m\}$ (see Appendix). Define u_ϵ as the deviate of the standardized normal curve for which

$$\epsilon = \int_{u_\epsilon}^{\infty} \frac{1}{\sqrt{(2\pi)}} e^{-\frac{1}{2}u^2} du \quad (\epsilon \leqslant \tfrac{1}{2}). \tag{13}$$

Then we can draw across the lattice diagram a significance level L_ϵ above and another L'_ϵ below* the diagonal $a/m = b/n$ such that

(i) all points (a, b) for which

$$\frac{(a + \frac{1}{2}) - \bar{a}}{\sigma_a} \leqslant -u_\epsilon \tag{14}$$

lie beyond, i.e. above, L_ϵ;

(ii) and all points (a, b) for which

$$\frac{(a - \frac{1}{2}) - \bar{a}}{\sigma_a} \geqslant u_\epsilon \tag{15}$$

lie beyond, i.e. below, L'_ϵ.

* The words 'above' and 'below' are used in the sense of Figs. 3 and 4.

If we wish to take special action either when a/m is significantly less than b/n or significantly greater, then we shall use both levels L_ϵ and L'_ϵ; if only, however, when $a/m < b/n$, then we use L_ϵ. The corresponding probability levels would be obtained by making ϵ for the second case twice its value for the first. Fig. 4 shows the 247 relative probabilities $P_1\{a \mid N, r, m\}$ for the case $m = 18$, $n = 12$. The unbroken, stepped lines are two contour levels determined in this way. Purely for convenience in drawing, the level with $\epsilon = 0{\cdot}05$ and $u_{0{\cdot}05} = 1{\cdot}6445$ has been put above the diagonal and that with $\epsilon = 0{\cdot}01$ and $u_{0{\cdot}01} = 2{\cdot}3263$ below.

36. If the normal approximation to the hypergeometric series were correct, it would follow that along every diagonal, $r =$ constant, the sum of the relative probabilities for points above L_ϵ would satisfy the inequality (11). Hence the inequality (12) for the complete area of the lattice above L_ϵ would hold, whatever the value of the common p. A similar result would hold for the area below L'_ϵ. Of course, the normal approximation will not hold precisely, particularly when r or s are small, but here we shall generally be on the safe side, in the sense that the hypergeometric distribution is flat-topped with abrupt ends so that the β_r of equation (11) will be considerably less than ϵ, and often zero.

37. It is interesting to examine the results set out in Fig. 4 with the help of the detailed calculations given in Table 4. Columns (2) and (3) give, for constant r, the mean and standard deviation of $P_1\{a \mid 30, r, 18\}$, while columns (4) (for $L_{0{\cdot}05}$) and (8) (for $L'_{0{\cdot}01}$) give the cut-off points defined by the normal approximation, i.e.

$$a_1 = \bar{a} - \tfrac{1}{2} - u_{0{\cdot}05} \times \sigma_a \quad \text{and} \quad a_2 = \bar{a} + \tfrac{1}{2} + u_{0{\cdot}01} \times \sigma_a. \tag{16}$$

The sums of the relative probabilities $P_1\{a \mid 30, r, 18\}$ for $a \leqslant a_1$ and $a \geqslant a_2$ are given in cols. (5) and (9) respectively. Thus, for example, for $r = 7$

$$a_1 = 4{\cdot}2 - 0{\cdot}5 - 1{\cdot}6449 \times 1{\cdot}1543 = 1{\cdot}80,$$

and the sum of the probabilities for $a = 0$ and 1 is

$$0{\cdot}0004 + 0{\cdot}0082 = 0{\cdot}0086.$$

These are the tail sums, termed β_r in equation (11). It is clear from an examination of cols. (5) and (9) that they are all less, and many of them very much less than 0·05 and 0·01. This is inevitable with a discrete distribution containing few terms. The contour levels have been drawn conventionally in Fig. 4 as steps passing through the half-integer points and not through the cut-off points of cols. (4) and (8). Clearly, whichever way they are drawn, they will separate off the same subset of the $(m+1)(n+1)$ points in the lattice diagram

38. The next question is this. If we were to use either of these levels, what in fact would be the chance of the sample doublet (a, b) falling beyond, if the null hypothesis were true? This will depend on the common value of p. The product sums

$$\sum_{r=0}^{N} [P_2\{r \mid p, N\} \times \beta_r] = \sum_{r=0}^{N} \left[\frac{N!}{r!\,s!} p^r (1-p)^s \times \beta_r\right] \tag{17}$$

obtained by multiplying the expressions in cols. (5) and (9) of Table 4 by the appropriate binomial terms are shown for a variety of values of p in Table 5, cols. (2) and (3). It is clear at once how far on the safe side we are in saying that these chances are $\leqslant 0{\cdot}05$ and 0·01 respectively. Similar calculations were carried out for a second example, taking $m = n = 10$,

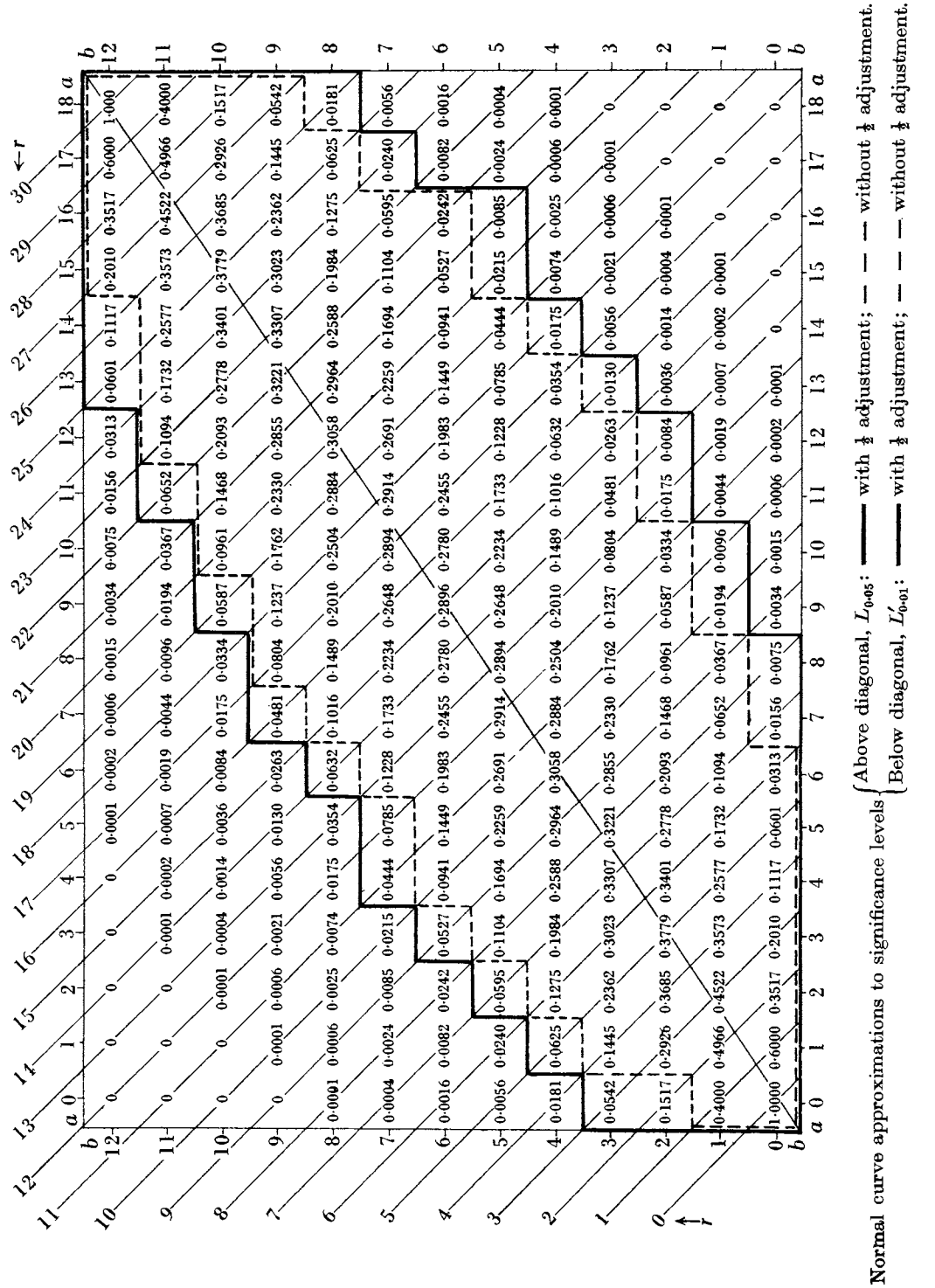

Fig. 4. Hypergeometric probabilities in lattice diagram for $m = 18$, $n = 12$.

Table 4. *Significance levels for case* $m = 18$, $n = 12$

r	$\bar{a}$	σ_a	Details for L_ϵ: $\epsilon = 0{\cdot}05$, $u_{0{\cdot}05} = 1{\cdot}6449$				Details for L'_ϵ: $\epsilon = 0{\cdot}01$, $u_{0{\cdot}01} = 2{\cdot}3263$				r
			Method 1		Method 2		Method 1		Method 2		
			Cut-off $\bar{a} - \frac{1}{2} - u_\epsilon \sigma_a$	Sum of terms beyond cut-off	Cut-off $\bar{a} - u_\epsilon \sigma_a$	Sum of terms beyond cut-off	Cut-off $\bar{a} + \frac{1}{2} + u_\epsilon \sigma_a$	Sum of terms beyond cut-off	Cut-off $\bar{a} + u_\epsilon \sigma_a$	Sum of terms beyond cut-off	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
0	0	0	−0·50	0	0	0	0·50	0	0	0	0
1	0·6	0·4899	−0·71	0	−0·21	0	2·24	0	1·74	0	1
2	1·2	0·6808	−0·42	0	0·08	0·1517	3·28	0	2·78	0	2
3	1·8	0·8187	−0·05	0	0·45	0·0542	4·20	0	3·70	0	3
4	2·4	0·9277	0·37	0·0181	0·87	0·0181	5·06	0	4·56	0	4
5	3·0	1·0171	0·83	0·0056	1·33	0·0681	5·87	0	5·37	0	5
6	3·6	1·0917	1·30	0·0256	1·80	0·0256	6·64	0	6·14	0	6
7	4·2	1·1543	1·80	0·0086	2·30	0·0681	7·39	0	6·89	0·0156	7
8	4·8	1·2069	2·31	0·0267	2·81	0·0267	8·11	0	7·61	0·0075	8
9	5·4	1·2507	2·84	0·0091	3·34	0·0618	8·81	0·0034	8·31	0·0034	9
10	6·0	1·2865	3·38	0·0241	3·88	0·0241	9·49	0·0015	8·99	0·0209	10
11	6·6	1·3152	3·94	0·0080	4·44	0·0524	10·16	0·0006	9·66	0·0102	11
12	7·2	1·3370	4·50	0·0197	5·00+	0·0982	10·81	0·0046	10·31	0·0046	12
13	7·8	1·3524	5·08	0·0414	5·58	0·0414	11·45	0·0020	10·95	0·0195	13
14	8·4	1·3615	5·66	0·0145	6·16	0·0777	12·07	0·0007	11·57	0·0091	14
15	9·0	1·3646	6·26	0·0301	6·76	0·0301	12·67	0·0038	12·17	0·0038	15
16	9·6	1·3615	6·86	0·0091	7·36	0·0572	13·27	0·0015	12·77	0·0145	16
17	10·2	1·3524	7·48	0·0195	7·98	0·0195	13·85	0·0060	13·35	0·0060	17
18	10·8	1·3370	8·10	0·0380	8·60	0·0380	14·41	0·0022	13·91	0·0197	18
19	11·4	1·3152	8·74	0·0102	9·24	0·0689	14·96	0·0080	14·46	0·0080	19
20	12·0	1·2865	9·38	0·0209	9·88	0·0209	15·49	0·0026	14·99	0·0241	20
21	12·6	1·2507	10·04	0·0401	10·54	0·0401	16·01	0·0006	15·51	0·0091	21
22	13·2	1·2069	10·71	0·0075	11·21	0·0727	16·51	0·0025	16·01	0·0025	22
23	13·8	1·1543	11·40	0·0156	11·90	0·0156	16·99	0·0086	16·49	0·0086	23
24	14·4	1·0917	12·10	0·0313	12·60	0·0313	17·44	0·0016	16·94	0·0256	24
25	15·0	1·0171	12·63	0	13·13	0·0601	17·87	0·0056	17·37	0·0056	25
26	15·6	0·9277	13·57	0	14·07	0·1117	18·26	0	17·76	0·0181	26
27	16·2	0·8187	14·35	0	14·85	0	18·60	0	18·10	0	27
28	16·8	0·6808	15·18	0	15·68	0	18·88	0	18·38	0	28
29	17·4	0·4899	16·09	0	16·59	0	19·04	0	18·54	0	29
30	18·0	0	17·50	0	18·00	0	18·50	0	18·00	0	30

and the results are shown in Table 5, cols. (6) and (7). In this case, the actual chances of (a, b) falling on or beyond the significance levels are even further below the nominal limits of 0·05 and 0·01. In fact, it becomes clear that in the case of small samples, at any rate, this method of introducing the normal approximation gives such an overestimate of the true chances of falling beyond a contour as to be almost valueless.

Table 5. *Showing the difference between nominal and actual significance levels*

p (if H_0 true)	1st example: $m = 18$, $n = 12$				2nd example: $m = 10 = n$				p (if H_0 true)
	Method 1		Method 2		Method 1		Method 2		
	True chance of falling on or beyond		True chance of falling on or beyond		True chance of falling on or beyond		True chance of falling on or beyond		
	$L_{0\cdot05}$	$L'_{0\cdot01}$	$L_{0\cdot05}$	$L'_{0\cdot01}$	$L_{0\cdot05}$	$L'_{0\cdot01}$	$L_{0\cdot05}$	$L'_{0\cdot01}$	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0·05	0·0010	0·0000	0·0478	0·0000	0·0000	0·0000	0·0069	0·0000	0·05
0·1	0·0054	0·0000	0·0602	0·0003	0·0005	0·0000	0·0251	0·0005	0·1
0·2	0·0141	0·0003	0·0483	0·0043	0·0037	0·0007	0·0455	0·0037	0·2
0·3	0·0174	0·0012	0·0490	0·0091	0·0058	0·0014	0·0495	0·0058	0·3
0·4	0·0204	0·0023	0·0542	0·0108	0·0062	0·0017	0·0546	0·0062	0·4
0·5	0·0219	0·0028	0·0498	0·0109	0·0062	0·0015	0·0572	0·0062	0·5
0·6	0·0221	0·0035	0·0437	0·0119	Repeat as for $1-p$				0·6
0·7	0·0204	0·0037	0·0431	0·0120					0·7
0·8	0·0126	0·0031	0·0459	0·0113					0·8
0·9	0·0019	0·0009	0·0282	0·0052					0·9
0·95	0·0001	0·0001	0·0058	0·0010					0·95

39. Before considering a second method, it will be useful to recapitulate certain characteristics of what I have termed Method 1. It provides for any nominal value of ϵ one systematic procedure of defining a critical boundary or significance level cutting off a region from the lattice diagram. Neither the subgroup of points cut off, nor the sum of the probabilities associated with them for a given p, will alter continuously with ϵ; they will change by discrete steps as the cut-off point, defined in para. 37, passes through a point (a, b). While we shall sometimes want to know whether the observed (a, b) falls beyond a level L_ϵ specified in advance, more often we shall ask what is the level on which (a, b) falls. This, using Method 1, we find by calculating

$$u = \frac{\bar{a} - (a + \frac{1}{2})}{\sigma_a} \quad \text{if} \quad a < \bar{a} \qquad \text{or} \qquad u = \frac{a - \frac{1}{2} - \bar{a}}{\sigma_a} \quad \text{if} \quad a > \bar{a}, \tag{18}$$

and finding ϵ from the normal integral of equation (13). In this way the nominal chance ϵ will be a little nearer the true upper limit than the figures in Table 5 suggest,* but not enough to modify the criticism expressed above.

* It will be seen from Table 4 that no point (a, b) gives a β_r in cols. (5) and (9) of exactly 0·05 or 0·01, respectively, so that no points actually lie on $L_{0\cdot05}$ or $L_{0\cdot01}$.

40. *Method* 2.- The introduction of the correction of $\frac{1}{2}$ for continuity is certainly appropriate in using the normal approximation to the hypergeometric series in Problem I, but I think it is not helpful in Problem II where we are concerned with a 2-dimensional experimental probability set. If instead of obtaining significance levels L_ϵ and L'_ϵ as in paras. 35–37, we obtain them from inequalities similar to (14) and (15) but with the correction of $\frac{1}{2}$ omitted, then there are several points to be noted:

(*a*) For the significance level L_ϵ, the expression

$$\beta_r = \sum_a [P_1\{a \mid N, r, m\}], \tag{19}$$

where the summation is for values of a on the diagonal, r = constant, for which

$$a \leqslant a_1 = \bar{a} - u_\epsilon \times \sigma_a \tag{20}$$

will be sometimes less and sometimes greater than ϵ. Hence, in the balance, it seems likely that the chance of the point (a, b) lying beyond L_ϵ or

$$\sum_{r=0}^{N} \left[\frac{N!}{r!\,s!} p^r (1-p)^s \times \beta_r\right] \tag{21}$$

will lie closer to ϵ than when the $\frac{1}{2}$ correction is used. The position will be the same for L'_ϵ.

(*b*) In drawing repeated samples of m and n from two populations in which there is a common chance, p, of an individual possessing character A, the ratio

$$u = \frac{a - \bar{a}}{\sigma_a} = \frac{a - rm/N}{\sqrt{\dfrac{mnrs}{N^2(N-1)}}} \tag{22}$$

has, whatever be p, (i) an expectation of zero, (ii) a unit standard deviation.* The shape of the distribution will, of course, depend on p, but, *faut de mieux*, we may not in the long run do too badly by assuming it to be normal. It is, of course, the weighted combination of a number of hypergeometric series whose shape depends on r.

41. Consider the result of applying this Method 2 to the case $m = 18$, $n = 12$ already discussed. The procedure for determining the 0·05 and 0·01 significance levels will be exactly as under Method 1, except that the continuity correction of $\frac{1}{2}$ is omitted. The resulting levels are shown as dashed, stepped lines in Fig. 4.† They fall, on the whole, inside the significance levels obtained by Method 1. Now turn to Table 4, where cols. (6) and (10) show the cut-off points a half unit further in towards the diagonal $a/m = b/n$. Cols. (7) and (11) give the values of β_r; some of these are considerably above the nominal values of $\epsilon = 0{\cdot}05$ and 0·01, others are still well below. But from the approach to Problem II that has been adopted, this is immaterial since the experimental probability set is the 2-dimensioned one of the lattice diagram and is not restricted to the diagonal r = constant on which the observed point (a, b) may happen to lie. What we are concerned with is the summed chance given by expression (21) and the value of this is given for eleven values of p in cols. (4) and (5) of Table 5. It will be seen that this true chance does sometimes exceed the nominal values of 0·05 and 0·01,

* Provided cases where r or s are zero, making the expression (22) indeterminate with $u = 0/0$, are excluded. Mr Barnard has pointed out that one way of avoiding this exclusion would be to lay down that, when $u = 0/0$, we assign to the ratio a value chosen at random from a population (say normal) with zero mean and unit variance.

† Again, for convenience the 5 % level is drawn above and the 1 % level below the diagonal.

but never by very much. Again, for the second example with $m = 10 = n$ (Table 5, cols. (8) and (9)) the true chance, while it sometimes exceeds the nominal value, is always considerably nearer it than using the significance levels of Method 1.

42. It is clear that no final conclusions can be based on two numerical examples, but it seems that the test of the null hypothesis in Problem II should be carried out as follows:

(*a*) When m, n, r or s are small, with the help of tables prepared on Barnard's lines, based on an ordered classification of the points in the lattice diagram, and giving the true upper bound of the chance that a point (a, b) falls on or beyond the level on which the observed result lies. The particular basis of his classification may, of course, be modified.

(*b*) When m, n, r and s are large, by assuming that the u of equation (22) is a normal deviate with unit standard deviation.

(vii) THE CLASSICAL APPROACH TO PROBLEM II

43. It has recently become customary to regard the test of significance applied to data given in a 2×2 table as the limiting case of a χ^2 test with one degree of freedom. But Problem II was originally answered in somewhat different terms. It was noted that if

$$p_1(A) = p_2(A) = p, \tag{23}$$

then the fractions a/m and b/n would both have expectations of p and variances of $p(1-p)/m$ and $p(1-p)/n$, respectively. Hence, if the null hypothesis were true, the difference

$$d = \frac{a}{m} - \frac{b}{n} \tag{24}$$

would have

$$\left.\begin{aligned} &\text{mean } d = 0 \\ &\sigma_d = \sqrt{\left[p(1-p)\left(\frac{1}{m}+\frac{1}{n}\right)\right]} \end{aligned}\right\}. \tag{25}$$

In large samples, therefore, it might be expected that

$$\frac{d}{\sigma_d} = \frac{a/m - b/n}{\sqrt{[p(1-p)\,(1/m+1/n)]}} \tag{26}$$

would be approximately normally distributed. Since by the nature of the problem the common value of p was unknown, an estimate was made from the sample, namely,

$$\hat{p} = \frac{a+b}{m+n} = \frac{r}{N}. \tag{27}$$

Substituting this into equation (26), we have

$$\frac{d}{s_d} = \frac{a/m - b/n}{\sqrt{[(r/N)(1-r/N)\,(1/m+1/n)]}} \tag{28·1}$$

$$= \frac{a - rm/N}{\sqrt{\left(\frac{mnrs}{N^3}\right)}}. \tag{28·2}$$

44. The form (28·2) is easily derived from (28·1), if we remember that $b = r-a$, $s = N-r$ and $m+n = N$.* It is seen that the ratio d/s_d is identical with the ratio u of equation (22), except for a factor $\sqrt{[(N-1)/N]}$ which is unimportant in large samples. Thus the classical test is practically identical with that suggested in paras. 40–42 above, though the two tests are differently derived.

* A third alternative form is, of course, $(ad-bc)\sqrt{N}/\sqrt{(mnrs)}$.

(viii) Problem III

45. This may be described as the test for the independence of two characters A and B. It is supposed that the probability that an individual selected at random will possess character A is $p(A)$ and that he will not possess it is $p(\bar{A}) = 1-p(A)$. The corresponding probabilities for character B are $p(B)$ and $p(\bar{B}) = 1-p(B)$. Four alternative combinations of the characters may occur, which may be denoted by AB, $A\bar{B}$, $\bar{A}B$ and $\bar{A}\bar{B}$. The various probabilities are set out in Table 6A. If the null hypothesis, H_0, specifying the independence of A and B is true, then

$$p(AB) = p(A)\times p(B), \quad p(A\bar{B}) = p(A)p(\bar{B}), \quad \text{etc.} \tag{29}$$

To test the hypothesis, we have a random sample of N observations with frequencies of occurrence of the combinations AB, $A\bar{B}$, etc., which may be classified in the 2×2 scheme of Table 6B. The sampling conditions are such that the probabilities of Table 6A are the same for all individuals selected, or, in conventional terms, the sample is drawn from an infinite population. Barnard calls this problem that of the double dichotomy.

Table 6A. *Probabilities*

	A	$\bar{A}$	Total
B	$p(AB)$	$p(\bar{A}B)$	$p(B)$
$\bar{B}$	$p(A\bar{B})$	$p(\bar{A}\bar{B})$	$p(\bar{B})$
Total	$p(A)$	$p(\bar{A})$	1

Table 6B. *Sample data*

	A	$\bar{A}$	Total
B	a	c	m
$\bar{B}$	b	d	n
Total	r	s	N

46. In Problem III there is only one application of a random process, the selection of N individuals, each one of which must fall into one or other of four alternative categories. If the random process were repeated and another sample of N drawn, not only are the frequencies a, b, c and d free to vary, but also *both* marginal totals, i.e. m may change as well as r. The experimental probability set will therefore contain results (a, b, c, d) restricted by the conditions (i) that none of the frequencies can be negative and (ii) that

$$a+b+c+d = N. \tag{30}$$

Geometrically, as Barnard points out, the set can be represented in 3 dimensions by points at unit intervals within a tetrahedron obtained by placing on top of one another the series of 2-dimensioned lattices of dimensions

$$0\times n, \quad 1\times(n-1), \quad 2\times(n-2), \quad \ldots, \quad (m-1)\times 1, \quad m\times 0. \tag{31}$$

47. We are again testing a composite hypothesis and should like to determine a family of critical surfaces to be used as significance levels, dividing the points within the tetrahedron in such a way that the chance of the sample point (a, b, c, d)* lying outside a given surface L_ϵ is equal to ϵ, whatever the values of the unknown probabilities $p(A)$ and $p(B)$. But again, as in Problem II, owing to the discontinuity in the set of points, there are no 'similar

* In view of the condition (30), the point can be defined by three co-ordinates, e.g. as (a, b, c), (a, b, m) or (a, r, m). In view of the form of equation (32), the last system of co-ordinates will be used.

regions'. We note that if H_0 is true, the probability of the observed result is a term of the multinomial expansion, viz.

$$\frac{N!}{a!\,b!\,c!\,d!}p(AB)^a p(A\bar{B})^b p(\bar{A}B)^c p(\bar{A}\bar{B})^d$$

$$= \frac{N!}{a!\,b!\,c!\,d!}p(A)^{a+b} p(B)^{a+c} p(\bar{A})^{c+d} p(\bar{B})^{b+d}$$

$$= \frac{N!}{m!\,n!}p(B)^m (1-p(B))^n \times \frac{N!}{r!\,s!}p(A)^r (1-p(A))^s \times \frac{m!\,n!\,r!\,s!}{a!\,b!\,c!\,d!\,N!}$$

$$= P_2\{m \mid p(B), N\} \times P_2\{r \mid p(A), N\} \times P_1\{a \mid N, r, m\}. \tag{32}$$

Here, the notation of para. 30 has been repeated.

48. Thus the probability of obtaining a sample represented by the triplet (a, r, m) may be regarded, if the characters A and B are independent, as the product of three terms:

(i) The probability of drawing m individuals with character B in a random sample of N, i.e. the probability that (a, r, m) falls in a horizontal section of the tetrahedron on which m = constant. This is the $(m+1)$th term in the expansion of the binomial

$$\{(1-p(B))+p(B)\}^N.$$

(ii) The probability of drawing r individuals with character A in a random sample of N, i.e. the probability that (a, r, m) falls on the vertical section of the tetrahedron on which r = constant. This is the $(r+1)$th term in the expansion of

$$\{(1-p(A))+p(A)\}^N.$$

(iii) The probability, given m and r, of the observed partition within the 2×2 table. This term represents the relative probability associated with the points lying along a straight line m = constant, r = constant; it is, of course, the same expression as has arisen in Problems I and II and is proportional to a term in the hypergeometric series $F(-r, -m, n-r+1, 1)$.

49. We are faced with a situation similar to that met under Problem II. Were it possible to cut off from each line on which m = constant, r = constant, a group of points such that

$$\sum_a [P_1\{a \mid N, r, m\}] = \epsilon, \tag{33}$$

then the subset of points within the tetrahedron composed of the sum of these groups for all possible combinations of m and r would have the property required of a 'critical region' in a significance test: i.e. the chance that the point (a, r, m) is included in the region, if H_0 is true, would be ϵ whatever values the irrelevant probabilities $p(A)$ and $p(B)$ assumed. However, (33) cannot be satisfied in general, and all that is possible is to define a family of significance contours such that the chance of a sample point falling beyond any one of them, say L_ϵ, is $\leqslant \epsilon$. By using the normal approximation to the sum of the hypergeometric tail-terms with the correction for continuity as described in paras. 35–39 for Problem II, we shall be very much on the safe side, i.e. the formal level of ϵ is likely to be much above the true chance of falling beyond the level, whatever be $p(A)$ or $p(B)$. The presence of the two binomial terms in equation (32) instead of the single term in equation (8·3), makes it likely that the overestimation of ϵ will be greater in Problem III than in II. It is to be expected, therefore, that any any rate when neither m, n, r or s are too small, the better approximation will be obtained by referring the u of equation (22) to the normal probability scale.

50. The handling of Problem III is discussed briefly by Barnard on p. 136 above. There is clearly room for further investigation. The general nature of the approximation

involved is of course that which arises in every χ^2 test for goodness of fit or for independence in an $h \times k$ table, where we replace a distribution consisting of a finite set of probabilities at discrete points in multiple space by a continuous distribution for which integration outside ellipsoidal contours is straightforward.

(ix) General comment

51. The duties of the statistician lie at many levels. He may be required merely to apply an established technique of analysis to an assembly of numerical data and this application may result in a statement, based on probability theory, of a 'level of significance' or a 'confidence interval', which will be used by others. Or he may be called on to share in planning the investigation or experiment which is to provide the data and then to draw conclusions from their analysis which will lead to further action. In this final role he needs to bring into play faculties which are no monopoly of his calling, the qualities of sound judgement which are the characteristics of a well trained, scientific mind. In the weighing of evidence, the result of the statistical analysis, expressed in one or more conventional probability figures, is only one factor in the summing up; as important, may be, is the question of whether the mathematical model is a fair counterpart to the happenings in the observational field. In addition, there will often be much information coming from outside the range of the immediate investigation, yet hardly expressible in numerical terms, which must influence decision.

52. It is perhaps hard experience gained in certain fields of war-time research, where decisions had to be reached on statistical data far less ample than could be wished, which has forced my own attention to this question: What weight do we actually give to the precise value of a probability measure when reaching decisions of first importance? One subject for examination falling under this inquiry is clearly the logical basis of the reasoning process by which judgement is influenced as a result of the application of a test of significance. This was the theme on which this paper opened. The approach illustrated in the pages which followed is a personal one and is set down, with no claim to be the best, in order to provoke thought and discussion. There appears no short route to a right answer in this matter; each individual who hopes to use his own judgement to the full in drawing conclusions from the statistical analysis of sampling data, must decide for himself what he requires of probability theory.

53. In the approach which I have followed and illustrated on the analysis of data classed in a 2×2 table, the appropriate probability set-up is defined by the nature of the random process actually used in the collection of the data. Consideration of this point forms the initial step in the determination of the appropriate test. On this score, what I have termed Problems I, II and III are differentiated. The difference is fundamental and lies at the bottom of the dilemma to which the Barnard-Fisher correspondence in *Nature* drew attention. It can be illustrated on the following data, given in Table 7, where I shall suppose that the effect we are interested in is that making a significantly greater than b.

54. If (a) the results have been obtained by random assignment of Treatment 1 to eighteen out of thirty individuals and Treatment 2 to the remaining twelve, and

(b) we merely ask whether the results are consistent with the hypothesis that the treatments are equivalent as far as these thirty individuals are concerned, so that the difference between the proportions 15/18 and 5/12 may reasonably be ascribed to a chance fluctuation,

(*c*) we are then concerned with Problem I, i.e. simply with the probabilities associated with the points $(a, 20-a)$ on the diagonal $r = 20$ of Fig. 4. The chance of getting $a \geqslant 15$, if the null hypothesis is true, is 0·0241,* or, using a common phrase, we can speak of the result being significant at the 2·5 % level.

55. On the other hand, if a sample of 18 has been drawn randomly from one population and a sample of 12 independently from a second and we wish to test whether $p_1(A) = p_2(A)$, then it seems to be an artificial procedure to restrict the experimental probability set to the 11 points on the line $r = 20$, i.e. to the values of a: 8, 9, ..., 18. A repetition of the double sampling process could give us a result (a, b) falling at any of the $19 \times 13 = 247$ points in the lattice diagram of Fig. 4. There will be a number of ways of defining a family of significance levels for this 2-dimensioned set; if we adopt that discussed in paras. 40–41, which

Table 7

For problem I	For problem II	Frequency of results		Total
		A	$\bar{A}$	
1st treatment	Sample from 1st population	$a = 15$	$c = 3$	$m = 18$
2nd treatment	Sample from 2nd population	$b = 5$	$d = 7$	$n = 12$
	Total	$r = 20$	$s = 10$	$N = 30$

gives as two of its members the dotted, stepped lines shown in Fig. 4, we can say that the chance of a result falling beyond the lower line is certainly less than 0·015.† The observed point, with $a = 15$, $b = 5$ falls beyond the line, so that the result is undoubtedly 'significant at the 1·5 % level'.

56. These two probabilities, 2·5 and 1·5 %, are not the same, but there is no inconsistency in their difference. The character of the two investigations is different and to treat Problem II as though it were Problem I seems to call for a probability set-up which is unnecessarily artificial, when a simpler one is available. Admittedly by getting what seems to me a closer relation between the probability set-up and the experimental procedure, we have sacrificed some simplicity in handling the 2×2 table. But this is only the case when dealing with small numbers. For large numbers the methods of handling Problems I, II and III become, practically, identical.

57. Consider again the heavy shell problem described in para. 7 above. If we are to introduce probability theory, it seems to me that we should regard the problem as one in which we have a sample of $m = 12$ from the possible output of shell made to one design or by one firm and of $n = 8$ from the possible output of a second. This sampling may be hypothetical in that these may be 'pilot' shell, the first off production; nevertheless, this construct is

* For the normal curve approximation, using the correction for continuity, we find

$$u = (15 - \tfrac{1}{2} - 12{\cdot}0)/1{\cdot}2865 = 1{\cdot}943.$$

The proportionate area under the normal curve beyond this deviation is 0·026.

† Table 5, col. (5) shows the largest value of this chance to be 0·0120 for $p = 0{\cdot}3$. This figure cannot be much exceeded for other p's though I have not determined the precise maximum. I give 0·015 as a safe-side limit.

clearly less artificial than one in which, on the null hypothesis, we regard the experiment as though it were made on twenty shells, to twelve of which has been randomly assigned the label 'Made by firm X' and to the other eight, 'Made by firm Y'.

58. It is clear that in the heavy shell problem there may be many reasons to doubt whether the rounds fired can be regarded as a random sample from future output. That is why I have emphasized that the exploration which the statistician makes in private will not necessarily be presented in figures at the conference table. In this example, the proportions of successful perforations were 2/12 and 5/8; these put us on the line, $r = 7$, of the lattice diagram for which the hypergeometric probabilities were shown in Fig. 2. The sum of the terms with $a \leqslant 2$ is 5·2 % (normal approximation, using the $\frac{1}{2}$-correction, 5·6 %). This is the chance of getting as great or a greater positive difference, $b-a$, if H_0 were true, treating the case as Problem I. Barnard's method has not yet been extended to cover this case, but if we were to use the large sample method for handling Problem II, described in my paras. 40–41, we should find from equation (22) that

$$u = (2-4{\cdot}2)/1{\cdot}072 = -2{\cdot}05,$$

which puts (a, b) outside the upper 2·5 % level.

59. Were the action taken to be decided automatically by the side of the 5 % level on which the observation point fell, it is clear that the method of analysis used would here be of vital importance. But no responsible statistician, faced with an investigation of this character, would follow an automatic probability rule. The result of either approach would raise considerable doubts as to whether the performance of the first type of shell was as good as that of the second, but without the whole background of the investigation it is impossible to say what the statistician's recommendation as to further action would be.

60. In the example of the proof of anti-tank shot discussed in para. 6, the chance of perforation, p, while varying from plate to plate and batch to batch, will almost certainly not range through the whole interval 0–1. The striking-velocity of the shot would also probably be adjusted so that for average proof-plate and batches, p was near $\frac{1}{2}$. Then the discriminating level (or levels*) set across the 13×13 lattice diagram would be fixed paying regard to the likely variation in p; thus a fairly close upper limit could be calculated to the true probability of (a, b) falling beyond the level if the fresh batch were of the same quality as the standard. This is the upper limit of the risk of segregating the batch wrongly.

61. Precisely similar problems arise for consideration in even more difficult form in the analysis of data arranged in a $h \times k$ table, where h or k or both are > 2. It has become common practice to speak of the solution of this problem in terms of 'fixed marginal totals', but it may be questioned whether the restriction in the experimental probability set implied is generally appropriate. The frequencies in a $h \times k$ table may have been obtained by many different sampling procedures for, as in the 2×2 problem, a single form of tabular presentation will follow from a variety of types of investigation. For most of these, a repetition of the random process of selection would give results with either one or both sets of marginal totals changed.

62. For convenience in solution we may, of course, start by considering the distribution of our test criterion, on the null hypothesis, within the sub-set of results for which the margins

* It is possible that two levels might be taken with the associated proof rules: (i) if (a, b) falls beyond the outer one, reject the batch; (ii) if between outer and inner, fire further rounds; (iii) if within the inner level, accept the batch.

are fixed. If this distribution were the same whatever these fixed values, then the overall, distribution for unrestricted sampling would be the same as that for variation subject to fixed margins. Thus, mathematically, the solution of the partial problem would be a step in the solution of the complete one. But when applying χ^2 analysis to an $h \times k$ table, this result is only true as a large-sample approximation.

63. If we use the mathematical model which it is suggested gives the most direct aid in reasoning from the observations, i.e. that which regards the experimental probability set as generated by a repetition of the random process of selection used in collecting the data, then in the majority of cases we cannot regard the marginal totals as fixed. Thus a rigorous treatment would lead, as in the case of the 2×2 table, to a differentiation into a number of solutions. It is to be hoped, however,* unless the numbers in the margins are very small, that the χ^2 approximation with its appropriate degrees of freedom† will give results which are not misleading. This approximation leads, of course, in the 2×2 table to the reference of the ratio u of equation (22) to the normal probability scale. Some aspects of the approximation in this more general case were discussed by Yates (1934, pp. 233–35).

64. In closing I should like again to acknowledge my indebtedness to Mr G. A. Barnard. Having had the good fortune to discuss these problems with him and see drafts of his work over a period of 2 or 3 years it is difficult to say how many of his ideas have been built unconsciously into my own earlier approach. But I am especially aware of the clarification which his emphasis on the distinction between Problems I, II and III brought to my survey. I am also very grateful to Mr M. G. Kendall, Dr R. C. Geary and Dr B. L. Welch for a number of helpful criticisms, and to Mrs Maxine Merrington for her extensive computing work, which has alone made possible the various numerical illustrations that I have given.

* From the point of view both of the exponents of the fixed marginal and unrestricted marginal approach.

† The statement that, for example, in applying the test of independence of two characters to an $h \times k$ table, the degrees of freedom are $(h-1) \times (k-1)$, does not of course mean that sampling is restricted by fixed marginal totals. All that is implied is that approximately the overall distribution of the χ^2 function of the observations used, is the same as that for sampling within the restricted sub-set; this is because the distribution within each sub-set is approximately independent of the particular marginal totals which define it.

REFERENCES

BARNARD, G. A. (1945*a*). *Nature, Lond.*, **156**, 177.
BARNARD, G. A. (1945*b*). *Nature, Lond.*, **156**, 783.
FISHER, R. A. (1935). *J. Roy. Statist. Soc.* **98**, 39.
FISHER, R. A. (1941). *Science*, **94**, 210.
FISHER, R. A. (1945*a*). *Nature, Lond.*, **156**, 388.
FISHER, R. A. (1945*b*). *Sankhyā*, **7**, 130.
KENDALL, M. G. (1943). *The Advanced Theory of Statistics*, **1**. London: Charles Griffin and Co. Ltd.
NEYMAN, J. & PEARSON, E. S. (1928*a*). *Biometrika*, **20A**, 195.
NEYMAN, J. & PEARSON, E. S. (1928*b*). *Biometrika*, **20A**, 263.
NEYMAN, J. & PEARSON, E. S. (1933). *Philos. Trans.* A, **231**, 289.
NEYMAN, J. & PEARSON, E. S. (1936). *Statist. Res. Mem.* **1**, 113.
NEYMAN, J. & PEARSON, E. S. (1938). *Statist. Res. Mem.* **2**, 25.
PEARSON, K. (1899). *Phil. Mag.* **47**, 236.
PRZYBOROWSKI, J. & WILENSKI, H. (1939). *Biometrika*, **13**, 313.
WILSON, E. B. (1941). *Science*, **93**, 557.
WILSON, E. B. (1942). *Proc. Nat. Acad. Sci., Wash.*, **28**, 94.
YATES, F. (1934). *J. Roy. Statist. Soc.* Suppl. **1**, 217.

APPENDIX

THE NORMAL CURVE APPROXIMATION IN PROBLEM I

1. The following Tables 8 and 9 (A), (B) and (C) show the order of accuracy which results from using the normal curve integral as an approximation to the tail sums in the series

$$P_1\{a \mid N, r, m\} = \frac{m!\, n!\, r!\, s!}{a!\, b!\, c!\, d!\, N!} \tag{34}$$

the terms of which are proportional to those in the hypergeometric series

$$F(-r, -m, N-m-r+1, 1).$$

Here a is a variable which can assume the range of positive, integral values indicated under (i), (ii) and (iii) in para. 20 above, while N, r and m are fixed. The relation between these quantities and b, c, d, n and s is given in Table 1, para. 17. The method of approximation, using the '$\frac{1}{2}$' correction for continuity, has been discussed in para. 25.

2. Table 8 takes the case of an equal partition, $m = n = \frac{1}{2}N$, and shows the sum of the terms in the expression (34) for which $a \geqslant a_1$ which is also the sum of terms for which $a \leqslant r - a_1$. For $m \neq n$, results are given in Table 9 for $m > n$ and for the following proportionate partitions of N:

(A) $m = \frac{3}{5}N$, $n = \frac{2}{5}N$; (B) $m = \frac{4}{5}N$, $n = \frac{1}{5}N$; (C) $m = \frac{9}{10}N$, $n = \frac{1}{10}N$.

Here sums of terms at both tails of the series are needed. The sums (or chances of $a \geqslant a_1$ or $\leqslant a_1$) have not been given for all possible values of a_1 but, broadly speaking, for those within the limits where significance is likely to be in question. Sums below 0·0010 have generally been omitted. In each case the true sum of the terms (34) is compared with the approximation from the normal integral.

3. In drawing conclusions from the comparison, we have to decide what degree of accuracy is called for. Clearly the normal integral does not give mathematically exact results to 4 decimal places. On the other hand, except for certain instances where the partition is very unequal ($m = \frac{4}{5}N$ and $\frac{9}{10}N$) and r is small, the order of the approximation may be said to follow that of the series closely. If decisions are made by rule of thumb, according to the side of the 5% or 1% significance level on which a falls, then there are a number of entries in the tables where the approximation would give a on the wrong side. But one may question whether judgement of significance based on a single experiment can in fact be made sensitive to a difference between, say, 0·06 and 0·04 (odds of 16 to 1 and 24 to 1) or between 0·012 and 0·008 (odds of 82 to 1 and 124 to 1) and, given such latitude in accuracy, the approximation will be found generally sufficient. These must be points, however, where personal opinions will differ. Whatever views are held, the tables are sufficiently extensive to make it possible to obtain from them a rough measure of the accuracy of approximation in a wide range of cases.

4. It will be noted that in the symmetrical case ($m = \frac{1}{2}N$) and also when $m = \frac{3}{5}N$ the normal approximation for the tail sum is almost invariably a little too large. Undoubtedly for the symmetrical case an improved approximation could be obtained by modifying the $\frac{1}{2}$ correction used in calculating the ratio of deviation to standard deviation. This second order term would, however need to vary with the probability level, thus complicating the procedure.

Table 8. *Case of equal partition, $m = n = \frac{1}{2}N$. Chance that $a \geqslant a_1$ = chance that $a \leqslant r - a_1$*

Partition		$m = n = 50$		$m = n = 30$		$m = n = 20$		$m = n = 15$		$m = n = 10$			
r	a_1	True	Normal approx.	True	Normal approx.	True	Normal approx.	True	Normal approx.	True	Normal approx.	a_1	r
30	17	0·2566	0·2574	0·2194	0·2212							17	30
	18	·1376	·1388	·0981	·1002							18	
	19	·0630	·0643	·0348	·0365							19	
	20	·0243	·0253	·0096	·0106							20	
	21	·0078	·0085	·0020	·0024							21	
	22	·0021	·0024									22	
20	12	0·2269	0·2278	0·2060	0·2076	0·1715	0·1745					12	20
	13	·1053	·1068	·0852	·0873	·0564	·0592					13	
	14	·0392	·0408	·0270	·0287	·0128	·0144					14	
	15	·0114	·0126	·0064	·0073	·0019	·0025					15	
	16	·0025	·0031	·0011	·0014							16	
15	9	0·2884	0·2887	0·2760	0·2772	0·2572	0·2595	0·2330	0·2364			9	15
	10	·1312	·1325	·1163	·1185	·0954	·0985	·0715	·0755			10	
	11	·0453	·0473	·0358	·0380	·0242	·0265	·0134	·0156			11	
	12	·0113	·0129	·0077	·0090	·0040	·0049	·0014	·0020			12	
	13	·0019	·0027	·0011	·0016							13	
10	7	0·1589	0·1599	0·1495	0·1514	0·1367	0·1397	0·1226	0·1266	0·0894	0·0955	7	10
	8	·0458	·0486	·0399	·0429	·0324	·0357	·0251	·0285	·0115	·0147	8	
	9	·0078	·0101	·0061	·0081	·0042	·0058	·0026	·0038	·0005	·0011	9	
	10	·0006	·0014	·0004	·0010							10	
7	5	0·2179	0·2177	0·2119	0·2126	0·2038	0·2056	0·1950	0·1980	0·1749	0·1804	5	7
	6	·0558	·0594	·0514	·0553	·0458	·0501	·0401	·0448	·0286	·0338	6	
	7	·0062	·0096	·0053	·0084	·0042	·0068	·0032	·0055	·0015	·0031	7	
5	4	0·1810	0·1806	0·1766	0·1771	0·1709	0·1735	0·1648	0·1677	0·1517	0·1571	4	5
	5	·0281	·0339	·0261	·0320	·0236	·0295	·0211	·0270	·0163	·0220	5	

Table 9. *Case of unequal partition. Chances that* $a \leqslant a_1$ *and* $a \geqslant a_1$

(A) $m = \frac{3}{5}N$, $n = \frac{2}{5}N$

Partition			$m = 60$, $n = 40$		$m = 36$, $n = 24$		$m = 24$, $n = 16$		$m = 18$, $n = 12$		$m = 12$, $n = 8$				
r	Chance that	a_1	True	Normal approx.	True	Normal approx.	True	Normal approx.	True	Normal approx.	True	Normal approx.	a_1	Chance that	r
30	$a \leqslant a_1$	11	0·0020	0·0019									11	$a \leqslant a_1$	30
		12	·0074	·0074	0·0016	0·0020							12		
		13	·0230	0231	·0084	·0093	—						13		
		14	·0601	·0604	·0320	·0337	0·0023	0·0050					14		
		15	·1330	·1339	·0936	·0957	·0270	·0329					15		
		16	·2512	·2531	·2148	·2165	·1311	·1348					16		
	$a \geqslant a_1$	20	0·2533	0·2531	0·2148	0·2165	0·1322	0·1348					20	$a \geqslant a_1$	
		21	·1323	·1339	·0936	·0957	·0318	·0329					21		
		22	·0580	·0604	·0320	·0337	·0045	·0050					22		
		23	·0209	·0231	·0084	·0093							23		
		24	·0061	·0074	·0016	·0020							24		
		25	·0014	·0019									25		
20	$a \leqslant a_1$	6	0·0027	0·0026	0·0010	0·0012							6	$a \leqslant a_1$	20
		7	·0114	·0112	·0060	·0063	0·0015	0·0021	—				7		
		8	·0381	·0378	·0255	·0262	·0112	·0128	0·0015	0·0033			8		
		9	·1019	·1021	·0816	·0829	·0526	·0555	·0290	·0260			9		
		10	·2211	·2232	·2005	·2028	·1665	·1695	·1170	·1218			10		
	$a \geqslant a_1$	14	0·2236	0·2232	0·2017	0·2028	0·1665	0·1695	0·1182	0·1218			14	$a \geqslant a_1$	
		15	·0994	·1021	·0798	·0829	·0526	·0555	·0241	·0260			15		
		16	·0341	·0378	·0233	·0262	·0112	·0128	·0026	·0033			16		
		17	·0086	·0112	·0048	·0063	·0015	·0021					17		
		18	·0015	·0026	·0006	·0012							18		
15	$a \leqslant a_1$	4	0·0053	0·0053	0·0032	0·0033	0·0013	0·0015					4	$a \leqslant a_1$	15
		5	·0236	·0233	·0171	·0173	·0098	·0106	0·0038	0·0052			5		
		6	·0776	·0775	·0650	·0657	·0481	·0499	·0301	·0335	—		6		
		7	·1948	·1968	·1804	·1827	·1588	·1618	·1317	·1358	0·0511	0·0616	7		
	$a \geqslant a_1$	11	0·1970	0·1968	0·1814	0·1827	0·1587	0·1618	0·1317	0·1358	0·0578	0·0616	11	$a \geqslant a_1$	
		12	·0734	·0775	·0614	·0657	·0458	·0499	·0301	·0335	·0036	·0051	12		
		13	·0188	·0233	·0138	·0173	·0082	·0106	·0038	·0052	—		13		
		14	·0029	·0053	·0018	·0033	·0008	·0015					14		
10	$a \leqslant a_1$	2	0·0088	0·0089	0·0067	0·0071	0·0045	0·0050	0·0026	0·0033	0·0004	0·0009	2	$a \leqslant a_1$	10
		3	·0457	·0453	·0395	·0398	·0318	·0329	·0241	·0260	·0099	·0131	3		
		4	·1538	·1549	·1447	·1464	·1322	·1348	·1182	·1218	·0849	·0910	4		
	$a \geqslant a_1$	8	0·1539	0·1549	0·1442	0·1464	0·1311	0·1348	0·1170	0·1218	0·0849	0·0910	8	$a \geqslant a_1$	
		9	·0386	·0453	·0334	·0398	·0270	·0329	·0209	·0260	·0099	·0131	9		
		10	·0044	·0089	·0034	·0071	·0023	·0050	·0015	·0033	·0004	·0009	10		
7	$a \leqslant a_1$	0	0 0012	0·0022	0·0009	0·0013	0·0006	0·0010	0·0004	0·0007			0	$a \leqslant a_1$	7
		1	·0156	·0189	·0134	·0140	·0109	·0118	·0086	·0097	0·0044	0·0059	1		
		2	·0884	·0956	·0827	·0832	·0756	·0770	·0681	·0704	·0521	·0564	2		
	$a \geqslant a_1$	6	0·1492	0·1587	0·1426	0·1450	0·1341	0·1378	0·1250	0·1300	0·1056	0·1127	6	$a \geqslant a_1$	
		7	·0241	·0385	·0216	·0306	·0186	·0269	·0156	·0232	·0102	0·0160	7		
5	$a \leqslant a_1$	0	0·0088	0·0099	0·0078	0·0090	0·0066	0·0080	0·0056	0·0070	0·0036	0·0051	0	$a \leqslant a_1$	5
		1	·0816	·0811	·0778	·0781	·0730	·0742	·0681	·0701	·0578	·0616	1		
	$a \geqslant a_1$	5	0·0725	0·0811	0·0690	0·0781	0·0646	0·0742	0·0601	0·0701	0·0511	0·0616	5	$a \geqslant a_1$	

Table 9 (*continued*)

(B) $m = \frac{4}{5}N$, $n = \frac{1}{5}N$

Partition			$m = 80$, $n = 20$		$m = 48$, $n = 12$		$m = 32$, $n = 8$	
r	Chance that	a_1	True	Normal approx.	True	Normal approx.	True	Normal approx.
30	$a \leqslant a_1$	18	0·0018	0·0014				
		19	·0084	·0073	0·0013	0·0020		
		20	·0306	·0288	·0106	·0125		
		21	·0884	·0874	·0521	·0548		
		22	·2046	·2078	·1667	·1685		
	$a \geqslant a_1$	26	0·2092	0·2078	0·1667	0·1685		
		27	·0824	·0874	·0521	·0548		
		28	·0227	·0288	·0106	·0125		
		29	·0039	·0073	·0013	·0020		
		30	·0003	·0014				
20	$a \leqslant a_1$	11	0·0040	0·0026	0·0013	0·0011	—	
		12	·0182	·0148	·0095	·0087	0·0016	0·0031
		13	·0638	·0600	·0460	·0448	·0218	·0255
		14	·1729	·1755	·1523	·1542	·1176	·1208
	$a \geqslant a_1$	18	0·1758	0·1755	0·1522	0·1542	0·1176	0·1208
		19	·0499	·0600	·0371	·0448	·0218	·0255
		20	·0066	·0148	·0041	·0087	·0016	·0031
15	$a \leqslant a_1$	7	0·0018	0·0009	0·0008	0·0004		
		8	·0107	·0074	·0064	·0049	0·0022	0·0024
		9	·0462	·0408	·0355	·0323	·0217	·0219
		10	·1470	·1480	·1329	·1338	·1115	·1133
	$a \geqslant a_1$	14	0·1453	0·1480	0·1294	0·1338	0·1079	0·1133
		15	·0262	·0408	·0206	·0323	·0141	·0219
10	$a \leqslant a_1$	4	0·0039	0·0019	0·0026	0·0013	0·0012	0·0008
		5	·0254	·0191	·0206	·0159	·0145	·0121
		6	·1095	·1068	·1012	·0988	·0893	·0882
	$a \geqslant a_1$	10	0·0951	0·1068	0·0868	0·0988	0·0761	0·0882
7	$a \leqslant a_1$	2	0·0033	0·0013	0·0024	0·0010	0·0015	0·0007
		3	·0282	·0203	·0246	·0181	·0201	·0155
		4	·1408	·1417	·1354	·1364	·1281	·1293
	$a \geqslant a_1$	7	0·1985	0·1910	0·1906	0·1848	0·1805	0·1776
5	$a \leqslant a_1$	1	0·0053	0·0022	0·0045	0·0021	0·0035	0·0016
		2	·0531	·0434	·0499	·0430	·0457	·0383
	$a \geqslant a_1$	5	0·3193	0·2841	0·3135	0·2835	0·3060	0·2776

(C) $m = \frac{9}{10}N$, $n = \frac{1}{10}N$

Partition			$m = 90$, $n = 10$	
r	Chance that	a_1	True	Normal approx.
30	$a \leqslant a_1$	22	0·0009	0·0006
		23	·0073	·0057
		24	·0388	·0352
		25	·1384	·1388
	$a \geqslant a_1$	29	0·1356	0·1388
		30	·0229	·0352
20	$a \leqslant a_1$	14	0·0039	0·0019
		15	·0254	·0191
		16	·1095	·1068
	$a \geqslant a_1$	20	0·0951	0·1068
15	$a \leqslant a_1$	9	0·0006	0·0001
		10	·0063	·0027
		11	·0408	·0316
		12	·1705	·1765
	$a \geqslant a_1$	15	0·1808	0·1765
10	$a \leqslant a_1$	5	0·0006	0·0001
		6	·0082	·0029
		7	·0600	·0486
		8	·2615	·2902
	$a \geqslant a_1$	10	0·3305	0·2902
7	$a \leqslant a_1$	3	0·0016	0·0003
		4	·0207	·0096
		5	·1442	·1492
	$a \geqslant a_1$	7	0·4667	0·3974
5	$a \leqslant a_1$	2	0·0067	0·0006
		3	·0769	·0538
	$a \geqslant a_1$	5	0·4163	0·5000