

1958

# Tests of hypotheses and related problems in 2X2 tables

---

<https://hdl.handle.net/2144/22050>

*Downloaded from OpenBU. Boston University's institutional repository.*

BOSTON UNIVERSITY

GRADUATE SCHOOL

Thesis

TESTS OF HYPOTHESES AND RELATED PROBLEMS IN  $2 \times 2$  TABLES

by

WARREN G. SPARKS  
(A.B., Boston University, 1957)

Submitted in partial fulfilment of the  
requirements for the degree of  
Master of Arts  
1958

11/11  
1961

Approved  
by

First Reader. . *Clues B. Knode* . . . .  
PROFESSOR OF MATHEMATICS

Second Reader. . *Francis Scheid* . . .  
PROFESSOR OF MATHEMATICS

## TABLE OF CONTENTS

<u>Topic</u>	<u>Page</u>
Introduction	i
Fisher's Exact Test	1
Barnard's "C.S.M." Test	4
a. Basis of Fisher's Test	5
b. Basis of C.S.M. Test	5
c. Double Dichotomy	6
d. Significance test for 2X2 comparative trial	7
Pearson's Solution	16
a. Problem I	16
b. Problem II	17
c. Problem III	25
d. Summary	28
Tocher's Method	31
a. 2X2 comparative trial	35
b. Double dichotomy	37
c. A common procedure	39
Power Function in a 2X2 Table	40
Summary	54

### Introduction

Very frequently in experimental work we deal with some characteristics or attributes that are not susceptible of accurate measurement, although it is possible to divide the population into two or more categories with reference to these attributes. The division into these categories produces a table which is called a contingency table. Suppose  $N$  objects are classified according as they possess one or both or neither of two qualitative traits or attributes which may, for convenience, be denoted by I and II. Such a classification will yield the following  $2 \times 2$  contingency table:

	Not II	II	Total
Not I	a	c	m
I	b	d	n
Total	r	s	N

where  $a + b + c + d = N$ , and the four classes being mutually exclusive but not necessarily exhaustive. The attributes may sometimes admit also of quantitative measurement, but we are considering only the case where they are classified in two classes, such as "tall" and "not tall", "male" and "female", "good" and "bad", etc.

Before any further discussion of contingency tables, let us consider a typical problem. Suppose that a gambler's

die is rolled 60 times and a record is kept of the number of times each face comes up.<sup>1</sup> If the die is an "honest" or "unbiased" die, each face will have the probability  $1/6$  of appearing in a single roll. Therefore, each face would be expected to appear 10 times in an experiment of this kind. Suppose that the experiment produced the following results, where the row labeled o represents the observed frequencies and the row labeled e represents the expected frequencies:

Face	1	2	3	4	5	6
o	15	7	4	11	6	17
e	10	10	10	10	10	10

As a measure of the compatibility of such observed and expected frequencies, it is customary to calculate a quantity called chi-square ( $\chi^2$ ), which is defined by

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

where  $k$  is the number of pairs of frequencies to be compared,  $o_i$  and  $e_i$  denote the  $i$ th pair of observed and expected frequencies, and  $\sum o_i = \sum e_i = N$ . In this

---

<sup>1</sup>Hoel, P.G., Introduction to Mathematical Statistics, p. 164.

problem  $k=6$  and

$$\begin{aligned} \chi^2 = & \frac{(15-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(4-10)^2}{10} + \frac{(11-10)^2}{10} \\ & + \frac{(6-10)^2}{10} + \frac{(17-10)^2}{10} = 13.6. \end{aligned}$$

A value of 0 would correspond to exact agreement with expectation, whereas increasingly large values of  $\chi^2$  may be thought of as corresponding to increasingly poor experimental agreement. If this experiment were repeated a large number of times with an unbiased die and each time the value of  $\chi^2$  were computed, a set of  $\chi^2_{15}$  would be obtained which could be classified into a relative frequency table of  $\chi^2_{15}$ . This relative-frequency table would tell one approximately in what percentage of such experiments various ranges of values of  $\chi^2$  could be expected to be obtained. Then one would be able to judge whether the value of  $\chi^2=13.6$  was unusually large as compared to the run of  $\chi^2_{15}$  that are obtained in experiments with an unbiased die. If the percentage of experiments for which  $\chi^2 > 13.6$  was very small, say less than 5 per cent, one would judge that the observed frequencies were not compatible with the frequencies expected for an unbiased die, and hence one would conclude that the gambler's die was biased. The following frequency function is known to approximate the frequency function of

$\chi^2$  very well when  $N$  is large, and is called the  $\chi^2$  frequency function.

$$f(\chi^2) = \frac{(\chi^2)^{\frac{v}{2}-1} e^{-\frac{\chi^2}{2}}}{2^{\frac{v}{2}} \Gamma(\frac{v}{2})}$$

This frequency function is for the continuous variable  $\chi^2$  and should not be confused with the unknown frequency function of the discrete variable  $\chi^2$  previously defined; it is only an approximation to the latter frequency function. The parameter  $v$  is called the number of degrees of freedom and is given by  $v = k - 1$ . The symbol  $\Gamma(x)$  denoted the gamma or factorial function of  $x$ . The remarkable feature of this frequency function is that it depends only upon  $k$ , the number of pairs of frequencies to be compared. Since the continuous frequency function is only an approximation to the discrete frequency function, care must be exercised that the  $\chi^2$  test is used only when the approximation is good. Experience and theoretical investigations indicate that the approximation is usually satisfactory provided that the  $e_i \geq 5$  and  $k \geq 5$ . If  $k < 5$ , it is best to have the  $e_i$  somewhat larger than 5.

Tables for the above  $\chi^2$  distribution are available and from these tables it will be found that, for a significance level of 5%,  $\chi^2 = 11.1$  for 5 degrees of freedom; hence, the value of  $\chi^2 = 13.6$  is significant at the 5% level and we conclude that the gambler's die is biased.

The  $\chi^2$  test possesses a remarkable property that permits it to be applied even when the cell probabilities depend upon unknown parameters. This property, although very difficult to prove, is very simple to state. It may be expressed as follows<sup>1</sup>: The  $\chi^2$  test is applicable when the cell probabilities depend upon unknown parameters, provided that the unknown parameters are replaced by their maximum likelihood estimates and provided that one degree of freedom is deducted for each independent parameter estimated.

A very useful application of the  $\chi^2$  test occurs in connection with testing the compatibility of observed and expected frequencies in contingency tables.

A contingency table is usually constructed for the purpose of studying the relationship between the two variables of classification. In particular, one may wish to know whether the two variables are related. By means of the  $\chi^2$  test, it is possible to test the hypothesis that the two variables are independent.

Let us consider a general contingency table containing  $r$  rows and  $c$  columns. Let  $p_{ij}$  be the probability that an individual selected at random from the population under consideration will be a member of the cell in the  $i$ th row

---

<sup>1</sup>Hoel, P.G., Introduction to Mathematical Statistics, p. 170.

and  $j$ th column of the contingency table. Let  $p_{i.}$  be the probability that the individual will be a member of the  $i$ th row and let  $p_{.j}$  be the probability that the individual will be a member of the  $j$ th column. Then the hypothesis that the 2 variables are independent can be written in the form  $H_0: p_{ij} = p_{i.} p_{.j}$  where  $i = 1, \dots, r$  and  $j = 1, \dots, c$ .

If a sample of  $n$  individuals is selected and  $n_{ij}$  of them are found in the cell in the  $i$ th row and  $j$ th column, then  $\chi^2$  will assume the form

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - np_{ij})^2}{np_{ij}}$$

But under the hypothesis  $H_0$ , this expression will become

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - np_{i.} p_{.j})^2}{np_{i.} p_{.j}}$$

Since the  $p_{i.}$  and  $p_{.j}$  are unknown, it is necessary to estimate them from the sample. If the estimates are maximum likelihood estimates, the  $\chi^2$  test may be applied, provided that 1 degree of freedom is deducted for each parameter so estimated. Since  $\sum_{i=1}^r p_{i.} = 1$  and  $\sum_{j=1}^c p_{.j} = 1$ , there are  $r-1+c-1 = r+c-2$  parameters to be estimated; hence the proper number of degrees of freedom for testing independence in a contingency table of  $r$  rows and  $c$  columns is given by  $v = k-1-(r+c-2) = (rc-1)-(r+c-2) = (r-1)(c-1)$ .

Now it is necessary to find the maximum likelihood estimates of the  $p_{i.}$  and  $p_{.j}$ . For this purpose let  $n_{i.}$  denote the sum of the frequencies in the  $i$ th row and let  $n_{.j}$  denote the sum of the frequencies in the  $j$ th column. Since the variables  $n_{ij}$  are discrete, the likelihood of the sample is the probability of obtaining the sample in the order in which it occurred. Thus, the likelihood of the sample will be given by

$$L = \prod_{i=1}^{\hat{r}} \prod_{j=1}^{\hat{c}} p_{ij}^{n_{ij}}$$

But, because of  $H_0$  and the definitions of  $n_{i.}$  and  $n_{.j}$ , this will reduce to

$$\begin{aligned} L &= \prod_{i=1}^{\hat{r}} \prod_{j=1}^{\hat{c}} (p_{i.} p_{.j})^{n_{ij}} \\ &= \prod_{i=1}^{\hat{r}} \prod_{j=1}^{\hat{c}} p_{i.}^{n_{ij}} \prod_{i=1}^{\hat{r}} \prod_{j=1}^{\hat{c}} p_{.j}^{n_{ij}} \\ &= \prod_{i=1}^{\hat{r}} p_{i.}^{\sum_{j=1}^{\hat{c}} n_{ij}} \prod_{j=1}^{\hat{c}} p_{.j}^{\sum_{i=1}^{\hat{r}} n_{ij}} \\ &= \prod_{i=1}^{\hat{r}} p_{i.}^{n_{i.}} \prod_{j=1}^{\hat{c}} p_{.j}^{n_{.j}} \end{aligned}$$

Before differentiating  $L$  with respect to  $p_{i.}$  for maximizing purposes, it is necessary to express one of the  $p_{i.}$ 's, say  $p_{r.}$ , in terms of the remaining ones through the relation  $\sum_{i=1}^{\hat{r}} p_{i.} = 1$ . If this is done,  $L$  will assume

the form

$$L = \left(1 - \sum_{i=1}^{r-1} p_{i.}\right)^{n_r} \cdot \prod_{i=1}^{r-1} p_{i.}^{n_{i.}} \cdot \prod_{j=1}^c p_{.j}^{n_{.j}}$$

Taking logarithms,

$$\log L = n_r \cdot \log \left(1 - \sum_{i=1}^{r-1} p_{i.}\right) + \sum_{i=1}^{r-1} n_{i.} \log p_{i.} + K$$

where  $K$  does not involve the variable  $p_{i.}$ . Now, differentiating with respect to  $p_{i.}$  and setting the derivative equal to 0 for a maximum,

$$\frac{\partial \log L}{\partial p_{i.}} = - \frac{n_r}{1 - \sum_{i=1}^{r-1} p_{i.}} + \frac{n_{i.}}{p_{i.}} = 0$$

Since  $1 - \sum_{i=1}^{r-1} p_{i.} = p_{r.}$ , this equation is equivalent to

$$p_{i.} = \frac{p_{r.}}{n_r} n_{i.} = \lambda n_{i.}$$

where  $\lambda$  does not depend upon the index  $i$ . Since this must hold for  $i = 1, 2, \dots, r$ , and since

$$1 = \sum_{i=1}^r p_{i.} = \lambda \sum_{i=1}^r n_{i.} = \lambda n$$

it follows that  $\lambda = 1/n$ , and hence that the maximum likelihood estimate of  $p_{i.}$  is

$$\hat{p}_{i.} = \frac{n_{i.}}{n}$$

By symmetry, the maximum likelihood estimate of  $p_{.j}$  is

$$\hat{p}_{.j} = \frac{n_{.j}}{n}$$

Therefore, if  $p_{i.}$  and  $p_{.j}$  are replaced by their maximum likelihood estimates,  $\chi^2$  will become

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{\left( n_{ij} - \frac{n_{i.} n_{.j}}{n} \right)^2}{\frac{n_{i.} n_{.j}}{n}}$$

This quantity may be treated as possessing a  $\chi^2$  distribution with  $(r-1)(c-1)$  degrees of freedom, provided that  $n$  is sufficiently large and  $H_0$  is true.

It is easily shown<sup>1</sup> that for a  $2 \times 2$  contingency table, using the cell frequencies of the table on page i,

$$\chi^2 = \frac{N(ad-bc)^2}{mnr s}$$

The approximation to  $\chi^2$  is something like the approximation of the discontinuous binomial distribution to a normal one, where the calculated frequency between two values of  $x$ , say  $a$  and  $b$  inclusive, is given by the area under the corresponding normal curve, not between  $a$  and  $b$  but between  $a - \frac{1}{2}$  and  $b + \frac{1}{2}$ . Similarly, as was suggested by Yates, the approximation to  $\chi^2$  is improved by replacing

---

<sup>1</sup>Kenney & Keeping, Mathematics of Statistics, Part II, p. 230.

one cell frequency, say  $d$ , by  $d \pm \frac{1}{2}$  according as  $ad \geq bc$ , and adjusting the others to keep the marginal totals unaltered. The effect is to replace  $ad-bc$  in the above formula by  $|ad-bc| - N/2$ . This is known as Yates' correction for continuity. It undoubtedly improves the estimate of significance for a  $2 \times 2$  table and should always be applied unless the cell frequencies are quite large. In using the  $\chi^2$  test for tables with small values of  $N$ , it should be borne in mind that the quantity  $N(ad-bc)^2/mnrs$  actually has the  $\chi^2$  distribution only in the limit as  $N$  tends to infinity. Even with the Yates correction, it cannot be assumed that for small values of  $N$  the probabilities calculated from  $\chi^2$  will be accurate.

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 been rounded off satisfactorily.

In 1945, Nature<sup>1</sup> published an exchange of correspondence

---

<sup>1</sup>Nature, v. 156, pp. 177, 388.

between G. A. Barnard and R. A. Fisher which 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.

Subsequent to the foregoing exchange of correspondence, different statisticians have published papers approaching the problem of the  $2 \times 2$  table from various viewpoints. This paper is a survey of these viewpoints.

## Fisher's "Exact" Test

It has been customary to regard the test of significance applied to data given in a 2X2 table as the limiting case of a  $\chi^2$  test with one degree of freedom. However, even after applying Yates' correction for continuity, it cannot be assumed that for small cell frequencies the probabilities calculated from  $\chi^2$  will be accurate.

Fisher<sup>1</sup> says that his "exact" treatment, although more laborious, should be applied whenever in doubt. This test is based on the following reasoning:

If  $p$  is the probability of any event, the probability that it will occur  $a$  times in  $(a+b)$  independent trials is the term of the binomial expansion

$$\frac{(a+b)!}{a! b!} p^a q^b$$

where  $q=1-p$ . The probability that in a sample of  $(c+d)$  trials that it will occur  $c$  times is

$$\frac{(c+d)!}{c! d!} p^c q^d$$

The probability of the observed frequencies  $a, b, c, d$  in a 2X2 table is the product

$$\frac{(a+b)!}{a! b!} \frac{(c+d)!}{c! d!} p^{a+c} q^{b+d}$$

and this in general must be unknown if  $p$  is unknown. The

---

<sup>1</sup>

Fisher, R.A., Statistical Methods for Research Workers, 10th ed., pp 96-97.

unknown factor involving  $p$  and  $q$  will, however, be the same for all tables having the same marginal frequencies  $a+c$ ,  $b+d$ ,  $a+b$ ,  $c+d$ , so that among possible sets of observations having the same marginal frequencies, the probabilities are in proportion to

$$\frac{1}{a! b! c! d!}$$

whatever may be the value of  $p$ , or, in other words, for all populations in which the four frequencies are in proportion.

The sum of the quantities

$$\frac{1}{a! b! c! d!}$$

for all samples having the same margins is

$$\frac{N!}{(a+b)! (c+d)! (a+c)! (b+d)!}$$

where  $N = a+b+c+d$ . Therefore, given the marginal frequencies, the probability of any observed set of entries is

$$p' = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{N! a! b! c! d!} \quad (1)$$

Fisher derives from this expression his "exact" test. This consists of computing the probability (1) of the observed distribution plus the probabilities of the "more extreme" or "less likely" distributions in the same direction, that is, for all values of  $d$  from 0 up to the observed value if  $d < \delta$  where  $\delta$  is the "expected" value, that is,

$\delta = (b+d)(c+d)/N$ . This probability  $P = p'_0 + p'_1 + \dots + p'_d$  corresponds to one tail of the distribution, and thus is

comparable with half of the probability calculated from  $\chi^2$ , since the latter corresponds to both tails of the distribution. If  $d > \delta$  the tail is from  $d$  up to  $(c+d)$  inclusive.

Example 1. The following table<sup>1</sup> exhibits a relationship between inoculation and immunity from attack among a population exposed to a certain disease.

	Inoculated	Not Inoculated	
Not attacked	3 = a	5 = b	8 = (a + b)
Attacked	10 = c	2 = d	12 = (c + d)
	13 = (a + c)	7 = (b + d)	20 = N

For this table  $\chi^2 = 4.43$ , corresponding to  $P = 0.035$ . With the Yates correction,  $\chi^2$  is reduced to 2.65, corresponding to  $P = 0.103$ . The probability of the observed distribution is  $(8! 12! 13! 7!)/(3! 5! 10! 2! 20!) = 0.0477$  and the probabilities of the two more extreme distributions corresponding to  $d = 1$  and  $d = 0$  are 0.0043 and 0.0001, so that Fisher's  $P = 0.052$ , or, for both tails, 0.104.

The chief objection to Fisher's exact test is the large amount of computation involved when the cell frequencies are large. Mainland<sup>2</sup> has published tables based on Fisher's

<sup>1</sup> Kenney and Keeping, Mathematics of Statistics, Part II, p. 231.

<sup>2</sup> Mainland, Herrera, & Sutcliffe, Tables for Use With Binomial Samples, Tables 3 and 4.

method for  $(a+b)=1,2,3,\dots,20$  and where  $(c+d)=1,2,3,\dots,20$ , thereby enabling us to estimate the significance of observed  $2 \times 2$  tables, derived from small samples, without going through the labor of computation.

### Barnard's "C. S. M." Test<sup>1</sup>

Suppose we are given two mass-production processes A and B, and we wish to test whether process A and process B are equally satisfactory in the sense that neither process is more likely to produce defective items than the other. For this purpose  $m$  articles made by process A and  $n$  made by process B are selected and tested under suitable conditions. We find that  $a$  out of the  $m$  articles defective while  $b$  out of the  $n$  articles are defective, a result which can be represented by the following  $2 \times 2$  contingency table:

	I (defective)	II (non-defective)	Total
Process A	a	c	m
Process B	b	d	n
Total	r	s	N

Table 1

On the facts stated above, Barnard shows that it is possible to form three different categories of abstract

<sup>1</sup>  
G. A. Barnard, "Significance Tests for  $2 \times 2$  Tables", Biometrika, v. 34, pp. 123-138.

pictures, any of which might be appropriate to the real case in question. We now turn to the discussion of these three categories.

a. The basis of Fisher's exact test. A simple abstract picture to which Fisher's exact test corresponds is one represented by the mathematical model of  $N$  similar balls in an urn,  $m$  marked A and  $n$  marked B. The balls are withdrawn in random order and placed, in order of their withdrawal, in a row of  $N$  receptacles,  $r$  of which are marked "I", the remainder being marked "II". Since this is a problem of sampling without replacement, the probability that  $a$  of the balls marked A are in receptacles marked "I" is given by the hypergeometric law and is

$$\binom{m}{a} \binom{n}{b} / \binom{N}{r} = \frac{m! n! r! s!}{a! b! c! d! N!} \quad (2)$$

The probability (2), plus those of all results less probable than that obtained, is the basis of Fisher's test. Barnard labels this category the "2X2 independence trial".

b. Basis of the C. S. M. test. Another abstract picture forms the basis of a test, to be developed later in this paper, which Barnard calls his "C.S.M." test. This picture is represented by the mathematical model of two urns A and B, each urn containing a large number of balls, all of which are labeled either "I" or "II". In urn A the proportion of "I"s is  $p_a$  while in urn B it is  $p_b$ . A random sample of  $m$

drawn from urn A contains a marked "I" and c marked "II". A random sample of n is then drawn from urn B and contains b marked "I" and d marked "II". The hypothesis to be tested is that  $p_a = p_b = p$ . If this hypothesis is true, the probability of the observed result is

$$\frac{m!}{a! c!} p^a (1-p)^c \times \frac{n!}{b! d!} p^b (1-p)^d = \frac{m! n!}{a! b! c! d!} p^r (1-p)^s \quad (3)$$

which is equal to Fisher's expression (2) multiplied by a factor  $N! p^r (1-p)^s / r! s!$ . Here, of course, the conditions are different because we are no longer insisting on constant column totals. In various repetitions of the experiment the column totals may vary, but the row totals remain fixed. Barnard labels this category the "2X2 comparative trial".

c. Another category. The third category is represented by the mathematical model of a single urn containing a large number of balls, each of which has two markings - one mark being either A or B, the other either "I" or "II". A random sample of N balls is drawn from the urn and their markings examined. If the proportion of balls marked "AI" is  $p_{a1}$  while  $p_{b1}$ ,  $p_{a2}$ ,  $p_{b2}$  similarly represent the proportion of the other markings in the urn, the probability associated with Table 1 is given by the multinomial theorem and is

$$\frac{N!}{a! b! c! d!} p_{a1}^a p_{b1}^b p_{a2}^c p_{b2}^d.$$

In this case the hypothesis tested, that the markings "I" and "II" and "A" and "B" are independent, may be put in the form

$$p_{a1}p_{b2} = p_{a2}p_{b1}$$

and, assuming that  $(p' = p_{a1} + p_{a2})$ ,  $(p_{b1} + p_{b2} = 1 - p')$ ,  $(p = p_{a1} + p_{b1})$ , and  $(p_{a2} + p_{b2} = 1 - p)$  do not vanish, the probability of our result, on the hypothesis tested is

$$\frac{N!}{a! b! c! d!} p^r (1-p)^s (p')^m (1-p')^n \quad (4)$$

which differs from (3) by a factor  $N! (p')^m (1-p')^n / m! n!$ .

This shows that (4) is related to (3) in much the same way as (3) is related to (2). Barnard labels this category the "double dichotomy".

d. Significance test for 2X2 comparative trial. After describing the foregoing categories of abstract pictures and asserting that Fisher's exact test applies only to one of the categories, Barnard proceeds to develop his significance test for the 2X2 comparative trial. For this test we are interested in the equality or otherwise of  $p_a$  and  $p_b$ . To say that  $p_a$  is greater than  $p_b$  will mean that process B is preferable, and conversely if  $p_b$  is greater than  $p_a$ , while to say that  $p_a$  and  $p_b$  are equal will mean that there is nothing to choose between the two processes, or, in other words, if processes A and B are both used, then it will be found that the

frequencies with which defectives appear in the two processes will, for practical purposes, be equal. Thus we shall assert that results in which the observed frequencies,  $a/m$  and  $b/n$ , differ widely are incompatible with the hypothesis  $p_a = p_b$ . The formulation of a test of significance then reduces to a formulation of what is meant by a "wide difference" in the frequencies  $a/m$  and  $b/n$ .

Suppose, for definiteness, we take  $m=8$  and  $n=6$ .

Then one result of Table 1 could be represented as follows:

Urn:	A	A	A	A	A	A	A	A	B	B	B	B	B	B
Mark:	II	I	II	II	I	II	II	II	I	I	II	I	I	I

Table 2

However, we must treat all results like Table 2, which give the same values to  $a$ ,  $b$ ,  $c$ ,  $d$ , in Table 1, as equivalent. Table 1, therefore, stands for  $m! n! / a! b! c! d!$  distinct, but equivalent, results, which we shall not distinguish from now on.

If we now take rectangular axes in a plane, we can represent Table 1 by the point whose coordinates are  $(a,b)$ . Thus "x" in Figure 1 represents the set of results equivalent to the results of Table 2. Therefore, all possible results of the experiment are represented by the points of the rectangle PQRS. We call this representation of possible results the lattice diagram. Our problem may now be regarded



be considered as indicating wider differences than (a,b) itself. Thus, referring to Fig. 1, the points immediately above and immediately to the left of "x" indicate wider differences than the point "x" itself. This condition implies that the set of points indicating differences as wide or wider than (a,b) will have a shape property related to convexity, so Barnard calls it the "C condition". It means that if we consider the table corresponding to Table 2 with cell frequencies

$$\begin{array}{cc} 2 & 6 \end{array}$$

$$\begin{array}{cc} 5 & 1 \end{array}$$

as significant evidence of difference, then we must also consider the tables

$$\begin{array}{ccc} 1 & 7 & \text{and} & 2 & 6 \end{array}$$

$$\begin{array}{ccc} 5 & 1 & & 6 & 0 \end{array}$$

as significant evidence of difference.

Geometrically, condition S implies that we can restrict our consideration to points in the triangle PRS. Condition C implies that, in this triangle, our "width of difference" must increase as we go upwards or to the left.

Any set of points in the lattice diagram, considered by some criterion agreeing with conditions S and C to indicate differences as wide or wider than those of a given result, will be associated with a probability P, on the

assumption  $p_a = p_b = p$ ; and this  $P$  will be a function of  $p$ ,

$$P(a,b;p) = \frac{m! n!}{a! b! c! d!} p^r (1-p)^s$$

rising from zero when  $p=0$  to a maximum in the neighborhood of  $p = \frac{1}{2}$ , and then falling again symmetrically (by condition S) to zero again at  $p=1$ . Our difficulty arises from the dependence of  $P$  on  $p$ . If the graph of  $P$  against  $p$  were a horizontal straight line, our difficulty would be overcome. What we propose, therefore, is to try to make the graph of  $P$  against  $p$  as near to a horizontal line as possible by suitably adapting our idea of what is meant by "width of difference". In making this adaptation, we shall insure that we do not violate the common-sense requirements as to the meaning of the term "width of difference" by requiring that conditions C and S should always be satisfied.

Condition C requires that the point of triangle PRS that indicates the "widest difference" be the point S at the corner (Fig. 1). The function  $P$  associated with this point, and its converse  $Q$ , is

$$P(0,6;p) = p^6 (1-p)^8 + p^8 (1-p)^6$$

and the maximum  $P_m$  occurs when  $p = \frac{1}{2}$ , therefore,

$$P_m(0,6) = 1/2^{13} = 1.22 \times 10^{-4}.$$

Condition C requires that the only points which might be considered as coming next after (0,6), in order of decreasing "width of difference", are (1,6) and (0,5). We have to adopt some principle to choose between these two. If (1,6) were taken next after (0,6), the function P associated with it would be

$$P(1,6;p) = P(0,6;p) + 16 [p^7(1-p)^7]$$

and  $P_m(1,6)$  would be  $9/2^{13} = 10.97 \times 10^{-4}$ . On the other hand, if (0,5) were chosen next,

$$P(0,5;p) = P(0,6;p) + 6 [p^9(1-p)^5 + p^5(1-p)^9]$$

and  $P_m(0,5)$  would equal  $8.58 \times 10^{-4}$ . Thus  $P_m(0,5)$  is smaller than  $P_m(1,6)$  and this lower maximum is associated with a flatter curve. Since a flat curve is our aim, we choose (0,5) to come next after (0,6). Having chosen (0,5) as the next "widest difference" point, condition C demands that we choose between points (1,6) and (0,4) as candidates for the next position. We then compare

$$P(1,6;p) = P(0,5;p) + 16 [p^7(1-p)^7]$$

with  $P(0,4;p) = P(0,5;p) + 15 [p^4(1-p)^{10} + p^{10}(1-p)^4]$

and the lower value of  $P_m$  as criterion shows that (1,6) is

now selected. Continuing in this manner, we can arrange the points of the lattice diagram in order.

The principle involved, which Barnard calls the "maximum condition", may be formally stated as follows:

Considering only points for which  $a/m$  is less than  $b/n$ , if the first  $(n-1)$  points  $(a_1, b_1), (a_2, b_2), \dots, (a_{n-1}, b_{n-1})$ , in order of decreasing "width of difference" have been chosen, and  $(a_{n-1}, b_{n-1})$  is associated with the function  $P(a_{n-1}, b_{n-1}; p)$ , then the  $n$ th point,  $(a_n, b_n)$  is that point, of all points permitted by the C condition, for which

$$P_m(a, b) = \max_{0 < p < 1} \left[ P(a_{n-1}, b_{n-1}; p) + \frac{m! \ n!}{a! \ b! \ c! \ d!} \left\{ p^r (1-p)^s + p^s (1-p)^r \right\} \right]$$

is least.  $(a_n, b_n)$  is then associated with the function

$$P(a_n, b_n; p) = P(a_{n-1}, b_{n-1}; p) + \frac{m! \ n!}{a! \ b! \ c! \ d!} \left[ p^r (1-p)^s + p^s (1-p)^r \right].$$

To complete the specification of the ordering we have to consider the case where there are several points giving the same value of  $P_m(a, b)$ , this value being less than that associated with any other permissible point. In this case we stipulate that all such points are to be given the same rank, and that the second term in the expression for  $P(a_n, b_n; p)$  be replaced by the corresponding sum over all these points. If there are  $k$  such points at any stage, then the next point after them will be denoted as the  $(n+k)$ th

point in the ordering.

Finally, the significance level to be attached to the point  $(a_n, b_n)$  will be

$$P_m(a_n, b_n) = \max_{0 < p < 1} P(a_n, b_n; p).$$

One of the major objections to Barnard's test is the great amount of computation involved and the non-availability of suitable tables. However, Barnard points out that, for large values of  $m$  and  $n$ , a test based on a normal approximation to the distributions involved would be quite adequate for practical purposes, and that tables are thus required only for small values of  $m$  and  $n$ . In the appendix to his article he gives specimen tables for the case where  $N=14$ . One of these tables, used in solving the following example, is reproduced below. The figures in parenthesis in Table 3 give significance levels on Fisher's "exact" test for  $2 \times 2$  independence trials, for comparison.

Example 2. Two boxes, each containing a large number of components, are to be tested for comparative quality measured by the respective proportions of defective components they contain. Two samples, each of seven components, are taken, at random, one from each box. One sample gives four defectives, the other, none. What is the significance of this result, in relation to the hypothesis that the boxes have the same quality?

Table for  $m=n=7$ 

7	0.012 (0.058)	0.18 (0.23)	0.7 (2.1)	2.4 (7.0)	7.5 (19)	20 (46)	--	--
6	0.18 (0.23)	1.3 (2.9)	5.7 (10)	13 (27)	--	--	--	--
5	0.70 (2.1)	5.7 (10)	21 (29)	--	--	--	--	20 (46)
4	2.4 (7.0)	13 (27)	--	--	--	--	--	7.5 (19)
3	7.5 (19)	--	--	--	--	--	13 (27)	2.4 (7.0)
2	20 (46)	--	--	--	--	21 (29)	5.7 (10)	0.70 (2.1)
1	--	--	--	--	13 (27)	5.7 (10)	1.3 (2.9)	0.18 0.23
0	--	--	20 (46)	7.5 (19)	2.4 (7.0)	0.70 (2.1)	0.18 (0.23)	0.012 (0.058)
	0	1	2	3	4	5	6	7

Table 3 (Figures indicate percentages)

Entering Table 3 at the point (0,4), we find the number 2.4. This means that the result is evidence against the hypothesis on the 2.4% level of significance. More precisely, what is asserted is that the maximum probability of getting a result not less significant than that obtained, is 0.024.

Barnard does not develop a separate test for the double dichotomy, but states that the C.S.M. test for the  $2 \times 2$

comparative trial would be a valid test if applied to double dichotomies. He concedes that the test would err somewhat on the side of "conservatism", but that the error does not appear to be large, except when the numbers involved are exceedingly small.

#### Pearson's Solution

In a paper published in Biometrika in 1947, E. S. Pearson<sup>1</sup> discusses Barnard's three categories, which Pearson calls Problems I, II and III, and shows how each of the problems could be solved using the normal approximation.

a. Problem I (2X2 independence trial). 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. 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  $X$ . The hypothesis tested is that there are  $r = a+b$  individuals who will react and  $s = c+d$  who will not, whatever the assignment of treatments.

The chance that  $a$  will react in  $m$  and  $b$  in  $n$  is, therefore, if the hypothesis is true, the hypergeometric probability

$$P(a|N, r, m) = \frac{m!n!r!s!}{a!b!c!d!N!}$$

<sup>1</sup>

E. S. Pearson, "Choice of Statistical Tests", Biometrika v. 34, pp. 139-167.

For this probability distribution, Kendall<sup>1</sup> has shown that

$$\text{mean of } a = \bar{a} = \frac{rm}{N} \quad (5)$$

and

$$\text{variance of } a = \sigma_a^2 = \frac{mnrs}{N^2(N-1)} \quad (6)$$

When dealing with small numbers, the calculation of the tail terms of the hypergeometric 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 (5) and (6) to represent the sum of the hypergeometric terms. As usual when approximating to the sum of the terms for  $x=a, a+1, a+2, \dots$ , etc., of a discrete probability distribution by the integral under a continuous curve, we take this integral from the point  $x=a-\frac{1}{2}$ . Pearson exhibits tables showing the approximation at various levels, 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.

b. Problem II (2X2 comparative trial). This may be described as the test whether the proportion of individuals bearing the character A is the same in two different populations, from each of which a random sample has been drawn,

---

<sup>1</sup>

M. G. Kendall, Advanced Theory of Statistics, v. 1, p. 127.

i.e., the test of the hypothesis that

$$p_1(A) = p_2(A) = p,$$

where  $p$  is some common, but unspecified proportion. 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  $A$ .

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  $(a,b)$  ( $0 \leq a \leq m, 0 \leq b \leq n$ ) which can be represented in the lattice diagram shown in Figure 2<sup>1</sup> for the special case  $m=12, n=8$ .

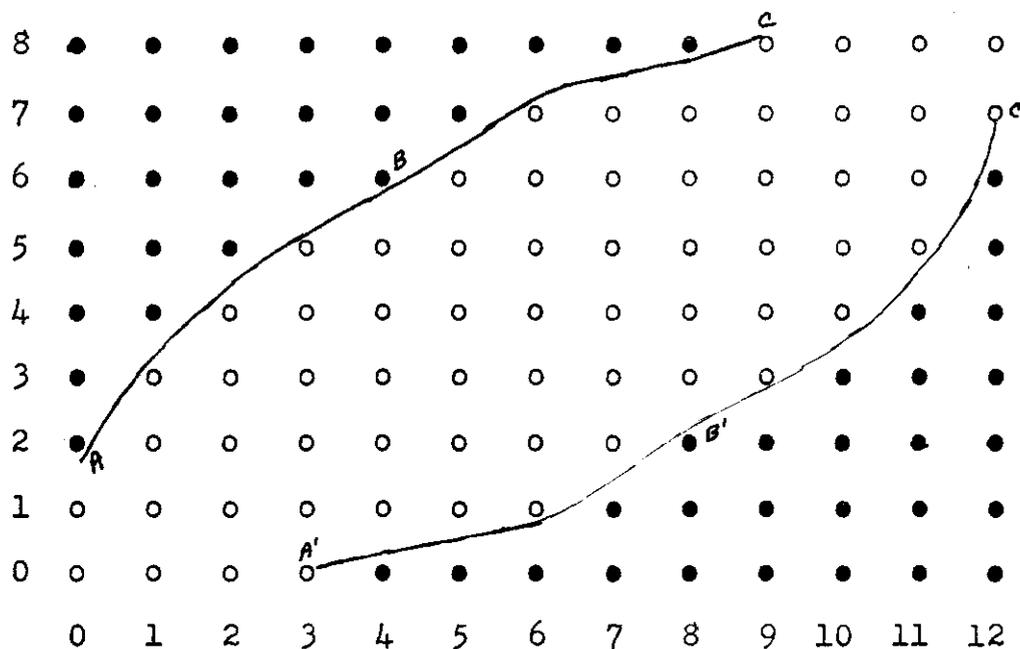


Fig. 2. The curves ABC and A'B'C' represent the significance contours  $L_\epsilon$  and  $L'_\epsilon$ .

<sup>1</sup>  
Biometrika, v. 34, p. 148

If the hypothesis is true, then the probability of the observed result may be written

$$\begin{aligned} P_2(a|p,m) \times P_2(b|p,n) &= \frac{m!}{a!c!} p^a(1-p)^c \times \frac{n!}{b!d!} p^b(1-p)^d \\ &= \frac{N!}{r!s!} p^r(1-p)^s \times \frac{m!n!r!s!}{a!b!c!d!N!} \\ &= P_2(r|p,N) \times P_1(a|N,r,m) \end{aligned}$$

where  $P_1$  denotes a hypergeometric probability and  $P_2$  a binomial probability.

If, now, it were possible to draw a boundary line  $L_\epsilon$  such as ABC shown in Fig. 2, cutting off at the end of each diagonal,  $r = \text{constant}$ , a group of points  $(a, r-a)$  such that

$$\sum_a [P_1(a|N,r,m)] = \epsilon$$

where  $\epsilon$  is a fraction between 0 and 1 chosen at will, we could then associate with this boundary line  $L_\epsilon$  the chance that, if the hypothesis is true, a result will occur in random sampling lying beyond this line. If the hypothesis were rejected when  $(a,b)$  fell beyond this boundary, the probability of doing so if the hypothesis were true, would be

$$\sum_{r=0}^N [P_2(r|p,N) \times \epsilon] = \epsilon \sum_{r=0}^N [P_2(r|p,N)] = \epsilon,$$

therefore it would be independent of the unknown common  $p$  of the hypothesis tested.

Unfortunately, this objective cannot be achieved because we are not dealing with continuous probability distributions and  $P_1(a|N,r,m)$  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 each end of each diagonal,  $r = \text{constant}$ , a group of points for which

$$\sum_a [P_1(a|N,r,m)] = \beta_\lambda \leq \epsilon. \quad (7)$$

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

$$\sum_{r=0}^N [P_2(r|p,N) \times \beta_\lambda] \leq \epsilon. \quad (8)$$

If the samples are large, the calculations of hypergeometric terms become laborious and we turn to the approximation using the normal curve. We define  $u_\epsilon$  as the deviate of the standardized normal curve for which

$$\epsilon = \int_{u_\epsilon}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \quad (\epsilon \leq 1/2).$$

Then we can draw across the lattice diagram a significance level  $L_\epsilon$  above and another  $L'_\epsilon$  below the diagonal  $a/m = b/n$  such that

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

$$\frac{(a + \frac{1}{2})}{\sigma_a} - \bar{a} \leq -u_\epsilon \quad (9)$$

lie above  $L_\epsilon$ , and

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

$$\frac{(a - \frac{1}{2}) - \bar{a}}{\sigma_a} \geq u_\epsilon \quad (10)$$

lie below  $L'_\epsilon$ .

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  $L_\epsilon$  and  $L'_\epsilon$ ; if only, however, when  $a/m < b/n$ , then we use  $L_\epsilon$ . The corresponding probability levels would be obtained by making  $\epsilon$  for the second case twice its value for the first. Figure 3 shows the 247 relative probabilities  $P_1(a|N,r,m)$  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  $\epsilon = 0.05$  and  $u_{0.05} = 1.6445$  has been put above the diagonal and that with  $\epsilon = 0.01$  and  $u_{0.01} = 2.3263$  below.

If the normal approximation to the hypergeometric series were correct, it would follow that along every diagonal,  $r = \text{constant}$ , the sum of the relative probabilities  $P_1(a|N,r,m)$  for points above  $L_\epsilon$  would satisfy the inequality (7). Hence the inequality (8) for the complete area of the lattice above  $L_\epsilon$  would hold, whatever the value of the common  $p$ . A similar result would hold for the area below  $L'_\epsilon$ . Of course, the normal approximation will not hold precisely, particularly when  $r$  or  $s$  is small, but we shall generally be on the safe side, in the sense that the hypergeometric

distribution is flat-topped with abrupt ends so that the  $\beta_\lambda$  of equation (7) will be considerably less than  $\epsilon$ , and often zero. The foregoing method is called Method 1 by Pearson.

Pearson asserts that the introduction of the  $\frac{1}{2}$  for continuity is certainly appropriate in using the normal approximation to the hypergeometric series in Problem I, but that it is not helpful in Problem II where we are concerned with a 2-dimensional experimental probability set. If, for Method 2, instead of obtaining significance levels  $L_\epsilon$  and  $L'_\epsilon$  as for Method 1, we obtain them from inequalities similar to (9) and (10) but with the correction of  $\frac{1}{2}$  omitted, then there are the following points to be noted:

(a) For the significance level  $L_\epsilon$ , the expression

$$\beta_\lambda = \sum_a P_1(a|N, r, m)$$

where the summation is for values of  $a$  on the diagonal,  $r = \text{constant}$ , for which

$$a \leq a_1 = \bar{a} - u_\epsilon \sigma_a$$

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

$$\sum_{\lambda=0}^N \left[ \frac{N!}{r! s!} p^r (1-p)^s \times \beta_\lambda \right]$$

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

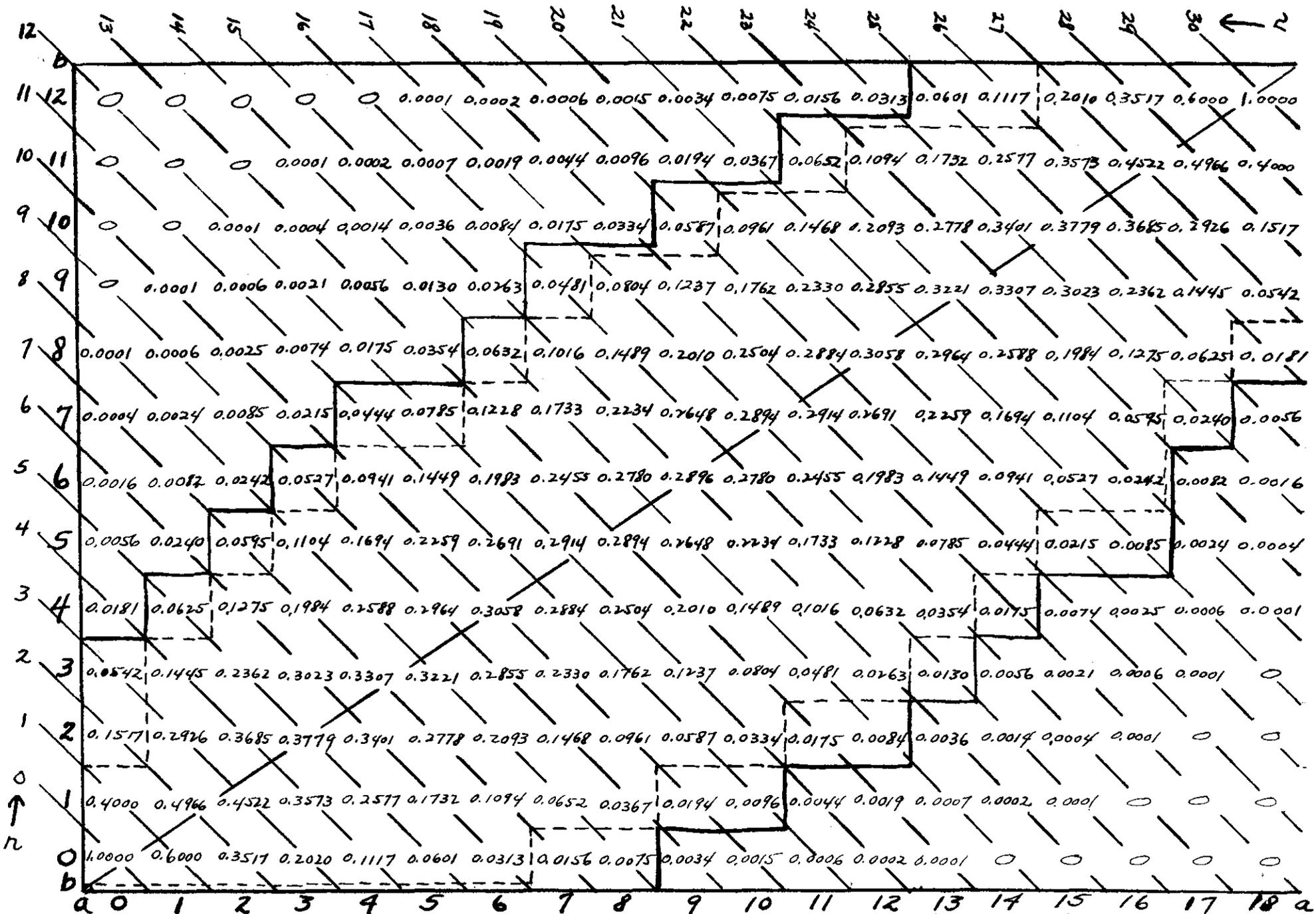


Fig. 3. Hypergeometric probabilities for  $m=18$ ,  $n=12$ . From Biometrika, v.34, p.153.

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

$$u = \frac{a - \bar{a}}{\sigma_a} = \frac{a - rm/N}{\sqrt{\frac{mnrS}{N^2(N-1)}}} \quad (11)$$

has, provided the cases where  $r$  or  $s$  are zero are excluded, whatever  $p$ , an expectation of zero and a unit standard deviation. The shape of the distribution will, of course, depend on  $p$ , but we may not in the long run do too badly by assuming it to be normal.

Consider the result of applying this Method 2 to the case where  $m=18$ ,  $n=12$  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 of Fig. 3. They fall, on the whole, inside the significance levels obtained by Method 1. Pearson exhibits tables showing that, for this example, the true probability does sometimes exceed the nominal values of 0.05 and 0.01, but never by very much. Also, for a second example with  $m=10=n$ , the true probability, while it sometimes exceeds the nominal value, is always considerably nearer it than when using the significance levels of Method 1.

In view of the foregoing, Pearson recommends that the test of the null hypothesis of Problem II should be carried out as follows:

(a) When  $m$ ,  $n$ ,  $r$  or  $s$  are small, by using 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 probability that a point  $(a,b)$  falls on or beyond the level on which the observed result lies.

(b) When  $m$ ,  $n$ ,  $r$  and  $s$  are large, by assuming that the  $u$  of equation (11) is a normal deviate with unit standard deviation.

c. Problem III (double dichotomy). 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 chosen 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 labeled  $AB$ ,  $A\bar{B}$ ,  $\bar{A}B$  and  $\bar{A}\bar{B}$ . If the hypothesis specifying the independence of  $A$  and  $B$  is true, then

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

To test the hypothesis, we have a random sample of  $N$

observations with frequencies of occurrence of the combinations  $AB$ ,  $A\bar{B}$ , etc., as noted in Table 4<sup>1</sup>.

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

Table 4

In problem III there is only one application of a random process, the selection of  $N$  individuals, each of which must fall into one of four alternative categories. If the random process were repeated and another sample  $N$  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 can be represented in 3 dimensions by points  $(a,r,m)$  at unit intervals within a tetrahedron obtained by placing on top of one another the series of 2-dimensioned lattices of dimensions  $0 \times n$ ,  $1 \times (n-1)$ ,  $2 \times (n-2)$ , ...,  $(m-1) \times 1$ ,  $m \times 0$ , where  $m+n=N$ .

The probability of the observed result, if the hypothesis is true, is a term of the multinomial expansion

<sup>1</sup>  
Biometrika, v. 34, p. 158.

$$\begin{aligned}
& \frac{N!}{a! b! c! d!} p(AB)^a p(A\bar{B})^b p(\bar{A}B)^c p(\bar{A}\bar{B})^d \\
&= \frac{N!}{a! b! c! d!} p(A)^{a+b} p(B)^{a+c} p(\bar{A})^{c+d} p(\bar{B})^{b+d} \\
&= \frac{N! p(B)^m [1-p(B)]^n}{m! n!} \times \frac{N! p(A)^r [1-p(A)]^s}{r! s!} \times \\
& \quad \frac{m! n! r! s!}{a! b! c! d! N!} \\
&= P_2(m|p(B), N) \times P_2(r|p(A), N) \times P_1(a|N, r, m). \quad (12)
\end{aligned}$$

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  $N$ .
- (ii) The probability of drawing  $r$  individuals with character A in a random sample of  $N$ .
- (iii) The probability, given  $m$  and  $r$ , of the observed partition within the  $2 \times 2$  table.

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

$$\sum_a P_1(a|N, r, m) = \epsilon \quad (13)$$

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 the hypothesis is true, would be  $\epsilon$  whatever values the irrelevant probabilities  $p(A)$  and  $p(B)$  assumed.

However, (13) cannot be satisfied in general, and all that is possible is to define a family of significance contours such that the probability of a sample point falling beyond any one of them, say  $L_\epsilon$ , is  $\leq \epsilon$ . By using the normal approximation to the sum of the hypergeometric tail-terms, with the correction for continuity, we shall be very much on the safe side, i.e. the formal level of  $\epsilon$  is likely to be much above the true probability of falling beyond the level, whatever be  $p(A)$  or  $p(B)$ . The presence of two binomial terms in equation (12) instead of the single term in the corresponding equation for Problem II (page 19), makes it likely that the overestimation of  $\epsilon$  will be greater in Problem III than in II. It is to be expected, therefore, that when neither  $m$ ,  $n$ ,  $r$  or  $s$  are too small, the better approximation will be obtained by referring the  $u$  of equation (11) to the normal probability scale.

d. Summary. In the foregoing approach to the analysis of data classed in a 2X2 table, the appropriate probability

set-up is defined by the nature of the random process actually used in the collection of the data. On this score, what Pearson has called Problems I, II and III are differentiated. The difference is fundamental and can be illustrated by the following data, given in Table 5<sup>1</sup>, where we shall suppose that the effect we are interested in is making  $a$  significantly greater than  $b$ .

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	$r = 20$	$s = 10$	$N = 30$

Table 5

If the results have been obtained by random assignment of Treatment 1 to 18 out of 30 individuals and Treatment 2 to the remaining 12, and 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 proportion  $15/18$  and  $5/12$  may reasonably be ascribed to a chance fluctuation, we are then concerned with Problem I, i.e. simply with the probabilities associated with the points  $(a, 20-a)$  on the

<sup>1</sup>  
Biometrika, v. 34, p. 160.

diagonal  $r=20$  of Fig. 3. The chance of getting  $a \geq 15$ , if the hypothesis is true, is 0.0241, or, we can speak of the result being significant at the 2.5% level.

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 values of  $a=8,9,\dots,18$ . A repetition of the double sampling process could give us a result  $(a,b)$  falling at any of the 247 points in the lattice diagram of Fig. 3. There will be a number of ways of defining a family of significance levels for this 2-dimensioned set, and if we adopt the method we discussed, which gives as two of its members the dashed, stepped lines shown in Fig. 3, we can say that the probability of a result falling beyond the lower line is certainly less than 0.015 (Pearson exhibits tables showing that the largest value of this probability to be 0.0120 for  $p=0.3$ . This figure cannot be much exceeded for other  $p$ 's though he has not determined the precise maximum. He gives 0.015 as a safe-side limit). The observed point  $(15,5)$  falls beyond the line, so that the result is undoubtedly significant at the 1.5% level.

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. Admittedly, by getting what seems to be a closer relation between the probability set-up and the experimental procedure, we have sacrificed some simplicity in handling the 2X2 table. However, 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.

#### Tocher's Method

In a paper published in Biometrika<sup>1</sup> in 1950, K. D. Tocher derives a modified version of the Fisher test. Tocher's test is based on the Neyman-Pearson likelihood ratio test for selecting the "best" test. The development of Tocher's test will not be given in detail, but we intend to give the basis for the test and then show how it could be applied to the 2X2 table.

We consider the case in which the possible events are the enumerable set  $E_1, E_2, \dots$  with a sample space consisting

---

<sup>1</sup>K. D. Tocher, "Extension of Theory of Tests to Discontinuous Variates", Biometrika, v. 37, pp. 130-144.

of the set of points  $P_1, P_2, \dots$ , the probability attached to the  $i$ th point being  $p_i(\theta)$ , where  $\theta$  is a collective symbol for the parameters of the distribution.

We require to test the hypothesis  $H_0, \theta = \theta_0$ , against the alternative  $H_1, \theta = \theta_1$ , and to use a region of bounded size  $\alpha$ . We shall adopt a method in common use of defining our region  $\alpha$ . This method is equivalent to attaching to each point in the sample space a number  $w$ , of value 1 if the point lies inside the region, and 0 otherwise. The number  $w$  associated with any point can then be regarded as the probability that the point lies in the region. An obvious generalization is to introduce a random process which allows the  $w$ 's to take on all values  $0 \leq w \leq 1$ . Of all such regions, that of maximum probability under the hypothesis  $H_1$  is required.

Put

$$\frac{p_i(\theta_1)}{p_i(\theta_0)} = \lambda_i \quad (14)$$

and suppose that the points can be ordered so that

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_n \geq \dots \quad (15)$$

and that if  $\lambda_i = \lambda_{i+1}$ , then  $p_i(\theta_0) \leq p_{i+1}(\theta_0)$ . (16)

If  $\lambda_i = \lambda_{i+1}$ , and  $p_i(\theta_0) = p_{i+1}(\theta_0)$ , then  $p_i(\theta_1) = p_{i+1}(\theta_1)$ .

In this case, the events  $E_i$  and  $E_{i+1}$  have the same probabilities under either hypothesis, so substituting one for the

other cannot influence any judgment about the two hypotheses. Assuming that it is desirable that the probabilities of such events should enter any test procedure symmetrically, this can be achieved by pooling such equivalent events as a composite event ( $E_i$  or  $E_{i+1}$ ) with probabilities  $2p_i(\theta_0)$  and  $2p_i(\theta_1)$ , respectively, under the two hypotheses. Larger groups of equivalent events can be dealt with similarly. With this convention the equality of (16) can never arise, and the points or events are completely ordered by (15) and (16).

Any set of probabilities  $w_i$  ( $i=1,2,3,\dots$ ) define a test procedure  $W(w_i)$  which rejects  $H_0$  in favor of  $H_1$  with probability  $w_i$  if the event  $E_i$  materializes in the trial or experiment.

We require to determine  $w_i$  to satisfy the following conditions:

$$0 \leq w_i \leq 1 \quad (i=1,2,3,\dots)$$

$$\sum_{i=1}^{\infty} w_i p_i(\theta_0) \leq \alpha \quad (17)$$

$$\sum_{i=1}^{\infty} w_i p_i(\theta_1) = \beta$$

where  $\alpha$  is the fixed size and  $\beta$  is maximized.

Define  $s$  by

$$\sum_{i=1}^s p_i(\theta_0) \leq \alpha \leq \sum_{i=1}^{s+1} p_i(\theta_0),$$

and consider the procedure  $W(\hat{w}_i)$  with  $\hat{w}_i$  given by

$$\left. \begin{aligned} \hat{w}_i &= 1 && (i = 1, 2, 3, \dots, s) \\ \hat{w}_{s+1} &= \frac{\alpha - \sum_{i=1}^s p_i(\theta_0)}{p_{s+1}(\theta_0)} \\ \hat{w}_i &= 0 && (i = s+2, s+3, \dots). \end{aligned} \right\}$$

Let  $\sum_{i=1}^{\infty} \hat{w}_i p_i(\theta_1) = \hat{\beta}$ .

Then for any other procedure of fixed size  $\alpha$

$$\hat{\beta} - \beta = \sum_{i=1}^s (1 - w_i) p_i(\theta_1) + \frac{\alpha - \sum_{i=1}^s p_i(\theta_0)}{p_{s+1}(\theta_0)} p_{s+1}(\theta_1) - \sum_{i=s+1}^{\infty} w_i p_i(\theta_1).$$

Eliminating  $\alpha$  with (17), using (14) and rearranging, we obtain

$$\hat{\beta} - \beta \geq \sum_{i=1}^s (1 - w_i) (\eta_i - \eta_{s+1}) p_i(\theta_0) + \sum_{i=s+2}^{\infty} w_i (\eta_{s+1} - \eta_i) p_i(\theta_0). \quad (18)$$

From (15) we deduce that every term on the right-hand side of (18) is non-negative and hence  $\hat{\beta} \geq \beta$ . The equality only holds if

(a)  $\sum_{i=1}^s w_i p_i(\theta_0) = \alpha$

(b) each term on the right-hand side of (18) vanishes,

i.e.

$$\begin{aligned} 1 - w_i &= 0 && (i = 1, 2, \dots, s), && \eta_i \neq \eta_{s+1} \\ w_i &= 0 && (i = s+2, s+3, \dots), && \eta_i \neq \eta_{s+1}. \end{aligned}$$

Thus equally powerful procedures only differ from  $W(\hat{w}_i)$  in

the  $w$ 's allotted to events of likelihood ratio  $\lambda_{s+1}$ .  $W(\hat{w}_i)$  really uses the random process only if the event  $E_{s+1}$  materializes, while all other procedures of equal power will involve it on some event of greater probability under  $H_0$ .

If  $H_0$  is to be tested against a class of alternatives  $H$  which give a common likelihood ratio ordering of the sample space, then  $W(\hat{w}_i)$  is a "best" procedure, using the word in the Neyman-Pearson sense.

We now see how this test can be applied to 2X2 tables.

a. 2X2 comparative trial. Suppose one trial of  $m$  members has  $a$  successes and  $c$  failures, while a second trial of  $n$  members has  $b$  successes and  $d$  failures. What is the best procedure for determining whether the success rates in the two trials are equal?

The possible events can be characterized by the number of successes in the two trials, viz.  $(a,b)$ .

If the success rates are  $p_1 (=p)$  and  $p_2 (= \xi p)$  the probability of the event  $(a,b)$  is

$$\begin{aligned} P(a,b) &= \binom{m}{a} \binom{n}{b} p_1^a (1-p_1)^c p_2^b (1-p_2)^d \\ &= \frac{\binom{m}{a} \binom{n}{b}}{\binom{N}{r}} \binom{N}{r} p^r (1-p)^s \xi^b \left( \frac{1-\xi p}{1-p} \right)^d \end{aligned}$$

where  $N = m+n$ ,  $r = a+b$ ,  $s = c+d = N-r$ .

On the hypothesis under test,  $p_1 = p_2$  or  $\xi = 1$ , the probabilities reduce to

$$P_0(a, b) = \pi_{\lambda, m}^{(a)} \binom{N}{r} p^r (1-p)^{N-r},$$

where

$$\pi_{\lambda, m}^{(a)} = \frac{\binom{m}{a} \binom{n}{b}}{\binom{N}{r}} = \frac{\binom{m}{a} \binom{N-m}{r-a}}{\binom{N}{r}}.$$

Put  $\prod_{\lambda, m}^{(a)} = \sum_{t=0}^a \pi_{\lambda, m}^{(t)}$ .

Note that in the example  $\xi$  is a parameter which is specified in the two hypotheses, while  $p$  is the "nuisance" parameter.

We have

$$\begin{aligned} \phi &= \frac{d \log P_0}{d p} = \frac{d}{d p} [r \log p + s \log (1-p)] \\ &= \frac{r}{p} - \frac{s}{1-p} \\ &= \frac{r - Np}{p(1-p)}. \end{aligned}$$

Thus the points of equal  $\phi$  have equal  $r$ , and this division of the points  $(a, b)$  into sets is invariant under change of  $p$ .

We have shown previously (page 19) that the conditional probability in the set of points  $a+b=r$  is given by the hypergeometric probability  $\pi_{\lambda, m}^{(a)}$ . The likelihood ratio is

$$\lambda(a, b) = \xi^b \left( \frac{1-\xi p}{1-p} \right)^d = \xi^a \left( \frac{1-\xi p}{1-p} \right)^{n-a} \left( \frac{1-\xi p}{\xi - \xi p} \right)^a.$$

11?

Thus within the set of points  $a+b=r$  the likelihood ratio ordering is that of  $a$ , increasing with  $a$  if  $\xi < 1$ , and decreasing with  $a$  if  $\xi > 1$ .

Therefore there is a common best procedure for testing against all alternatives  $p_1 < p_2$  ( $\xi > 1$ ). This is defined by  $w(a,b)$ , where

$$\left. \begin{aligned} w(a, r-a) &= 1 && (a < \bar{a}_{r,m}) \\ w(\bar{a}_{r,m}, r-\bar{a}_{r,m}) &= \left[ \alpha - \prod_{r,m} (\bar{a}_{r,m} - 1) \right] / \prod_{r,m} (\bar{a}_{r,m}) \\ w(a, r-a) &= 0 && (a > \bar{a}_{r,m}) \end{aligned} \right\} \quad (18)$$

and  $\bar{a}_{r,m}$  defined by

$$\prod_{r,m} (\bar{a}_{r,m} - 1) \leq \alpha < \prod_{r,m} (\bar{a}_{r,m}). \quad (19)$$

Similarly, there is a common best procedure for the other class of one-sided alternatives  $\xi < 1$ . This is most easily obtained by interchanging success and failure,  $a$  with  $c$  and  $b$  with  $d$ , and applying the above procedure.

b. Double Dichotomy. Suppose a sample of  $N$  is classified according to two characteristics, and the numbers in the four resulting classes are  $a, b, c, d$ . What is the best procedure for determining whether the characteristics are independent?

The possible events can be characterized by the triad  $(a,r,m)$ . If the probabilities associated with the four

classes are  $p_{11}$ ,  $p_{12}$ ,  $p_{21}$ ,  $p_{22}$ , and we define

$$\begin{aligned} p &= p_{11} + p_{12} & q &= 1-p \\ p' &= p_{11} + p_{21} & q' &= 1-p' \\ \xi &= p_{11} - pp' \end{aligned}$$

then we have

$$\begin{aligned} p_{11} &= pp' + \xi & p_{12} &= pq' - \xi \\ p_{21} &= qp' - \xi & p_{22} &= qq' + \xi . \end{aligned}$$

The hypothesis of independence is  $\xi=0$ , and  $p$ ,  $p'$  are the "nuisance" parameters.

The probability of event  $(a, r, m)$  can be written

$$\begin{aligned} P(a, r, m) &= \frac{N!}{a!b!c!d!} p_{11}^a p_{12}^b p_{21}^c p_{22}^d \\ &= \pi_{\lambda, m}(a) \left[ \binom{N}{r} p^r q^{N-r} \right] \left[ \binom{N}{m} p'^m q'^{N-m} \right] \\ &\quad \times \left( \frac{1-\xi}{pq'} \right)^r \left( \frac{1-\xi}{qp'} \right)^m \left( \frac{1+\xi}{qq'} \right)^{N-m} \left[ \frac{(1+\xi/pp')(1+\xi/qq')}{(1-\xi/pq')(1-\xi/qp')} \right]^a \quad (20) \end{aligned}$$

In particular, when  $\xi=0$ ,

$$P_0(a, r, m) = \pi_{\lambda, m}(a) \left[ \binom{N}{r} p^r q^{N-r} \right] \left[ \binom{N}{m} p'^m q'^{N-m} \right].$$

We have

$$\begin{aligned} \frac{d}{dp} (\log P_0) &= \frac{r-Np}{pq} \\ \frac{d}{dp'} (\log P_0) &= \frac{m-Np'}{p'q'} . \end{aligned}$$

Thus we divide the points into sets of equal  $r$  and  $m$ ,

and the conditional probability in these sets is  $\pi_{r,m}(a)$ . From (20) the likelihood ratio within each set of constant  $r$  and  $m$  is a monotonic function of  $a$ , increasing when  $\xi > 0$  and decreasing when  $\xi < 0$ . Thus, a best common procedure for the one-sided alternatives  $\xi < 0$  is given by the same system of  $w$ 's as in (18) and (19).

Since Barnard's "2X2 independence trial" (Pearson's Problem I and Fisher's "exact" test) is also based on the hypergeometric distribution  $\pi_{r,m}(a)$ , it follows that a common procedure can be used in each of the three situations.

The formal procedure consists of the following steps:

- (a) Fix a value of  $\alpha$ ;
- (b) Perform the experiment and record the result  $(a, r, m)$ ;
- (c) Calculate  $\pi_{r,m}(a)$  and  $\pi_{r,m}(a-1)$ ;
- (d) Reject the hypothesis if  $\pi_{r,m}(a) \leq \alpha$ . Accept the hypothesis if  $\pi_{r,m}(a-1) > \alpha$ .

(e) Otherwise, take a single sample at random from a distribution uniform over the interval  $(0,1)$ . If this value be  $\xi$ , compare this with the calculated quantity

$$\hat{\xi} = \frac{\alpha - \pi_{r,m}(a-1)}{\pi_{r,m}(a) - \pi_{r,m}(a-1)}$$

Reject the hypothesis if  $\hat{\xi} > \xi$ . Accept the hypothesis if  $\hat{\xi} < \xi$ .

Tocher asserts that the slight modification of the

Fisher test derived here produces a better test than the more elaborate changes suggested by Barnard. For very large size tables, the numerical difficulties of this test can be avoided by noting that the effect of the introduction of the random variable becomes negligible, so an ordinary "exact" test can be applied. For the large sample sizes this can be replaced by the normal approximation advocated by Pearson.

#### The Power Function in a 2X2 Table

Where there is no doubt about the most appropriate test and no sequential scheme of sampling is possible, the power function may play a useful part in indicating, before the data are collected, how large the samples should be to avoid an inconclusive result.

If we define  $u_\alpha$  as the deviate of the standardized normal curve for which  $\alpha = \int_{u_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$

where

$$\frac{\frac{a-rm/N}{\sqrt{\frac{mnrs}{N^2(N-1)}}}} = u_{\frac{1}{2}\alpha}$$

then, using the two-tailed test, we should reject the hypothesis that  $p_1 = p_2$  (using the two-dimensioned sample space as in Pearson's Problem II) at the significance

level  $\alpha$  when  $|u| > u_{\frac{1}{2}\alpha}$ ; in the case of the one-tailed test, we should reject the hypothesis that  $p_1 \leq p_2$  at the same significance level when  $u > u_\alpha$ .

We form the boundaries of the critical region in the same manner as we did when we were discussing Pearson's method. If  $p_1 \neq p_2$ , then the power of the test of  $H_0$  with regard to the alternative hypothesis  $H_1$  is the probability that the point (a,b) falls in the critical region when sampling from populations with proportions  $p_1$  and  $p_2$ , i.e.

$$P(|u| > u_{\frac{1}{2}\alpha} \mid p_1, p_2) \quad (\text{for two-tailed test})$$

or  $P(u > u_\alpha \mid p_1, p_2) \quad (\text{for one-tailed test}).$

This is the total probability density at all the discrete points (a,b) included in the critical region. If this is expressed in a readily calculable form, two types of application are evident:

(1) When the decision has been made to take two samples of, say, 50, or when the available data happen to consist in samples of this size, we may ask, "what is the chance that the test described will show a difference in observed proportions significant at the 5% level when in fact,  $p_1$  and  $p_2$  are as different, say, as 0.50 and 0.65?".

(2) On the other hand, we may use the theory to ask in advance how large the samples should be so that the risk of failing to detect a given difference between  $p_1$  and  $p_2$  which

is considered to be of importance, shall be acceptably small. For example, we may ask, "what sizes of samples should we take so that in applying our test we may have a high, say a 90%, chance of detecting that the proportions are not equal when, in fact, they are as different as 0.50 and 0.65?".

If the above procedure is to be easily applied, a ready means must be available of calculating the power of the test for a given significance level and sample size. In a paper published in Biometrika<sup>1</sup> in 1948, P. B. Patnaik presented in a simple, though approximate, form a means of determining the power function of the test for the difference between two proportions. One of Mr. Patnaik's approximations will be developed and an example given showing the application of the power function approximated by this method.

It is clear that for given  $m$ ,  $n$  and  $\alpha$  the power of the test will be constant on certain contours in the  $p_1, p_2$  space such as those shown in Fig. 4. We shall examine the approximate form of these contours and show that in the important case when  $m = n = \frac{1}{2}N$ , the family of contours is independent of  $N$  and  $\alpha$ , although the power associated with a particular contour will be a function of  $N$  and  $\alpha$ , which has been tabled. Throughout the investigation, approximations are made of the type involved in representing binomial or

---

<sup>1</sup>

P. B. Patnaik, "The Power Function in a 2X2 Table",  
Biometrika, v. 35, pp. 157-173.

hypergeometric distributions by normal distributions.

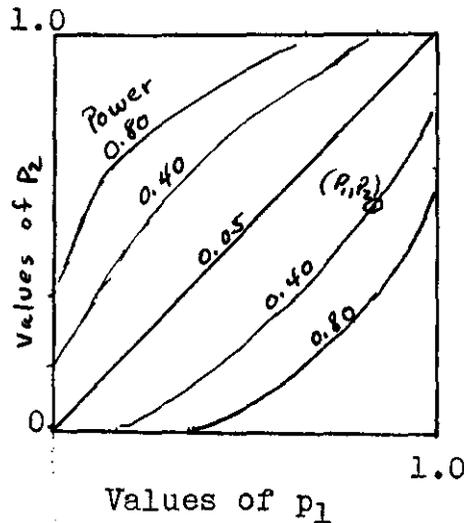


Fig. 4. Power contours  
( $m = n = 50$ ).

We now consider an approximation to the distribution of  $a$  under the hypothesis  $H_1$ , under which the population proportions are  $p_1$  and  $p_2$ .

$$p(a, r) = \frac{N!}{r!s!} p_2^r q_2^s \left(\frac{q_1}{q_2}\right)^m \left[ \frac{m!n!r!s!}{N!a!b!c!d!} \right] \left(\frac{p_1 q_2}{p_2 q_1}\right)^a.$$

If we replace the hypergeometric term in the brackets by the ordinate of a normal curve having the mean and standard deviation of the hypergeometric series, then

$$p(a, r) = \frac{N!}{r!s!} p_2^r q_2^s \left(\frac{q_1}{q_2}\right)^m \frac{1}{\sqrt{2\pi} \sqrt{\frac{mnr s}{N^2(N-1)}}} \exp\left[-\frac{(a-rm)^2}{\frac{2mnr s}{N^2(N-1)}}\right] \left(\frac{p_1 q_2}{p_2 q_1}\right)^a.$$

Writing  $\left(\frac{p_1 q_2}{p_2 q_1}\right)^a$  as  $\exp\left[a \log\left(\frac{p_1 q_2}{p_2 q_1}\right)\right]$ , collecting the terms

containing  $a$  and making a perfect square, we obtain

$p(a, r) = p(r)p(a|r)$ , where

$$p(r) = \frac{N!}{r!s!} p_2^r q_2^s \left(\frac{q_1}{q_2}\right)^m \exp\left[\frac{rm}{N} \log \frac{p_1 q_2}{p_2 q_1} + \frac{mnrs}{2N^2(N-1)} \left(\log \frac{p_1 q_2}{p_2 q_1}\right)^2\right] \quad (21)$$

$$\text{and } p(a|r) = \frac{1}{\sqrt{2\pi} \sqrt{\frac{mnrs}{N^2(N-1)}}} \exp\left[-\frac{\left(\frac{a-rm}{N} - \frac{mnrs}{N^2(N-1)} \log \frac{p_1 q_2}{p_2 q_1}\right)^2}{2 \frac{mnrs}{N^2(N-1)}}\right] \quad (22)$$

Thus (22) is the approximate conditional distribution of  $a$  on the diagonal  $r=a+b$  and is seen to be normal with mean  $= rm/N + mnrs/N^2(N-1) \log(p_1 q_2/p_2 q_1)$  and standard deviation  $= \sqrt{mnrs/N^2(N-1)}$ . Defining

$$u = \frac{a-rm/N}{\sqrt{mnrs/N^2(N-1)}}$$

equation (22) becomes

$$\begin{aligned} p(u|r) &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(u - \sqrt{mnrs/N^2(N-1)} \log \frac{p_1 q_2}{p_2 q_1}\right)^2\right] \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(u-h(r))^2} \end{aligned} \quad (23)$$

$$\text{where } h(r) = \sqrt{mnrs/N^2(N-1)} \log \frac{p_1 q_2}{p_2 q_1}, \quad (24)$$

and is a function of  $r$  only, since  $s(=N-r)$  and the other quantities are given.

If  $p_1$  and  $p_2$  are equal, then (23) reduces to the distribution of  $u$  under  $H_0$ ,

$$p(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}, \quad (25)$$

which is the normal approximation used in obtaining the test criterion. This distribution is independent of  $r$ , but the distribution of  $u$  under  $H_1$ , given by (23) is not independent of  $r$ . It is normal, with the same standard deviation as for (25), but with its mean shifted by  $h(r)$ .

What may be termed the conditional power, for  $r$  fixed, with regard to  $H_1 (p_1 \neq p_2)$  is then, for the two-tailed test,

$$\begin{aligned} P(|u| > u_{\frac{1}{2}\alpha} | r) &= \int_{-\infty}^{-u_{\frac{1}{2}\alpha}} p(u|r) du + \int_{u_{\frac{1}{2}\alpha}}^{\infty} p(u|r) du \\ &= 1 - \frac{1}{\sqrt{2\pi}} \int_{-u_{\frac{1}{2}\alpha} - h(r)}^{u_{\frac{1}{2}\alpha} - h(r)} e^{-\frac{1}{2}u^2} du \end{aligned} \quad (26)$$

and, for the one-tailed test,

$$P(u > u_{\alpha} | r) = \int_{u_{\alpha}}^{\infty} p(u|r) du = \frac{1}{\sqrt{2\pi}} \int_{u_{\alpha} - h(r)}^{\infty} e^{-\frac{1}{2}u^2} du. \quad (27)$$

Since  $p(u) = p(u|r)p(r)$ , the "over-all" power function is

$$\int_{-\infty}^{\infty} P(|u| > u_{\frac{1}{2}\alpha} | r) p(r) dr \quad (28)$$

for the two-tailed test with a similar expression for the one-tailed test. The labor involved in calculating the

over-all power would in general be prohibitive, therefore some approximation for  $p(r)$ , expression (21), is required. The simplest approximation is obtained by assuming  $r$  to be normally distributed. Since  $a$  and  $b$  are distributed binomially with means  $mp_1$  and  $np_2$ , and standard deviations  $mp_1q_1$  and  $np_2q_2$ ,  $r = a + b$  can be considered as distributed normally with mean  $= mp_1 + np_2$  and standard deviation  $= (mp_1q_1 + np_2q_2)^{\frac{1}{2}}$ . That is,

$$p(r) = \frac{1}{\sqrt{2\pi} \sqrt{(mp_1q_1 + np_2q_2)}} \exp \left[ -\frac{\{r - (mp_1 + np_2)\}^2}{2(mp_1q_1 + np_2q_2)} \right]. \quad (29)$$

Hence, the expression (28) for the over-all power function becomes

$$\frac{1}{\sqrt{2\pi} \sqrt{(mp_1q_1 + np_2q_2)}} \int_{-\infty}^{\infty} P(|u| > u_{\frac{1}{2}\alpha} | r) \exp \left[ -\frac{\{r - (mp_1 + np_2)\}^2}{2(mp_1q_1 + np_2q_2)} \right] dr \quad (30)$$

for the two-tailed test with a similar expression for the one-tailed test.

The over-all power is a function of  $p_1$  and  $p_2$ , for given  $m$ ,  $n$ , and  $\alpha$  and may be written as  $\beta(p_1, p_2 | m, n, \alpha)$ . Similarly, the conditional power function may be written as  $\beta(p_1, p_2 | m, n, \alpha, r)$ . They will be denoted by  $\beta$  and  $\beta(r)$  respectively.

Suppose  $\mu_1 = mp_1 + np_2$ ,  $\sigma^2 = \mu_2 = mp_1q_1 + np_2q_2$  and  $\mu_3, \mu_4,$

etc. equal the higher moments of the normal distribution (29). Then, from (30)

$$\beta = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{(r-\mu_1)^2}{2\sigma^2}\right] \beta(r) dr.$$

Expanding  $\beta(r)$  by Taylor's series,

$$\beta(r) = \beta(\mu_1) + (r-\mu_1)\beta'(\mu_1) + \frac{(r-\mu_1)^2}{2!}\beta''(\mu_1) + \dots$$

and substituting in above, we obtain

$$\begin{aligned} \beta &= \beta(\mu_1) \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{(r-\mu_1)^2}{2\sigma^2}\right] dr \\ &+ \beta'(\mu_1) \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{(r-\mu_1)^2}{2\sigma^2}\right] (r-\mu_1) dr \\ &+ \frac{\beta''(\mu_1)}{2!} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{(r-\mu_1)^2}{2\sigma^2}\right] (r-\mu_1)^2 dr + \dots \\ &= \beta(\mu_1) + \frac{\beta''(\mu_1)}{2!} \mu_2 + \frac{\beta^{(4)}(\mu_1)}{4!} \mu_4 + \dots \end{aligned}$$

It follows that an approximation to the over-all power  $\beta$  is

$$\beta(\mu_1) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-u_{\frac{\alpha}{2}} - h(\mu_1)}^{u_{\frac{\alpha}{2}} - h(\mu_1)} e^{-\frac{u^2}{2}} du \quad (31)$$

for the two-tailed test, or

$$\beta(\mu_1) = \frac{1}{\sqrt{2\pi}} \int_{u_{\alpha} - h(\mu_1)}^{\infty} e^{-\frac{u^2}{2}} du \quad (32)$$

for the one-tailed test, substituting  $\mu_1$  for  $r$  in the expressions (26) and (27).

Suppose  $m = n = \frac{1}{2}N$ . Then  $\mu_1 = n(p_1 + p_2)$ . Substituting in (24) and replacing  $N^2(N-1)$  by  $N^3 (= 8n^3)$ , the error being negligible when  $n$  is not too small, we find

$$h(\mu_1) = (n)^{\frac{1}{2}} 2^{\frac{1}{2}} \left[ (p_1 + p_2) (2 - p_1 - p_2) \right]^{\frac{1}{2}} \log \frac{p_1(1-p_2)}{p_2(1-p_1)}.$$

In the case of the one-tailed test, where the alternative is  $p_1 > p_2$ ,  $h(\mu_1)$  is positive. In the other case with alternatives  $p_1 < p_2$  or  $p_1 > p_2$ ,  $h(\mu_1)$  is negative or positive. But from the expression (31), for the approximate power, it is seen that the sign of  $h(\mu_1)$  is immaterial. So, putting  $h = |h(\mu_1)|$ ,

$$k = k(p_1, p_2) = \sqrt{\frac{2}{4}} \sqrt{(p_1 + p_2)(2 - p_1 - p_2)} \left| \log \frac{p_1(1-p_2)}{p_2(1-p_1)} \right|, \quad (33)$$

we have  $h = kn^{\frac{1}{2}}$ . (34)

From (31) and (32) it follows that the approximation, for given  $n = \frac{1}{2}N$  and  $\alpha$ , the power,  $\beta$ , is a function of  $k$  only, i.e.

$$\beta(p_1, p_2 | n, n, \alpha) = \beta(k | n, \alpha).$$

From (33) Patnaik calculated  $k$  for

$$p_1, p_2 = 0.05(0.05)0.95,$$

some values of which are given in Table 6. Thus for  $m = n$  the values of  $p_1$ ,  $p_2$  and the power of the test may be linked up as follows:

- (i) Table 6 relates  $p_1, p_2$  to  $k$ .
- (ii) Equation (34) gives  $h$  in terms of  $k$  and  $n$ .
- (iii) The normal integrals (31) and (32) give the power in terms of  $h$  and the significance level  $\alpha$  employed in the test.

The power as a function of  $h = kn^{\frac{1}{2}}$  and  $\alpha$  is given in Table 7. For the two-tailed test we require the integral (31). Denoting this by  $P$ , i.e.

$$P = 1 - \frac{1}{\sqrt{2\pi}} \int_{-u_{\frac{1}{2}\alpha} - h}^{u_{\frac{1}{2}\alpha} - h} e^{-\frac{u^2}{2}} du \quad (35)$$

the value of  $P$  is given in columns 2 and 3 of Table 7. As  $h$  increases, the contribution to this integral from one tail rapidly becomes negligible, so that (35) approximates to

$$\frac{1}{\sqrt{2\pi}} \int_{u_{\frac{1}{2}\alpha} - h}^{\infty} e^{-\frac{1}{2}u^2} du.$$

From (32) we see that this integral is the power of the one-tailed test, applied at the significance level  $\frac{1}{2}\alpha$ .

Example 3.<sup>1</sup> Two equal groups of seeds were allowed to germinate in dishes containing filter papers soaked respectively in rain water and in water allowed to seep through loam before use.

---

<sup>1</sup>Biometrika, v. 35, p. 167.

	Germinated	Ungerminated	Total
Loam water	37	13	50
Rain water	32	18	50
Total	69	31	100

To test if the type of water affects germination,  $u$  has been calculated to be 1.11 and referred to the normal probability scale. It is seen that there is no significant difference at the 5% level.

It might be asked what magnitude of difference could we hope to detect using two samples of 50. Suppose, for example, that for these populations 80%, say, of seeds will germinate in loam water and 60% in rain water; what would have been the chance of establishing significance at 5% level?

To obtain this, we find from Table 6 the value of  $k$  for  $p_1 = 0.8$  and  $p_2 = 0.6$ , which is seen to be 0.318. Then we enter Table 7 in the column of  $n = 50$  and find that this value of  $k$  lies between the tabulated values 0.311 and 0.325. They correspond to the figures 0.5949 and 0.6331 in the column of  $P$  for  $\alpha = 0.05$ . So, the approximation to the power lies between these values and by linear interpolation is found to be 0.61. If the level chosen for the test is  $\alpha = 0.01$ , we find that the power is only 0.37. This



P of equation (35)		Values of k according to h and n						
h \ $\alpha$	0.05	0.01	10	15	50	100	150	n \ h
0.1	0.0511	0.0104	0.032	0.026	0.014	0.010	0.008	0.1
	Values for h=0.2(0.1)2.1 are omitted.				Values of n=20(5)45 are omitted.			
2.2	0.5949	0.3535	0.696	0.568	0.311	0.220	0.180	2.2
2.3	0.6331	0.3913	0.727	0.594	0.325	0.230	0.188	2.3
	Values for h=2.4(0.1)3.0 are omitted.				Values of n=60(10)90 are omitted.			
3.2	0.8925	0.7337	1.012	0.826	0.453	0.320	0.261	3.2
3.4	0.9251	0.7951	1.075	0.873	0.481	0.340	0.278	3.4
	Values for h=3.6(0.2)5.0 are omitted.							

Table 7. Relating Power and k, n, n and  $\alpha$ . This table is an excerpt from a table in Biometrika, v. 35, p.175.

indicates that with two samples of 50, there is a very considerable risk of failing to establish significance when the difference in chances of germination is of this order.

Suppose now we asked how large the samples should have been to give a probability of 0.9 of establishing significance when the true percentages germinating are 80 in loam water and 60 in rain water? We then proceed as follows: In Table 7, entering the column of P under  $\alpha = 0.05$ , we find that 0.90 lies between 0.8925 and 0.9251 and following the rows of these figures we see that  $k = 0.318$  lies between the figures in the columns of  $n = 100$  and  $n = 150$ . As the interval is too wide for interpolation we find  $h$  in column 1 corresponding to  $P = 0.90$  in column 2 and then from the relation  $n = h^2/k^2$  we obtain  $n$  to be nearly 105. If the level chosen is  $\alpha = 0.01$ , the samples should be of size 150 to give the same power.

### Summary

We have investigated the viewpoints from which four different authors have approached the problem of the  $2 \times 2$  table, and have developed the significance test for the  $2 \times 2$  table that was proposed by each of these authors. No attempt has been made to distinguish which, in our opinion, is the "best" test. We have presented these tests from the viewpoint of the respective author; therefore, it is up to the reader to choose, if he so desires, which test he believes to be the "best" test.

Finally, we have shown how the power function may play a useful part in indicating, before the data are collected, how large the sample should be to avoid an inconclusive result, and developed one of Patnaik's methods for approximating the power function.

## BIBLIOGRAPHY

- Fisher, R. A., Statistical Methods for Research Workers, 10th Edition, Hafner, New York, 1948.
- Hoel, P. G., Introduction to Mathematical Statistics, 2nd Edition, Wiley, New York, 1954.
- Kendall, M. G., Advanced Theory of Statistics, volume 1, Charles Griffin Co., London, 1947.
- Kenny, J. F., and Keeping, E. S., Mathematics of Statistics, Part II, 2nd Edition, D. Van Nostrand, New York, 1951.
- Mainland, D., Herrera, L., and Sutcliffe, M.I., Tables for Use With Binomial Samples, NYU College of Medicine, New York, 1956.
- \_\_\_\_\_, Nature (A Weekly Journal of Science), Macmillan, New York.
- \_\_\_\_\_, Biometrika, (A Journal for the Statistical Study of Biological Problems), University Press, Cambridge, England.