A NOTE ON THE PRACTICAL INTERPRETATION OF 2×2 TABLES

BY

J. H. EDWARDS

From the Department of Social Medicine, University of Birmingham

There have been intensive studies of the problem of whether the proportions into which any set is doubly dichotomized could reasonably be due to chance, and numerous tests and tables are available, although there is still disagreement about some fundamental assumptions underlying most test procedures. In practice, this problem should not commonly arise, since, outside genetics, there is almost always an *a priori* expectation of some interaction, and, in such cases, a significance test is merely a method of demonstrating whether the size of the sample is sufficient to exhibit this interaction. In most cases the experimenter or observer is mainly interested in the extent of the interaction, and its confidence limits.

There are three situations in which data are commonly assembled into 2×2 tables.

- (1) The classical situation of paired attributes.
- (2) Two populations, defined by some attribute, further dichotomized by the more or less arbitrary division of some continuously distributed variate.
- (3) The dichotomizing of two continuously distributed variates.

CASE No. 1

In the case of paired attributes, it is necessary to distinguish between two distinct situations, which may be termed simple sampling and compound sampling.

In simple sampling, a supposedly random sample is drawn from some population, and then classified into a 2×2 table by two pairs of exhaustive and exclusive attributes.

In this case, taking the classification

a	b
c	d

,

the most meaningful expression of any difference is that between the proportions

$$\frac{a}{a+b} \text{ and } \frac{c}{c+d} \text{ or } \frac{b}{a+b} \text{ and } \frac{d}{c+d}.$$

Any difference in proportion may be most meaningfully expressed as either a ratio or a difference depending on circumstances.

Consider the case when a difference is relevant (for example, the cure rate between two remedies).

If a, b, c, d represent

	NOT CURED	CURED
NOT TREATED	a	b
TREATED	c	d

,

the appropriate estimator is, of course,

$$\frac{d}{c+d} - \frac{b}{a+b},$$

with variance

$$\frac{ab}{(a+b)^2} + \frac{cd}{(c+d)^2}.$$

On the other hand, if the ratio is relevant (for example in assessing the value of an immunization procedure against a disease), when a, b, c, d refer to

	NOT ATTACKED	ATTACKED
NOT IMMUNIZED	a	b
IMMUNIZED	c	d

,

the appropriate estimator is

$$\frac{\frac{d}{c+d}}{\frac{b}{a+b}},$$

the sampling variance of which is not simple.

In practice, b and d are often so small compared with a and c that the approximation:

$$\frac{\frac{d}{c+d}}{\frac{b}{a+b}} \approx \frac{d/c}{b/a} = \frac{ad}{bc}$$

will allow a simple solution with trivial bias.

In compound sampling, a subset of a population specified by some attribute, for example a disease, is classified in terms of another pair of exhaustive and exclusive attributes, and then the parent population is sampled and classified similarly. The situation considered by Woolf (1955) is that of blood groups and disease. When the subset represents an extreme minority (as is usual in disease), simple and compound sampling differ trivially.

In this case, the most meaningful expression of the interaction is the liability of one attribute to be related to another relative to the liability of the opposite of the first attribute: a relationship specified by the ratio of the products of the diagonals (the cross-ratio). For example, if one considers a random sample of p_t patients with some disease who smoke, and p_o who do not, compared with q_t and q_o controls respectively, randomly selected from the parent population, the relative liability of smokers to this disease compared with that of non-smokers is

$$\frac{p_t/p_o}{q_t/q_o} = \frac{p_t q_o}{q_t p_o}.$$

Woolf (1955) pointed out that the natural logarithm of this ratio is a maximum likelihood estimator with sampling variance

$$\frac{1}{p_t} + \frac{1}{q_t} + \frac{1}{p_o} + \frac{1}{q_o}.$$

Haldane (1956) has shown that an almost least biased estimator is given by

$$\log_e \frac{(p_t + \frac{1}{2})(q_o + \frac{1}{2})}{(q_t + \frac{1}{2})(p_o + \frac{1}{2})},$$

with variance

$$\frac{1}{p_t + 1} + \frac{1}{q_t + 1} + \frac{1}{p_o + 1} + \frac{1}{q_o + 1}.$$

This provides an elegant and meaningful interpretation of such data, the confidence limits can be calculated directly, and also if desired the critical ratio on the null hypothesis. Being efficient, the unbiased combination of data from various sources is practicable, and this does not depend on any equi-proportion in the expectation of the marginal totals: nor is the relevance of the subset of samples of identical marginal totals invoked.

The classical method of assessing significance is by the estimator

$$\frac{ad - bc}{N^2}$$

which has, on the null hypothesis, an expectation of zero and a sampling variance

$$\frac{m^4}{N^3},$$

where m is the geometric mean of the marginal totals. Although efficient in the limiting case of no interaction, this estimator has no intelligible meaning in most applications, and, in all three cases considered above, its use in assessing the significance of sets of data from various sources will lead to spurious heterogeneity unless the marginal totals are equi-proportional.

If these three methods of estimation are compared on one sample, and the significance of any discrepancy from the null hypothesis assessed, they will be found to give results which differ only trivially.

For example, consider the 2×2 table:

750	250
700	300

ESTIMATOR	VALUE	CRITICAL RATIO
$\frac{ad - bc}{N^2} \pm \left(\frac{m}{N}\right)^2 \frac{1}{\sqrt{N}}$	$+ 0.01250 \pm 0.004992$	2.504
$\log_e \frac{ad}{bc} \pm \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$	$+ 0.2513 \pm 0.1004$	2.504
$\frac{b}{a+b} - \frac{d}{c+d} \pm \sqrt{\frac{ab}{(a+b)^2} + \frac{cd}{(c+d)^2}}$	$- 0.0500 \pm 0.01994$	2.508

CASE No. 2

When some variate, common to two populations defined by attributes, is dichotomized, the most meaningful expression of the difference is the difference in units of standard deviation between the means of the population (assuming that the standard deviation is the same in both populations). If the dichotomy is not into grossly unequal parts, a close approximation to this difference may be computed, with confidence limits, by making use of the approximation of the first derivative of the logistic function

$$\left(\frac{1}{2(1 + \text{Cosh } x)} \right)$$

to the normal curve.

The method is essentially the comparison of medians, a usually less efficient but simpler alternative to the comparison of means. Although relatively inefficient in the case in which the parent distribution is normal, it may be more efficient in leptokurtic distributions and has the particular advantage of being less influenced by either the inclusion or the rejection of outliers, a feature which may be of great advantage in biology, and particularly in the biology of man.

In the case of random sampling from two populations with a normally distributed variate, the variances being equal, but the means differing by t units of standard deviation, the 2×2 table may be considered in the following terms where the numerical values of a, b, c, d are proportional to the areas in the Figure.

	-	+
POPULATION 1	a	b
POPULATION 2	c	d

In Population 1, the dichotomizing line is distant say t_1 from the median where

$$\frac{a}{b} = \frac{-\int_{-\infty}^{t_1} \frac{e^{-\frac{1}{2}t^2}}{\sqrt{2\pi}} dt}{\int_{t_1}^{+\infty} \frac{e^{-\frac{1}{2}t^2}}{\sqrt{2\pi}} dt}$$

t_1 has the sampling variance of

$$\frac{ab}{(a+b)f_1^3}$$

where f_1 is the ordinate t_1 units from the median.

Now, to a close approximation,

$$t_1 = \frac{\sqrt{2\pi}}{4} \log_e \frac{a}{b},$$

FIGURE.—Interpretation of a 2×2 table as an arbitrary dichotomy of two normal distributions.

$$\text{and } f_1 = \frac{4}{\sqrt{2\pi}} \frac{ab}{(a+b)}.$$

So that the variance of the estimate is:

$$\frac{ab}{(a+b) \left(\frac{4}{\sqrt{2\pi}} \frac{ab}{(a+b)} \right)^2} = \frac{\pi}{8} \left(\frac{1}{a} + \frac{1}{b} \right).$$

Similarly,

$$t_2 = \sqrt{\frac{\pi}{8}} \log_e \frac{c}{d}, \text{ with variance } \frac{\pi}{8} \left(\frac{1}{c} + \frac{1}{d} \right).$$

So that the difference between the medians, $t = t_2 - t_1$,

$$\begin{aligned} &= \sqrt{\frac{\pi}{8}} \left(\log_e \frac{c}{d} - \log_e \frac{a}{b} \right), \\ &= \sqrt{\frac{\pi}{8}} \left(\log_e \frac{bc}{ad} \right), \end{aligned}$$

with variance

$$\frac{\pi}{8} \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right).$$

This, of course, differs only in a constant term from Woolf's method for estimating the interaction of attributes. It gives a direct answer in meaningful terms and, provided the disproportions are not too extreme (e.g. that the smallest tail is not dichotomized much more than one standard deviation from its mean), the bias is slight and the efficiency high enough to be useful.

The Efficiency Compared with the Estimates of the Mean

Given a single dichotomy, say into numbers a and b , the variance of the mean compared with that of the median is not less than:

$$\frac{1}{\frac{a+b}{2} \left(\frac{1}{a} + \frac{1}{b} \right)},$$

which reaches a maximum of $\frac{2}{\pi}$ (63.7 per cent.).

The efficiency is demonstrated in Appendix A, which also shows the bias.

The simplicity of the analysis, and of the collection and storage of the relevant data, will often outweigh the inefficiency and slight bias. For in experimental and observational data it will often be far easier to subdivide a continuous distribution into two than into many classes, especially when some "sieving" process can be employed, as in cutting a chromatogram, sorting eggs, diagnosing benign hypertension, exposing children to an 11+ examination, etc.; and where the amount of data collected can be several-fold increased by such measurement or classification a more precise estimate may be made of the difference between the medians than would be practicable for the difference between the means.

The method is particularly suited to the analysis of punch-card data, when the four sub-groups may be counted mechanically, and very large numbers handled. As in the previous case, a form of significance test in which the difference must be both appreciable and significant can easily be made by defining a "zone of triviality" which must be excluded from the confidence range.

Haldane's corrections* should be employed in estimating both the difference and its variance.

i.e.

$$\sqrt{\frac{\pi}{8} \log_e \frac{(b + \frac{1}{2})(c + \frac{1}{2})}{(a + \frac{1}{2})(d + \frac{1}{2})}},$$

and

$$\frac{\pi}{8} \left(\frac{1}{a+1} + \frac{1}{b+1} + \frac{1}{c+1} + \frac{1}{d+1} \right).$$

CASE NO. 3

In the third case, a 2×2 table may represent a bivariate population arbitrarily or otherwise dichotomized into quadrants: the most relevant question in relation to such data is the coefficient of correlation, and, because of the closeness of the normal curve to the first derivative of the logistic curve in any bivariate normal population, the cross-ratio is almost independent of the dichotomizing planes, providing they do not dichotomize into grossly unequal proportions, and this cross-ratio is therefore almost uniquely specified by the correlation coefficient. The natural logarithm of this cross-ratio is almost linearly related to Fisher's transformation of the correlation coefficient*.

$$\frac{\pi}{8} \log_e \frac{bc}{ad} \approx z,$$

where

$$z = \frac{1}{2} \log_e \frac{1+r}{1-r},$$

and has a sampling variance

$$\left(\frac{\pi}{8} \right)^2 \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right).$$

This estimate of z has a bias of the order $0.02 z^2$, which is of trivial consequence except when r is large, when the method is also unsuited through inefficiency.

The Efficiency of Estimates Compared with Complete Analysis

The efficiency, compared with a product-moment analysis of completely specified data is, in a large sample, not less than

$$\frac{1}{\frac{a+b+c+d}{2} \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right)},$$

*In Haldane's opinion (personal communication) the term continuity correction is inappropriate in this context.

* A non-rigorous proof in a limiting condition is given in Appendix B.

or, in proportions when

$$a' = \frac{a}{a+b+c+d}, \quad b' = \frac{b}{a+b+c+d}, \quad \text{etc.,}$$

$$\frac{64}{\pi^2 \left(\frac{1}{a'} + \frac{1}{b'} + \frac{1}{c'} + \frac{1}{d'} \right)}.$$

The maximum efficiency (in the case $a = b = c = d$) is $\frac{4}{\pi^2}$ or 40.5 per cent. Although low, its simplicity gives it some advantages where vast data have to be extracted or coded with limited facilities or analysed with limited labour. Where data are only available in a 2×2 table, the method is far simpler than by the interpolation of tetra-choric functions and not less efficient.

The bias under various conditions, and also the cross-ratio when the dichotomies are through the means are shown in Appendix B. The calculations are based on Karl Pearson's tables for the volumes of quadrants of the bivariate normal surface.

As in the previous case, it is simple to test for "appreciable and significant" correlation by defining a zone of triviality. The cross-ratios, and their differences, tabulated in Appendix B, provide a rough method of estimating the correlation coefficient from the cross-ratio. Haldane's corrections should be used.

APPENDIX A

The bias and efficiency of calculating the difference between medians from the cross-ratio in a 2×2 table, one entry of which refers to a dichotomized normally distributed variate, are given below.

" δ " refers to the difference between the medians (or means) of the parent populations (in units of

SUMMARY

The conventional method of testing for interaction in a 2×2 table utilizes an estimator which has no direct meaning, so that it is not possible to dissociate the distinct questions whether an interaction is important and whether it is significant; nor is it usually possible to combine data from different sources without loss of information and bias.

In observational data, it is not usually reasonable to ask the question "is interaction present" (obvious exceptions occur with genetic linkage, blood groups and disease, and differences in response to *equally reasonable* remedies). In practice, interaction can usually be assumed, and its direction is usually obvious, and the only way in which statistics can assist in extracting meaningful information is by a meaningful estimator to which confidence limits may be calculated.

Some common situations in which data are acquired, or extracted, in a 2×2 form are considered, and some meaningful estimators, and methods of deriving their confidence limits, are discussed. One estimator provides a simple approximate method of estimating the regression coefficient from a 2×2 table pertaining to a normal bivariate population.

The concept of a "zone of triviality", as an aid to the detection of differences which are both significant and important, is introduced.

REFERENCES

- Haldane, J. B. S. (1956). *Ann. hum. Genet.*, 20, 309.
Sheppard, W. F. (1897). *Proc. roy. Soc.*, 62, 170.
Woolf, B. (1955). *Ann. hum. Genet.*, 19, 251.

δ	Antimodal		Modal	
	d	η per cent.	d	η per cent.
0.2	0.2001	63.3	0.2004	62.8
0.4	0.4007	62.1	0.4029	60.4
0.6	0.6024	60.1	0.6098	56.4
0.8	0.8058	57.5	0.8235	51.0
1.0	1.0114	54.3	1.0454	44.3

standard deviation), " d " the estimated difference, and " η " the efficiency. Two sets of calculations are made, one in the most efficient dichotomy midway between the means, (called the antimodal), the other through one mean (called the modal).

APPENDIX B

The approximation

$$z \simeq \frac{\pi}{8} \log_e \frac{bc}{ad}$$

may be demonstrated as an equality in the limiting case of a bivariate normal surface of correlation r (when $r \rightarrow 0$) dichotomized through the means.

Consider the relative volumes:

a	b
c	d

$$(a = d, b = c).$$

Then $r = \text{Cosine} \left(\frac{a}{a+b} \right) \pi$ (Sheppard, 1897).

$$= \text{Cosine} \left(\frac{1}{\frac{b}{a} + 1} \right) \pi$$

$$\text{Let } \frac{bc}{ad} = \left(\frac{b}{a} \right)^2 = w.$$

When $r \rightarrow 0$, $w \rightarrow 1$, $1 - \sqrt{w} \rightarrow 0$, $z \rightarrow r$,

and neglecting second order terms

$$\begin{aligned} &= \text{Cosine} \left(\frac{1}{2 - (1 - \sqrt{w})} \right) \pi \\ &= \text{Cosine} \left(\frac{1}{2} + \frac{1}{4} (1 - \sqrt{w}) \right) \pi \\ &= -\text{Sine} \frac{1}{4} (1 - \sqrt{w}) \pi \\ &= -\frac{\pi}{4} (1 - \sqrt{w}) \\ &= \frac{\pi}{4} \log_e \sqrt{w} \\ &= \frac{\pi}{8} \log_e w \\ &= \frac{\pi}{8} \log_e \frac{bc}{ad}. \end{aligned}$$

The cross-ratio of the evenly dichotomized bivariate normal surface, and estimates of r , based on the method described above, of a bivariate normal surface, of correlation r dichotomized (0, 0), (0, 0.5) and (0.5, 0.5) units of standard deviation from the means, are tabulated below.

r	Cross-Ratio (0, 0)	Estimates of r		
		(0, 0)	(0.5, 0.0)	(0.5, 0.5)
0.0	1.00 0.29	0.000	0.000	0.000
0.1	1.29 0.38	0.100	0.103	0.106
0.2	1.67 0.52	0.199	0.206	0.210
0.3	2.19 0.73	0.299	0.310	0.313
0.4	2.92 1.08	0.398	0.412	0.414
0.5	4.00 1.70	0.496	0.514	0.513
0.6	5.70 3.00	0.594	0.616	0.610
0.7	8.70 6.37	0.691	0.718	0.705
0.8	15.07 20.51	0.788	0.820	0.799
0.9	35.58	0.886	0.924	0.893