Med. J. Malaysia Vol XXX No. 2. Dec. 1975

Analysis of 2-category data from two matched samples

By: JAMES LEE
B.Sc. (Brit. Col.), M.Sc. (Brit. Col.), Ph.D. (Manit.), M.I.S.,
Department of Social Medicine and Public Health, University of Singapore, Outram Hill, Singapore 3.

Introduction

In medical and epidemiological investigations, matching of similar subjects is commonly used to minimize the effects of extraneous factors that may confound the factor which is under study (Mainland, 1963; MacMahon and Pugh, 1970). In the case of individual matching (i.e. one control is matched with one case in a case-control study), the analysis is due to McNemar (1947). Pike and Morrow (1970) developed a test which handles any number of controls which are matched to a single case (Appendix 1). However, not infrequently an investigator may want to match a group of controls with a group of cases which need only show a certain degree of similarity in the matching variable(s). In more general terms, a number of subjects in the first group is matched with a number of subjects in the second group. For example, in a clinical trial to compare the effects of two drugs, there may not be enough subjects available which exhibit sufficient similarity in the relevant matching variables, such as sex, age, weight and so on, to enable pairing (individual matching). In an observational study which is based on available records, individual matching may be achieved only at the great expense of data attrition. A good example of data attrition is shown by Christiansen's matched sample of high school graduates with high school dropouts. He had to discard over 96% of his completed interviews in order to finally achieve a matched sample of 46 cases (Chapin, 1955).

In the foregoing situations, group-matching (sometimes called "stratification") can be a good compromise. The principles underlying matched sampling and the various group-matching methods (i.e. interval, frequency or quantile matching and the like) are discussed in MacMahon and Pugh (1970), Althauser and Rubin (1970), and Rubin (1973).

There are several statistical methods which can be adapted for the analysis of dichotomous data from two grouped-matched samples. However, some of these methods are presented technically in specialized statistical literature and therefore are not generally known to the medical investigator.

The objectives of this article are as follows:

(1) To survey some statistical methods that are relevant for the analysis of dichotomous data from two grouped-matched samples.

(2) By means of a numerical example, to illustrate Berkson's (1968) minimum logit chi-square method. This method is, in my opinion most useful in terms of computational simplicity and accuracy.

(3) To facilitate computations, I have written a FORTRAN program of the Berkson analysis. The program, along with detailed user's instructions, is available upon request.

Concepts, Terminology and Examples

It is useful at this point to briefly clarify a few terminologies which will be used repeatedly in the subsequent sections: Factor under study, matching variable, extraneous variables and response variable. In a study of possible association between variables, the factor under study is the variable (i.e. smokers vs. non-smokers) which the investigator suspect to have some influence (but not necessarily a direct

cause) on the outcome of the response variable (cancer/not cancer). A matching variable is a possible confounding variable (i.e race, sex, age, etc..). That is, a matching variable is one that is related to the response variable and for this reason, must be made equivalent or comparable between the two groups (smokers and non-smokers). Imagine if age were related to cancer and one is comparing an older group of smokers with a younger group of non-smokers! Matching will make such "incomparable" comparisons unlikely. In effect, matching will produce similar distributions of the confounding variable (age) in the two groups (smokers and non-smokers) and will therefore make the two groups comparable, so far as this confounding variable is concerned. (For the proper selection of matching variables, see pages 253–256 in MacMahon and Pugh). The extraneous variables are all of the suspected and not suspected confounding variables that go unmatched for various reasons. Therefore, in any investigation, the two groups are seldomly, if ever, perfectly comparable.

The following are a few additional examples which are two group-matched samples with a dichotomous response variable:

(1) A therapeutic trial to compare the effects of amplicillin (a derivative of penicillin) with a standard penicillin (factor under study) on curing urinary tract infection (response variable) in pregnant women. Say after seven days of administration of the antibiotic, a negative culture would be considered "cured" while a positive culture, "not cured". The possible matching variables would be race, age and the physiological state of the woman.

(2) In a survey to compare the working and non-working married women on their attitude towards children, a question such as this could be use: "Having children is the most important function of marriage." The response variable would be agree/disagree. The possible matching variables are the ethnic group, age, religion and educational background of the woman.

(3) Sometimes, an investigator may want to group a quantitative response variable into a dichotomy. Thus, in a survey to compare the haemoglobin levels (response variable) between children living in high altitude and children living in low altitude (factor under study). Haemoglobin level might be grouped into two classes, those with $\leq$ 12 gm% and those with $>$ 12 gm%.

(4) In a prospective epidemiological study, the factor under study is typically the groups that are exposed and not exposed to the suspected "risk" of the disease. The response variable is the presence/absence of the disease. Thus, in a prospective study of the possible association between the use of oral contraceptives and thromboembolic disease in women or reproductive age. The factor under study would be the users (exposed to risk) and non-users (not exposed to risk) of oral contraceptives. Presence/absence of thrombosis would be the response variable. Possible matching variables are ethnic group and age. But note that if the same investigation were to be carried out as a case control retrospective study, then the two groups ("factor under study"), would be the cases (thrombosis) and controls (non-thrombosis) and the "response variable" would be users and non-users of oral contraceptives.

In all the studies of this type, the questions of interest are:

(1) Is the response variable (i.e. percent of the patients cured) differ significantly (in the statistical sense) between the two groups (i.e. amplicillin vs penicillin)?

(2) Is the response variable differ significantly among the levels of the matching variable (i.e. the age of the patients) on the whole?

(3) Does the response variable exhibit significant interaction between the factor under study (amplicillin vs penicillin) and the matching variable (age)? For instance, supposing amplicillin were more 'effective' in curing patients of the younger age group while penicillin, older age group, this would constitute one form of interaction between factor under study and matching variable.

Statistical Background

The basic data generated from such a study might be arranged as follows:

Level of the matching variable	Group 1			Group 2		
1	n_{11}	t_{11}	p_{11}	n_{12}	t_{12}	p_{12}
2	n_{21}	t_{21}	p_{21}	n_{22}	t_{22}	p_{22}
⋮		⋮			⋮	
i	n_{il}	t_{il}	p_{il}	n_{i2}	t_{i2}	p_{i2}
⋮		⋮			⋮	
r	n_{rl}	t_{rl}	p_{rl}	n_{r2}	t_{r2}	p_{r2}

Notations n_{ij} is the number of subjects in the ij^{th} cell; t_{ij} is the number of "positive" (i.e. cures) in the ij^{th} cell; $p_{ij} = (t_{ij}/n_{ij}) \times 100$ is the % "positives" in the ij^{th} cell.

Now, if the response variable had been on the quantitative scale (i.e. body weight), the data would be subjected to a regular analysis of variance (a 2-way classification with unequal subclass numbers). But here, the response variable is dichotomous, resulting in a single value (p_{ij}) for each cell. The linear statistical model associated with this design might be represented as $p_{ij} = P+S_i+G_j+(SG)_{ij}+e_{ij.}$, where S_i is the effect of the i^{th} level of the matching variable, G_j is the effect of the j^{th} group (j=1 or 2) and $(SG)_{ij}$ is the interaction (lack of independence) between S_i and G_j, and $e_{ij.}$ is the random error associated with p_{ij}. The p_{ij} values should not be subjected to the variance analysis as the variance of p_{ij} is a function of p_{ij} (maximum at $P_{ij}=0.5$ and decreases as p_{ij} deviates from 0.5 in either direction). Consequently, the variance of p_{ij} will vary from cell to cell depending on the p_{ij} values in the respective cells. One of the assumptions associated with the analysis of variance model is that the error variance is homogeneous (i.e. the variance of $e_{ij.}$ is constant, within statistical limits, among cells). It is noted that the variance of $e_{ij.}$ is a direct consequence of p_{ij}. For example, for a given n_{ij}, if p_{ij} is closed to 0.5, e_{ij} would be subjected to a larger sampling variance than if p_{ij} were closed to 1 or to 0. To overcome this difficulty, the arcsine transformation of p_{ij} (i.e. $\bar{y}_{ij} = \text{arcsine } (p_{ij})^{0.5}$) might be used. It is known that this transformation will result in the variance of $\bar{y}_{ij}$ approximately equal to $821/(n_{ij}+0.5)$, whatever the p_{ij} value may be. Furthermore, the distribution of $\bar{y}_{ij}$ tends to normal as n_{ij} increases with mean equal to arcsine $(p_{ij})^{0.5}$ and variance within ± 6% $821/(n_{ij} + 0.5)$ for almost all binomial distributions with $n_{ij}p_{ij}$ equal to or greater than one (Keeping, 1962). Treating y_{ij} as a normally distributed continuous variable, the linear statistical model is $\bar{y}_{ij} = \mu + S_i + G_j + (SG)_{ij} + e_{ij.}$, where $e_{ij.}$ is the random error associated with $\bar{y}_{ij}$, the variance of which is $821/(n_{ij} + 0.5)$. This model can be analysed by the variance analysis. To handle the unequal sample size in the various cells, Yate's (1934) method of weighted squares of means can be used. However, the method suggested can only be regarded as approximate. It may be sufficiently accurate for practical purposed providing that (i) n_{ij} in each cell is not too small, (ii) n_{ij} is not grossly different among cells, and (iii) the p_{ij} values are not too far from 0.5 in either direction.

Alternatively, the data may be analysed as a series of 2 x 2 contingency tables (i.e. one table for each level of the matching variable) and then use Cochran's test (Snedecor and Cochran, 1967). However, Cochran's test does not provide answer to the question concerning interaction between factor under study and matching variable.

Grizzle (1961) developed a method to analyse data of this type by the use of maximum likelihood for the estimation of the various 'effects' and used Pearson's chi-square for hypothesis testing. Since Grizzle's maximum likelihood method involves iterative procedures and is therefore computationally

cumbersome, Berkson (1968) introduced the minimum logit chi-square method. Both the Berkson and Grizzle procedures produce numerically similar results but the Berkson method is considerably simpler computationally. The relative merits of the two methods is discussed by Berkson.

Grizzle, Starmer and Koch (1969) and a series of articles that published subsequently dealt with the analysis of categorical data by linear models. They have noted that the Berkson method is a special case (i.e. 2 x 2 x r) of their much more generalised method and therefore both methods produced numerically identical results. However, the Grizzle-Starmer-Koch analysis requires not only vigorous matrix manipulations but also demands some familarity on the understanding of linear models on the part of the user.

Therefore, from the medical investigator's point of view, Berkson's method is, I think, the most logical choice because of computational simplicity and yet produce results similar to the more cumbersome methods.

Numerical Example

Let us suppose that a case-control retrospective study was undertaken to investigate the possible association between the use of oral contraceptives and thromboembolic disease. Married women patients with thrombosis (cases) were matched for age within a 5-year interval with married women without thrombosis (controls) (Appendix 2). Each subject was then ascertained whether or not she had use oral contraceptives. The basic data generated from this investigation is presented in Table 1. Notations, intermediate statistics and the computation of the three Berkson minimum logit chi-square values are shown in Tables II, III and IV, respectively, while the chi-square test of significance is summarized in Table V.

From Table V, it is found that (i) the percent of oral contraceptive users is significantly different ($p < 0.005$) between the case and control groups, (ii) the percent of oral contraceptive users is not significantly different among different age groups on the whole (Appendix 4), and (iii) there is no statistical significant interaction between the groups (cases and controls) and the age levels of thewomen on the percent of oral contraceptive users (Appendix 5).

TABLE I

Basic Data from a Case-Control Study of Association Between the Use of Oral Contraceptives & Thromboembolic Disease

Age Group	Patients	Oral contraceptive pill user?		Total
		Yes	No	
16-20	Cases	3 (20·0)	12	15
	Controls	14 (18·4)	62	76
21-25	Cases	4 (25·0)	12	16
	Controls	12 (17·6)	56	68
26-30	Cases	6 (26·1)	17	23
	Controls	16 (17·4)	76	92
31-35	Cases	12 (40·0)	18	30
	Controls	22 (20·8)	84	106
36-40	Cases	11 (45·8)	13	24
	Controls	20 (23·8)	64	84
41-45	Cases	7 (46·7)	8	15
	Controls	10 (21·7)	36	46

Values within brackets are percentages

TABLE II

Notation Used for the Computation of Berkon's Minimum Logit Chi-squares

K	Group	Response Variable	
		+	−
1	1	3 (a_1)	12 (b_1)
	2	14 (c_1)	62 (d_1)
2	1	4 (a_2)	12 (b_2)
	2	12 (c_2)	56 (d_2)
3	1	6 (a_3)	17 (b_3)
	2	16 (c_3)	76 (d_3)
4	1	12 (a_4)	18 (b_4)
	2	22 (c_4)	84 (d_4)
5	1	11 (a_5)	13 (b_5)
	2	20 (c_5)	64 (d_5)
6	1	7 (a_6)	8 (b_6)
	2	10 (c_6)	36 (d_6)

$K = 1, 2, \ldots\ldots, r$. In this example; $r = 6$

TABLE III

Intermediate Statistics Needed for Berkson's Chi-squares

K	C_{1k}	C_{2k}	$\bar{W}_k$	l_{1k}	l_{2k}	B_k	$\bar{W}_k B_k$	$\bar{W}_k B_k^2$	B'_k	$\bar{W}_k B'_k$	$\bar{W}_k B'^2_k$
1	0·41667	0·08756	1·98322	−1·38629	−1·48808	0·10179	0·20187	0·02055	2·87437	5·70051	16·38537
2	0·33333	0·10119	2·30139	−1·09861	−1·54044	0·44183	1·01682	0·44926	2·63905	6·07348	16·02823
3	0·22549	0·07583	3·31873	−1·04145	−1·55814	0·51669	1·71475	0·88600	2·59959	8·62734	22·42754
4	0·13889	0·05736	5·09554	−0·40546	−1·33977	0·93431	4·76081	4·44808	1·74523	8·89289	15·52014
5	0·16783	0·06562	4·28357	−0·16705	−1·16315	0·99610	4·26686	4·25022	1·33020	5·69800	7·57949
6	0·26786	0·12778	2·52755	−0·13353	−1·28093	1·14740	2·90011	3·32759	1·41446	3·57512	5·05686
Σ	1·55007	0·51534	19·51000	−4·23239	−8·37051	–	14·86120	13·38170	–	38·56734	82·99763

$C_{1k} = \frac{1}{a_k} + \frac{1}{b_k}$ $\quad l_{1k} = l_n(a_k) - l_n(b_k)$

$C_{2k} = \frac{1}{c_k} + \frac{1}{d_k}$ $\quad l_{2k} = l_n(c_k) - l_n(d_k)$

(Appendix 3)

$\bar{W}_k = \frac{1}{C_{1k} + C_{2k}}$ $\quad B_k = l_{1k} - l_{2k}$ $\quad B'_k = l_{1k} + l_{2k}$

TABLE IV

Computation of Berkson's Chi-square Values

Interaction

$$\chi^2 = \sum_{k=1}^{r} \bar{W}_k B_k^2 - \hat{\beta}^2 \sum_{k=1}^{r} \bar{W}_k$$

Where

$$\hat{\beta} = \sum_{k=1}^{r} \bar{W}_k B_k \Big/ \sum_{k=1}^{r} \bar{W}_k$$

$= 14·8612/19·51000 = 0·76172$

$$\therefore \chi^2 = 13·38170 - \left[(0·76172)^2 \times 19·51000\right]$$

$= 2·06$

With $(r-1) = 5$ degrees of freedom

Cases vs. Controls

$$\chi^2 = 0·25\,(\bar{C}_t \lambda^2)$$

Where

$$\bar{C}_t = \sum_{k=1}^{r} C_{1k} + \sum_{k=1}^{r} C_{2k}$$

$= 1·55007 + 0·51534$

$= 2·06541$

$$\lambda = 2\left(\sum_{k=1}^{r} l_{1k} - \sum_{k=1}^{r} l_{2k}\right) \Big/ \bar{C}_t$$

$= 2\left[-4·23239 - (-8·37051)\right] \Big/ 2·06541$

$= 4·00707$

$$\chi^2 = 0·25 \times \left[2·06541 \times (4·00707)^2\right]$$

$= 8·29$ with 1 degree of freedom

Age

$$\chi^2 = \sum_{k=1}^{r} \bar{W}_k B'^2_k - \hat{\beta}'^2 \sum_{k=1}^{r} \bar{W}_k$$

Where

$$\hat{\beta}' = \left(\sum_{k=1}^{r} \bar{W}_k B'_k\right) \Big/ \sum_{k=1}^{r} \bar{W}_k$$

$= 38·56734/19·51000$

$= 1·97680$

$$\therefore \chi^2 = 82·99763 - \left[(1·97680)^2 \times 19·51000\right]$$

$= 6·76$

With $(r-1) = 5$ degrees of freedom

TABLE V

Summary Table Showing the Chi-square Tests of Significance

Source	Degrees of Freedom	Chi-square
Cases vs Control	1	$8 \cdot 29^{**}$
Age	5	$6 \cdot 76^{ns}$
Interaction	5	$2 \cdot 06^{ns}$

** = Highly significant ($p < 0 \cdot 005$)

ns = Not significant ($p > 0 \cdot 05$)

APPENDIX

(1) The Pike-Morrow method should be very useful in the case-control retrospective study of rare diseases (i.e. the cases are difficult to find but the matched-controls are readily available). Thus, to study the possible association between the exposure to diagnostic X-ray during the pregnancy of the mother and leukaemia in children, the cases (laukaemic children) would be first located, and a number of controls (children without leukaemia) would then be matched to each case. The matching variables might be the sex and age of the child. Note that the number of controls need not be the same for each case. From each of the case and control children, we then ascertain whether or not the mother had been exposed to X-ray during her pregnancy.

The McNemar test is a special case of this method. That is, when one control is matched with one case, both methods produced identical numerical results. I have worked out some numerical examples of the Pike-Morrow test.

(2) In any case-control study, the investigator can use either one of two sampling schemes, depending on the availability of the cases. If the cases are relatively difficult to locate, the investigator may want to select a sample of cases first, then the control subjects are selected in order to match with the cases. But if both the cases and controls are readily available, the investigator might select both samples simultaneously and then construct the matched sample with the subjects so selected. The relative merit of these two sampling schemes is briefly discussed in Althauser and Rubin (1970).

I might also point not that matching does not necessarily have to be done at the sampling stage. Matching (or stratification) can also be done at the analysis stage (i.e. after the data have been collected). A possible danger of this approach is that if the distribution of the matching variable in the two samples obtained is grossly different, the samples cannot be matched. For example, if the age of the 'cases' selected ranges from 31 to 42 while that of the 'controls' ranges from 20 to 30, then obviously we cannot match the two samples for age.

(3) When a value (i.e. a_k) is zero, its natural logarithm (1_n) is $-\infty$ and therefore the 1_{jk} value cannot be obtained. In such a case, replace the zero value by 0.5.

(4) Obviously this finding has no relevance to the objective of the investigation. However, had the investigation been a prospective one, then the finding would be of relevance (i.e. effect of age on the percent of cases).

(5) For a more thorough discussion on the meaning of statistical interaction, see Snedecor and Cochran (1967) or Armitage (1971). But briefly, it is this: If the percent of oral contraceptive pill users (Y-axis) is plotted against the age groups (X-axis) for the cases and for the controls, then interaction would result in two non-parallel lines.

Now, upon looking at the following table, the reader may wonder why interaction was not statistically significant.

It is apparent that the two lines diverge in the direction of older age groups. However, if we consider the 95% confidence limits (i.e. allowing for sampling errors) of the respective percentages, we see that most of them overlapped.

	CASE						CONTROL					
Age	16–	21–	26–	31–	36–	41–	16–	21–	26–	31–	36–	41–
% BC pill users	20.0	25.0	26.1	40.0	45.8	46.7	18.4	17.6	17.4	20.8	23.8	21.7
95% CL	4.4– 47.6	7.3– 52.0	10.5 47.8	23.0– 59.0	26.0– 67.0	21.0– 73.0	10.8– 28.6	9.5– 28.7	10.8– 26.4	13.5– 29.6	15.2– 34.4	10.8– 36.2

REFERENCES

1. Althauser, R.P and Rubin, D.B. (1970). The computerized construction of a matched sample. Am. J. Sociology 76: 325–346.
2. Armitage, P. (1971). Statistical methods in medical research. Blackwell, Oxford.
3. Berkson, J. (1968). Application of minumum logit chi square estimate to a problem of Grizzle with a notation on the problem of 'no interaction'. Biometrics 24: 75–95.
4. Chapin, F.S. (1955). Experimental designs in sociological research, Harper, New York.
5. Grizzle, J.E. (1961). A new method of testing hypotheses and estimating parameters for the logistic model. Biometrics 17. 372–385.
6. Grizzle J.E. Starmer, C.F., and Koch, G. G. (1969). Analysis of categorical data by linear models. Biometrics 25: 489–504.
7. Keeping, E.S. (1962). Statistical inference. Van Hostrand, New York.
8. MacMahon, B. and Pugh, T.F. (1970). Epidemiology: Principles and methods. Little, Brown and Co., Boston.
9. McNemar, Q (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12: 153.
10. Mainland, D (1963). Elementary medical statistics. 2nd Ed., Saunders. Philadelphia.
11. Pike, M.C. and Morrow, R.H. (1970). Statistical analysis of patient – control studies in epidemiology. Brit. J. Prev. & Soc. Med. 24: 42–44.
12. Rubin, D.B. (1973). Matching to remove bias in observational studies. Biometrics 29: 159–183.
13. Snedecor. G.W. and Cochran, W.G. (1967). Statistical methods. 6th Ed. Iowa Stat University Press, Ames.
14. Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different subclasses, J. Am. Stat. Assoc. 29: 51–66.