International Journal of Innovation in Science and Mathematics
Volume 6, Issue 1, ISSN (Online): 2347–9051

Matched Pairs Models

Dr. Fatma Abdel-Aty
Professor of Applied Statistic
email id: fabdelaty2010@gmail.com

Areej Abdul Azeem*
Lecturer in the Statistics Dept. at King Abdul Aziz University
*email id: aabdulazim@kau.edu.sa

*Corresponding author

Date of publication (dd/mm/yyyy): 11/01/2018

***Abstract* – Matched pairs data can always arise from measuring a response at two variables on different level and it presented in square contingence table way, where the variables have the same category levels. This paper presents a non-convention approach for the analysis of two dimensional tables, when responses are ordered categories. Two variables are considered, and the analysis is done using data compiled from student records, college of science, King AbdulAziz University (2009/2010). The models of symmetry and marginal homogeneity are examined. To link the two models, the concepts of quasi-symmetry, ordinal quasi-symmetry are introduced. The parametric representation of the models is discussed. The expected cell values under the four models are estimated by maximum likelihood methods. A way of looking on the notion of marginal homogeneity conditional on the models of symmetry and quasi-symmetry, and on the models symmetry and ordinal quasi-symmetry being true are presented. Various relationships under models are reached. The related asymptotic distribution under marginal homogeneity is chi-square distribution.**

***Keywords* – Matched pairs models, Marginal homogeneity, Quasi Symmetry, McNemar like test statistic, Ordinal Quasi-Symmetry.**

I. Introduction

Matched pairs data can always arise from measuring a response at two variables or questions on different levels and it's represented in contingency table way. Contingency table is a type of tables which often used to record and analyze the relation between two or more categorical variables. For the two categorical variables, one variable determines the row categories; the other variable determines the column categories. In other word the contingency table contain observations (subjects) from two samples on two response variables and several categories. The two samples responses are called matched pairs data because each observation from one response variable in the first sample pairs with one and only one observation from the other response variable in the second sample and the tow response variables are dependent. [1].

There is one way to represent the samples of matched pairs. Table (1-1) summarizes the data in contingency table that has same categories response for both classifications.

Where n_{ij} represent the number of subjects who respond in category i for the first variable and category j for the second variable, with corresponding probability π_{ij}.

In this paper the concept of comparing marginal proportion for 2X2 tables and for a square IXI tables is presented in section (II). In section (III) the idea of symmetry and quasi-symmetry models for square tables is introduced along with an ordinal quasi-symmetry model. The empirical study on students' data is presented in section (IV).

Table (1-1). IXI contingency table

First variable	Second variable				Total
	Category 1	Category 2	………	Category I	
Category 1	n_{11} π_{11}	n_{12} π_{12}	………	n_{1I} π_{1I}	n_{1+} π_{1+}
Category 2	n_{21} π_{21}	n_{22} π_{22}	………	⋮	⋮
⋮	⋮	⋮		⋮	⋮
Category I	n_{I1} π_{I1}	n_{I2} π_{I2}	………	⋮	⋮
Total	n_{+1} π_{+1}	…..	………	….	n 1

II. Marginal Proportion and Marginal Homogeneity

Marginal proportion can be defined as the proportions of rows or columns. It can be calculated by dividing either total of rows or columns by the total sample size n. to compare marginal proportion, tests of marginal homogeneity are used. Marginal homogeneity can be arise when the probability of a specific outcome of question one are identical with the probability of the same outcome of question two [2].

Assuming marginal homogeneity, $\pi_{i+} = \pi_{+i}$ which implies in case of 2X2 tables: $\pi_{1+} = \pi_{+1}$ and $\pi_{2+} = \pi_{+2}$ and accordingly $\pi_{12} = \pi_{21}$.

A. Statistical Inference of Marginal Homogeneity for 2X2 Tables

The marginal homogeneity is equivalent to symmetry of probability across the main diagonal.

For testing marginal homogeneity with H_0: $\pi_{12} = \pi_{21}$ in case of small sample size, $n_{12} \sim$ bin ($\pi_{12}+\pi_{21}$, 0.5) with P-value can be obtained from the binomial table [3]. For large sample size the statistical test can be calculated as follows:

$$Z=\frac{n_{12}-n_{21}}{\sqrt{n_{12}+n_{21}}} \qquad (1)$$

With $E(n_{12}) = 0.5$ ($\pi_{12}+\pi_{21}$) and $Var(n_{12}) = 0.25$ ($\pi_{12}+\pi_{21}$). $Z^2 \sim \chi^2_1$ which called MacNemar test introduced by MacNemar in [1947] [3 or 4]. The confidence interval takes the form

$$(p_{1+} - p_{+1}) \pm z_{\alpha/2} \text{ (SE)} \qquad (2)$$

Where $p_{1+} - p_{+1}$ is the difference between the sample marginal proportions which estimate the true difference in population $\pi_{1+} - \pi_{+1}$, and

$$SE = \sqrt{(n_{12}+n_{21}) - (n_{12}-n_{21})^2/n} \; /n$$

B. Comparing Marginal Proportion for a Square IXI Tables

For an IXI tables, let {Y$_1$, Y$_2$} be the observation for a randomly selected subject and {n$_{ij}$} is the cell count of possible outcome (i, j). Let π_{ij} = p(Y$_1$ = i, Y$_2$ = j), so the marginal homogeneity assumption requires:

$$P(Y_1 = i) = p(Y_2 = i) \text{ for I} = 1,\ldots.,I \qquad (3)$$

This means that each row marginal probability equals the corresponding column marginal probability [5].

In case of ordinal classification, ordinal data categorical data where there is a logical ordering of the categories. The model with ordinal classification uses cumulative probabilities. Let the response Y = 1, 2,…,I, where the order is relevant. The associated probabilities are $\{\pi_1, \pi_2, \ldots, \pi_I\}$ where π_I is the probability that the response variable Y take the value i. the cumulative probability of a response less than or equal to i is:

$$P(Y \leq i) = \pi_1 + \pi_2 + \ldots\ldots + \pi_i \qquad (4)$$

Then a cumulative logit, where logit can be defined as logit p = log (p/1-p), is:

$$logit[p(Y \leq i)] = log\left(\frac{P(Y \leq i)}{P(Y > i)}\right) = log\left(\frac{P(Y \leq i)}{1 - P(Y \leq i)}\right) = log\left(\frac{\pi_1 + \cdots + \pi_i}{\pi_{i+1} + \cdots + \pi_i}\right) \qquad (5)$$

Which describes the log-odds of two cumulative probabilities, that are how likely is the response to be category i or below i versus a response to be a category higher than i. this model is a more common way to approach the modeling of an I-categories response variable through the use of cumulative logits as:

$$L_1 = \log\left(\frac{\pi_1}{\pi_2 + \cdots + \pi_I}\right)$$

$$L_2 = \log\left(\frac{\pi_1 + \pi_2}{\pi_3 + \cdots + \pi_I}\right) \qquad (6)$$

$$\vdots$$

$$L_i = \log\left(\frac{\pi_1 + \pi_2 + .. + \pi_i}{\pi_{i+1} + \cdots + \pi_I}\right)$$

Where L$_i$ is the log-odds of falling into or below category i versus falling above it. The model for L$_i$ takes the form:

$$L_i = \alpha_i + \beta \qquad (7)$$

To compare the marginal matched pairs with two ordinal response variables, the model of ordinal logits such as logits of cumulative probabilities take the form:

$$\text{logit } [P(Y_{i1} \leq j)] = \alpha_{ij} + \beta, \text{ logit } [P(Y_{i2} \leq j)] = \alpha_{ij}, j = 1, 2,\ldots.., I - 1, i = 1,\ldots,n \qquad (8)$$

Taking the exponentials of both sides of (8) we get

$$\left(\frac{P(Y_{i1} \leq j)}{P(Y_{i1} > j)}\right) = e^{\alpha_{ij}} e^{\beta}$$

With (9)

$$\left(\frac{P(Y_{i2} \leq j)}{P(Y_{i2} > j)}\right) = e^{\alpha_{ij}}$$

Which means, for matched pairs, that the odds that observation 1 falls in category j or below instead of above category j are e^{β} times the odds for observation 2. Marginal homogeneity assumption imply that $\beta = 0$.

III. Symmetry and Quasi-Symmetry Models for Square Table

The probabilities in a square table satisfy symmetry if

$$\pi_{ij} = \pi_{ji} \qquad (10)$$

For all pairs of cells. Cell probabilities on one side of the main diagonal are a minor image of those on the other side. When symmetry holds, necessarily marginal homogeneity also holds. When I>2, though, marginal homogeneity can occur without symmetry.

A. Symmetry as a Logistic Model and Log-Linear Model

The symmetry condition has the simple logistic form:

$$\text{Log }\left(\frac{\pi_{ij}}{\pi_{ji}}\right) = 0, \text{ for all i and j} \qquad (11)$$

The symmetry has log-linear for expected frequencies μ$_{ij}$ = nπ_{ij} as

$$\text{Log } (\mu_{ij}) = \mu + \lambda_i + \lambda_j + \lambda_{ij} \qquad (12)$$

Where,

Log (μ$_{ij}$) = is the log of the expected cell frequency of the cases for cell (i, j) in the contingency table.

μ = is the overall mean of the natural log of the expected frequencies.

i and j = refer to the categories within the variables.

Therefore:

λ_i = The main effect for variable 1 at level i.

λ_j = The main effect for variable 2 at level j.

λ_{ij}= The interaction effect term.

The Maximum Likelihood fit of the symmetry model has expected frequency estimates

$$\widehat{\mu_{ij}} = \frac{n_{ij} + n_{ji}}{2} \qquad (13)$$

The fit satisfies $\widehat{\mu_{ij}} = \widehat{\mu_{ji}}$. It has $\widehat{\mu_{ii}} = n_{ii}$, a perfect fit on the main diagonal. The residual df for chi-square goodness-of -fit tests equal I(I-1)/2. The standardized residuals for the symmetry model equal

$$r_{ij} = \frac{n_{ij} - n_{ji}}{(n_{ij} + n_{ji})^{1/2}} \qquad (14)$$

The sum of squared standardized residuals, one for each pair of categories, equal χ^2 for testing the model fit [3].

B. Quasi-Symmetry as Loistic and Log-Linear Models

Marginal heterogeneity can be accommodated by the quasi-symmetry (QS) model,

$$\text{Log }\left(\frac{\pi_{ij}}{\pi_{ji}}\right) = \beta_i - \beta_j, \quad \text{for all i and j} \qquad (15)$$

The QS model can be written as log- linear model as:

$$\text{Log } (\mu_{ij}) = \mu + \lambda_i^x + \lambda_j^y + \lambda_{ij}^{xy} \qquad (16)$$

The quasi-symmetry unlike the symmetry does not imply the marginal homogeneity. The likelihood equations are

$$\widehat{\mu_{i+}} = n_{i+,} \quad i = 1, \ldots\ldots, I$$

$$\widehat{\mu_{+j}} = n_{+j,} \quad j = 1, \ldots\ldots, I \qquad (17)$$

$$\widehat{\mu_{ij}} + \widehat{\mu_{ji}} = n_{ij} + n_{ji}$$

Fitting the QS model requires iterative methods [4]. The fitted marginal totals equal the observed totals. Its residual df = (I-1) (I-2)/2.

C. AN Ordinal Quasi-Symmetry Model

The symmetry and quasi-symmetry models treat the classification as nominal. A special case of quasi-symmetry often is useful when the categories are ordinal. Let $u_1 \leq$

$u_2 \le \cdots \le u_I$ denote the ordered score for both the row and column categories. The ordinal quasi-symmetry model is:

$$\text{Log}\left(\frac{\pi_{ij}}{\pi_{ji}}\right) = \beta\,(u_j - u_i) \qquad (18)$$

This is a special case of the quasi-symmetry model (QS) in which $\{\beta_i\}$ have a linear trend. The symmetry model is the special case $\beta = 0$.

The greater the value of $|\,\beta\,|$, the greatest the difference between π_{ij} and π_{ji} and accordingly between the marginal distributions. The fitted marginal have the same means as the observed marginal counts. For the chosen category scores $\{u_i\}$, the sample mean for the row variable is $\sum_i u_i\, p_{i+}$. This equals the row mean $\sum_i u_i\, \hat{\pi}_{i+}$ for the fitted value. A similar equality holds for the column means [3].

D. *Testing Marginal Homogeneity*

Testing marginal homogeneity is equivalent to test the null hypothesis that the symmetry (S) model hold against the alternative hypothesis of quasi-symmetry (QS). The likelihood-ratio tests compare the G^2 goodness of fit statistic:

$$G^2(S\,|\,QS) = G^2\,(S) - G^2\,(QS) \qquad (19)$$

With (I-1) df.

Marginal homogeneity can be tested by H$_0$: symmetry (S) holds against H$_A$: ordinal quasi-symmetry (OQS) holds, with

$$G^2(S\,|\,OQS) = G^2\,(S) - G^2\,(OQS) \qquad (20)$$

with one df.

IV. The Empirical Study

The data used in the analysis are of two dimensions, with the variable for rows having the same categories as the variable for columns. The two variables are students' grades regarding the third and fourth terms of academic year of 2009, college of science, king AbdulAziz University, Saudi Arabia.

The analysis of the data is done by fitting models of symmetry, quasi-symmetry, ordinal quasi-symmetry and quasi-independence model. The observed and expected values, given in table (4-1), under the four models are estimated by the maximum likelihood method. The tests of the models are done along with comparative study. Also, 5X5 table has been partitioned into all different 2X2 type tables, with expected values are estimated by MLE. Interpretation and comparisons among them are performed.

Table (4-1): observed student records of term- Grading data, with expected values under four models.

Table (4-1): students grade Data

The third term	The forth term					Total
	Excellent	very good	Good	Passed	Fail	
Excellent	24 24a 24b 24c 24d	1 3 1 1.46 0.2	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0	25
very good	5 3 5 4.53 0.56	142 142 142 142 142	18 29.5 17.9 14.38 20.87	0 0.5 0.1 0.1 1.56	0 0 0 0 0	165
Good	0 0 0 0 3.65	41 29.5 41.1 44.62 35.2	270 270 270 270 270	8 20.5 7.9 9.99 10.16	0 0 0 0 0	319
Passed	0 0 0 0 0.71	1 0.5 0.9 0.9 6.85	33 20.5 33.1 31 26.43	53 53 53 53 53	0 2 0 .97 0	87
Fail	0 0 0 0 0.08	0 0 0 0 0.76	0 0 0 0 2.9	4 2 0 3.02 0.22	0 0 0 0 0	4
Total	29	185	321	65	0	600

a: Symmetry model.
b: Quasi-symmetry model.
c: Ordinal quasi symmetry model.
d: Quasi—Independent model.

A. *Symmetry Model*

The model states that $\mu_{ij} = \begin{cases} \mu_{ji} & \text{for i } \neq \text{j} \\ n_{ii} & \text{for } i = j \end{cases}$

MLE of μ_{ij} from the symmetry model:

$$\widehat{\mu_{ij}} = \widehat{\mu_{ji}} = \frac{n_{ij} + n_{ji}}{2}$$

Goodness of fit statistics χ^2 and G^2 are 31.8767 and 35.416 respectively with p-value = 0.0004, indicate the symmetry model do not fit the table well.

B. *Quasi-Symmetry model*

Since symmetry is so restrictive, the restriction can be removed. So, to fit the quasi–symmetry model (QS), methods of fitting symmetry must be modified. In SAS program. The calculate values of $^2 = 0.2$ and $G^2 = 0.1$ with 6 df attained the p-value of 0.9998, which indicated that the QS model fit the data very well.

C. *An Ordinal Quasi-Symmetry Model*

With ordered score for both row and column categories, the fitted ordinal quasi-symmetry model result in Goodness of fit statistics $\chi^2 = 3.31$ and $G^2 = 4.34$, which with 9 df have p-value of 0.9507. The model fit the data of the contingency table very well.

D. *Quasi–Independence Model*

The quasi-independence model (QI) is given as:

$$\log(\mu_{ij}) = \lambda + \lambda_i^{term1} + \lambda_j^{term2} + \delta_i I(i = j)$$

Fitting quasi-independence as a generalized linear model has been done using SAS program. The values of $\chi^2 = 122.339$ and $G^2 = 62.5$ with 11 df refereed to poor fit of the model to data set.

E. *Marginal Homogeneity*

The test is given as:

G^2 (marginal homogeneity) = G^2 (symmetry) - G^2 (quasi symmetry) with $df = I - 1$.

Table (4-2) shows the summary of the tests of the different models with different df.

Model	df	G^2
symmetry	10	35.416
Quasi-symmetry	6	0.1
Ordinal Quasi-symmetry	9	4.34
Quasi Independence	11	62.5
Marginal homogeneity	4	35.316

Table (4-2): Goodness of fit G^2 of the different models

It can be included from table (4-2) that symmetry model (p-value = 0.0004) do not fit the data well. Quasi-symmetry model (p-value = 0.9998) and an ordinal quasi-symmetry model (p-value = 0.9507) fit the data very well. Quasi-independence model (p-value = 0.001) do not fit the data. The test of marginal homogeneity (p-value = 0.0001) indicated different distribution for row and column.

Partitioning the original table into many 2X2 tables with each grade classified against the lower one, resulted in the following conclusion

1) *Excellent Versus all Lower Grades*

a. A very good grade: MacNemar test = 2.667 (p-value = 0.102) indicate with odds ratio of 681.6>1, a very strong association between the two term grades.

b. Good, passed and fail grades: MacNemar test = 0 which indicate no association, i.e., student change their grade to two or more less grade.

2) *Very Good Versus all Lower Grades*

MacNemar test = 8.96 (p-value = 0.0028) with odds ratio = 51.95 for good grade MacNemar test = 9.6 (p-value = 0.0019) with odds ratio = 69.17 for passed and fail together. In both case, the null hypothesis is rejected which mean that the student grade of the third term associate with student grade of the fourth term.

V. Conclusion

An approach to analyze and test of marginal homogeneity, symmetry, quasi-symmetry and ordinal quasi-symmetry for tables of matched pairs data is presented in this paper, with categorical response from two dependent samples. The method of iterative proportional fitting to compute estimated expected cell values variable direction to the other variable to see the presence of consistency and compatibility in the proportions of the margins is reviewed.

The variables used in the study are "Third term grade" and " Forth term grade". The models of quasi-symmetry and ordinal quasi-symmetry are found to be the most appropriate models. The MacNemar-like test statistic is computed.

For 5X5 table, the quasi-symmetry model fits well. The simpler ordinal quasi-symmetry also fits well, with significant test of marginal homogeneity.

For 2X2 different tables comparing excellent grade with all other grades, the student tend to keep their excellent grade from the third term to the fourth term, this also is confirmed by the very high odds ratio values.

References

[1] Abdel-Aty, F. A. (1990). Symmetry and Marginal Homogeneity of an IxI Contingency Table: a Log-Linear Model Approach. *PROC. INT. CONF. ST., COMP. SC., SOC. RES. and DEM.,* Ain Shams University.

[2] Abdel-Aty, F. A. (1993). Analyzing Agreement among Multiple Observers for Categorical Data. *The Egyptian Journal for comme--rce Studies, Faculuty of commerce, Mansoura University.*

[3] Agresti, A. (2002). *Categorical Data Analysis*, 2nd end. New Jersey: John Wiley.

[4] Agresti, A. (2007). An *Introduction to Categorical Data Analysis*, 2nd edn. New Jersey: John Wiley.

[5] Bishop, Y. M. M, Fienberg, S. E., and Holland, P. W. *(*2007). *Discrete multivariate analysis: theory and practice.* New York: MIT Press.