Parametric Test of Equality of Two Frequency Distributions or Matrices

Silvia Irin Sharna[1]+ and Mian Arif Shams Adnan[2]* and Md Shamsuddin[3]
[1,2]Department of Statistics, Jahangirnagar University, Dhaka 1342, Bangladesh
[3]Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia.

ABSTRACT

The parametric test for testing the equality of two frequency distributions has been developed. The tests of the equality of two contingency tables or two frequency matrices and two transition frequency matrices have also been developed. Examples are cited for all cases.

*To whom correspondence should be addressed at julias284atyahoo.com

Key words: Chi-square matrix, *p*-value matrix, Multi-level multi-variate vector.

1. INTRODUCTION

A frequency table is defined as a summarized grouping of data divided into mutually exclusive classes and the number of occurrences in all cells. Each cell in the table contains the frequency or count of the occurrences of values within a particular interval or group. Multivariate joint frequency distributions are often presented as (multi-way) contingency tables. As for example, a bivariate joint frequency distribution is presented as a two way contingency table or matrix where the total row and total column report the marginal frequencies or marginal distributions and each cell of the body of the table does report the joint frequencies.

The term contingency table was first coined by Karl Pearson (1904). He proposed chi-square goodness of-fit test for the analysis of a 2×2 contingency table. Fisher (1935) inaugurated the randomization of two-factor association using the extended hypergeometric distribution. Testing the independence in a 2 × 2 table was due to Fisher (1925, 1935) and Neyman and Pearson (1928). Barnard (1945, 1949) discussed the Convexity-Symmetry-Maximum (CSM) triple-condition test based on the sample space of the two independent binomial models. Another classic unconditional test was proposed in the 1950 which was a mixture of the exact conditional tests (Bennet and Hsu (1960)). In certain designs of experiments, a random sample is often selected from the entire population to assess the odds of having the attribute A in the two subpopulations (e.g., Lehmann (1986, Sec. 4.7)). An information theoretic approach to the evaluation of 2×2 contingency tables was proposed by Cheng *et al* (2008). By investigating the relationship between the Kullback-Leibler divergence and the maximum likelihood estimator, information identities are established for testing hypotheses, in particular, for testing independence. Klugkist, I et al (2010) proposed a test of equality of constrained hypotheses for contingency tables using Bayesian analysis. Agresti A. et al (2005) demonstrated multivariate tests comparing binomial probabilities using a pooled variance

Wald test statistic. It offers the checking of comparing probabilities but not the all individual and marginal probabilities and is not proposed for multiple multivariate samples. Yalonetz, G. (2009) authored a working paper with heterogeneity indicies with the ratio of Pearson's goodness of fit statistics.

Therefore, no parametric test is developed so far (by numerous authors referred in the reference) for testing the equality of two contingency tables or two joint frequency distributions or two marginal frequency distributions. The authors aim to develop new parametric test statistics for checking the similarity or dissimilarity between the individual (cell) frequencies, marginal frequencies and overall discrepancy of two populations.

However, a stochastic process or random process is a collection of random variables that represents the evolution of some physical process through the change of time, state or space. There are several (often infinitely many) directions in which the process may evolve. In case of discrete time, a stochastic process amounts to a sequence of random variables known as Markov chain. And the other is a random field, whose domain is a region of space, or random function whose arguments are drawn from a range of continuously changing values. One approach to stochastic processes treats them as functions of one or several deterministic arguments whose values (outputs) are random variables: non-deterministic (single) quantities which have certain probability distributions. Random variables corresponding to various times (or points, in the case of random fields) may be completely different. Although the random values of a stochastic process at different times may be independent random variables, in most commonly considered situations they exhibit complicated statistical correlations. Assessing these correlations can be evaluated by means of knowing transitions which express the changes of state of the system and the probabilities associated with various state-changes are called transition probabilities. Markov chain, due to Andrey Markov, is a mathematical system that undergoes transitions from one state to another, between a finite or countable number of possible states. It is a random process characterized as memoryless stating the conditional probability distribution for the sequence in the system at the next step (and in fact at all future steps) depending only on the current stat, and not additionally on the state at previous steps. So, a Markov Chain is completely characterized by the set of all states and transition probabilities. By convention, we assume all possible states and transitions have been included in the definition of the Markov processes in such a way that there is always a next state and the process goes on forever. Thus, Markov chains have many applications as statistical models of real-life processes.

Checking the discordance of two Markov Chains is a preliminary step of finding the mobility of any system over the change of time or place or other dimension(s). Muse *et al* (1992) proposed a likelihood ratio test for the equality of evolution rates. Tan *et al* (2002) developed a Markov-chain-test for time dependence and homogeneity using likelihood ratio test statistic. Dannemann *et al* (2007) proposed a method of testing the equality of transition parameters based on transition probabilities and likelihood ratio test statistic that simply gives the significant dissimilarity of the total transition but not that of the individual transition. Falay, B. (2007) described intergenerational income mobility by testing the equality of opportunity due to knowing the comparison of East and West Germany using a transition matrix having positive and negative elements. Bartolucci, F. *et al* (2010) demostarted the use of a multidimensional extension of the latent Markov model using a multidimensional two parameter logistic model where they developed likelihood ratio test based on log of the ratio of transition probabilities. Cho, J. S *et al*

(2012) expresses a test of equality of two unknown positive definite matrices with an application of information matrix testing. Hillary, R. M. (2011) proposed a Bayesian method of estimation the growth transition matrices. Altug, S *et al* (2011) showed the cyclical dynamics of industrial production and employment over developed and developing countries. Recently a new statistical method of Pair-wise sequence alignment has been developed by Adnan *et al* (2011). It accomplishes not only an overall decision of the significant similarity/dissimilarity but also the similarity/dissimilarity of all possible individual and group wise transitions that help the biotechnologists to quickly identify the portion of the total infrastructure of the entire transitions that is significantly differing from that of the other sequence and detect the core fact(s) for possible differences between bio-organisms.

The present study aims to improve the comparison method of two transition probability matrices considering more analysis of transition probabilities of the two sampled transition probability matrices. The author introduces an idea of using the difference of pair wise transition probabilities of the two transition probability matrices which will ensure three advantages at least. Firstly, it will find the degree of disorderness between all possible individual and groupwise transition probabilities of states of two Markov chains; and secondly, will reduce the incompleteness of comparison between the two chains from the two unknown populations. Thirdly, it clearly identifies the portion of the total infrastructure of the entire transition that is significantly differing from that of the other chain.

2. METHODS AND METHODOLOGY

2.1 Test of Equality of Two Contingency Tables and Test of Equality of Two Frequency Distributions

With an aim of finding a test for comparing two contingency tables, let us demonstrate our method assuming that we have two population contingency tables or matrices and let the hypothesis be

$$H_0: N = M$$

$$\Rightarrow H_0: \begin{pmatrix} N_{11} & N_{12} & \cdots & N_{1c} \\ N_{21} & N_{22} & & N_{2c} \\ & \vdots & & \\ N_{r1} & N_{r2} & & N_{rc} \end{pmatrix} = \begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1c} \\ M_{21} & M_{22} & & M_{2c} \\ & \vdots & & \\ M_{r1} & M_{r2} & & M_{rc} \end{pmatrix}$$

$$\therefore H_0: P = Q$$

$$\Rightarrow H_0: \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1c} \\ p_{21} & p_{22} & & p_{2c} \\ & \vdots & & \\ p_{r1} & p_{r2} & & p_{rc} \end{pmatrix} = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1c} \\ q_{21} & q_{22} & & q_{2c} \\ & \vdots & & \\ q_{r1} & q_{r2} & & q_{rc} \end{pmatrix}.$$

where, the N and M are two population frequency matrices or contingency tables; P and Q are the two population probability matrices or contingency tables such that $P =$

$(p_{ij})_{r\times c}$, . $Q = (q_{ij})_{r\times c}$, where $p_{ij} = \frac{N_{ij}}{N_{..}}$, $q_{ij} = \frac{M_{ij}}{M_{..}}$ whereas N_{ij} and M_{ij} are the population frequencies of the $(i,j)^{th}$ element of the population frequency matrices N and M of the first population and second population respectively and $N_{..} = \sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij}$; $M_{..} = \sum_{i=1}^{r}\sum_{j=1}^{c} M_{ij}$; $\forall\, i = 1,2,\dots,r; j = 1,2,\dots,c.$

k pairs of sample contingency tables from two population joint frequency distributions (a total of *k* samples are collected from each population) have been collected and on the basis of these samples we want to test whether they come from the same population. After collecting *k* sample-frequency matrices or tables from each of the two populations, the maximum likelihood estimators of the probability matrices are obtained as $\hat{P} = (\hat{p}_{ij})_{r\times c}$ where $\hat{p}_{ij} = \frac{n_{ij}}{n_{..}}$ whereas n_{ij} is the average frequency of the $(i,j)^{th}$ element of the average frequency matrix n constructed from *k* sample-frequency tables drawn from the 1st population and $\hat{Q} = (\hat{q}_{ij})_{r\times c}$ where $\hat{q}_{ij} = \frac{m_{ij}}{m_{..}}$ whereas m_{ij} is the average frequency of the $(i,j)^{th}$ element of the average frequency matrix m constructed from *k* sample-frequency matrices drawn from the 2nd population. Here, $n_{..} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$; $m_{..} = \sum_{i=1}^{r}\sum_{j=1}^{c} m_{ij}$; $\forall\, i = 1,2,\dots,r; j = 1,2,\dots,c.$

For large $n_{..}$ and $m_{..}$ the asymptotic distribution of each element of average relative frequency matrices, according to the Central Limit Theorem, is normal such that

$$\hat{p}_{ij} \sim N\left(p_{ij}, \frac{p_{ij}\ (1-p_{ij})}{kn_{..}}\right) \textit{and} \text{ and } \hat{q}_{ij} \sim N\left(q_{ij}, \frac{q_{ij}\ (1-q_{ij})}{km_{..}}\right).$$

$$\therefore (\hat{p}_{ij} - \hat{q}_{ij}) \sim N\left[(p_{ij} - q_{ij}), \quad \frac{1}{k}\left(\frac{p_{ij}\ (1-p_{ij})}{n_{..}} + \frac{q_{ij}\ (1-q_{ij})}{m_{..}}\right)\right];$$
$$\forall\, i = 1,2,\dots,r; j = 1,2,\dots,c.$$

Therefore, $\begin{bmatrix} \hat{p}_{11} - \hat{q}_{11} \\ \vdots \\ \hat{p}_{rc} - \hat{q}_{rc} \end{bmatrix}$ is a multivariate (rc variate) vector such that

$$\begin{bmatrix} \hat{p}_{11} - \hat{q}_{11} \\ \vdots \\ \hat{p}_{rc} - \hat{q}_{rc} \end{bmatrix}$$
$$\sim N\left(\begin{bmatrix} p_{11} - q_{11} \\ \vdots \\ p_{rc} - q_{rc} \end{bmatrix}, \frac{1}{k}\begin{bmatrix} \frac{p_{11}\ (1-p_{11})}{n_{..}} + \frac{q_{11}\ (1-q_{11})}{m_{..}} & \cdots & -\left(\frac{p_{11}\ p_{rc}}{n_{..}} + \frac{q_{11}\ q_{rc}}{m_{..}}\right) \\ \vdots & \ddots & \vdots \\ -\left(\frac{p_{rc}p_{11}}{n_{..}} + \frac{q_{rc}q_{11}}{m_{..}}\right) & \cdots & \frac{p_{rc}\ (1-p_{rc})}{n_{..}} + \frac{q_{rc}\ (1-q_{rc})}{m_{..}} \end{bmatrix}\right)$$

Although the concern proofs are very much trivial, are available from the author. However, after dividing each element of the difference matrix by their respective standard error, we obtain an element-standardized-matrix *Z* of the following form

$$Z = \begin{pmatrix} \dfrac{\hat{p}_{11} - \hat{q}_{11}}{\sqrt{\frac{1}{k}\left(\frac{p_{11}\ (1-p_{11})}{n_{..}} + \frac{q_{11}\ (1-q_{11})}{m_{..}}\right)}} & \cdots & \dfrac{\hat{p}_{1c} - \hat{q}_{1c}}{\sqrt{\frac{1}{k}\left(\frac{p_{1c}\ (1-p_{1c})}{n_{..}} + \frac{q_{1c}\ (1-q_{1c})}{m_{..}}\right)}} \\ \vdots & \ddots & \vdots \\ \dfrac{\hat{p}_{r1} - \hat{q}_{r1}}{\sqrt{\frac{1}{k}\left(\frac{p_{r1}\ (1-p_{r1})}{n_{..}} + \frac{q_{r1}\ (1-q_{r1})}{m_{..}}\right)}} & \cdots & \dfrac{\hat{p}_{rc} - \hat{q}_{rc}}{\sqrt{\frac{1}{k}\left(\frac{p_{rc}\ (1-p_{rc})}{n_{..}} + \frac{q_{rc}\ (1-q_{rc})}{m_{..}}\right)}} \end{pmatrix}$$

$$= \begin{pmatrix} Z_{11} & \cdots & Z_{1c} \\ \vdots & \ddots & \vdots \\ Z_{r1} & \cdots & Z_{rc} \end{pmatrix}$$

Now squaring each element of the *Z* matrix, the matrix of chi-squares is obtained as below

$$\chi 2 = \begin{pmatrix} Z_{11}^2 & \cdots & Z_{1c}^2 \\ \vdots & \ddots & \vdots \\ Z_{r1}^2 & \cdots & Z_{rc}^2 \end{pmatrix},$$

$$\therefore \chi 2 = \begin{pmatrix} \chi_{11}^2 & \cdots & \chi_{1c}^2 \\ \vdots & \ddots & \vdots \\ \chi_{r1}^2 & \cdots & \chi_{rc}^2 \end{pmatrix}.$$

The above matrix can also be called as element-chi-square-matrix since each of its elements is an individual chi-square. Using this matrix we can test four types of hypotheses which are as follows:

(*i*) $H_0: p_{ij} = q_{ij}$; or, the hypothesis of testing the equality of the each population probabilities-pair of the two population probability matrices *P* and *Q*.

(*ii*) $H_0: (p_{i1} \quad p_{i2} \quad \cdots \quad p_{ic}) = (q_{i1} \quad q_{i2} \quad \cdots \quad q_{ic})$; or, the hypothesis of checking the equality of the i-th row probability vector or frequency distribution of the 1st population probability matrix or table and that of the 2nd population probability matrix or table. Actually, it tests the equity of the frequentness of the ith variable of the first category over all intervals of the second category of two population contingency tables. Indeed the equality of the frequency distribution of the ith variable of the 1st category is tested over two populations. That is, two (types of) frequency distributions are being tested whether equal or not for same variable. So, over a variable the equity of two frequency distributions drawn from two populations is being tested.

(*iii*) $H_0: [p_{1j} \quad p_{2j} \quad \cdots \quad p_{rj}] = [q_{1j} \quad q_{2j} \quad \cdots \quad q_{rj}]$; or, the hypothesis of checking the equality of the j-th column vector of the 1st population probability matrix and that of the 2nd population probability matrix. Actually, it tests the equity of the frequentness of the jth variable of the second category over all variables of the first category of two population contingency tables. The frequency distribution of the jth variable of the 2nd category is tested whether equal or not over two populations.

(*iv*) $H_0: P = Q$; or the hypothesis of testing the equity of the total contingency table or matix for one population is significantly varying to that of the other population. It tests the similarity of two populations where each of the two populations has joint frequency distributions over *rc* cells or whether the two types of sample-joint frequency distributions or matrices or tables are drawn from same population.

For the aforementioned tests for two populations, the concern test statistics are given below respectively

(i) **Test of equality of two [(i,j)th] cell frequencies**: Comparing each χ^2_{ij} with the tabulated $\chi^2_{(1,\propto)}$ of 1 degree of freedom,

(ii) **Test of equality of two [ith variable's] marginal frequency distributions**: Comparing each $\sum_j \chi^2_{ij}$ with the tabulated $\chi^2_{(c,\propto)}$ of *c* degrees of freedom,

(iii) **Test of equality of two [jth variable's] marginal frequency distributions**: Comparing each $\sum_i \chi^2_{ij}$ with the tabulated $\chi^2_{(r,\propto)}$ of *r* degrees of freedom,

(iv) **Test of equality of two joint frequency distributions:** Comparing Chi-squares' matrix sum $= \chi^2_{11} + \cdots + \chi^2_{1c} + \cdots + \chi^2_{r1} + \cdots + \chi^2_{rc}$ with the tabulated $\chi^2_{(rc-1,\propto)}$ of (*rc*-1) degrees of freedom.

2.2 Test of Equality of Two Transition Probability Matrices

For developing a test procedure of the equality of two transition probability matrices or two evolutionary rates from two Markov chains or two sequences, let us demonstrate our method assuming that we have two population transition frequency matrices or two population transition probability matrices or two Markov chains having *r* states and let the hypothesis be

$$H_0: N = M$$

$$\Rightarrow H_0: \begin{pmatrix} N_{11} & N_{12} & \cdots & N_{1r} \\ N_{21} & N_{22} & & N_{2r} \\ & \vdots & & \\ N_{r1} & N_{r2} & & N_{rr} \end{pmatrix} = \begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1r} \\ M_{21} & M_{22} & & M_{2r} \\ & \vdots & & \\ M_{r1} & M_{r2} & & M_{rr} \end{pmatrix}$$

$$H_0: P = Q$$

$$H_0: \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & & p_{2r} \\ & \vdots & & \\ p_{r1} & p_{r2} & & p_{rr} \end{pmatrix} = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1r} \\ q_{21} & q_{22} & & q_{2r} \\ & \vdots & & \\ q_{r1} & q_{r2} & & q_{rr} \end{pmatrix}$$

where, the N and M are two population transition frequency matrices; P and Q are the two population transition probability matrices such that $P = (p_{ij})_{r\times r}, .\ Q = (q_{ij})_{r\times r}$, where $p_{ij} = \frac{N_{ij}}{N_{i.}}$, $q_{ij} = \frac{M_{ij}}{M_{i.}}$ whereas N_{ij} and M_{ij} is the population transition frequency

of the $(i,j)^{th}$ element of the population transition frequency matrices N and M of the first population and second population respectively and $N_{i.} = \sum_{j=1}^{r} N_{ij}\,;\; M_{i.} = \sum_{j=1}^{r} M_{ij}\,; \forall\, i = 1,2,\ldots,r.$

k pairs of sample sequences from two populations (a total of *k* sample-sequences are collected from each population) have been collected and on the basis of these samples we want to test whether they come from the same population. After collecting *k* sample-sequences we obtain *k* transition frequency matrices from each of the two populations. The maximum likelihood estimators of the transition relative frequency or probability matrices are obtained as $\hat{P} = (\hat{p}_{ij})_{r\times r}$ where $\hat{p}_{ij} = \frac{n_{ij}}{n_{i.}}$ whereas n_{ij} is the average frequency of the $(i,j)^{th}$ element of the average transition frequency matrix n constructed from *k* sample-transition frequency matrices drawn from the 1st population and $\hat{Q} = (\hat{q}_{ij})_{r\times r}$ where $\hat{q}_{ij} = \frac{m_{ij}}{m_{i.}}$ whereas m_{ij} is the average frequency of the $(i,j)^{th}$ element of the average transition frequency matrix m constructed from *k* sample-transition frequency matrices drawn from the 2nd population. Here, $n_{i.} = \sum_{j=1}^{r} n_{ij}\,;\; m_{i.} = \sum_{j=1}^{r} m_{ij}\,; \forall\, i = 1,2,\ldots,r.$

Let the difference matrix is *D* such that

$$\widehat{D} = \hat{P}_{r\times r} - \hat{Q}_{r\times r}$$

$$= \begin{pmatrix} \hat{p}_{11} & \hat{p}_{12} \cdots \hat{p}_{1r} \\ \hat{p}_{21} & \hat{p}_{22} \quad \hat{p}_{2r} \\ & \vdots \\ \hat{p}_{r1} & \hat{p}_{r2} \quad \hat{p}_{rr} \end{pmatrix} - \begin{pmatrix} \hat{q}_{11} & \hat{q}_{12} \cdots \hat{q}_{1r} \\ \hat{q}_{21} & \hat{q}_{22} \quad \hat{q}_{2r} \\ & \vdots \\ \hat{q}_{r1} & \hat{q}_{r2} \quad \hat{q}_{rr} \end{pmatrix}$$

$$= \begin{pmatrix} \hat{p}_{11} - \hat{q}_{11} \cdots \hat{p}_{1r} - \hat{q}_{1r} \\ \hat{p}_{21} - \hat{q}_{21} \cdots \hat{p}_{2r} - \hat{q}_{2r} \\ \vdots \\ \hat{p}_{r1} - \hat{q}_{r1} \cdots \hat{p}_{rr} - \hat{q}_{rr} \end{pmatrix}$$

For large $n, n_i,\ m, m_i$; the asymptotic distribution of each element of the estimated transition probability matrices, according to the Central Limit Theorem, is normal such that

$$\hat{p}_{ij} \sim N\left(p_{ij}, \frac{p_{ij}\ (1-p_{ij})}{kn_i}\right) \text{ and } \hat{q}_{ij} \sim N\left(q_{ij}, \frac{q_{ij}\ (1-q_{ij})}{km_i}\right).$$

$$\therefore (\hat{p}_{ij} - \hat{q}_{ij}) \sim N\left[(p_{ij} - q_{ij}),\ \frac{1}{k}\left(\frac{p_{ij}\ (1-p_{ij})}{n_i} + \frac{q_{ij}\ (1-q_{ij})}{m_i}\right)\right].$$

Therefore, $\begin{bmatrix} \hat{p}_{i1} - \hat{q}_{i1} \\ \vdots \\ \hat{p}_{ir} - \hat{q}_{ir} \end{bmatrix}$ is a ith level's multivariate (*r* variate) vector such that

$$\begin{bmatrix} \hat{p}_{i1} - \hat{q}_{i1} \\ \vdots \\ \hat{p}_{ir} - \hat{q}_{ir} \end{bmatrix}$$

$$\sim N\left(\begin{bmatrix} p_{i1} - q_{i1} \\ \vdots \\ p_{ir} - q_{ir} \end{bmatrix}, \frac{1}{k} \begin{bmatrix} \frac{p_{i1}\ (1-p_{i1})}{n_i} + \frac{q_{i1}\ (1-q_{i1})}{m_i} & \cdots & -\left(\frac{p_{i1}\ p_{ir}}{n_i} + \frac{q_{i1}\ q_{ir}}{m_i}\right) \\ \vdots & \ddots & \vdots \\ -\left(\frac{p_{i1}\ p_{ir}}{n_i} + \frac{q_{i1}\ q_{ir}}{m_i}\right) & \cdots & \frac{p_{ir}\ (1-p_{ir})}{n_i} + \frac{q_{ir}\ (1-q_{ir})}{m_i} \end{bmatrix} \right)$$

$\forall\ i = 1, 2, \ldots, r$. Although the concern proofs are very much trivial, are available from the author if required. However, after dividing each element of the difference matrix by their respective standard error, we obtain an element-standardized-matrix *Z* of the following form

$$Z = \begin{pmatrix} \frac{\hat{p}_{11} - \hat{q}_{11}}{\sqrt{\frac{1}{k}\left(\frac{p_{11}\ (1-p_{11})}{n_1} + \frac{q_{11}\ (1-q_{11})}{m_1}\right)}} & \cdots & \frac{\hat{p}_{1r} - \hat{q}_{1r}}{\sqrt{\frac{1}{k}\left(\frac{p_{1r}\ (1-p_{1r})}{n_1} + \frac{q_{1r}\ (1-q_{1r})}{m_1}\right)}} \\ \vdots & \ddots & \vdots \\ \frac{\hat{p}_{r1} - \hat{q}_{r1}}{\sqrt{\frac{1}{k}\left(\frac{p_{r1}\ (1-p_{r1})}{n_r} + \frac{q_{r1}\ (1-q_{r1})}{m_r}\right)}} & \cdots & \frac{\hat{p}_{rr} - \hat{q}_{rr}}{\sqrt{\frac{1}{k}\left(\frac{p_{rr}\ (1-p_{rr})}{n_r} + \frac{q_{rr}\ (1-q_{rr})}{m_r}\right)}} \end{pmatrix}$$

$$= \begin{pmatrix} Z_{11} & \cdots & Z_{1r} \\ \vdots & \ddots & \vdots \\ Z_{r1} & \cdots & Z_{rr} \end{pmatrix}$$

Now squaring each element of the *Z* matrix, a matrix $\chi 2$ each of which matrix is an individual chi-square of the following form is obtained as the matrix of chi-squares,

$$\chi 2 = \begin{pmatrix} Z_{11}^2 & \cdots & Z_{1r}^2 \\ \vdots & \ddots & \vdots \\ Z_{r1}^2 & \cdots & Z_{rr}^2 \end{pmatrix},$$

$$\therefore \chi 2 = \begin{pmatrix} \chi_{11}^2 & \cdots & \chi_{1r}^2 \\ \vdots & \ddots & \vdots \\ \chi_{r1}^2 & \cdots & \chi_{rr}^2 \end{pmatrix}.$$

The above matrix of chi-squares can also be called as element-chi-square-matrix. From this matrix we basically can test three types of hypotheses which are as follows:

(*i*) $H_0: p_{ij} = q_{ij}$; or, the hypothesis of testing the equality of the each population transition probabilities-pair of the two population transition probability matrices *P* and *Q*.

(*ii*) $H_0: (p_{i1} \quad p_{i2} \quad \cdots \quad p_{ir}) = (q_{i1} \quad q_{i2} \quad \cdots \quad q_{ir})$; or, the hypothesis of checking the equality of the i-th row vector of the 1st population transition probability matrix and that of the 2nd population transition probability matrix. Actually, it tests the equity of the

frequentness of the transition of the random movement of two population sequences or Markov chain from each state to all states.

(*iii*) $H_0: P = Q$; or the hypothesis of testing the equity of the total transitions for one population Markov chain or sequence is significantly varying to that of the other population Markov chain or sequence. It tests the similarity of two types population Markov chains or sequences; or whether the two sample sequences are drawn from same population Markov chain.

For the aforementioned tests the concern test statistics are given below respectively.

(i) Comparing each χ^2_{ij} with the tabulated $\chi^2_{(1,.\propto)}$ of 1 degree of freedom,

(ii) Comparing each $\sum_j \chi^2_{ij}$ with the tabulated $\chi^2_{(r-1,.\propto)}$ of (r-1) degrees of freedom,

(iii) Comparing Chi-squares' matrix sum $= \chi^2_{11} + \cdots + \chi^2_{1r} + \cdots + \chi^2_{r1} + \cdots + \chi^2_{rr}$ with the tabulated $\chi^2_{(r(r-1),.\propto)}$ of r(r-1) degrees of freedom.

3. REAL LIFE EXAMPLES

3.1 An Application of the Test for the Equality of Two Contingency tables or Frequency distributions in Environmetrics

Now we are considering two 5×5 average contingency tables obtained from joint frequency distribution of rainfall and temperature for the last 30 years in Dhaka station and Chittagong station of Bangladesh. The distance between the two stations is about 250 miles. Dhaka is in the center and Chittagong is in the south eastern coastal region of Bangladesh. For each of the last 30 years, the days which had rainfall have been considered along with the information of the corresponding amount of rainfall and temperature for each rainy day. So, average numbers of days rained per year were 39 and 42 for Dhaka and Chittagong respectively. The average contingency frequency and relative frequency matrices for Dhaka and Chittagong are respectively as follows:

Frequency Matrix for Dhaka

		Temperature (°C)				
		<27	**27-28**	**28-29**	**29-30**	**30+**
Rainfall (mm)	**1-11**	1.73	4.13	6.87	6.37	2.50
	11-21	1.13	1.80	2.07	1.27	0.67
	21-31	0.67	0.63	1.37	0.70	0.17
	31-41	0.70	0.73	0.47	0.37	0.13
	41+	1.93	1.30	1.03	0.47	0.13

Frequency Matrix for Chittagong

		Temperature (°C)				
		<27	**27-28**	**28-29**	**29-30**	**30+**
Rainfall (mm)	1-11	3.03	4.43	6.33	2.73	0.77
	11-21	1.87	1.70	1.17	0.63	0.27
	21-31	2.10	1.40	0.73	0.20	0.00
	31-41	1.70	0.53	0.80	0.17	0.00
	41+	6.73	2.73	1.63	0.40	0.07

Relative Frequency Matrix (P) for Dhaka

		Temperature ($^{\circ}C$)				
		<27	**27-28**	**28-29**	**29-30**	**30+**
Rainfall (mm)	**1-11**	0.04	0.11	0.17	0.16	0.06
	11-21	0.03	0.05	0.05	0.03	0.02
	21-31	0.02	0.02	0.03	0.02	0.00
	31-41	0.02	0.02	0.01	0.01	0.00
	41+	0.05	0.03	0.03	0.01	0.00

Relative Frequency Matrix (Q) for Chittagong

		Temperature ($^{\circ}C$)				
		<27	**27-28**	**28-29**	**29-30**	**30+**
Rainfall (mm)	**1-11**	0.07	0.11	0.15	0.06	0.02
	11-21	0.04	0.04	0.03	0.02	0.01
	21-31	0.05	0.03	0.02	0.00	0.00
	31-41	0.04	0.01	0.02	0.00	0.00
	41+	0.16	0.06	0.04	0.01	0.00

We want to infer whether the joint distribution of amount of rainfall and temperature for Dhaka and Chittagong are significant dissimilar or not. The chi-square and *p*-value matrices are as follows

$$\text{Chi-square matrix} = \begin{pmatrix} 8.67 & 0.00 & 2.65 & 59.26 & 33.07 \\ 4.14 & 0.44 & 9.93 & 7.93 & 6.09 \\ 20.29 & 7.39 & 7.33 & 9.55 & 5.38 \\ 10.90 & 1.43 & 2.02 & 2.71 & 4.30 \\ 81.22 & 13.20 & 3.01 & 0.33 & 0.81 \end{pmatrix}$$

$$p\text{-value matrix} = \begin{pmatrix} 0.00 & 0.99 & 0.10 & 0.00 & 0.00 \\ 0.04 & 0.51 & 0.00 & 0.00 & 0.01 \\ 0.00 & 0.01 & 0.01 & 0.00 & 0.02 \\ 0.00 & 0.23 & 0.15 & 0.10 & 0.04 \\ 0.00 & 0.00 & 0.08 & 0.57 & 0.37 \end{pmatrix}$$

The tabulated value of Chi – square at 1% level of significance with 1 degree of freedom is 6.634897. There is one calculated value for each of the 25 chi-square test statistics for 25 types of cells in the matrix of chi-squares. The resultant decision matrix for the 25 various cell frequencies is given below:

$$\text{the resultant decision matrix} = \begin{pmatrix} DS & S & S & DS & DS \\ S & S & DS & DS & S \\ DS & DS & DS & DS & S \\ DS & S & S & S & S \\ DS & DS & S & S & S \end{pmatrix}.$$

Moreover, the calculated value of overall chi – square, the sum of all individual chi-squares of the chi-squares' matrix sum, is obtained as 302. Therefore, the null hypothesis $H_0: P_{5\times5} = Q_{5\times5}$ of the equality of joint probability matrix of two population joint

probability distribution is rejected with p-value 0.00. So, the two population joint distributions do not belong to the same joint distributions of rainfall and temperature. All marginal frequencies for rainfall and temperature for both locations are not similar with *p*-value 0.00. The dissimilarity between the all row-wise marginal probabilities, column-wise marginal probabilities and maximum cell probabilities of the two joint frequency matrices are the acute evidence of the dissimilarity between two bivariate populations of rainfall and temperature pattern of two geographically distant regions.

3.2 An Application of the Test for the Equality of Two Transition probability matrices in Pairwise Sequence Alignment

Since it is stated that in a Markov process all possible states and transitions have been assumed in such a way that there is always a next state and the process goes on forever; the characteristics of the DNA, the basic genetic material in living organisms and having a double standed-helical structure each of which is consisting of very long sequence from four letters/alphabets (nucleotides), *a, g, c,* and *t* (for adenine, guanine, cytosine, and thymine, respectively), sequence that undergoes the change within any population over the course of many generations, as random mutations arise and become fixed in the population can easily be treated as a Markov Chain. If two sequences from different organisms are similar, there may have been a common ancestor sequence, and the sequences are then defined as being homologous. The alignment indicates the changes that could have occurred between the two homologous sequences and a common ancestor sequence during evolution. So, a common gauge is to check whether the two sequences show significant similarity, to assess, for example, whether they have a remote common ancestor. As a result, sequence alignment is one of the most important techniques to analyze biological system.

Suppose we have two small DNA sequences such as those in the book of 'Statistical Methods in Bioinformatics' by Ewens, W. *et al* (2004), 30 pairs of sample sequences from same species have been considered. The average transition raltive frequencies or estimated probabilities for first and second sample sequences are estimated as follows:

$$\hat{P} = \begin{array}{c} \\ a \\ t \\ c \\ g \end{array} \begin{array}{c} \begin{array}{cccc} a & t & c & g \end{array} \\ \begin{pmatrix} 0.200 & 0.213 & 0.292 & 0.295 \\ 0.260 & 0.201 & 0.186 & 0.358 \\ 0.305 & 0.342 & 0.267 & 0.086 \\ 0.220 & 0.161 & 0.229 & 0.389 \end{pmatrix} \end{array},$$

$$\hat{Q} = \begin{array}{c} \\ a \\ t \\ c \\ g \end{array} \begin{array}{c} \begin{array}{cccc} a & t & c & g \end{array} \\ \begin{pmatrix} 0.193 & 0.228 & 0.294 & 0.286 \\ 0.203 & 0.197 & 0.240 & 0.361 \\ 0.270 & 0.341 & 0.225 & 0.134 \\ 0.198 & 0.154 & 0.262 & 0.386 \end{pmatrix} \end{array}$$

From the transition probability graphs of the matrix $\hat{P}$ and $\hat{Q}$ we observe that all staes are recurrent. Second and third eigen values of the 1st matrix, on the other hand all of the eigen values of the 2nd matrix, are complex numbers. Eigen vectors for the both first and second matrices, consisting of complex numbers, also indicate the comparability in the two matrices. Determinant of the first matrix is -0.00022 and for the second matrix -

0.00021. The ranks of them are same (= 4) which is a sign of justification of comparing the two matrices.

Here the number of transition in each sample sequence is n = 19. The chi-square and the p-value matrices have been obtained as

$$\text{chi square matrix} = \begin{matrix} \\ a \\ t \\ c \\ g \end{matrix} \begin{matrix} \begin{matrix} a & t & c & g \end{matrix} \\ \begin{pmatrix} 0.02090 & 0.08567 & 0.00084 & 0.02894 \\ 1.11023 & 0.00744 & 1.04373 & 0.01582 \\ 0.37403 & 0.00034 & 0.00000 & 1.53070 \\ 0.28819 & 0.02776 & 0.52639 & 0.00456 \end{pmatrix} \end{matrix}$$

$$\text{p value matrix} = \begin{matrix} \\ a \\ t \\ c \\ g \end{matrix} \begin{matrix} \begin{matrix} a & t & c & g \end{matrix} \\ \begin{pmatrix} 0.885 & 0.960 & 0.956 & 0.956 \\ 0.957 & 0.959 & 0.958 & 0.951 \\ 0.954 & 0.952 & 0.956 & 0.969 \\ 0.966 & 0.969 & 0.964 & 0.960 \end{pmatrix} \end{matrix}.$$

The tabulated value of each element of the chi – square matrix at 5% level of significance and 1 $d.f$ is 3.84. So, the Decision matrix for individual chi square test is

$$Decision\ Matrix = \begin{matrix} \\ a \\ t \\ c \\ g \end{matrix} \begin{matrix} \begin{matrix} a & t & c & g \end{matrix} \\ \begin{pmatrix} S & S & S & S \\ S & S & S & S \\ S & S & S & S \\ S & S & S & S \end{pmatrix} \end{matrix}$$

where "S" refers to the acceptance of the null hypothesis $H_0: p_{ij} = q_{ij}$. The calculated value of overall chi – square is, the sum of all individual chi-squares, obtained as 5.04463. Finally, the null hypothesis of equality of transition probability matrix of two sequences is accepted at 5 % level of significance (since the tabulated value of the sum of chi-square with 12 degree of freedom is 21.03) which means that the two sequences are similar that is the sequences come from same origin. The row wise chi square statistics are obtained as 0.02937, 0.01221, 0.01147 and 0.00775 respectively for the 1st , 2nd , 3rd and 4th row along with p values 0.9986, 0.9996, 0.9996, and 0.9998 as well. Obviously, the accuracy of our decision is evident from the original two sequences taken from the same population species. Hence the performance of the proposed test seems better.

4. ADVANTAGES

The credence of the proposed tests for the equality of two frequency distributions or two joint frequency distributions is evident from the given real life example. The p values of the proposed tests for the equality of the marginal row frequency distributions or column frequency distributions over two populations are 0. The results seem to be appreciating since for two geographically distant locations the joint distributions of rainfall and temperature should be dissimilar. Besides, maximum of the cell frequencies vary between two populations. There are acute differences between row-wise marginal probability distributions and the column wise marginal probability distributions.

Again for the given 2nd example, the *p* - value of the proposed test is 1 (because *p*-values are 0.968, 0.993, 1.00 respectively) referring to strong acceptance of the equality of the transition pattern for two matrices since the samples were truly collected from same species. Therefore, the performance of the proposed method is well enough.

The proposed approach for comparing two transition probability matrices gives not only an overall decision of the significant similarity or dissimilarity of the individual paired transitions but also the significant similarity or dissimilarity of all possible transitions. It clearly identifies the possible dissimilarity between two population sequences. The current method specifically detects for which transition(s) the overall dissimilarity for the two population Markov chain is being evident. This idea of more specification can help the biotechnologist to quickly detect the core fact of the possible difference between bio-organisms more easily and more efficiently.

CONCLUDING REMARKS

Frequency distributions, Contingency tables and Transition Probability Matrices have been widely being studied by numerous authors since the childhood of statistics. Unfortunately, the discordance of them have not yet been studied so far with parametric tests. These tests ensembles the individual, group wise and overall pattern of the frequencies of one population whether significantly differing from those of the other population. Advanced multiple test for the equality of any univariate or bivariate or transition frequency distributions for the several populations can be the further scope of the proposed heuristics. Any inquiry and prove(s) of the mathematical development of the tests can be accessible from the authors on demand. The author is also preparing the more interesting issues including the test of equality of two Relative Risks or two Odds Ratios even as an aid for the advanced Mantel Haenszel test.

REFERENCES

Adnan M. A. S., Moinuddin, M, Roy, S, *et al*. (2011). An Alternative Approach of Pair-wise sequence Alignment. Proceedings. JSM 2011, American Statistical Association, p2941 - 2951.

Agresti, A. (1990) (2nd ed., 2002). Categorical Data Analysis. Wiley, New York.

Agresti, A. (2005) Multivariate tests comparing binomial probabilities with application to safety studies for drugs. Web.

Barnard, G. A. (1945). A new test for tables. Nature 156, 177.

Barnard, G. A. (1947). Significance tests for tables. Biometrika 34, 123-138.

Barnard, G. A. (1949). Statistical inference. J. Roy. Statist. Soc. Ser. B 11, 115-139.

Barnard, G. A. (1979). In contradiction to J. Berkson's dispraise: conditional tests can be more efficient. J. Statist. Plann. Inference 3, 181-187.

Bartlett, M. S. (1984). Discussion on tests of significance for 2x2 contingency tables (by F. Yates). J. Roy. Statist. Soc. Ser. A 147, 453.

Bartolucci, F et al (2009) Multidimensional latent Markov models in a development study of inhibitory control and attentional flexibility in the early childhood. Web.

Behseta, S et al (2005). Testing equality of two functions using BARS. Statistics in Medicine. Doi: 10. 1002/sim.2195.

Bennett, B. M. and Hsu, P. (1960). On the power function of the exact test for the contingency table. Biometrika, 47, 393-398 (correction 48 (1961), 475).
Berger, R. L. and Boos, D. D. (1994). P-values maximized over a confidence set for the nuisance parameter. J. Amer. Statist. Assoc. 89, 1012-1016.
Berger, R. L. (1996). More powerful tests from confidence interval values. Amer. Statist. 50, 314-318.
Berkson, J. (1978). In dispraise of the exact test. J. Statist. Plann. Inference 2, 27-42.
Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). J. Amer. Statist. Assoc. 57, 269-326.
Boschloo, R. D. (1970). Raised conditional level of significance for the 2 × 2 table when testing the equality of probabilities. Statistica Neerlandica 24, 1-35.
Cheng et al (2008) Identification identities and testing hypotheses: Power analysis for contingency tables.
Chernoff, H. (1954). On the distribution of the likelihood ratio. Ann. Math. Statist. 25, 573-578.
Cho, J. S et al (2011). Testing the equality of two positive definite matrices with application to information matrix testing. Web.
Cox, D. R. and Snell, E. J. (1989). The Analysis of Binary Data. 2nd Edition. Chapman and Hall, London.
Dannemann, J And Holzmann, H (2007). The likelihood ratio test for hidden Markov models in two-sample problems. Comp. Stat. & Dtat Analysis. V 52, P:1850 -1859.
Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Statist. 11, 427-444.
Falay, B. (2007) Intergenerational income mobility: Equality of Opportunity: A comparison of East and West Germany. EKONOMI YUKSEK LISANS PROGRAMI, Istanbul Bilgi University. 2007
Feller, W. (1968). An Introduction to Probability Theory and its Applications. 3rd Edition. Wiley, New York.
Fisher, R. A. (1922). On the interpretation of _2 from contingency tables, and the calculation of P. J. Roy. Statist. Soc. 85, 87-94.
Fisher, R. A. (1925) (5th ed., 1934; 10th ed., 1946). Statistical Methods for Research Workers, Oliver & Boyd, Edinburgh.
Fisher, R. A. (1935). The logic of inductive inference. J. Roy. Statist. Soc. Ser. A 98, 39-54.
Fisher, R. A. (1962). Confidence limits for a cross-product ratio. Austral. J. Statist. 4, 41.
Gail, M. and Gart, J. J. (1973). The determination of sample sizes for use with the exact conditional test in comparative trials. Biometrics 29, 441-448.
Gokhale, D. V. and Kullback, S. (1978). The Information in Contingency Tables. Marcel Dekker, New York.
Goodman, L. A. (1984). The Analysis of Cross-Classified Data Having Ordered Categories. Harvard University Press, Cambridge.
Gray, R. M. (1990). Entropy and Information Theory. Springer-Verlag, New York.
Greenland, S. (1991). On the logical justification of conditional tests for two-by-two contingency tables. Amer. Statist. 45, 248-251.
Grizzle, J. E. (1967). Continuity correction in the test for tables. Amer. Statist. 21, 28-32.
Haber M. (1986). An exact unconditional test for the comparative trial. Psychol. Bull. 99, 129-132.
Johnson, N. L. and Kotz, S. (1969). Discrete Distributions. Wiley, New York.
Kempthorne, O. (1978). Comments on J. Berkson's paper "In Dispraise of the Exact Test". J. Statist. Plann. Inference 3, 199-213.

Kendall, M. G. and Stuart, A. (1979). The Advanced Theory of Statistics. Vol. 2, 4th edition. Charles Griffin, London.
Klugkist, I et al (2010). Bayesian Evaluation of Inequality and Equality Constrained Hypotheses for contingency Tables. Web. NOW-VICI-453-05-002.
Kou, S. G. and Ying, Z. (1996). Asymptotics for a table with fixed margins. Statist. Sinica 6, 809-829.
Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. Ann. Math. Statist. 22, 79-86.
Lancaster, H. O. (1949). The combination of probabilities arising from data in discrete distri- butions. Biometrika 36, 370-382, Corrig. 37, 452.
Lancaster, H. O. (1969). The Chi-squared Distributions. Wiley, New York.
Lehmann, E. L. (1986). Testing Statistical Hypotheses. 2nd Edition. Wiley, New York.
Lindley, D. V. (1956). On a measure of the information provided by an experiment. Ann. Math. Statist. 27, 986-1005.
Little, R. J. A. (1989). Testing the equality of two independent binomial proportions. Amer. Statist. 43, 283-288.
Mehta, C. R. and Patel, N. R. (1980). A network algorithm for the exact treatment of the contingency table. Comm. Statist. Ser. B 9, 649-664.
Muse, S. V. et al (1992) Testing the equality of evolutionary rates. Genetics. 1322: 269-276.
Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 20, 263-274.
Pearson, E. S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a table. Biometrika 34, 139-167.
Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil. Mag. Series 50, 157-175.
Pearson, K. (1904). Mathematical contributions to the theory of evolution XIII: On the theory of contingency and its relation to association and normal correlation. Draper's Co. Research
Memoirs, Biometric Series, no. 1. (Reprinted in Karl Pearson's Early Papers, ed. E. S. Pearson, Cambridge: Cambridge University Press, 1948.)
Plackett, R. L. (1964). The continuity correction in tables. Biometrika 51, 327-337.
Plackett, R. L. (1977). The marginal totals of a table. Biometrika 64, 37-42.
Rao, C. R. (1973). Linear Statistical Inference and Its Applications. 2nd edition. Wiley, New York.
Santner, T. J. and Duffy, D. E. (1989). The Statistical Analysis of Discrete Data. Springer- Verlag, New York.
Suissa, S. and Shuster, J. (1985). Exact unconditional sample sizes for the binomial trial. J. Roy. Statist. Soc. Ser. A 148, 317-327.
Upton, G. J. G. (1982). A comparison of alternative tests for the 2 × 2 comparative trial. J. Royal Statist. Soc. A 145, 86-105.
Wilks, S. S. (1935). The likelihood test of independence in contingency tables. Ann. Math. Statist. 6, 190-196.
Yates, F. (1934). Contingency tables involving small numbers and the test. J. Royal Statist. Soc. Suppl. 1, 217-235.
Yates, F. (1984). Tests of Significance for contingency tables (with discussion). J. Royal Statist. Soc. A 147, 426-463.
Yule, G. U. (1911). An Introduction to the Theory of Statistics. Griffin, London.
Wikipedia: Frequency Distribution, Contingency Table, Markov chain.