Tests of Equality of Several High Dimensional Contingency Tables

Silvia Irin Sharna[1], Mian Arif Shams Adnan[2], Asif Shams Adnan[3] and Rahmatullah Imon[4]
[1]Bowling Green State University, Bowling Green, OH 43402.
[2]Bowling Green State University, Bowling Green, OH 43402.
[3]East West University, Dhaka 1212, Bangladesh.
[4]Ball State University, Muncie, IN 47306.

Abstract
Tests of equality of several high dimensional contingency tables has been developed. Here the number of dimensions is equal to or greater than three.

Key Words: Matrix of p-values.

1. Introduction

A frequency distribution shows us a summarized grouping of data divided into mutually exclusive classes and the number of occurrences in each class. 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 (2x2) contingency table.

Adnan (2015) and Sharna *et al* (2012) developed a class of new statistical tests for checking the similarity or dissimilarity among the individual (cell) frequencies, marginal frequencies and total frequencies of several univariate or joint probability distributions.

Later Adnan *et al* (2016) demonstrated an idea of building three-dimensional volume matrix.

The aim of the present paper is to develop a class of multiple comparison test statistics for several high dimensional contingency tables.

2. Several Multiple Comparison Tests

Let we have *m n*-dimensional contingency space tables or matrices from *m* populations and let the hypothesis be

$$H_0\colon N_1 = N_2 = \cdots N_m$$

$$H_0\colon \left(N_{ij\ldots k1}\right)_{a\times b\times\ldots\times n} = \left(N_{ij\ldots k2}\right)_{a\times b\times\ldots\times n} = \cdots = \left(N_{ij\ldots km}\right)_{a\times b\times\ldots\times n}$$

$$\therefore H_0\colon P_1 = P_2 = \cdots . = P_m$$

$$H_0\colon \left(P_{ij\ldots k1}\right)_{a\times b\times\ldots\times n} = \left(P_{ij\ldots 2}\right)_{a\times b\times\ldots\times n} = \cdots = \left(P_{ij\ldots km}\right)_{a\times b\times\ldots\times n}$$

where, the N_l ($\forall\, l = 1,2,\dots,m$) is the high dimensional population frequency matrix or high dimensional population contingency table of the l^{th} population, P_l ($\forall\, l = 1,2,\dots,m$) is the population high dimensional probability region matrix or high dimensional contingency space table of the l^{th} population such that P = $\left(P_{ij\dots kl}\right)_{a\times b\times\dots\times n}$, where $P_l = (p_{ij\dots kl})_{a\times b\times\dots\times n}$, $p_{ij\dots kl} = \frac{N_{ij\dots kl}}{N_{\dots\dots l}}$ where $N_{ij\dots kl}$ is the population frequency of the $(i,j,\dots,k)^{\text{th}}$ cell or element of the population space contingency table. There are n variables named variable A, variable B,…, variable N. Variables A ,B, …, N have total number of categories a, b,…, n respectively.

2.1 Notation

$N_{\dots\dots l} = \sum_{i=1}^{a}\sum_{j=1}^{b}\dots\sum_{k=1}^{n} N_{ij\dots kl}\ \forall\, i = 1,2,\dots,a; j = 1,2,\dots,b;\dots; k = 1,\ 2,\ \dots, n.$

q samples are collected from each of m populations. On the basis of these q samples, we want to test whether future samples came from the same population or not. So, q sample contingency tables from each of the m population joint frequency distributions are available. The maximum likelihood estimators of the relative frequency contingency region tables are obtained as $\hat{P}_l = (\hat{p}_{ij\dots kl})_{a\times b\times\dots\times n}$ *where* $\hat{p}_{ij\dots kl} = \frac{n_{ij\dots kl}}{n_{\dots\dots l}}$ *whereas* $n_{ij\dots kl}$ *is the average frequency of the (i,j,…,k)th cell or element of the average frequency space table or matrix* $n_{\dots\dots l}$ *constructed from q sampled space frequency tables drawn from the* l^{th} *population. Here,* $n_{\dots\dots l} = \sum_{i=1}^{a}\sum_{j=1}^{b}\dots\sum_{k=1}^{n} n_{ij\dots kl}\ \forall\, i = 1,2,\dots,a; j = 1,2,\dots,b;\dots; k = 1,\ 2,\ \dots, n.$

2.2 Theorem

Each Element of the Decision Matrix is a chi-square test statistic.

Proof: For large $\mathbf{n}_{\dots\dots l}$ the asymptotic distribution of each element of transition probability matrices, according to the Central Limit Theorem, is distributed as normal such that

$$\hat{p}_{ij\dots kl} \underset{\sim}{\overset{n_{\dots\dots l}\to\infty}{}} N\left(p_{ij\dots kl},\ \frac{p_{ij\dots kl}\ (1-p_{ij\dots kl})}{q n_{\dots\dots l}}\right).$$

$$\therefore \sum_{l=1}^{m} \frac{\left(\hat{p}_{ijkl}-\bar{p}_{ijk.}\right)^2}{\frac{\bar{p}_{ijk.}(1-\bar{p}_{ijk.})}{q n_{\dots l}}} \sim \chi^2_{(m-1)}\ \forall\, i = 1,2,\dots,a; j = 1,2,\dots,a;\dots;\ k = 1,2,\dots,n$$

Where, $\bar{p}_{ijk.} = \frac{n_{ij\dots k1}\hat{p}_{ij\dots k1}+\dots+n_{ij\dots km}\hat{p}_{ij\dots km}}{n_{ij\dots k1}+\dots+n_{ij\dots km}}$; $\forall\, i = 1,2,\dots,a; j = 1,2,\dots,b;\dots;$

$$k = 1,2,\dots,n.$$

We obtain an element-chi-square space matrix $\chi 2$ in the following form. However, the sum of correlated chi-squares is also a chi square statistic (Joarder and Omar 2013).

$$\chi 2 = \left(\sum_{l=1}^{m} \frac{\left(\hat{p}_{ij\dots kl}-\bar{p}_{ij\dots k.}\right)^2}{\frac{\bar{p}_{ij\dots k.}\left(1-\bar{p}_{ij\dots k.}\right)}{q n_{\dots\dots l}}}\right)_{a\times b\times\dots\times n} = \left(\chi^2_{ij\dots k}\right)_{a\times b\times\dots\times n}.$$

2.3 Explicit Hypotheses for Multiple Comparison Tests

Several hypotheses of the several multiple comparison tests have been addressed in the following.

2.3.1 Element Comparison and Matrix Region Comparison Tests

(i) $$H_0: p_{ij\ldots k1} = \ldots = p_{ij\ldots km};$$

or, the hypothesis of testing the equality of the each $(i,j,\ldots,k)^{th}$ individual cell probabilities of the m n-dimensional population probability space tables or matrices $(P_{ij\ldots k1})_{a\times b\times\ldots\times n}, (P_{ij\ldots k2})_{a\times b\times\ldots\times n}, \ldots, (P_{ij\ldots km})_{a\times b\times\ldots\times n}$.

(ii) $$H_0: (p_{ij\ldots k1})_{b\times\ldots\times n} = \cdots = (p_{ij\ldots km})_{b\times\ldots\times n};$$

or, the hypothesis of checking the equality of the i^{th} row region probability matrix or probability distribution for all populations.

(iii) $$H_0: (p_{ij\ldots k1})_{a\times c\times\ldots\times n} = \cdots = (p_{ij\ldots km})_{a\times c\times\ldots\times n}$$

or, the hypothesis of checking the equality of the j^{th} column region probability matrix or probability distribution for all populations.

(iv) $$H_0: (p_{ij\ldots k1})_{a\times b\times\ldots\times m} = \cdots = (p_{ij\ldots km})_{a\times b\times\ldots\times m}$$

or, the hypothesis of checking the equality of the k^{th} layered region probability matrix or probability distribution for all populations.

2.3.2 Multiple Diagonal Matrix Region Comparison Tests

(v) $$H_0: (p_{ij\ldots k1})_{b\times\ldots\times n} = \cdots = (p_{ij\ldots km})_{b\times\ldots\times n}; j = \ldots = k$$

or, the equality of the diagonal region probability matrix or distribution for all (m) populations where the upper left corner is the $(1,1,\ldots,1)^{th}$ individual cell probability and the lower right corner is the $(a, b, \ldots,n)^{th}$ individual cell probability of the $b \times \ldots \times n$ diagonal region probability matrix or probability distribution for all populations. As for example, for 3 D Contingency Tables, a = b = n = 4;

$$H_0: \begin{pmatrix} p_{1111} & p_{1221} & p_{1331} & p_{1441} \\ p_{2111} & p_{2221} & p_{2331} & p_{2441} \\ p_{3111} & p_{3221} & p_{3331} & p_{3441} \\ p_{4111} & p_{4221} & p_{4331} & p_{4441} \end{pmatrix}_{b\times n} = \cdots = \begin{pmatrix} p_{111m} & p_{122m} & p_{133m} & p_{144m} \\ p_{211m} & p_{222m} & p_{233m} & p_{244m} \\ p_{311m} & p_{322m} & p_{333m} & p_{344m} \\ p_{411m} & p_{422m} & p_{433m} & p_{444m} \end{pmatrix}_{b\times n}$$

or, the equality of the diagonal probability plate or matrix or distribution for m populations where the upper left element is the $(1,1,1)^{th}$ individual probability and the lower right

element is the (4, 4, 4)th individual probability of the 4 × 4 diagonal probability plate of each population.

2.3.3 Multiple Vector Comparison Tests

(vi) $$H_0: \left(p_{ij\ldots k1}\right)_{1\times b} = \cdots = \left(p_{ij\ldots km}\right)_{1\times b};$$

or, the hypothesis of checking the equality of the ith row probability vector or probability distribution for the kth category of variable K of the nth dimension for all populations. For a 3 D contingency tables, a = b = n = 4; $(p_{i1k1} \quad p_{i2k1} \quad p_{i3k1} \quad p_{i4k1})_{1\times 4} = \cdots = (p_{i1km} \quad p_{i2km} \quad p_{i3km} \quad p_{i4km})_{1\times 4}$
or, the equality of the ith row probability vector or distribution for the kth category of variable K of the nth dimension for all populations.

(vii) $$H_0: \left(p_{ij\ldots k1}\right)_{1\times a} = \cdots = \left(p_{ij\ldots km}\right)_{1\times a};$$

or, the hypothesis of checking the equality of the jth column probability vector or probability distribution for the kth category of variable K over nth dimension for all populations.

(viii) $$H_0: \left(p_{ij\ldots k1}\right)_{1\times n} = \cdots = \left(p_{ij\ldots km}\right)_{1\times n};$$

or, the hypothesis of checking the equality of the kth layer probability vector or probability distribution for the ith category of variable A over first dimension for all populations.

2.3.4 Hypotheses for Multiple Diagonal Vector Comparison and Overall Comparison Tests

(ix) $$H_0: \left(p_{ij\ldots k1}\right)_{1\times n} = \cdots = \left(p_{ij\ldots km}\right)_{1\times n};$$

or, the hypothesis of checking the equality of the row diagonal probability vector or probability distribution from the ith category of the variable A to the kth category of variable K for all populations. For a 3 D contingency tables, a = b = n = 4;

$$(p_{1111} \quad p_{2221} \quad p_{3331} \quad p_{4441})_{1\times 4} = \cdots = (p_{111m} \quad p_{222m} \quad p_{333m} \quad p_{444m})_{1\times 4}$$

or, the equality of the row diagonal probability vector or probability distribution from the 1st category of the variable A to the 4th category of variable C for all populations.

(x) $$H_0: \left(P_{ij\ldots k1}\right)_{a\times b\times\ldots\times n} = \left(P_{ij\ldots k2}\right)_{a\times b\times\ldots\times n} = \cdots = \left(P_{ij\ldots km}\right)_{a\times b\times\ldots\times n}$$

or, the hypothesis of testing the equity of the whole n-dimensional contingency space table or space matrix for one population is varying from that of the other populations. It tests the similarity of m populations where each of the m populations has joint probability space (volume for 3 dimensions) distributions over a × b × … × n cells.

3. Test Statistics for Several Multiple Comparison Tests

(i). Test of equality of m [(i,j,…,k)th] cell frequencies: Comparing each $\chi^2_{ij\ldots k}$ with the tabulated $\chi^2_{(m-1,\propto)}$ of $(m-1)$ degree of freedom,

(ii). Test of equality of m [ith variable's] row marginal plate/matrix region probability distributions: Comparing each $\sum_{j\ldots k}\chi^2_{ij\ldots k}$ with the tabulated $\chi^2_{(b\times\ldots\times n\times(m-1),\propto)}$ of $b\times\ldots\times n\times(m-1)$ degrees of freedom,

(iii). Test of equality of m [jth variable's] column marginal frequency plate/matrix region probability distributions: Comparing each $\sum_{i\ldots k}\chi^2_{ij\ldots k}$ with the tabulated $\chi^2_{(a\times c\times\ldots\times n\times(m-1),\propto)}$ of $a\times c\times\ldots\times n\times(m-1)$ degrees of freedom,

(iv) Test of equality of m [kth variable's] layer marginal frequency plate/matrix region distributions: Comparing each $\sum_{i\ldots(k-1)}\chi^2_{ij\ldots k}$ with the tabulated $\chi^2_{(a\times b\times\ldots\times m\times(m-1),\propto)}$ of $a\times b\times\ldots\times m\times(m-1)$ degrees of freedom,

(v). Test of equality of m diagonal region probability matrix or distributions: Comparing each $\sum_{j=\cdots=k}\chi^2_{ij\ldots k}$ with the tabulated $\chi^2_{(b\times\ldots\times n\times(m-1),\propto)}$ of $b\times\ldots\times n\times(m-1)$ degrees of freedom,

(vi). Test of equality of m ith row probability vectors or probability distributions: Comparing each $\sum_j\chi^2_{ij\ldots k}$ with the tabulated $\chi^2_{(b\times(m-1),\propto)}$ of $b\times(m-1)$ degrees of freedom,

(vii). Test of equality of m jth column probability vectors or probability distributions: Comparing each $\sum_i\chi^2_{ij\ldots k}$ with the tabulated $\chi^2_{(a\times(m-1),\propto)}$ of $a\times(m-1)$ degrees of freedom,

(viii). Test of equality of m kth layer probability vectors or probability distributions: Comparing each $\sum_k\chi^2_{ij\ldots k}$ with the tabulated $\chi^2_{(n\times(m-1),\propto)}$ of $n\times(m-1)$ degrees of freedom,

(ix). Test of equality of m row diagonal probability vector or probability distributions from the ith category of the variable A to the kth category of variable K: Comparing each $\sum_{i=j=\cdots=k}\chi^2_{ij\ldots k}$ with the tabulated $\chi^2_{(n\times(m-1),\propto)}$ of $n\times(m-1)$ degrees of freedom,

(ix). Test of the equity of m n-dimensional whole contingency space table or space matrix or space probability distributions: Comparing Chi-squares' matrix sum

$$= \chi^2_{11\ldots1}+\cdots+\chi^2_{1b\ldots1}+\cdots+\chi^2_{a11}+\cdots+\chi^2_{ab1}{}^{+}\ldots{}^{+}\chi^2_{11\ldots2}+\cdots+\chi^2_{1b\ldots2}+\cdots+\chi^2_{a1\ldots2}+\cdots+\chi^2_{ab\ldots2}{}^{+}\ldots{}^{+}\chi^2_{11\ldots n}+\cdots+\chi^2_{1b\ldots k}+\cdots+\chi^2_{a1\ldots n}+\cdots+\chi^2_{ab\ldots n}$$

with the tabulated $\chi^2_{(a\times b\times\ldots\times n\times(m-1),\propto)}$ of degrees of freedom $a\times b\times\ldots\times n\times(m-1)$.

4. Proposed Multiple Comparison Tests for a Real-life Problem

After observing 30 pairs of tri-variate samples (30 tri-variate samples from each tri-variate population) from two tri-variate populations we have obtained the two average frequency volumes or average frequency volume matrices. So, the tri-variate average volume frequency or matrices are we have two contingency $2 \times 2 \times 2$ tables as given below:

Lower Level	Not a biter	Mild biter	Flagrant biter
Mice	20	16	24
Guinea pigs	19	11	50

Higher Level	Not a biter	Mild biter	Flagrant biter
Mice	25	21	29
Guinea pigs	24	16	55

Higher Level	Not a biter	Mild biter	Flagrant biter
Mice	100	56	44
Guinea pigs	19	11	50

Lower Level	Not a biter	Mild biter	Flagrant biter
Mice	120	76	49
Guinea pigs	29	20	58

4.1 Computation for Multiple Comparison Tests

The chi square matrix for Lower Level $= \begin{pmatrix} 502 & 121 & 6 \\ 158 & 90 & 451 \end{pmatrix}$ and

the chi square matrix for Higher Level $= \begin{pmatrix} 583 & 187 & 21 \\ 117 & 68 & 440 \end{pmatrix}$

The resultant decision matrix for Lower Level $= \begin{pmatrix} DS & DS & S \\ DS & DS & DS \end{pmatrix}$ and the resultant

decision matrix for Higher Level $= \begin{pmatrix} DS & DS & DS \\ DS & DS & DS \end{pmatrix}$. Here, 'DS' means dissimilar.

(i). H_0: $p_{\text{mice, not a biter, lower level}} = q_{\text{mice, not a biter, lower level}}$
is rejected at 1 percent level of significance with p value 4.09×10^{-11}.

(ii). The sum of chi- squares for the 1st, 2nd and 3rd lower leveled-columns are calculated as 660, 211 and 457 respectively. The tabulated value of the column wise sum of chi-squares with 4 degree of freedom is 13.28 at 1 % level of significance. Again, the sum of chi-squares for the 1st, 2nd and 3rd higher leveled-columns are calculated as 701, 255 and 461 respectively. So, all columns are dissimilar for the two populations' joint distributions,

(iii). H_0: $P_{2\times3\times2} = Q_{2\times3\times2}$ is rejected at 1 % for calculated value of chi-square 2745.

5. Further Applications of Multiple Comparison Tests for Plates and Diameters of 3D Spheres and their Advantages

If several boxes are inscribed in sphere(s), various types of equalities like diameters, diameter plates, etc can be tested for several spheres from multiple populations.

The credence of the proposed tests for the equality of two frequency distributions or two joint frequency distributions are evident from the given real-life examples. The *p* values of the proposed tests for the equality of the marginal row frequency distributions or column frequency distributions over two populations are very much near 0. The results seem to be appreciating since maximum of the cell frequencies vary between populations. Besides, the test for comparing the two joint frequency distributions prescribed a *p* value of 0 which means that two population joint distributions are not similar and is supported by the given samples. This certain difference is very much due to the difference between/among row-wise marginal probability distributions, the column wise marginal probability distributions, and so on for the p dimensional marginal probability distributions.

Conclusion

Frequency distributions or contingency tables have been widely being studied by numerous authors since the childhood of statistics. Unfortunately, the discordance of them has not yet been studied so far with parametric tests. This paper also aided the test of equality of several frequency distributions. The tests presented in this paper ensemble the individual, group wise and overall pattern of the frequencies of one population whether significantly differing from those of other populations. Developing the mathematical and graphical representation of several types of Partial Linear and Non-Linear Correlation Plate and Logistic Regression Plate for the 3D case is the further scope of this paper.

References

Adnan, M. A. S., Crouch, S. Islam, K, and Zhu, J. (2016). Three Dimensional Contingency Tables, Measures of Association and Correlation. JSM Proceedings. Statistical Computing Section. Alexandria, VA: American Statistical Association, 211-221.

Adnan, M. A. S. (2015). Parametric Tests of Equality of Several Univariate Frequency Distributions/Several Contingency Tables and Several Markov Chains/Several Transition Frequency Matrices. JSM Proceedings. Government Statistics Section. Alexandria, VA: American Statistical Association, 28-41.

Joarder, A. H. and Oar. M. H. (2013). Exact distribution of the sum of two correlated chi-square. Kuwait Journal of Science. 40(2), 61-81.

Sharna, S.I., Adnan, M.A.S. and Shamsuddin, M. (2012). Parametric Test of Equality of Two Frequency Distributions/Matrices. JSM Proceedings. Inference from Combined Data Sets. Government Statistics Section. Alexandria, VA: American Statistical Association, 2025-2039.