Three Dimensional Contingency Tables, Measures of Association and Correlation

Mian Adnan[1], Shannon Crouch[2], Khairul Islam[3] and Judy Zhu[4]
[1,2]Ball State University, 2000 W University Avenue, Muncie, IN 47304
[3]Eastern Michigan University, 515B Pray-Harrold, Ypsilanti, MI 48197
[4]Apple IST, 100 Zhongshan Avenue, Xuanwu, Nanjing, Jiangsu, China 210000

Abstract
Contingency Tables, Measures of Association, Correlation, etc need some theoretical developments for three dimensional cases. Attempts have been made here to have statistical analyses for three dimensional contingency tables, three-dimensional relative risk and odds ratio, three-dimensional correlation coefficient, and the test of equality of several three dimensional contingency tables.

Key words: Concentration, Space, Volume Contingency Table.

1. Introduction

Contingency Tables, Measures of Associations and Correlation are being used for presenting a between relationship for two or more than two variables. We can term the aforesaid tools as the 2D (two dimensional) Measures of Associations. Attempts have been made here to develop the three dimensional measures of associations including Three Dimensional Contingency Tables, Three-Dimensional Correlation Coefficient, Three-Dimensional Odd's Ratio, Three-Dimensional Relative Risk, and the test of equality of several three dimensional contingency tables.

2. Three Dimensional Correlation

Let $(x_{111}, y_{111}, z_{111})$, $(x_{222}, y_{222}, z_{222})$, …, $(x_{nnn}, y_{nnn}, z_{nnn})$ be a set of *n* observations each of which represent the realized value of three different variables X, Y, Z. We have the data set according to the following format.

Number of observations	X_i	Y_i	Z_i
1	x_1	y_1	z_1
2	x_2	y_2	z_2
…	…	…	…
n	x_n	y_n	z_n

The 3 dimensional correlation co-efficient among these 3 variables can be presented as below

$$\text{3 D Correlation Co-efficient, } r_{xyz} = \frac{\sum_{i=1}^{n}[(x_i-\bar{x})(y_i-\bar{y})(z_i-\bar{z})]}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2(z_i-\bar{z})^2\right]^{\frac{1}{3}}}.$$

$$= \frac{Cov(x, y, z)}{sd(x)sd(y)sd(z)}$$

The 3D correlation coefficient ranges from $-\infty$ to $+\infty$. And there are prefect correlation when $r_{xyz} = 0$. Since each 3D space has 4 biggest diagonals, there are 4 types of perfect linear correlations. But these 4 types of prefect linear correlation lines are the bigger diagonals of four 2D diagonal plates which may have different angels with base surface. Moreover, these correlation lines for 3D space has been introduced to demonstrate the inherent infrastructure and their visualizations of the associations of 3 variables from several look/dimensions. Three variables may have overall linear relationship among them, but two variables may have nonlinear relationship having various different local linear relationship for the change of the levels of another variable. Various types of relationships among three variables have been tabulated to get the value of the 3D linear correlation coefficient.

x	**y**	**z**	$\boldsymbol{r_{xyz}}$
1	1	1	0
2	2	2	
3	3	3	
1	1	3	0
2	2	2	
3	3	1	
1	3	1	0
2	2	2	
3	1	3	
1	3	3	0
2	2	2	
3	1	1	
1	1	1	0.48
2	2	2	
4	4	4	
1	1	4	-0.17
2	2	2	
4	4	1	
1	1	1	12.80
2	2	2	
40	40	40	
1	1	40	-6.17
2	2	2	
40	40	1	
1	1	1	30.65

10	10	10	
100	100	100	
1	1	100	-13.50
10	10	10	
100	100	1	
1	1	1	331.42
10	10	10	
1000	1000	1000	
1	1	1000	-163.50
10	10	10	
1000	1000	1	
1	1	1	0.13
2	2	2	
3	3	4	
1	1	1	-0.24
2	2	2	
3	3	2	
1	1	1	-0.50
2	2	2	
3	3	0	
1	10	100	0
2	20	200	
3	30	300	
1	1	2	undefined
2	2	2	
3	3	2	
1	1	1	2.40
2	2	2	
3	3	1000	
1	1	10	1.56
2	10	1	
3	0	0	
1	1	10	1.56
2	10	1	
3	11	11	
1	1	10	7.21
2	10	1	

3	100	100	
1	1	1	-1.41
4	5	6	
6	5	4	

3. Three Dimensional Contingency Tables, Relative Risk and Odd's Ratio

Let we have a contingency table of 8 observations x_{ijk} ∨ $i, j, k = 1, 2$. The layout of the data is described as below.

Absent	**Absent**	**Present**	**Total**
Absent	$\boldsymbol{x_{111}}$	$\boldsymbol{x_{121}}$	$x_{1.1}$
Present	$\boldsymbol{x_{211}}$	$\boldsymbol{x_{221}}$	$x_{2.1}$
Total	$x_{.11}$	$x_{.21}$	$x_{..1}$

Present	**Absent**	**Present**	**Total**
Absent	$\boldsymbol{x_{112}}$	$\boldsymbol{x_{122}}$	$x_{1.2}$
Present	$\boldsymbol{x_{212}}$	$\boldsymbol{x_{222}}$	$x_{2.2}$
Total	$x_{.12}$	$x_{.22}$	$x_{..2}$

For a $2 \times 2 \times 2$ contingency table, the relative risk will be

$$\text{3 D Relative Risk, RR} = \frac{\left[\frac{\left(\frac{x_{111}}{x_{1.1}}\right)}{\left(\frac{x_{112}}{x_{1.2}}\right)}\right]}{\left[\frac{\left(\frac{x_{211}}{x_{2.1}}\right)}{\left(\frac{x_{212}}{x_{2.2}}\right)}\right]}.$$

Moreover, the odd's ratio will be

$$\text{3 D Odd's Ratio, OR} = \frac{\left[\frac{\left(\frac{x_{111}}{x_{121}}\right)}{\left(\frac{x_{112}}{x_{122}}\right)}\right]}{\left[\frac{\left(\frac{x_{211}}{x_{221}}\right)}{\left(\frac{x_{212}}{x_{222}}\right)}\right]}$$

If all the four components of the one level of the third variable are same or four components remain same to all the levels of the third factor, then the three dimensional relative risk and odd ratio measures tend the two dimensional relative risk and odd ratio measures.

Therefore, if x_{11k}, x_{12k}, x_{21k}, x_{22k} remain same irrespective of k at different level of the third variable then

2 D Odd's Ratio = 3 D Odd's Ratio
2 D Relative Risk = 3 D Relative Risk

4. Test of Equality of Several Volume Contingency Tables

Adnan (2015) and Sharna *et al* (2012) developed a class of new parametric test statistics 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.

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

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

$$\Rightarrow H_0: \left(N_{ijk1}\right)_{r\times c\times l} = \left(N_{ijk2}\right)_{r\times c\times l} = \cdots = \left(N_{ijkm}\right)_{r\times c\times l}$$

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

$$\Rightarrow H_0: \left(P_{ijk1}\right)_{r\times c\times l} = \left(P_{ijk2}\right)_{r\times c\times l} = \cdots = \left(P_{ijkm}\right)_{r\times c\times l}$$

where, the N_p $(\forall\, p = 1,2,\dots,m)$ is the population frequency volume matrix or contingency volume table of the pth population; P_p is the population probability matrix or contingency table of the pth population such that $P = (p_{ijkl})_{r\times c\times l}$, , where $p_{ijkl} = \frac{N_{ijkl}}{N_{\dots}}$ whereas N_{ijkl} is the population frequency of the (*i,j,k*)th element of the population frequency volume matrix N_p of the *l*th population and $N_{\dots l} = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} N_{ijkl}\,;\forall\, i = 1,2,\dots,r; j = 1,2,\dots,c; k = 1,2,\dots,l.$ q sample contingency tables from each of the *m* population joint frequency distributions (a total of q 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 n sample-frequency matrices or tables from each of the m populations, the maximum likelihood estimators of the probability matrices are obtained as $\hat{P}_l = (\hat{p}_{ijkl})_{r\times c\times l}$ where $\hat{p}_{ijkl} = \frac{n_{ijkl}}{n_{\dots l}}$ whereas n_{ijkl} is the average frequency of the (i,j,k)th element of the average frequency volume matrix $n_{\dots l}$ constructed from n sample-frequency tables drawn from the *l*th population. Here, $n_{\dots l} = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} n_{ijkl}\,;\forall\, i = 1,2,\dots,r; j = 1,2,\dots,c; k = 1,2,\dots,l.$

For large $n_{\dots l}$ the asymptotic distribution of each element of transition probability matrices, according to the Central Limit Theorem, are distributed as normal such that

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

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

where $\bar{p}_{ijk.} = \frac{n_{ijk1}\hat{p}_{ijk1}+\cdots+n_{ijkm}\hat{p}_{ijkm}}{n_{ijk1}+\cdots+n_{ijkm}}$; $\forall\ i = 1,2,\ldots,r; j = 1,2,\ldots,c;\ k = 1,2,\ldots,l.$

However, we obtain an element-chi-square volume matrix $\chi 2$ of the following form

$$\chi 2 = \left(\sum_{l=1}^{m} \frac{\left(\hat{p}_{ijkl} - \bar{p}_{ijk.}\right)^2}{\frac{\bar{p}_{ijk.}\left(1-\bar{p}_{ijk.}\right)}{qn_{\ldots l}}} \right)_{r\times c\times l}$$

$$\therefore \chi 2 = \left(\chi_{ijk}^2\right)_{r\times c\times l}.$$

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

(i)

$$H_0: p_{ijk1} = \ldots = p_{ijkm}\ ;$$

or, the hypothesis of testing the equality of the each $(i,j,k)^{\text{th}}$ individual probabilities of the m population probability volume matrices $\left(P_{ijk1}\right)_{r\times c\times l}, \left(P_{ijk2}\right)_{r\times c\times l}, \ldots, \left(P_{ijkm}\right)_{r\times c\times l}$.

(ii)

$$H_0: \left(p_{ijk1}\right)_{c\times l} = \cdots = \left(p_{ijkm}\right)_{c\times l};$$

or, the hypothesis of checking the equality of the i^{th} row probability matrix or frequency distribution for all populations. Actually, it tests the equity of the frequentness of the i^{th} variable of the first category over all cells of the second and third categories of m population contingency volume tables. Indeed the equality of the frequency matrix distribution of the i^{th} variable of the 1st category is tested over m populations. That is, m (types of) frequency matrix distributions are being tested whether equal or not for same variable. So, over a variable the equity of m frequency matrix distributions drawn from m populations is being tested.

(iii)

$$H_0: \left(p_{ijk1}\right)_{r\times l} = \cdots = \left(p_{ijkm}\right)_{r\times l}$$

or, the hypothesis of checking the equality of the jth column matrix for all populations. Indeed, it tests the equity of the frequentness of the jth variable of the second category over all cells/variables of the 1st and 3rd categories of m population contingency volume tables. The frequency matrix distribution of the jth variable of the 2nd category is tested whether equal or not over m populations.

(iv)

$$H_0: \left(p_{ijk1}\right)_{r\times c} = \cdots = \left(p_{ijkm}\right)_{r\times c};$$

or, the hypothesis of checking the equality of the i^{th} layer probability matrix or frequency distribution for all populations. Actually, it tests the equity of the frequentness of the kth variable of the 3rd category over all cells of the first and second categories of m population contingency volume tables. Indeed the equality of the frequency matrix distribution of the kth variable of the 3rd category is tested over m populations.

(v)

$$H_0: P_1 = \ P_2 \ \ = \cdots . = P_m;$$

or the hypothesis of testing the equity of the total contingency volume table or volume matix for one population is significantly varying to that of the other populations. It tests the similarity of m populations where each of the m populations has joint frequency volume distributions over rcl cells or whether the m types of sample-joint volume frequency distributions or volume matrices or volume tables are drawn from same population.

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

(i) Test of equality of m [(i,j,k)th] cell frequencies: Comparing each χ^2_{ijk} 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 frequency plate/matrix distribution**s**: Comparing each $\sum_{jk}\chi^2_{ijk}$ with the tabulated $\chi^2_{(ck(m-1),.\propto)}$ of $ck(m-1)$ degrees of freedom,

(iii) Test of equality of m [jth variable's] column marginal frequency plate/matrix distributions: Comparing each $\sum_{ik}\chi^2_{ij}$ with the tabulated $\chi^2_{(rk(m-1),.\propto)}$ of $rk(m-1)$ degrees of freedom,

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

(v) Test of equality of m joint frequency distributions**:** Comparing Chi-squares' matrix sum $= \chi^2_{111} + \cdots + \chi^2_{1c1} + \cdots + \chi^2_{r11} + \cdots + \chi^2_{rc1} + \ldots + \chi^2_{112} + \cdots + \chi^2_{1c2} + \cdots + \chi^2_{r12} + \cdots + \chi^2_{rc2} + \ldots + \chi^2_{11k} + \cdots + \chi^2_{1ck} + \cdots + \chi^2_{r1k} + \cdots + \chi^2_{rck}$ with the tabulated $\chi^2_{(rck(m-1),.\propto)}$ of $rck(m-1)$ degrees of freedom.

Suppose we have two contingency 2 × 3 × 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

and

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

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

The problem is to gauge whether the two 3-D contingency tables show significant dissimilarity, to assess, for example, whether they have a common joint distribution or tri-variate distribution that is whether two tri-variate samples come from same tri-variate distribution. If each of the samples were generated at random, we like to use our proposed statistical method for assessing the similarity of two joint frequency distributions. Due to a quick unavailability of the replicates of two types of tri-variate samples, we are assuming that 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 frequency volumes or volume matrices are

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

and

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

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

Therefore, the average relative frequency volume tables or average probability volume tables or matrices are as below

Lower Level	*Not a biter*	*Mild biter*	*Flagrant biter*
Mice	0.06	0.05	0.08
Guinea pigs	0.06	0.04	0.16

Higher Level	*Not a biter*	*Mild biter*	*Flagrant biter*
Mice	0.08	0.07	0.09
Guinea pigs	0.08	0.05	0.18

and

$Lower\ Level$	$Not\ a\ biter$	$Mild\ biter$	$Flagrant\ biter$
$Mice$	0.16	0.09	0.07
$Guinea\ pigs$	0.03	0.02	0.08

$Higher\ Level$	$Not\ a\ biter$	$Mild\ biter$	$Flagrant\ biter$
$Mice$	0.19	0.12	0.08
$Guinea\ pigs$	0.05	0.03	0.08

The averages transition probability volume matrices result as follows

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

and

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

So, p value matrix for lower level $= \begin{pmatrix} 4.09 \times 10^{-111} & 2.98 \times 10^{-28} & 0.02 \\ 3.02 \times 10^{-36} & 2.78 \times 10^{-21} & 4.1 \times 10^{-100} \end{pmatrix}$

and

p value matrix for higher level $= \begin{pmatrix} 6.81 \times 10^{-129} & 1.12 \times 10^{-42} & 4.38 \times 10^{-6} \\ 2.43 \times 10^{-27} & 1.77 \times 10^{-16} & 9.27 \times 10^{-98} \end{pmatrix}$

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 12 chi-square test statistics for 12 types of cells in the volume matrix of chi-squares. For the first cell (mice, not a biter), the calculated value (= 502) of chi-square test statistic is greater than the tabulated value (= 6.634897) which means the null hypothesis

$$H_0\colon p_{mice,\ not\ a\ biter,\ lower\ level} = q_{mice,\ not\ a\ biter,\ lower\ level}$$

is rejected at 1 percent level of significance with p value 4.09$\times 10^{-11}$. So, we conclude that the joint probability of two populations for the joint occurrence of mice with not a biter at lower level is dissimilar and we denote the dissimilarity by a notation "DS".

Again for the joint frequentness (mice and flagrant biter), the null hypothesis

$$H_0\colon p_{mice,\ flagrant\ biter,\ lower\ level} = q_{mice,\ flagrant\ biter,\ lower\ level}$$

is not rejected at the same level of significance. It can be inferred that the frequentness of contemporarily happening of mice with no biter for two population joint distributions is similar and we denote similarity by a notation "S". So the resultant decision matrix for the 12 various cells is given below:
\

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

and

the resultant decision matrix for Higher Level = $\begin{pmatrix} DS & DS & DS \\ DS & DS & DS \end{pmatrix}$

Moreover, the calculated value of overall chi – square, the sum of all individual chi-squares of the chi-squares' matrix sum, is obtained as 2745. Therefore, the null hypothesis

$$H_0: P_{2\times3\times2} = Q_{2\times3\times2}$$

of the equality of joint probability matrix of two populations' joint probability distribution is rejected at 1 % level of significance (since the tabulated value of the chi-squares matrix sum with 12 degrees of freedom is 26.22). So, with an overall point of view, it can be concluded that the two population joint distributions are dissimilar or do not belong to the same tri-variate distributions.

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, that is, 1st column of the one category and that of the same category over same level for the two populations are dissimilar and so forth.

The sum of chi- squares for the 1st and 2rd lower-leveled-rows are calculated as 629, 792 respectively. The tabulated value of the column wise sum of chi-squares with 6 degree of freedom is 16.81 at 1 % level of significance. Again, the sum of chi- squares for the 1st and 2rd higher-leveled-rows are calculated as 699, 625 respectively. So, all rows are dissimilar for the two population joint distributions, that is, 1st row of the one category and that of the same category over same level for the two populations are dissimilar and so forth.

The sum of chi- squares for the upper layered mice-row plate is calculated as 1421. The tabulated value of the upper layered mice-row-plated sum of chi-squares with 6 degree of freedom is 16.81 at 1 % level of significance. Again, the sum of chi- squares for the lower layered pig-row plate is calculated as 1324. So, all row-plates are dissimilar for the two population joint distributions, that is, layer plated row of the one category and that of the same category for the two populations are dissimilar and so forth.

The sum of chi- squares for the Not a Bitter-layered column plate is calculated as 1360. The tabulated value of the layered column plated 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 Mild Bitter-layered column plate is calculated as 467 and that of Flagrant Bitter plate is 918. So, all column-plates are dissimilar for the two population joint distributions, that is, layer plated column of the one category and that of the same category for the two populations are dissimilar and so forth.

So, the marginal frequencies of one category over various row(s) or column(s) or layer(s) or plates in one population is dissimilar to those of the same category over the same row(s) or column(s) or layer(s) or plates in the another population. The dissimilarity between the all row-wise marginal probabilities, column-wise marginal probabilities, row-plate-wise marginal probabilities, column-plate-wise marginal probabilities and all most all cell probabilities of the two joint frequency volume-matrices is also a potential evidence of ensuring the conclusion that the two tri-variate populations are dissimilar.

Conclusion

We are trying to develop the mathematical and graphical representation of several types of Partial Linear and Non Linear Correlation Plate for the three dimensional case.

Reference

(i). 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.

(ii). Islam, K. and Adnan, M. A. S. (2016). New Approach of Mathematical Operations for Volume Matrices. Unpublished manuscript presented in the 2016 Joint Mathematical Meetings organized by American Mathematical Association, held in Seattle, Washington, USA during January 6-9, 2016.

(iii). 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.