Open Science Journal of Statistics and Application
2015; 3(3): 25-28
Published online September 18, 2015 (http://www.openscienceonline.com/journal/osjsa)

Analyzing Transportation Preference Using Contingency Table

Friday Zinzendoff Okwonu

Department of Mathematics and Computer Science, Delta State University, Abraka, Nigeria

Emailaddress

fzokwonu_delsu@yahoo.com

To cite this article

Friday Zinzendoff Okwonu. Analyzing Transportation Preference Using Contingency Table. *Open Science Journal of Statistics and Application.* Vol. 3, No. 3, 2015, pp. 25-28.

Abstract

We investigate transportation preference among students using contingency table by comparing between two profile variables (cheap and fast). The comparison is done by investigating students' patronage of either mini buses or kekenapep by determining if the latter is preferred over the former due to cost (cheap) and fast. We applied the chi square test, likelihood ratio chi square test, Mantel-Haenszel chi-square and the phi coefficient to determine if the above factors influence transportation system patronage. The computed results indicated that the null hypothesis is rejected is rejected in favor of the alternate hypothesis. This implies that students prefer kekenapap over mini buses due to few sitting space and easy of charter. The phi coefficient based on 0-1 scale indicated that the dependency of these profile variables is moderate.

Keywords

Contingency Table, Keke Napep, Mini Buses, Chi Square, Likelihood Ratio Chi Square

1. Introduction

The term contingency table was first used by Pearson in 1904; the emphasis was based on the theory of association (Pearson 1904). The contingency table is also called the frequency table, a table of count (Barak 2007). Contingency table is applied to organize categorical variables and testing hypothesis using the chi-square test to determine independence. In general, the contingency table describes the relationship between two categorical variables. Contingency table helps to determine the effectiveness of a system under study or if the effectiveness of the system is based on certain profile variables (Albrecht 2013). The chi-square test is hypothesis testing technique that produces statistics that is approximately distributed as the chi square distribution (Howell 2012). It is advisable to use contingency table and chi-square test before decisions are taken, this is due to the information it reveals about the profile variables. The Chi squared and the likelihood ratio chi squared statistics are applied to investigate the existence and non-existence of a relationship between objects. Though, the contingency table is computationally complex when $k = 4$ which can though be computed since the model is not expressed in multiplicative pattern and the maximum likelihood estimates of the expected value cannot be written as a closed form expression of the observed values (Tomizawa 1993; McCullagh 1978; Gomez-Villegas 2005; Lauritzen 2002). The contingency table have been applied to different discipline to determine the association and non-association of the factors been compared or treatment and response. The interpretation of the contingency table and the measures of association was extensively discussed by (Simpson 1951), this concept has extensive applications. Different tests procedure have been advocated to test the efficacy of the contingency tables, one of such test procedure is the Fisher exact test coined after Sir Ronald Aylmer Fisher. This test was applied to test homogeneity of proportion of the contingency table (West 2002; Fisher 1954).West and Hankin, generalized the Fisher exact testto the cases whenzeros appears in some cells of the contingency tables (West 2002). The 2 by 2 contingency tables have been discussed extensively by (Howard 1998). In this paper, transportation preference based on cost and comfort (fast) is analyzed using contingency table. The Chi-square procedure is applied to determine if cost and fast are independent with respect to transportation system.

In this consideration, transportation system simply refers to the means of transportation by Delta state university students, say mini buses and kekenapep. Our goal is to infer

while students prefer either of these. Specifically, the term fast simply mean the number of sitting space per the transport systems (mini buses or kekenapep) and the time of departure/arrival and cost refers to per space fare. Based on this, the null and alternative hypotheses are described as follows;

H_0 : Students prefers mini buses because it is cheap,

H_1 : Students preferkekenapep because it is fast.

Since the Chi-square value is unbounded, its modification known as the Phi coefficient (Fleiss 2003) is used to measure association of transportation system rely strictly on cost and fast as the preferred profile variables.

The rest of this paper is organized as follows. The general form of the contingency table is contained in Section Two and the 2 x 2 contingency table for transportation data are described in Section Three. The likelihood chi square test is contained in Section Four. Section Five contains results and discussion while conclusion is contained in Section Six.

2. The Contingency Table

A contingency table, though depending on the number of rows and columns can be denoted and defined in a matrix form, say

Table 1. *General Form of Contingency Table.*

A/B	1	C	Σ
1	N_{11}	N_{1C} nd	$N_{1.}$
R	N_{R1}	N_{RC}	$N_{R.}$
Σ	$N_{.1}$	$N_{.C}$	N

The values of N_{ij} denotes the observed values, where $N_{i.}$ and $N_{.j}$ denotes the row and column sums or simply called the marginal counts (Andel 2002; Lauritzen 2002; Gomez-Villegas 2005). The probabilities of the observed values based on each cell is computed and denoted by

$$p_{ij} = N_{ij} / N, N = \sum_{ij=1}^{k} N_{ij}. \tag{2.1}$$

Equation (2.1) implies that all the cells in Table 1, say all the N objects are stochastically independent. Equation (2.1) satisfies the following conditions:

1. $p_{ij} \geq 0,$
2. $\sum_{ij} p_{ij} = 1.$

From the above, Equation (2.1) is unrestricted and as such, $p_{ij} = p_{i.} \times p_{.j}$

Is independence (Lauritzen 2002). Conventionally, this procedure describes relationships or non-relationships between factors been considered.

3. 2 x 2 Contingency Table for Transportation Data

This paper investigates student transportation preference using the kekenapep and the mini buses as a case study in Delta state university, Abraka. This survey was carried out for the period of three months (March to June, 2014). The objective is to investigate the reason while some categories of students preferkekenapep over the mini bus and also to determine otherwise. The data reported in this paper are based on extensive questionnaire survey in Abraka, Delta State. Precisely, about 1535 students were selected for this survey of which 667 prefer kekenapep and 871 prefer mini buses. In this survey, the profile variables used are cheap and fast. This implies that students may prefer either kekenapep because it is fast due to few sitting space or mini buses because it is cheap. Based on the information of the survey, we apply 2 by 2 contingency table to determine if transportation systems (kekenapep or mini buses) and the profile variables influence transport system patronage. As noted in (Ingersoll G.M. 2010) a 2 x 2 contingency tableis used to conceptualize, organize and report data. The null hypothesis tested with a chi-square test based on a 2 x 2 contingency table is considered as test of independence (Ingersoll G.M. 2010). It is a well-established fact that the expected frequency should be at least five to enable the application of the chi-square meaningful otherwise other techniques are applied. Detail of the survey is reported in Table 2 below.

Table 2. *Observed value for the transportation data.*

Transportation systems	Cheap	Fast	Total
Kekenapep	212	452	664
Mini buses	556	315	871
Total	768	767	1535

In this consideration, we consider the test of independence by using the above data set. In this case, the Chi square distribution involves using the sample data to test for independence of the two variables. In this case we consider whether preference for transportation depends on cheap or fast. If preference for transportation depends on cost, we will further research on how best we could devise cheap transportation means. If otherwise, we will improve on fast transportation means among students. For the 2x2 contingency table, the Chi square test value based on the observed and expected frequencies is computed based on the following formula

$$\chi^2 = \sum_i \sum_j (OV_{ij} - EF_{ij})^2 / EF_{ij}. \tag{3.1}$$

Where OV denote the observed value and EF denote the expected frequency. Note that Equation (3.1) is approximately distributed as chi square on $(n-1)(k-1)$ degrees of freedom, where n and k denotes the rows and columns, respectively. Observe that if the value of the observed value is greater than the expected frequency then

the chi-square value will be small. On the contrary, if the expected frequency is greater than the observed value the chi square value tends to be large. In this case, since the expected frequency is greater than five for all cases, it is straight forward to apply the Chi square test statistic.

Table 3. *Expected frequency for the transportation data.*

Transportation systems	Cheap	Fast	Total
Kekenapep	332.2	331.8	664
Mini buses	435.8	435.2	871
Total	768	767	1535

The chi square table value at 5% level of significant with $(n-1)(k-1)=(2-1)(2-1)=1$ degrees of freedom is equal to 3.84. Since the computed value is greater than the table value, we reject the null hypothesis and conclude that student prefer keke nape over mini buses because kekenapep is fast due to few sitting spaces. Many reasons exist for this:

I. kekepnapep carry maximum of three passengers whereas the mini buses can take about 18 passengers (main problem: ticketing period, that is sourcing for passengers),
II. A passenger may decide to charter the kekenapep due to few sitting space but same is not true for mini buses, etc.
III. Departure, stoppage and arrival is advantageous tokekenapepuser over mini bus userand patronage of narrow route e.t.c.

Therefore mobility is independent of the transportation systems and the profile variables. We conclude that they are dependent.

Table 4. *Computed Chi square value for transportation data.*

Observed value (OV)	Expected frequency (EF)	D=(OV-EF)	D^2/EF
212	332.2	-120.2	43.492
452	331.8	120.2	43.544
556	435.8	120.2	33.153
315	435.2	-120.2	33.199
			153.4

4. The Likelihood Ratio Chi Square

In this section, the objective is to compute and compare the Likelihood ratio Chi square value with the Chi square table value in other to infer more information. The likelihood ratio Chi square is defined as

$$G^2 = 2\sum\left[O_{ij}\log(O_{ij}/E_{ij}\right]. \quad (4.1)$$

Though, this formula is applicable for a large dimensional table which can be decomposed into smaller components (Howell, 2012), this allows its application to this data set valuable. The likelihood Chi square can be applied to investigate if there is a significant difference between observed and the expected frequency (Olmus 2012). Using the formula above, the G^2 is equal to 156.3. This implies that $G^2 > \chi^2$. Suppose, we are to compare the likelihood ratio chi square value with the chi square table value at 0.05 level of significant. The conclusion will corroborate with the rejection of the null hypothesis. This result revealed that both techniques converge as the sample sizes increases.

It is evident that the larger the G^2 value the higher the probability of rejecting the null hypothesis. For large sample size, G^2 has a chi square null distribution with $k-1$ degree of freedom. If the null hypothesis is true, both the Pearson chi square χ^2 and the likelihood chi square G^2 have asymptotic chi square distributions with $k-1$ degree of freedom (Agresti 2002). On the other hand, they are asymptotically equivalent, if $\chi^2 - G^2$ converges in probability to zero. For fixed k and increasing sample size, the distribution of the Chi square converges to Chi square faster than the likelihood ratio chi square. The Chi square approximation is poor for G^2 if $n/k < 5$ (Agresti 2002). Observe that the precise distribution of the Chi square and the Likelihood Chi square is intractable (Lauritzen 2002), as such, any applicable inference of the Chi square and the Likelihood Chi square distributions are based on the asymptotic distribution which is the Chi square itself with degrees of freedom (Lauritzen 2002).

The phi coefficient was recently applied to investigate the dependency of locations and facilities on cigarette consumption in Penang Island, Malaysia (Okwonu 2014). In this paper, we will apply the method adduced by Okwonu et al., 2014 to investigate the preference of transportation system over cost and fast in the next section using the phi coefficient and Mantel - Haenszel chi square.

5. Results and Discussion

Although the null hypothesis was rejected indicating that there is a relationship between transportation system and the profile variables, the chi square value of 153.4 will produce moderate Phi coefficient to indicate a moderate relationship between these two profile variables. The Phi coefficient for this data set is 0.3162. The Mantel-Haenszel chi-square value for this data set is 153.3 is equivalent to the Pearson chi square value and the likelihood ratio chi square value, respectively. The moderate Phi coefficient revealed that there is moderate association between transportation system and the profile variables.

In addition to concluding that the transportation system and the profile variables have a relationship, we can also infer on the differences in the proportions in the 2 x 2 table (Utts 2006; Walpole 2011). They are: Observation 1 (Keke napep): students prefer keke to mini buses because of few sitting space and easy of charter. Observation 2 (Mini bus):

most student prefer mini buses to keke because it is cheaper. The analysis revealed students preference in terms of transportation system which depends strictly on the profile variables. Based on this analysis, the transport managers can infer quality information on how to improve the transportation system in Abraka, Delta State.

6. Conclusions

Results based on the Chi Square, the Likelihood ratio Chi square, and the Mantel-Haenszel Chi-square value revealed that the null hypothesis is rejected which implies that transportation system is dependent on the profile variables. The Chi square value of 153.4 and Mantel-Haenszel Chi-square value of 153.3 will produce moderate phi coefficient to indicate a moderate relationship between these two profile variables. The phi coefficient, however, on a scale of zero to one indicated that the dependency of these variables is moderate. Conclusively, the analysis revealed that students prefer kekenapep over mini buses due to the easy of charter, time and few sitting space.

References

[1] Agresti, A., 2002, Categorical data analysis (John Wiley & Sons, INC., Hoboken, New Jersey, Hoboken).

[2] Albrecht, C., 2013, Contingency table and the chi square statistics: Interpreting computer printouts and constructing tables, *http://extension.usu.edu/evaluation/files/uploads/Start%20Your%20Engine/Study%20the%20Route/Analyze%20the%20Data/Interpreting_Chi_Square_Printouts.pdf*.

[3] Andel, J., 2002, Zaklady matematicke statistiky, *MFF UK Praha*, 283-317.

[4] Barak, B., Chaudhuri, K., Dwork, C.,Kale, S., McSherry, F., and Talwar, K., 2007, Privacy, accuracy and consistency too. A holistic solution to contingency table release(http://cseweb.ucsd.edu/~kamalika/pubs/bcdkmt07.pdf, ACM 978-1-59593-685-01/07/0006, Beijing, China).

[5] Fisher, R. A., 1954, Statistical methods for research workers, *Oliver and Boyd*.

[6] Fleiss, J. L., Levin, B., and Paik, M.C., 2003, Statisyical Methods for Rates and Proportions (John Wiley & Sons Inc, Hoboken, NJ.

[7] Gomez-Villegas, M. A., and Perez, B. G., 2005, Bayesian analysis of contingency tables, *Communication in Statistics-Theory and Methods* 34, 1743-1754.

[8] Howard, J. V., 1998, The 2x2 table: A discussion from a Bayesian viewpoint, *Statistical Science* 13, 351-367.

[9] Howell, D. C., 2012, Chi-square test-analysis of contingency tables, *http://www.uvm.edu/~dhowell/methods8/supplement*.

[10] Ingersoll G.M., 2010, Analysis of 2x2 contingency tables in educational research and evaluation, *International Journal for Research in Education*, 1-14.

[11] Lauritzen, S. L., 2002, Lectures on contingency tables, *http://www.stats.ox.ac.uk/~steffen/papers/cont.pdf*.

[12] McCullagh, P., 1978, A class of parametric models for the analysis of square contingency tables with ordered categories, *Biometrika* 65, 413-418.

[13] Okwonu, F. Z., Othman, A. R., and Dieng, H., 2014, Using contingency table to analyze cigarrete consumption in Sungai Dua and Batu Uban area of Penang Island, Malaysia, *Tecnica Vitivinicola*, 297-301, 2014.

[14] Olmus, H., and Erbas, S., 2012, Analysis of traffic accidents caused by drivers by using log-linear model, *Promet- Traffic and Transportation* 24, 495-504.

[15] Pearson, K., 1904, On the theory of contingency and its relation to association and normal correlation, the analysis of contingency table. Forward to Draper company research memoirs, *Biomteric*.

[16] Simpson, E. H., 1951, Interpretation ofinteraction in contingency tables, *Journalof the Royal Statistical Society:Series B (Methodological)* 13, 238-241.

[17] Tomizawa, S., 1993, A simple statistics to test generalized Palindromic symmetry model in a 4 x 4 contingency table, *Questiio* 17, 33-38.

[18] Utts, J. M., Heckard, R.F., 2006, Mind on Statistics (Cengage Learning, Belmont, CA.

[19] Walpole, R. E., Myers, R.H., Myers, S.L., Ye, K.E.,, 2011, Probability and Statistics for Engineers and Scientists (Pearson, New York.

[20] West, L. J., and Hankin, R. K. S., 2002, Exact tests for two way contingency tables with structural zeros, *http://cran.r-project.org/web/packages/aylmer/vignettes/fishervig.pdf*.