Robust Mantel-Haenszel Test using Probabilistic Approach

Awopeju K.A[1*], Umeh E.U[2], and Ajibade B.F[3].

[1,2]Department of Statistics, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria.

[3]Petroleum Training Institute, Warri, Effurun, Delta State, Nigeria

ak.awopeju@unizik.edu.ng, abidemiawopeju@gmail.com

Abstract

Mantel-Haenszel test statistic is one of the common test statistics for test of significance variation between/among factors and its application is similar to One-way Analysis of Variance and Kruskal-Wallis test statistics. The method can be categorized as non-parametric and robust in nature. It has been used over time by researchers for test of significance variation among factors. Critical look at the test statistic reveals it weakness which is inability to remove variation among factors in terms of sample size or weight. To remove biasness in the test of hypothesis with Mantel-Haenszel test as the statistic, there is need for proper and appropriate modification. This paper addressed the noticed short fall of the test statistic with illustrative example for easy computation by users. Similar data used by researchers in the past was also used in the study using the propose method called modified Mantel-Haenszel test statistic.

Keywords: Variation, Non-parametric, Probability, Sample Space, Hyper-geometric

1.0 Introduction

In research, researchers often encounter problems of comparing factors with similar characteristics which may not necessarily come from the same group such as financial factors of countries, human trait, etc. there are numerous test statistics that can be used which include One-way or Two-way Analysis of Variance; depending on the design of the factors, Student T test, Mann-Whitney, Kruskal-Wallis test etc. Also, Mantel-Haenszel can be used to test for significance variation among factors (Mantel and Haenszel, 1959; Peto abd Peto, 1972; John, 2013).

Considering the test statistics suggested, all have certain characteristic or assumptions that must be met before usage. Critical look at the Mantel-Haenszel test statistic, the method was derived from Hyper-geometric distribution and close look at the application (see Paul and Brani, 2007), it is not compulsory for the factors to have equal length in the comparison but the observations must be from the same factor of interest. This gives room for comparison of unequal observations which can contribute to higher chance of type-I or type-II error. In the example used by Paul and Brani, the total number of objects considered for first group is 508, 213 for second group and 250 for the third group. Comparing the groups, there is already significant variation in the number of observations for the groups which implies test of significant variation using the data may result to wrong choice of hypothesis (Williams, 1954).

The primary objective of the research is to correct the anomalies in the existing method; Mantel-Haenszel Test, giving the standardized method capable of equalizing the variation in the observations and at the same time compare for significant variation among factors of interest.

2.0 Existing Method

According to Mantel and Haenszel (1959), the test statistic can be used to test if proportions of factor are significantly different or not. Mathematically, the test is

$$T = \frac{\sum_{i=1}^{k} x_i - \sum_{i=1}^{k} \frac{r_i c_i}{n_i}}{\sqrt{\sum_{i=1}^{k} \left[\frac{r_i c_i (n_i - r_i)(n_i - c_i)}{n^2 (n_i - 1)} \right]}} \quad (1)$$

where k is number of independent classifications, r_i is row total, c_i as column total and n_i as grand total. This implies the observations must be arranged in row-column wise. As observed from Equation (1), the test statistic is derived from hyper-geometric distribution as the denominator is the standard deviation of hyper-geometric distribution and part of the numerator is mean of hyper-geometric distribution (Mantel, 1963).

Considering hyper-geometric distribution,

$$\frac{\binom{r_i}{x_i}\binom{n_i - r_i}{c_i - x_i}}{\binom{n_i}{c_i}}$$

with mean $\frac{r_i . c_i}{n_i}$

and variance $\frac{r_i . c_i (n_i - r_i)(n_i - c_i)}{n_i^2 (n_i - 1)}$.

From the expressions, n is the grand total for each group and r and c are row and column total for each group.

As seen in Paul and Brani (2007; Page 170);

Table 1: Data used by Paul and Brani (2007) for Mantel- Haenszel Test Statistic

	Zhengzhou			Taiyuan			Nanchang		
Cancer Diagnosis	Yes	No	Total	Yes	No	Total	Yes	No	Total
Smoker	183	156	338	60	99	159	104	89	193
Non-smoker	72	98	170	11	43	54	21	36	57
Total	254	254	508	71	142	213	125	125	250

Conclusion reached in the example, is that the test statistic value is 3.95 and the null hypothesis rejected. From Table 1, the total numbers of people from each location vary significantly; therefore, meaningful comparison may not be reached as the numbers (Totals) are unequal.

Proportion as a method of standardizing the observations can be used to normalize the natural variation among the location such that probability of having the disease given that the person smokes cigarette is used across the locations.

3.0 Proposed Method

Given a "2X2" observations, the entries can be arranged thus;

Table 2: Row and Column Observations

	C_1	C_2	C_j
R_1	R_1C_1	R_1C_2	R_1C_j
R_2	R_2C_1	R_2C_2	R_2C_j
R_i	R_2C_i	R_iC_2	R_iC_j

where R_iC_j is grand total. The proportions for each entry are;

$$P(R_1C_1)=\frac{R_1C_1}{R_iC_j},\ P(R_1C_2)=\frac{R_1C_2}{R_iC_j},\ \ldots,\ P(R_iC_2)=\frac{R_iC_2}{R_iC_j}.$$

Table 2 entries will now become;

Table 3: Row and Column Observations

	C_1	C_2	C_j
R_1	$P(R_1C_1)$	$P(R_1C_2)$	$P(R_1C_j)$
R_2	$P(R_2C_1)$	$P(R_2C_2)$	$P(R_2C_j)$
R_i	$P(R_2C_i)$	$P(R_iC_2)$	$P(R_iC_j)=1$

From Table 3, the total population becomes total proportion for the group which is 1. Marginal total for row is $P(R_i)$ and marginal total for column is $P(C_j)$. Then, the Probability Mass Function (P.M.F) of the distribution can be written as;

$$P(x_i)=\frac{\binom{P(R_i)}{x_i}\binom{1-P(R_i)}{P(C_j)-x_i}}{\binom{1}{P(C_j)}},\ i=1,2,3,\ldots,n\ and\ \ j=1,2,3,\ldots,n.$$

In the probability function, P.M.F is used since the distribution is discrete in nature; countable observations. Restricting the P.M.F to 2×2 design, we have;

$$P(x_i) = \frac{\begin{pmatrix} P(R_i) \\ x_i \end{pmatrix}\begin{pmatrix} 1-P(R_i) \\ P(C_j)-x_i \end{pmatrix}}{\begin{pmatrix} 1 \\ P(C_j) \end{pmatrix}}, i = 1, 2. \text{ and } j = 1, 2..$$

Generalized approach for more than 2×2 contingency table can also be derived using similar approach.

The test statistic becomes;

$$T = \frac{k.\sum_{i=1}^{k} p_i - \sum_{i=1}^{k} \frac{p(r_i).p(c_i)}{1}}{\sqrt{\sum_{i=1}^{k}\left[\frac{p(r_i)\,p(c_i)(1-p(r_i))(1-p(c_i))}{1}\right]}} \quad (2)$$

But $\sum_{i=1}^{k} p_i = 1$,

Therefore, Equation 2 can be written as;

$$T = \frac{k - \sum_{i=1}^{k} p(r_i).p(c_i)}{\sqrt{\sum_{i=1}^{k} p(r_i)\,p(c_i)(1-p(r_i))(1-p(c_i))}} \quad (3)$$

Using Equation 3, the variations in the observations that can lead to increase in type I error has been removed.

Illustrative Example:

Using the data on Table 1, and the proposed method, Table 1 can be restructured to have;

Table 4: Modification of Data used by Paul and Brani (2007)

	Zhengzhou			Taiyuan			Nanchang		
Cancer Diagnosis	Yes	No	Total	Yes	No	Total	Yes	No	Total
Smoker	0.3602	0.3071	0.6654	0.2817	0.4648	0.7465	0.416	0.356	0.772
Non-smoker	0.1417	0.1929	0.3346	0.0516	0.2019	0.2535	0.084	0.144	0.228
Total	0.5	0.5	1	0.3333	0.6667	1	0.5	0.5	1

Hypothesis of interest

Null hypothesis: the total numbers of people from each location vary insignificantly

Alternative hypothesis: the total numbers of people from each location vary insignificantly

Decision Rule

Accept the null hypothesis if the calculated value is less than the table value (critical region). Otherwise, reject.

Calculated Value using Proposed Method (Equation 3):

$$T = \frac{k - \sum_{i=1}^{k} p(r_i).p(c_i)}{\sqrt{\sum_{i=1}^{k} p(r_i)\,p(c_i)(1-p(r_i))(1-p(c_i))}} \quad (3)$$

$$\sum_{i=1}^{3} p(r_i).p(c_i) = p(r_{11}).p(c_{11}) + p(r_{21}).p(c_{21}) + \ldots + p(r_{23}).p(c_{23})$$

$$= 1.4178169 \approx 1.4178$$

Therefore, the numerator of the expression is;

$$k - \sum_{i=1}^{k} p(r_i).p(c_i) = 3 - 1.4178 = 1.5822$$

For the denominator,

$$\sum_{i=1}^{3} p(r_i)\,p(c_i)(1-p(r_i))(1-p(c_i)) = 0.111322 + 0.0841 + 0.088008 = 0.28343$$

$$T = \frac{1.5822}{\sqrt{0.28343}} = \frac{1.5822}{0.53238} = 2.97193 \approx 2.9719$$

Calculated value using the test statistic is 2.9719.

T is approximately normal and in the previous example, the test was done at level of significance of 0.10 which implies the table value is 1.65 which was compared with calculated value of 3.95 and since the calculated value is greater than the table value, the null hypothesis was rejected.

Using the same approach to create bases for comparison, using level of significance of 10%, the table value remains 1.65 and the calculated value is 2.97 which is greater than the table value. This implies there is enough evidence to equally reject the null hypothesis and conclude that the distribution of cancer patients among the cities is significantly different.

The conclusion reached using the propose method is the same as the conclusion reached using the earlier formulated expression but the propose method is better used for such hypothesis testing as it is free from biasness in the choice of sample size from each location which could affect the decision/conclusion reached.

3.0 Summary of Findings and Conclusion

Mantel-Haensel test statistic was proposed by two researchers in 1959 and was published with a reputable journal. Critical look at the test statistic formulated reveals its weakness as it fails to consider the number of subjects selected from locations considered as it can be used for n-locations. Recall that the test statistic is used to the determination of presence of significance difference among factors but the bases for comparison should

be that the number of subjects or sample sizes considered from each location or factor is uniform. Neglecting this fact could increase type I error of the test statistic thereby resulting to erroneous conclusion.

In this paper, the test statistic is modified to correct the anomalies that may set-in as a result of non-uniformity of sample size as shown in the illustrative example. The proposed method does not restrict choice of sample size in the selection of subjects but normalize the data by using probabilistic approach considering the fact that irrespective of the location, the P(S) is 1.

For validation of propose method, earlier used example was used in the paper and the conclusion reached as the same as the conclusion of the previous researchers.

Conclusively, the modified Mantel-Haensel test statistic is highly recommended for test of significant variation among factors irrespective of sample size variation among the factors.

References

1. John M. Lachin (2013) Power of the Mantel-Haenszel and Other Tests for Discrete or Grouped Time-to-event Data Under a Chained Binomial Model. Stat Med.; 32(2): 220–229. doi:10.1002/sim.5480.
2. Mantel N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother. Rep.; 50:163–170.
3. Nathan Mantel and William Haenszel (1959). "Statistical aspects of the analysis of data from retrospective studies of disease". Journal of the National Cancer Institute 22 (4): 719–748
4. Nathan Mantel (1963). "Chi-Square Tests with One Degree of Freedom, Extensions of the Mantel-Haenszel Procedure". Journal of the American Statistical Association 58 (303): 690–700
5. Peto R, Peto J. (1972). Asymptotically efficient rank invariant test procedures (with discussion) J. Roy. Statist. Soc. A.; 135:185–206
6. Wallenstein S, Wittes J. (1993). The power of the Mantel-Haenszel test for grouped failure time data. Biometrics; 49:1077–1087.
7. William G. Cochran (1954). "Some Methods for Strengthening the Common χ^2 Tests". Biometrics 10 (4): 417–451