The Development of a Correction Method for Ensuring a Continuity Value of The Chi-square Test with a Small Expected Cell Frequency

Kajita Matchima[1]*, Jumlong Vongprasert[2] and Nipaporn Chutiman[3]

[1]Department of Educational Research and Evaluation, Faculty of Education, Ubon Ratchathani Rajabhat University, Ubon Ratchathani, 34000.

[2]Department of Statistics, Faculty of Science, Ubon Ratchathani Rajabhat University, Ubon Ratchathani, 34000.

[3]Department of Mathematics, Faculty of Science, Mahasarakham University, Mahasarakham, 44150.

* Corresponding Author. E-mail address: kajita.m@ubru.ac.th

Received: 13 March 2017; Accepted: 6 June 2017

Abstract

Using the chi-square test with a small expected cell frequency is an important problem in generally survey and experimental research because it cannot control type-I error led to amiss conclude the result in our work. The purposes of this work were first to develop a correction method for ensuring a continuity value of the chi-square test and secondly to compare its efficiency with other methods, namely; Yate's correction and William's correction by using simulation data. The comparisons were made with the following condition; two significant levels of 0.01 and 0.05, six contingency table sizes (2x2, 2x3, 2x4, 3x3, 3x4 and 4x4), a small expected cell frequency up to 30% of the total cell and a sample size between 5 to 10 times that of the total cell.

We found that type I error in chi-square test with developed correction and significant level is similar values (can control type I error). The similarity values are higher than chi-square test without correction, Yate's correction and William's correction. Larger sample sizes resulted is better control type I error at both levels of significance. For the contingency table size 2x2 to 4x4, chi-square test with developed correction can control type I error better than chi-square test without correction and William's correction at both 0.01 and 0.05 significant levels. The correction method used to control the type-I error was obtained using a developed correction in every condition.

Keywords: Test of independent, Chi-Square Test, Correction method, Type-I error

Introduction

In social and behavioural science research, surveys and experiments use qualitative or categorical measurement methods to determine the results rather than quantitative methods; that is, a quality or characteristic is measured for each experimental unit. We can summarize this type of data by creating a list of the categories or characteristics and report a count of the number of measurements that fall into each category. These are some of the many situations in which the data set has characteristics appropriate for the multinomial experiment. The statistics test created for the multinomial experiment was derived by a British statistician named Karl Pearson in 1900 and is called the 'Chi-square statistic test'. (Mendenhall, Beaver, & Beaver, 2013)

The Chi-square test is only an approximate large-sample test and it is recommended that it not used when one (or more) of the expected frequencies is less than five (Freund, 2004). When the sample sizes are too small, you should not use chi square test or *G*-test. However, how small is "too small"? The conventional rule of thumb is that if all of the expected numbers are greater than 5, it's acceptable to use the chi square or *G*-test; if an expected number is less than 5, you should use an alternative (McDonald, 2014). At large sample sizes, many asymptotic properties of test statistics derived for independent sample comparison are still applicable in adaptive randomization provided that the patient allocation ratio converges to an appropriate target asymptotically. However, the small sample properties of commonly used test statistics in response-adaptive randomization are not fully studied (Gu & Lee, 2010). The researchers recommended that we should not use chi-square test when the expected numbers are small because it will lead to erroneous results of a research study, meaning the conclusion of the research cannot be rightly interpreted. Yate's correction (Yates, 1934) and William's correction methods (McDonald, 2014) are used to test independence of events in a cross table. It is done by reducing the difference between each observed value and its expected value. These tests are commonly used when expected frequencies are less than ten.

In this article, we present our own developed correction method to maintain a continuity value to be used when small expected cell frequencies on chi-square test for independence exist in the research data. The objectives of this study are to compare our developed correction method's efficiency of control of a type-I error with Yate's correction and William's correction methods. The simulation data used the Monte Carlo technique with R programming language in different situations; significance levels, contingency table sizes, sample sizes and the number of small cells. Our powerful correction method will control a type-I error more than the other correction methods.

Materials and Methods

Chi-Square Test

Many experiments result in measurements that are qualitative or categorical rather than quantitative. In these instances, a quality or characteristic is identified for each experimental unit. Data associated with such measurement can be summarized by providing the count of the number of measurements that fall into each of the distinct categories associated with the variable.

In 1900 Karl Pearson proposed the following test statistic which is a function of the squares of the deviations of the observed counts from their expected values weighted by the reciprocals of their expected values:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left[n_i - E\left(n_i\right)\right]^2}{E\left(n_i\right)} = \sum_{i=1}^{k} \frac{\left[n_i - np_i\right]^2}{np_i} \qquad (1)$$

Where there are k categories with probabilities p_i and n_i is sample size in each categories.

Yate's Correction

For the test of independent creating a 2x2 contingency table that used chi-square test, the Yate's correction is usually recommended especially if more than 20% of the expected cell frequencies are below 5. The chi-square formula equation is below (2), Where f_O is observed frequencies, f_e is expected frequencies.

$$\chi^2_{Yates} = \sum^{k} \frac{\left(\left|f_O - f_e\right| - 0.5\right)^2}{f_e} \qquad (2)$$

William's Correction

For the independent test in contingency table with R (row) and C (column), the William's correction for contingency to compute q to divide the chi-square test are as follows (3 and 4).

Where f_O is observed frequencies and f_e is expected frequencies.

$$\chi^2_{William} = \left(\sum^{k} \frac{(f_O - f_e)^2}{f_e} \right) \Big/ q \qquad (3)$$

Where q is defined as (4)

$$q = 1 + \frac{\left(n\{[1/(raw1total)] + ... + [1/(rawRtotal)]\} - 1\right)\left(n\{[1/(column1total)] + ... + [1/(columnCtotal)]\} - 1\right)}{6(R-1)(C-1)} \qquad (4)$$

Where n is samples size, contingency table with R rows and C columns.

Developed Correction

A major limitation of the test of independent with chi-square test is its inability to control a type-I error when an expected frequency is small. An accurate test of independent was needed when the type-I error and significant level had similar values.

After test of independent by classical chi-square (without correction for continuity) we consider type-I error (number rejection of null hypothesis divided by 10,000) and significant level.

Where the type-I error is greater than the significant level, the chi-square test equation to be used is as follows (5)

$$\chi^2 = \sum^{k} \frac{(|f_O - f_e| - C)^2}{f_e} \qquad (5)$$

Where the type-I error is less than the significant level, the chi-square test is (6)

$$\chi^2 = \sum^{k} \frac{(|f_O - f_e| + C)^2}{f_e} \qquad (6)$$

Where the chi-square tests in both situations are defined in (5) and (6), we can define the developed correction of chi-square test by the following (5)

Where C is developed correction value. It was computed in two case as follows; if the type-I error is greater than the significant level we try to replace the value C into the equation (5) start from 0.01, 0.02, 0.03,.... .

If the type-I error is less than the significant level we try to replace the value C into the equation (5) start from -0.01, -0.02, -0.03,.... . After we replace value C and computed type-I error then to compared with significant level. Developed correction value (C) is the value which get type-I error and significant level are very similar values.

Simulation Study

The performances of the three corrections for continuity (Yate's Correction, William's Correction and our Developed Correction) were evaluated using a simulation study with a pattern of data set at a significant level of 0.05 and 0.01. Contingency Tables were generated between 2x2 to 4x4 cells. The numbers of small expected cell frequencies up to 30% of total cell were used. Sample sizes were generated at 5 to 10 times the total cell size. The data was simulated using R programming language on the Monte Carlo technique. The data was simulated 10,000 times for each pattern. For comparison, the accuracy of the three correction methods was evaluated. For each pattern, the simulation was used to find the correction values that best controlled the type-I error. The results were tabulated to display the relationships between the contingency table's pattern, the significant levels and the correction values.

Results

The corrections for continuity include Yate's correction (Y), William's correction (W), our Developed correction (D) and the chi-square without corrections for continuity (N). Their accuracy was compared using a simulation study classified by contingency table size, the number of small expected cell frequencies (No. Sc.), the sample sizes (Ss.) and the significant level. **Table 1-12** indicates the type-I error and developed correction value (C) in each situation.

From table 1-12 found that when we used our developed correction value (C), type-I error was nearer the significant level than the other correction methods reviewed in every situation and its was control type I error better when sample sizes was increased at both levels of significance. The result showed that developed correction can control type-I error better than other methods.

Table 1 Type-I error of 2x2 contingency table in 0.01 significant level.

No.Sc.	Ss.	Type I Error				
		N	Y	W	D	C
1	20	0.0011	0.0304	0.0052	0.0101	-0.76
	30	0.0028	0.0226	0.0016	0.0101	-2.29
	40	0.0038	0.0302	0.0027	0.0100	-1.28

Table 2 Type-I error of 2x2 contingency table in 0.05 significant level.

No.Sc.	Ss.	Type I Error				
		N	Y	W	D	C
1	20	0.0176	0.0638	0.0127	0.0500	-0.22
	30	0.0216	0.0578	0.0084	0.0500	-1.71
	40	0.0324	0.0681	0.0126	0.0500	-0.71

Table 3 Type-I error of 2x3 contingency table in 0.01 significant level.

No.Sc.	Ss.	Type I Error			
		N	W	D	C
1	30	0.0034	0.0013	0.0100	-0.86
	40	0.0039	0.0013	0.0097	-1.00
	50	0.0018	0.0013	0.0103	-2.75
2	30	0.0015	0.0013	0.0103	-0.94
	40	0.0014	0.0010	0.0101	-1.12
	50	0.0014	0.0008	0.0100	-1.80

Table 4 Type-I error of 2x3 contingency table in 0.05 significant level.

No.Sc.	Ss.	Type I Error			
		N	W	D	C
1	30	0.0115	0.0046	0.0500	-0.45
	40	0.0200	0.0015	0.0528	-0.37
	50	0.0034	0.0012	0.0502	-2.21
2	30	0.0018	0.0021	0.0505	-0.58
	40	0.0017	0.0018	0.0503	-0.53
	50	0.0013	0.0016	0.0501	-1.30

Table 5 Type-I error of 2x4 contingency table in 0.01 significant level.

No.Sc.	Ss.	Type I Error			
		N	W	D	C
1	40	0.0009	0.0004	0.0101	-1.42
	60	0.0020	0.0007	0.0103	-1.71
	80	0.0015	0.0006	0.0100	-2.12
2	40	0.0050	0.0017	0.0103	-1.73
	60	0.0090	0.0013	0.0101	-1.32
	80	0.0143	0.0008	0.0100	0.13

Table 6 Type-I error of 2x4 contingency table in 0.05 significant level.

No.Sc.	Ss.	Type I Error			
		N	W	D	C
1	40	0.0011	0.0022	0.0501	-1.02
	60	0.0013	0.0015	0.0501	-1.15
	80	0.0009	0.0010	0.0500	-1.51
2	40	0.0468	0.0024	0.0507	-1.32
	60	0.0489	0.0034	0.0503	-0.84
	80	0.0539	0.0055	0.0500	1.09

Table 7 Type-I error of 3x3 contingency table in 0.01 significant level.

No.Sc.	Ss.	Type I Error			
		N	W	D	C
1	50	0.0023	0.0007	0.0103	-1.36
	70	0.0028	0.0008	0.0101	-1.34
	90	0.0035	0.0014	0.0100	-2.33
2	50	0.0076	0.0019	0.0103	-1.81
	70	0.0120	0.0029	0.0102	0.18
	90	0.0076	0.0021	0.0100	-1.23
3	50	0.0056	0.0015	0.0101	-1.87
	70	0.0086	0.0023	0.0101	-0.44
	90	0.0127	0.0036	0.0100	0.09

Table 8 Type-I error of 3x3 contingency table in 0.05 significant level.

No.Sc.	Ss.	Type I Error			
		N	W	D	C
1	50	0.0018	0.0011	0.0502	-0.87
	70	0.0020	0.0015	0.0502	-0.75
	90	0.0027	0.0021	0.0500	-1.78
2	50	0.0371	0.0007	0.0501	-1.28
	70	0.0560	0.0040	0.0501	0.07
	90	0.0440	0.0032	0.0500	-0.63
3	50	0.0380	0.0021	0.0501	-1.36
	70	0.0471	0.0037	0.0501	-0.15
	90	0.0576	0.0047	0.0500	0.52

Table 9 Type-I error of 3x4 contingency table in 0.01 significant level.

No.Sc.	Ss.	Type I Error			
		N	W	D	C
1	60	0.0020	0.0011	0.0100	-0.70
	90	0.0028	0.0021	0.0099	-1.76
	120	0.0051	0.0034	0.0100	-1.87
2	60	0.0015	0.0009	0.0100	-0.87
	90	0.0031	0.0019	0.0099	-5.63
	120	0.0048	0.0038	0.0100	-1.79
3	60	0.0013	0.0021	0.0101	-0.95
	90	0.0113	0.0028	0.0102	4.50
	120	0.0126	0.0052	0.0100	0.51
4	60	0.0057	0.0014	0.0102	-0.39
	90	0.0078	0.0027	0.0101	-1.55
	120	0.0074	0.0043	0.0100	-0.85

Table 10 Type-I error of 3x4 contingency table in 0.05 significant level.

No.Sc.	Ss.	Type I Error			
		N	W	D	C
1	60	0.0027	0.0021	0.0501	-0.32
	90	0.0019	0.0010	0.0499	-1.35
	120	0.0037	0.0022	0.0500	-1.43
2	60	0.0022	0.0016	0.0501	-0.53
	90	0.0013	0.0013	0.0501	-5.18
	120	0.0024	0.0021	0.0500	-1.35
3	60	0.0058	0.0076	0.0500	-0.68
	90	0.0521	0.0117	0.0501	3.92
	120	0.0534	0.0148	0.0500	0.14
4	60	0.0534	0.0048	0.0501	-0.05
	90	0.0479	0.0062	0.0500	-1.29
	120	0.0479	0.0074	0.0500	-0.43

Table 11 Type-I error of 4x4 contingency table in 0.01 significant level.

No.Sc.	Ss.	Type I Error			
		N	W	D	C
1	80	0.0025	0.0012	0.0101	-0.75
	110	0.0016	0.0016	0.0102	-0.50
	140	0.0029	0.0025	0.0102	-0.98
2	80	0.0020	0.0014	0.0100	-0.32
	110	0.0090	0.0029	0.0098	-0.54
	140	0.0070	0.0034	0.0100	-1.20
3	80	0.0070	0.0022	0.0101	-1.02
	110	0.0070	0.0015	0.0102	-0.50
	140	0.0082	0.0027	0.0100	-0.35
4	80	0.0012	0.0021	0.0102	-1.42
	110	0.0032	0.0029	0.0101	-1.21
	140	0.0080	0.0043	0.0100	-1.61
5	80	0.0155	0.0012	0.0102	0.08
	110	0.0061	0.0018	0.0100	-0.56
	140	0.0030	0.0036	0.0100	-1.42

Table 12 Type-I error of 4x4 contingency table in 0.05 significant level.

No.Sc.	Ss.	Type I Error			
		N	W	D	C
1	80	0.0081	0.0041	0.0500	-0.45
	110	0.0140	0.0075	0.0502	-0.87
	140	0.0147	0.0116	0.0500	-1.30
2	80	0.0043	0.0045	0.0500	-0.04
	110	0.0415	0.0031	0.0502	-0.48
	140	0.0134	0.0059	0.0500	-0.82
3	80	0.0053	0.0050	0.0500	-0.74
	110	0.0174	0.0028	0.0503	-0.87
	140	0.0614	0.0039	0.0500	0.06
4	80	0.0101	0.0011	0.0501	-1.19
	110	0.0381	0.0031	0.0500	-1.03
	140	0.0388	0.0045	0.0500	-1.31
5	80	0.0516	0.0045	0.0501	0.38
	110	0.0285	0.0145	0.0500	-0.28
	140	0.0171	0.0295	0.0500	-1.11

Conclusion and Discussion

We applied three imputation methods to treat the problem of small expected cell frequency when using the chi-square test. We reviewed and provided technical details of the different methods used, including Yate's correction, William's correction and our Developed correction.

As depicted in Table 1-12, all methods led to an improvement in accuracy, as measured by type-I error for each situation. The method which outperformed the control of type-I error was the developed correction method in all condition.

Association in 2×2 tables traditionally has been tested using the chi-square test for larger samples. There are Yate's correction for aiming to improve the small expected cell frequency. For William's correction used to continuity chi-square test for independence when contingency table lager than 2×2. In this study to developed continuity corrections for contingency table between 2×2 to 4×4, we found that type I error in chi-square test with developed correction and significant level is similar values. And it is similarly values more than chi-square test without correction, Yate's correction and William's correction. When sample sizes was increased the resulted is better control type I error at both levels of significance.

For the size of contingency table 2x2 to 4x4, chi-square test with developed correction can control type I error better than chi-square test without correction and William's correction at both 0.01 and 0.05 significant levels. It outperformed to control type I error with C value better than other correction in all condition. Especially 4×4 table, there is similar values of type I error and significant level.

A correction value for chi-square test depends on the pattern of a contingency table. The appropriate correction value is not necessarily equal to Yates' correction value of 0.5. When a contingency table with small expected frequencies is used as an input, the test procedure is to identify the table's pattern and use the appropriate correction value for that pattern. The appropriate correction values for the patterns associated with 2x2 to 4x4 contingency tables are tabulated in this article.

References

Freund, J. E. (2004). *Modern Elementary Statistics*. New Jersey: Pearson Education, Inc.

Mcdonald, J. H. (2014). *Handbook of Biological Statistics*. Maryland: Sparky House Publishing.

Yates, F. (1934). Contingency tables involving small numbers and the χ 2 test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217-235.

Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2013). *Introduction to Probability and Statistics*. Australia: Boosk/Cole, Cengage Learning.

Gu, X., & Lee, J. J. (2010). A simulation study for comparing testing statistics in response-adaptive randomization. *BMC medical research methodology*, *10*, 48.