Journal of the Nigerian Association of Mathematical Physics,
Volume 8 (*November 2004*)

A simplification of the likelihood ratio test statistic for testing hypothesis about goodness of fit of multinomial probabilities.

K. Ayinde and **D. B. Adekanmbi,**
Department of Pure and Applied Mathematics

Ladoke Akintola University of Technology
Ogbomoso, Oyo State, Nigeria.

Abstract

The traditional likelihood ratio test statistic for testing hypothesis about goodness of fit of multinomial probabilities in one, two and multi – dimensional contingency table was simplified. Advantageously, using the simplified version of the statistic to test the null hypothesis is easier and faster because calculating the expected cell frequency becomes absolutely unnecessary. The conclusions of the numerical examples considered to illustrate their usage agreed perfectly with that of the traditional and simplified method of the Pearson chi–squared statistic even when the observed cell frequencies of some cells are small.

pp 305 - 310

1.0 Introduction

Various methods are now available for testing hypothesis about goodness of fit of multinomial probabilities. Among them are the Karl Pearson's and Neyman's chi–square statistics introduced in 1900 and 1949 respectively [1]. The use Likelihood ratio test was also introduced by Neyman and Pearson in 1928[2]. These statistics are distributed as chi – square (χ_v^2) distribution in large samples, where v is the degree of freedom [2]. Returning to the underlying χ^2 approximation to each of these statistics, it has been suggested that approximation is only valid when the expected values are large and that the approximation ceases to be appropriate if any of the expected cell frequencies becomes too small. Their asymptotic equivalence can be found in the work of Bishop et al [3].

The simplified version of the Pearson chi-squared statistic in one and two-dimensional contingency table was provided by Ayinde and Iyaniwura [4]. Ayinde [5] also gave the simplified version of the statistic in multi- dimensional contingency table. The simplified version of the Neyman chi-squared statistic in one, two and multi- dimensional contingency table was provided by Ayinde and Ayinde [6]. The results of these simplified versions showed that testing hypothesis about goodness–of–fit of multinomial probabilities could be done without calculating the expected cell frequencies.

In this paper we therefore provided the simplified versions of the traditional likelihood ratio test statistics in one, two and multi–dimensional contingency table; and at the same time give numerical examples to illustrate their usages even when the observed cell frequencies is less than 5.

2.0 Materials and methods

The traditional likelihood test statistic to test hypothesis about goodness-of-fit of multinomial probabilities demands that Y^2 be greater than the $\chi^2_{tab.}$ before the null hypothesis (H_o) can be rejected [3].

Consequently, for the likelihood test statistic the decision rule is reject H_o if

$$Y^2 = 2\sum_{i=1}^{k} o_i \, log\left(\frac{o_i}{e_i}\right) > \chi^2_{tab.} \qquad (2.1)$$

with (k - 1) degree of freedom if it is a one dimensional table, where $e_i = np_i$ and p_i=probability of each cell. In a two dimensional contingency table, the decision rule is reject H_O if

$$Y^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} o_{ij}\ log\left(\frac{o_{ij}}{e_{ij}}\right) > \chi^2_{tab.} \qquad (2.2)$$

with (r - 1)(c - 1) degree of freedom if it is a two dimensional contingency table, where $e_{ij} = np_{ij}$ and $p_{ij} = \frac{n_{i.}}{n} \times \frac{n_{.j}}{n}$. (Independent of the factors). Also, in a multi–dimensional (d) contingency table, the decision rule is reject H_O if

$$Y^2 = 2\sum_{ijk}^{rcm} O_{ijk\ldots}\ log\left(\frac{O_{ijk\ldots}}{e_{ijk\ldots}}\right) > \chi^2_{tab.} \qquad (2.3)$$

with (r x c x m x…) - (r + c + m +…) + (d - 1) degree of freedom, where $e_{ijk\ldots} = np_{ijk\ldots}$, $p_{ijk\ldots} = \frac{n_{i\ldots..}}{n} \times \frac{n_{.j\ldots.}}{n} \times \frac{n_{..k\ldots}}{n} \times \ldots$ (Independent of the factors) and $\ldots$ implies the continuation of the factors.

2.1 Derivation of the simplified likelihood ratio statistic in one-dimensional table.

$Y^2 = 2\sum_{i=1}^{k} PO_i\ log\left(\frac{O_i}{e_i}\right) > \chi^2_{tab.} \Rightarrow 2\sum_i O_i\left[log\ O_i - log\ e_i\right] > \chi^2_{tab.}$ but $e_i = np_i$, therefore we have

$Y^2 = 2\sum_i O_i\left[log\ O_i - \left(log\ n + log\ p_i\right)\right] > \chi^2_{tab.} \Rightarrow 2\sum_i O_i\left[log\left(\frac{O_i}{p_i}\right) - log\ n\right] > \chi^2_{tab.} \Rightarrow 2\sum_i O_i\ log\left(\frac{O_i}{p_i}\right)$ $- 2\sum_i O_i\ log\ n > \chi^2_{tab.}$. Also, noting that $\sum_{i=1}^{k} O_i = n$

$$Y^2 = 2\sum_i O_i\ log\left(\frac{O_i}{p_i}\right) - 2n\ log\ n > \chi^2_{tab.} \qquad (2.4)$$

This is the simplified likelihood test statistic, which can be used to test the same hypothesis in one-dimensional table.

2.2 Derivation of the simplified likelihood ratio statistic in two-dimensional table.

$Y^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}\ log\left(\frac{O_{ij}}{e_{ij}}\right) > \chi^2_{tab.} \Rightarrow 2\sum_i\sum_j O_{ij}\left[log\ O_{ij} - log\ e_{ij}\right] > \chi^2_{tab.}$ but $e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$, therefore we have $Y^2 = 2\sum_i\sum_j O_{ij}\left[log\ O_{ij} - \left(log\ n_{i.} + log\ n_{.j} - log\ n\right)\right] > \chi^2_{tab.}$

$$\Rightarrow 2\sum_i\sum_j O_{ij}\left[log\left(\frac{O_{ij}}{n_{i.} \times n_{.j}}\right) + log\ n\right] > \chi^2_{tab.}$$

$\Rightarrow 2\sum_i\sum_j O_{ij}\ log\left(\frac{O_{ij}}{n_i \times n_{.j}}\right) + 2\sum_i\sum_j O_{ij}\ log\ n > \chi^2_{tab.}$. Also, noting that $\sum_{i=1}^{r}\sum_{j=}^{c} O_{ij} = n$

$$Y^2 = 2\sum_i\sum_j O_{ij}\ log\left(\frac{O_{ij}}{n_i \times n_{.j}}\right) + 2n\ log\ n > \chi^2_{tab.} \qquad (2.5)$$

This is the simplified likelihood test statistic, which can be used to test the same hypothesis in two - dimensional table.

2.3 Derivation of the simplified likelihood ratio statistic in multi-dimensional table.

$Y^2 = 2\sum_{ijk\ldots}^{rcm\ldots} O_{ijk\ldots}\ log\left(\frac{O_{ijk\ldots}}{e_{ijk\ldots}}\right) > \chi^2_{tab.} \quad \Rightarrow \quad 2\sum_{ijk\ldots} O_{ijk\ldots}\left[log\ O_{ijk\ldots} - log\ e_{ijk\ldots}\right] > \chi^2_{tab.}$ but

$e_i = \frac{n_{i\ldots..} \times n_{..j\ldots.} \times n_{..k\ldots} \times \ldots}{n^{d-1}}$, (where *d* = number of the dimension of the table), therefore we have

$$Y^2 = 2\sum_{ijk\ldots} O_{ijk\ldots}\left[log\, O_{ijk\ldots} - \left(log\, n_{i\ldots\ldots} + log\, n_{.j\ldots.} + log\, n_{..k\ldots} + \ldots - log\, n^{d-1}\right)\right] > \chi^2_{tab.}$$

$$\Rightarrow 2\sum_{ijk\ldots} O_{ijk\ldots}\left[log\left(\frac{O_{ijk\ldots}}{n_{i\ldots\ldots}\times n_{.j\ldots.}\times n_{..k\ldots}\times\ldots}\right) + (d-1)\,log\, n\right] > \chi^2_{tab.}$$

$$\Rightarrow 2\sum_{ijk\ldots} O_{ijk\ldots}\; log\left(\frac{O_{ijk\ldots\ldots}}{n_{i\ldots\ldots}\times n_{..j\ldots\ldots}\times n_{..k\ldots}\times\ldots}\right) + 2\sum_{ijk\ldots} O_{ijk\ldots}(d-1)\,log\, n > \chi^2_{tab.}$$

Also, noting that $\sum_{ijk\ldots}^{rcm\ldots} O_{ijk\ldots} = n$

$$Y^2 = 2\sum_{ijk\ldots.} O_{ijk\ldots}\; log\left(\frac{O_{ijk\ldots}}{n_{i\ldots\ldots}\times n_{.j\ldots.}\times n_{..k\ldots}\times\ldots}\right) + 2n(d-1)\,log\, n > \chi^2_{tab.} \qquad (2.6)$$

This is the simplified likelihood test statistic, which can be used to test the same hypothesis in multi – dimensional (*d*) table. Now if *d* = 1, equation (2.6) becomes $Y^2 = 2\sum_i O_i\; log\left(\frac{O_i}{n_i}\right) > \chi^2_{tab.}$. But $n_i = e_i = np_i$, therefore we have

$$Y^2 = 2\sum_i O_i\left[log\left(\frac{O_i}{p_i}\right) - log\, n\right] > \chi^2_{tab.} \Rightarrow 2\sum_i O_i\; log\left(\frac{O_i}{p_i}\right) - 2\sum_i O_i\; log\, n > \chi^2_{tab.}$$

Also, noting that $\sum_{i=1}^{k} O_i = n$

$$Y^2 = 2\sum_i O_i\; log\left(\frac{O_i}{p_i}\right) - 2n\,log\, n > \chi^2_{tab.} \qquad (2.7)$$

This is the same as that of equation (2.4). If *d* = 2, equation (2.6) becomes

$$Y^2 = 2\sum_i\sum_j O_{ij}\; log\left(\frac{O_{ij}}{n_i \times n_{.j}}\right) + 2n\,log\, n > \chi^2_{tab.} \qquad (2.8)$$

This is the same as (2.5) above. If *d* = 3, we obtain equation (2.9) from (2.6) as

$$\sum_{ijk} O_{ijk}\; log\left(\frac{O_{ijk}}{n_{i\ldots}\times n_{.j.}\times n_{..k}}\right) + 4n\,log\, n > \chi_{tab}{}^2 \qquad (2.9)$$

If *d* = 4, equation (2.6) gives (210) as

$$\sum_{ijkl} O_{ijkl}\; log\left(\frac{O_{ijkl}}{n_{i\ldots.}\times n_{.j\ldots}\times n_{..k.}\times n_{\ldots l}}\right) + 6n\,log\, n > \chi^2_{tab} \qquad (2.10)$$

This can continue for any number of contingency table.

3.0 Numerical Examples

Example 1: The table below shows the numerical example considered by Ayinde and Iyaniwurwa [4] in their paper.

Table 1: The number of heads obtained when 4 coins are tossed 120 times.

Number of heads (*x*)	0	1	2	3	4
Number of times (*f*)	15	35	40	20	10

Test the hypothesis that the coins are fair at α = 0.1, 0.05 0.03 and 0.01; and compare your results with that of Pearson Chi-squared Statistic.

HINT: $P_0 = \frac{1}{16}, P_1 = \frac{4}{16}, P_2 = \frac{6}{16}, P_3 = \frac{4}{16}, P_4 = \frac{1}{16}$.

Solution

This is a one-dimensional problem.

1. ***Using the traditional Likelihood method***

The traditional Likelihood test statistic for a one-dimensional table as given in equation (1) is

$$Y^2 = 2\sum_{i=1}^{k} O_i \; log\left(\frac{O_i}{e_i}\right) > \chi^2_{tab.}$$

$$Y^2 = 2\sum_{i=1}^{k} O_i \; log\left(\frac{O_i}{e_i}\right) = 2\left[15\,log\left(\frac{15}{7.5}\right) + 35\,log\left(\frac{35}{30}\right) + ... + 10\,log\left(\frac{10}{7.5}\right)\right] = 11.69736$$

Using the compute tool of SPSS 10.0, the tabulated values are obtained as: At α = 0.1, $\chi^2_{0.9,4} = 7.7794$; α = 0.05, $\chi^2_{0.95,4.} = 9.4877$; α = 0.03, $\chi_{0.97,4} = 10.7119$; α = 0.01, $\chi^2_{0.99} = 13.2767$

Decision rule

Reject H_O if $Y^2 > \chi^2_{tab.}$. Hence H_O is only accepted at 0.01 level of significance since 11.6968 < 13.2767.

Conclusion

The coins are fair at 0.01 level of significance only.

(2) **Using the simplified likelihood method**

The simplified likelihood test statistic for a one dimensional problem as given in equation (3.1) demands

$$Y^2 = 2\sum_{i} O_i \; log\left(\frac{O_i}{p_i}\right) - 2n\,log\,n > \chi^2_{tab.} \qquad (3.1)$$

$$2\sum_{i=1}^{k} O_i \; log\left(\frac{o_i}{p_i}\right) - 2n\,log\,n = 2\left[15\,log\left(\frac{15\times 16}{1}\right) + 35\,log\left(\frac{35\times 16}{4}\right) + ... + 10\,log\left(\frac{10\times 16}{1}\right)\right] - 2\times 120\times log\,120$$

$= 11.69729$

Decision rule

Reject H_O if $Y^2 > \chi^2_{tab.}$. Hence H_O is only accepted at 0.01 level of significance since 11.69729 < 13.2767.

Conclusion

The coins are fair at 0.01 level of significance only.

The results from a computer program are summarized in Table 2.

Example 2

The table below shows a study of relationship among race, blood type and sex in a country.

	BLOOD TYPES								
	O		A		B		AB		
	SEX								
Races	M	F	M	F	M	F	M	F	$n_{i..}$
Race1	3	3	30	62	20	26	25	25	194
Race2	45	36	28	2	3	2	18	12	146
Race3	38	32	40	12	22	23	3	10	180
Race4	8	2	10	10	7	8	16	12	73
Total	94	73	108	86	52	59	62	59	
$n_{.j.}$	167		194		111		121		593

Test the hypothesis that (*i*) race and blood group are independent (*ii*) race, blood group and sex are completely independent at α = 0.005 level of significance; and compare your results with that of Pearson Chi-squared Statistic.

Solution

(*i*) This is a two - dimensional contingency table and the results obtained from computer programs are summarized in Table 3. Using the compute tool of SPSS 10.00 the obtained tabulated value at α = 0.005 is is $\chi^2_{0.995,9.} = 23.5894$.

(*ii*) By this exercise, apart from testing the required hypothesis we also intend to make some comments on the behaviour of traditional and simplified form of the statistics when observations in some cells are less than 5. This three - dimensional contingency table and the results obtained from computer programs are summarized in Table 4. T Using the compute tool of SPSS 10.00 the obtained tabulated value at α = 0.005 is $\chi^2_{0.995,24.} = 45.5585$.

Comment

When some of the observed cells frequencies are less than 5, the result of the traditional and the simplified version of the statistics could still be considered to be the same since the differences are negligible (Table 4).

4.0 **Conclusion**

This modified Likelihood ratio test statistic method of testing hypothesis about multinomial probabilities in contingency tables has some advantages over the traditional method. It is easier and faster because there is no need of calculating the expected cell frequencies before the hypothesis can be tested. Furthermore, the risk of committing either type 1 or type 11 errors is minimized since the problem of figure approximation is frequently reduced.

Table2: Summary of the results of the Pearson’s and the Likelihood method

Methods		Value Obtained	Level of significance	Decision	Conclusion
Pearson	Traditional	13.0555	0.1	Reject H_0	The coins are not fair
			0.05	Reject H_0	The coins are not fair
			0.03	Reject H_0	The coins are not fair
			0.01	Accept H_0	The coins are fair
	Modified	13.0555	0.1	Reject H_0	The coins are not fair
			0.05	Reject H_0	The coins are not fair
			0.03	Reject H_0	The coins are not fair
			0.01	Accept H_0	The coins are fair
Likelihood	Traditional	11.69736	0.1	Reject H_0	The coins are not fair
			0.05	Reject H_0	The coins are not fair
			0.03	Reject H_0	The coins are not fair
			0.01	Accept H_0	The coins are fair
	Modified	11.69729	0.1	Reject H_0	The coins are not fair
			0.05	Reject H_0	The coins are not fair
			0.03	Reject H_0	The coins are not fair
			0.01	Accept H_0	The coins are fair

Table3: Summary of the results of the Pearson and the Likelihood method.

Methods		Value Obtained	Decision	Conclusion
Pearson	Traditional	169.9814	Reject H_0	Race and Blood type are not independent
	Simplified	169.9815	Reject H_0	Race and Blood type are not independent
Likelihood	Traditional	200.2103	Reject H_0	Race and Blood type are not independent

	Simplified	200.2107	Reject H_0	Race and Blood type are not independent

Table4: Summary of the results of the Pearson and the Likelihood method.

Methods		Value Obtained	Decision	Conclusion
Pearson	Traditional	223.0238	Reject H_0	Race, Blood type and Sex are not independent
	Simplified	223.0239	Reject H_0	Race, Blood type and Sex are not independent
Likelihood	Traditional	264.0141	Reject H_0	Race, Blood type and Sex are not independent
	Simplified	264.0151	Reject H_0	Race, Blood type and Sex are not independent

References

[1] Lindeman, R. H., Merenda,P.F. and Gold,R.Z (1980). Introduction to Bivariate and Multivariate Analysis.1ar edition .Scott, Foresman and Company,England.

[2] Sanni, O.O.M and Jolayemi, E.T (1998). Robustness of some Categorical test Statistics in small sample situations. Journal of the Nigerian Statisticians. Vol 2, 29 – 35.

[3] Bishop Y.M.M, Finenbery,S.E and Holland,P.W (1975).Discrete Multivariate Analysis.Theory and Practice.Mass: M.I.T Press.Especially 14:7 – 14.

[4] Ayinde k. and Iyaniwura J.O. (2001). A modification of chi- square statistics to test hypotheses about multinomial probabilities in one–dimensional and two-dimensional contingency table. Journal of applied Sciences. 4(1): 1749-1758.

[5] Ayinde K. (2003):Modified chi–square statistic to test hypothesis about multinomial probabilities in multi – dimensional contingency table. An international Journal of Biological and Physical Sciences (Science Focus).Vol. No.2: 28 - 31

[6] Ayinde,K and Ayinde O.E. (2003): Modified Neyman's chi-square statistic for testing hypothesis about goodness of fit of multinomial probabilities. Zuma Journal of Pure Applied Science 5 (2):123-127