FULL LENGTH RESEARCH ARTICLE

SIMPLIFIED FREEMAN-TUKEY TEST STATISTICS FOR TESTING PROBABILITIES IN CONTINGENCY TABLES

*K. AYINDE[1] & A. O. ABIDOYE[2]

[1]*Department of Pure and Applied Mathematics*
Ladoke Akintola University of Technology
P. M. B. 4000
Ogbomoso Oyo State Nigeria
[2]*Department of Statistics*
University of Ilorin
P.M.B 4000
Ilorin Kwara State Nigeria
*Corresponding author
bayoayinde@yahoo.com

ABSTRACT
This paper presents the simplified version of the Freeman-Tukey test statistic for testing hypothesis about multinomial probabilities in one, two and multi-dimensional contingency tables that does not require calculating the expected cell frequencies before test of significance. The simplified method established new criteria of collapsing cells whose frequency are less than 5. Illustrated examples compared favourably the new method with Pearson, Neyman and Likelihood ratio chi- squared statistics. Apart from being faster, the simplified version provides more accurate result since the problem of figure approximation is reduced.

***Keywords*:** Freeman-Tukey statistic, dimension, contingency table, multinomial probabilities, expected cell frequencies.

INTRODUCTION
Hypothesis testing about multinomial probabilities can be done using different methods. Among the most frequently used methods are the Pearson (1900), Neyman (1949) and the Likelihood ratio test chi-squared statistics given by West & Kempthrone (1972).

Another commonly used method is the Freeman-Tukey test statistics introduced by Freeman & Tukey (1950). These statistics are distributed as chi-square (χ_d^2) distribution in large samples, where d is the degree of freedom (Sanni & Jolayemi 1998). Their asymptotic equivalence can be found in the work of Bishop *et al.* (1975). 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 (Lawal 2003; Adegboye 2004). The comparative accuracies of some of these statistics have been investigated (Lawal 2003; Larntz 1978; Kochler & Larntz 1980; West & Kemphorne 1972).

Recently, the simplified form of the Pearson, Neyman and the Likelihood ratio test chi-squared statistics in one, two and multi-dimensional (P) contingency table was provided (Ayinde & Adekanmbi 2004; Ayinde & Ayinde 2003; Ayinde 2003; Ayinde & Iyaniwura 2001). These simplified versions do not only allow hypothesis to be tested without calculating the expected cell frequencies but also make hypothesis testing easier and faster.

Consequently, in this paper we have made effort to provide the simplified form of the Freeman-Turkey test statistic for testing hypothesis about multinomial probabilities in one, two and multi - dimensional contingency table; and established a new condition for the simplified version of the statistic when expected cell frequencies of any of the cells are less than 5. Furthermore, we gave some numerical examples to illustrate their usages.

MATERIALS AND METHODS

The traditional Freeman-Tukey (1950) statistic to test hypothesis about multinomial probabilities is

$$T^2 = 4\sum_{i=1}^{k}\left(\sqrt{o_i}-\sqrt{e_i}\right)^2 \ldots \qquad (1)$$

with (k-1) degree of freedom for a one dimensional table, where $e_i = np_i$ and p_i =probability of each cell. In a two dimensional contingency table, we have

$$T^2 = 4\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\sqrt{o_{ij}}-\sqrt{e_{ij}}\right)^2 \ldots \qquad (2)$$

with (r-1) (c-1) degree of freedom, where $e_{ij} = np_{ij}$ and $p_{ij} = \frac{n_{i.}}{n}\times\frac{n_{.j}}{n}$ (Independence of the factors). Also in a multi-dimensional (P) contingency table, to test the hypothesis that the (P) factors are not unconditionally independent (i.e. the factors are completely independent), we have

$$T^2 = 4\sum_{a_1a_2\ldots a_p}^{A_1A_2\ldots A_P}\left(\sqrt{o_{a_1a_2\ldots a_p}}-\sqrt{e_{a_1a_2\ldots a_p}}\right)^2 \ldots \qquad (3)$$

with $\prod_{i=1}^{p} A_i - \sum_{i=1}^{p} A_i + (p-1)$ degree of freedom, p>1, where $e_{a_1 a_2 \ldots a_p} = np_{a_1 a_2 \ldots a_p}$,

$p_{a_1 a_2 \ldots a_p} = \frac{n_{a_1 \ldots\ldots}}{n} \times \frac{n_{.a_2 \ldots.}}{n} \times \ldots \times \frac{n_{\ldots..a_p}}{n} = \frac{1}{n^p}\left[n_{a_1 \ldots\ldots} \times n_{.a_2 \ldots.} \times \ldots \times n_{\ldots..a_p}\right]$ (Lawal 2003; Lindeerman *et al.* 1980).

Simplification of the Freeman – Tukey statistics in one - dimensional table.

$$T^2 = 4\sum_{i=1}^{k}\left(\sqrt{o_i} - \sqrt{e_i}\right)^2$$

$$= 4\sum_{i=1}^{k}\left[o_i - 2\sqrt{e_i}\sqrt{o_i} + e_i\right]$$

$$= 4\left[\sum_{i=1}^{k} o_i - 2\sum_{i=1}^{k}\sqrt{e_i o_i} + \sum_{i=1}^{k} e_i\right]$$

$$= 4\left[\sum_{i=1}^{k} o_i - 2\sum_{i=1}^{k}\sqrt{e_i o_i} + \sum_{i=1}^{k} e_i\right]$$

But $\sum_{i=1}^{k} o_i = \sum_{i=1}^{k} e_i = n$ and $e_i = np_i$ Therefore,

$$T^2 = 4\left[2n - 2\sum_{i=1}^{k} \sqrt{np_i o_i}\right]$$

$$= 8\left[n - n^{\frac{1}{2}}\sum_{i=1}^{k}\left(o_i p_i\right)^{\frac{1}{2}}\right] \ldots \quad (4)$$

This is the simplified Freeman-Tukey test statistic which can be used to test the same hypothesis in one-dimensional table. The contribution of each cell to the simplified version above is no more e_i as in the traditional method (equation (1)) but rather p_i, thus the new condition for cells to be collapsed now becomes

$$e_i < 5$$

$$\Rightarrow np_i < 5$$

$$\Rightarrow p_i < \frac{5}{n} \ldots \quad (5)$$

Simplification of the Freeman-Tukey statistic in two-dimensional contingency table.

$$T^2 = 4\sum_{i=1}^{r}\sum_{j=1}^{c} \left(\sqrt{o_{ij}} - \sqrt{e_{ij}}\right)^2$$

$$= 4\sum_{i=1}^{r}\sum_{j=1}^{c} \left[o_{ij} - 2\sqrt{e_{ij}}\sqrt{o_{ij}} + e_{ij}\right]$$

$$= 4\left[\sum_{i=1}^{r}\sum_{j=1}^{c} o_{ij} - 2\sum_{i=1}^{r}\sum_{j=1}^{c} \sqrt{e_{ij} o_{ij}} + \sum_{i=1}^{c}\sum_{j=1}^{c} e_{ij}\right]$$

But $\sum_{i=1}^{r}\sum_{j=1}^{c} o_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{c} e_{ij} = n$, and $e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$. Therefore,

$$T^2 = 4\left[2n - 2\sum_{i=1}^{c}\sum_{j=1}^{c} \sqrt{\frac{n_{i.} \times n_{.j}}{n} o_{ij}}\right]$$

$$= 8\left[n - n^{-\frac{1}{2}}\sum_{i=1}^{r}\sum_{j=1}^{c} \left(o_{ij} \times n_{i.} \times n_{.j}\right)^{\frac{1}{2}}\right] \ldots \qquad (6)$$

This is the simplified Freeman-Tukey test statistic which can be used to test the same hypothesis in two-dimensional contingency table.

Similarly, the contribution of each cell to the simplified version above is no more e_{ij} as in the traditional method (equation (2)) but rather $n_{i.} \times n_{.j}$, thus the new condition for cells to be collapsed now becomes

$$e_{ij} < 5$$

$$\Rightarrow \frac{n_{i.} \times n_{.j}}{n} < 5$$

$$\Rightarrow n_{i.} \times n_{.j} < 5n \ldots \qquad (7)$$

Simplification of the Freeman-Tukey test statistic in multi-dimensional (P) contingency table.

$$T^2 = 4\sum_{a_1a_2\ldots a_p}^{A_1A_2\ldots A_p}\left(\sqrt{o_{a_1a_2\ldots a_p}} - \sqrt{e_{a_1a_2\ldots a_p}}\right)^2$$

$$= 4\sum_{a_1a_2..a_p}^{A_1A_2\ldots A_p}\left[o_{a_1a_2\ldots a_p} - 2\sqrt{e_{a_1a_2\ldots a_p}}\sqrt{o_{a_1a_2\ldots a_p}} + e_{a_1a_2\ldots a_p}\right]$$

$$= 4\left[\sum_{a_1a_2\ldots a_p}^{A_1A_2\ldots A_p} o_{a_1a_2\ldots a_p} - 2\sum_{a_1a_2\ldots a_p}^{A_1A_2\ldots A_p}\sqrt{e_{a_1a_2\ldots a_p} o_{a_1a_2\ldots a_p}} + \sum_{a_1a_2\ldots a_p}^{A_1A_2\ldots A_p} e_{a_1a_2\ldots a_p}\right]$$

But $\sum_{a_1a_2\ldots a_p}^{A_1A_2..A_p} o_{a_1a_2\ldots a_p} = \sum_{a_1a_2\ldots a_p}^{A_1A_2..A_p} e_{a_1a_2\ldots a_p}$ = n and $e_{a_1a_2\ldots a_p} = \frac{1}{n^{p-1}}\left[n_{a_1\ldots\ldots} \times n_{.a_2\ldots.} \times \ldots \times n_{\ldots\ldots a_p}\right]$

Therefore,

$$T^2 = 4\left[2n - 2\sum_{a_1a_2\ldots a_p}^{A_1A_2\ldots A_p}\sqrt{\frac{1}{n^{p-1}}\left[n_{a_1\ldots\ldots} \times n_{.a_2\ldots.} \times n_{\ldots\ldots a_p}\right] o_{a_1a_2\ldots a_p}}\right]$$

$$= 8\left[n - n^{-\frac{1}{2}(p-1)}\sum_{a_1a_2\ldots a_p}^{A_1A_2\ldots A_p}\left[(n_{a_1\ldots\ldots} \times n_{.a_2\ldots} \times \ldots \times n_{\ldots\ldots a_p}) o_{a_1a_2\ldots a_p}\right]^{\frac{1}{2}}\right] \ldots \qquad (8)$$

This is the simplified Freeman-Tukey test statistic which can be used to test the same hypothesis in P- dimensional contingency table. Similarly, the contribution of each cell to the simplified version above is no more $e_{a_1a_2\ldots a_p}$ as in the traditional method (equation (3)) but rather $\left[n_{a_1.....} \times n_{.a_2....} \times ... \times n_{.....a_p}\right]$, thus the new condition for cells to be collapsed now becomes

$$e_{a_1 a_2 ... a_p} < 5$$

$$\Rightarrow \frac{1}{n^{p-1}} \left[n_{a_1.....} \times n_{.a_2....} \times ... \times n_{.....a_p}\right] < 5$$

$$\Rightarrow \left[n_{a_1.....} \times n_{.a_2....} \times ... \times n_{.....a_p}\right] < 5n^{p-1} ... \qquad (9)$$

Now if P = 1, equation (8) becomes

$$T^2 = 8\left[n - \sum_{a_1=1}^{A_1} \left[n_{a_1} \times o_{a_1}\right]^{\frac{1}{2}}\right]$$

$$= 8\left[n - \sum_{i=1}^{k} \left[n_i \times o_i\right]^{\frac{1}{2}}\right]$$

But $n_i = np_i$ .Therefore,

$$T^2 = 8\left[n - n^{\frac{1}{2}} \sum_{i=1}^{k} \left(o_i p_i\right)^{\frac{1}{2}}\right] ... \qquad (10)$$

This is the same as that of equation (4) with (k-1) degree of freedom. Also if P = 1 in equation (9) becomes

$$n_{a_i} < 5$$

$$\Rightarrow n_i < 5$$

But $n_i = np_i$. Therefore,

$$\Rightarrow np_i < 5$$

$$\Rightarrow p_i < \frac{5}{n} ... \qquad (11)$$

This is the same as equation (5) above. If P = 2, equation (8) above becomes

$$T^2 = 8\left[n - n^{-\frac{1}{2}} \sum_{a_1 a_2}^{A_1 A_2} \left(o_{a_1 a_2} \times n_{a_1.} \times n_{.a_2}\right)^{\frac{1}{2}}\right]$$

$$= 8\left[n - n^{-\frac{1}{2}} \sum_{i=1}^{r} \sum_{j=1}^{c} \left(o_{ij} \times n_{i.} \times n_{.j}\right)^{\frac{1}{2}}\right] ... \qquad (12)$$

This is the same as (6) above with r x c - (r +c) +1 = (r-1) (c-1) degree of freedom. Also if P = 2 in equation (9), the new condition for collapsing the cells becomes

$$n_{a_1} \times n_{a_2} < 5n$$

$$\Rightarrow n_i \times n_{.j} < 5n ... \qquad (13)$$

This is the same as equation (7) above. If P = 3, we obtain equation (14) from (8) as

$$T^2 = 8\left[n - n^{-1} \sum_{a_1 a_2 a_3}^{A_1 A_2 A_3} \left(o_{a_1 a_2 a_3} \times n_{a_1..} \times n_{.a_2.} \times n_{..a_3}\right)^{\frac{1}{2}}\right]$$

$$=8\left[n-n^{-1}\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{m}\left(o_{ijk}\times n_{i..}\times n_{.j.}\times n_{..k}\right)^{\frac{1}{2}}\right]\ldots \tag{14}$$

with r x c x m - (r +c +m) +2 degree of freedom (Complete independence of the three factors). Also if P = 3 in equation (9), the new condition for collapsing the cells becomes

$$n_{a..}\times n_{.a_2.}\times n_{..a_3} < 5n^2$$

$$\Rightarrow n_{i..}\times n_{.j.}\times n_{..k} < 5n^2\ldots \tag{15}$$

If P = 4, equation (8) gives (16) as

$$T^2=8\left[n-n^{-\frac{3}{2}}\sum_{a_1a_2a_3a_4}^{A_1A_2A_3}\left(o_{a_1a_2a_3a_4}\times n_{a_1\ldots}\times n_{.a_2..}\times n_{..a_3.}\times n_{\ldots a_4}\right)^{\frac{1}{2}}\right]$$

$$=8\left[n-n^{-\frac{3}{2}}\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{m}\sum_{l=1}^{t}\left(o_{ijkl}\times n_{i\ldots}\times n_{.j..}\times n_{..k.}\times n_{\ldots l}\right)^{\frac{1}{2}}\right]\ldots \tag{16}$$

with r x c x m x t - (r +c +m + t) +3 degree of freedom (Complete independence of the four factors). This can continue for any P-dimensional contingency table. Also if P = 4 in equation (9), the new condition for collapsing the cells becomes

$$n_{a_1\ldots}\times n_{.a_2..}\times n_{..a_3.}\times n_{\ldots a_4} < 5n^3$$

$$\Rightarrow n_{i\ldots}\times n_{.j..}\times n_{..k.}\times n_{\ldots k} < 5n^3\ldots \tag{17}$$

This can also continue for any P-dimensional contingency table.

NUMERICAL EXAMPLES

Example 1: Table 1 below shows the numerical example considered by Ayinde & Iyaniwura (2001).

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 and compare your results with that of Pearson, Neyman and the Likelihood ratio test chi-squared statistics. **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. A computer programme was written to handle the computation while the compute of SPSS 10.0 was used to obtain the P-valve. The summary of the results is shown in Table 7.

Example 2: A random sample of 40 students in one of the Nigerian University was cross-classified according to their sex and mode of entry. The table 2 below shows the data. This is the example considered by Adegboye (2004).

TABLE 2: CROSS-CLASSIFICATION OF STUDENTS BASED ON THEIR SEX AND MODE OF ENTRY.

		Mode of Entry			
		JAMB	**Pre-NCE**	**Others**	$n_{i.}$
Sex	**Male**	4	8	6	**18**
	Female	2	13	7	**22**
	$n_{.j}$	**6**	**21**	**13**	**40**

Test the hypothesis that student's sex is independent of mode of entry. Use α = 0.05.

Solution: The expected cell frequencies are calculated and shown in the table 3 below.

TABLE 3: THE EXPECTED FREQUENCIES OF TABLE 2.

		Mode of Entry			
		JAMB	Pre-NCE	Others	$n_{i.}$
Sex	Male	2.7	9.45	5.85	**18**
	Female	3.3	11.55	7.15	**22**
	$n_{.j}$	**6**	**21**	**13**	**40**

From the table above the cell frequencies of the first column are less than 5, thus we need to collapse the first and second columns together by adding the frequencies of the columns together. The result is presented table 4 below. The expected frequencies are in the parenthesis.

TABLE 4: RE-CLASSIFICATION OF THE OBSERVED AND EXPECTED FREQUENCIES.

	Mode of Entry	
	JAMB & Pre - NCE	Others
Male	12 (12.15)	6 (5.85)
Female	15 (14.65)	7 (7.15)

The computation using various traditional methods is done and the results are presented in Table 6. Similarly, using the simplified method with the new condition established when the expected cell frequencies are less than 5, the results of the computation are also shown in Table 6.

Example 3: The Table below showed a study of the relationship among race, blood group and sex in a country. This is the example was taken from Ayinde (2003).

TABLE 5: A STUDY OF RELATIONSHIP AMONG RACE, BLOOD TYPE AND SEX IN A COUNTRY.

	BLOOD GROUP								
	O		A		B		AB		
	SEX								
Race	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	$n_{i..}$
Race1	40	49	30	62	20	26	25	25	277
Race2	45	36	28	20	30	24	18	12	213
Race3	38	32	40	12	22	23	8	10	185
Race4	8	7	10	10	7	8	16	12	78
Total	**131**	**124**	**108**	**104**	**79**	**81**	**67**	**59**	
$n_{.j.}$	**255**		**212**		**160**		**126**		**753**

$n_{..1}$ = 131+108+79+67 = 385 $n_{..2}$ = 124+104+81+59 =368

Test the hypothesis that race, blood group and sex are completely independent and compare your results with that of Pearson, Neyman and the Likelihood chi-squared statistics.

Solution: This is a three-dimensional problem. Similarly, a computer programme was written to handle the computation while the compute of SPSS 10.0 was used to obtain the P-valve. The summary of the results is shown in Table 6.

TABLE 6: SUMMARY OF THE RESULTS ON THE ANALYSIS IN EXAMPLES 1, 2 AND 3.

STATISTICS	METHOD	EXAMPLE 1			EXAMPLE 2			EXAMPLE 3		
		Cal Value	DF	Sig	Cal Value	DF	Sig	Cal Value	DF	Sig
Pearson	Traditional	13.05555	4	.011	.01036	2	.995	76.70828	24	.000
	Simplified	13.05555	4	.011	.01036167	2	.995	76.70827	24	.000
Neyman	Traditional	10.71428	4	.030	.01033928	2	.995	79.13822	24	.000
	Simplified	10.71429	4	.030	.01033691	2	.995	79.13819	24	.000
Likelihood	Traditional	11.69736	4	.120	.01034883	2	.995	73.41119	24	.000
	Simplified	11.69729	4	.120	.01035111	2	.995	74.41493	24	.000
Freeman-Tukey	Traditional	11.26509	4	.024	.01035237	2	.995	73.32258	24	.000
	Simplified	11.26507	4	.024	.01038074	2	.995	73.32337	24	.000

Thus at α =0.05 ,we conclude that the coins are not fair in Example 1, Sex and Mode of entry are independent in Example 2 and that Race, Blood group and Sex are completely dependent in Example 3 .

Without lost of generality, the simplified version has some advantages over the traditional ones. Apart from the fact that it is easier and faster because calculating the expected cell frequencies is not necessary, the method also provides more accurate result since the problem of figure approximations is considerably reduced thereby minimizing the risk of committing either type 1 or 11 error.

REFERENCES

Adegboye, A. 0. 2004. Introductory statistics, probability and test of hypotheses. *1st Edition. Ilorin Kola Success Press.*

Ayinde, K. 2003. Modified chi-squared statistic to test hypothesis about multinomial probabilities in multi-dimensional contingency table. *An International Journal of Biological and Physical Sciences (Science Focus).2: 28-31.*

Ayinde, K. & Ayinde, O. E. 2003. Modified Neyman chi- squared statistic for testing hypothesis about goodness-of-fit of multinomial probabilities. *Zuma Journal of Pure and Applied Sciences. 5(2):123-127.*

Ayinde, K. and Adekanmbi, D. B. 2004. A simplification of the likelihood ratio test statistic for testing hypothesis about goodness of-fit-of multinomial probabilities**.** *Journal of the Nigerian Association of Mathematical Physics. 8: 305-310.*

Ayinde, K. & Iyaniwura, J. O. 2001. A modification of chi-squared statistics to test hypotheses about multinomial probabilities in one-dimensional and two-dimensional contingency table. *Journal of Applied Sciences. 4(1): 1749-1758.*

Bishop, Y. M. M.; Fienberg, S. E & Holland, P. W. 1975. *Discrete Multivariate Analysis*. MIT Press.

Freeman, M. F., & Tukey, J. W. 1950. Transformation related to the angular and square root.. *Annals of Mathematical. Statistics. 27:601-611.*

Koehler, K. and Larntz, K (1980). Empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of. American. Statistical. Association. 73:253-263.*

Larntz, K. 1978. Small-sample comparisons of exact levels of for chi-square goodness-of-fit statistics. *Journal of American. Statistical Association 76:253-263.*

Lawal, H. B. 2003. *Categorical Data Analysis with SAS and SPSS Application*. Lawrence Erlbaum Associates Inc., Publishers Mahwah, New Jersey London.

Lindeman, R. H.; Merenda, P. F. & Gold, R. Z. 1980. *Introduction to Bivariate and Multivariate Analysis.* 1st edn Scott, Foresman and Company,England.

Neyman, J. 1949. Contribution to the theory of the χ^2 test. *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability, 239-293.*

Pearson, K. 1900. On a criterion that a given system of deviations from probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philological. Magicine series 5(50): 157-175.*

Sanni, O. O. M & Jolayemi, E. T. 1998.Robustness of some Categorical test Statistics in small sample situations. *Journal of the Nigerian Statisticians. 2:29-35.*

West, E. N. & Kempthorne, O. 1972. A comparison of the χ^2 and the Likelihood ratio tests for composite alternatives. *Journal of Statistics and Computer Simulation 1:-33.*