THE INTERNATIONAL JOURNAL OF HUMANITIES & SOCIAL STUDIES

Use of Contingency Tables to Investigate the Association of Gender, Type of Primary School, KCPE Results and KCSE Results in Mathematics

Mackenzie M'mata Evans
Assistant Tutorial Lecturer, Department of Mathematics and Actuarial Science, Kisii University, Kenya
Titus Kibua
Senior Lecturer, Department of Statistics and Actuarial Science, Kenyatta University, Kenya
Richard Tinega
Assistant Lecturer, Department of Mathematics and Actuarial Science, Kisii University, Kenya

Abstract:
In this study, the log-linear model with four categorical variables is used to investigate the association between the results of two stages in the Kenya Education system. The source of data was the 2010 KCPE mathematics results and the corresponding 2014 KCSE results in mathematics. The study sample consisted of 1161 candidates, 603 males and 558 females who graduated within the period under study (2010-2014). The instrument used for data collection was the records of candidates' results that made up the sample. In particular, factors such as gender of candidate, type of primary school attended, KCPE Grade and KCSE Grade in mathematics were examined for presence or absence of association. The first two variables had each two categories and the last two variables had each three categories. A four dimensional 2×2×3×3 contingency table was considered. The null hypothesis of mutual independence among these variables was subjected to test. Data was analyzed using the R version 3.2.1 software. KCSE mathematics grade was found to be associated with the KCPE grade but not associated with the type of primary school or gender.

Keywords: *Hierarchical, Log linear, Grade, interaction, categorical.*

1. Introduction

Many situations in Education research result in the collection of data that may be conveniently represented by means of multi-dimensional contingency table. Recent years have seen the emergence of a number of procedures that examine such structures holistically permitting the examination of various higher-order interactions. One of such approach is that of log-linear analysis.

Essentially, analysis of categorical data using log-linear techniques requires the creation of contingency tables. Each variable in the table has a number of categories. The major emphasis is to obtain a log-linear model that is linear in the logarithms of expected frequencies or fits the associations and interactions that exist in the original frequency table (Wrigley 1985). The convectional log-linear models comprise of both saturated and unsaturated log-linear models. This study utilizes conventional log linear models which are hierarchical and considers main effects and interactions.

In analyzing contingency tables, we are interested in the associations among categorical variables or the interpretations of the parameters describing the model structure.

The literature on categorical data analysis is now vast and there are many different strands involving alternative models and methods. Our focus is on the development of log-linear models for four-way contingency tables and the use of chi-square test statistic and likelihood ratio test of goodness of fit to establish if there is any association among the four categorical variables.

Log linear model is a special case of the generalized linear model for Poisson distributed data and is more commonly used for analyzing multidimensional contingency tables that involve more than two variables. It can also be used to analyze two-way contingency tables too (Jeansonne, 2002). A log linear model is similar to the more familiar ANOVA model except that it is applied to the natural logarithm of the expected frequencies (Jibasen, 2004; Lawal, 2003). Response observations in ANOVA are assumed to be continuous normal while in log linear modeling; observations are counts having Poisson distribution (Lawal, 2003).

Until recently the statistical methods available for analyzing associations among more than two categorical variables were extremely limited. The literature on categorical data analysis date back with Bartlett's (1935) work beginning with an article by Roy and Kastenbaum (1956) that formed the basis of the log-linear model approach to contingency tables.

From 1970's, revolution in contingency table analysis has swept through the social sciences casting aside most of the older forms for determining associations among variables. From then the analysis of contingency tables changed quite dramatically with the publication of a series of papers on log-linear models by L.A. Goodman (1970, 1971).Bishop, Fienberg, and Holland (1975), Everitt (1977), Heberman (1978), Agresti (1996, 2002), Knoke and Burke (1980). Through their contributions analysis of associations among more than two categorical variables may be modeled by a log-linear model fitted to the cell frequencies. The appropriateness of log-

linear models for identifying interactions underlying categorical data in higher education research has been established by several authors.

1.1. Estimation of Parameters

Two approaches are used in literature to estimate the parameters in the log-linear models; (1) minimum modified chi-square approach attributed to Grizzle et al (1969) and (2) minimum discrimination information approach. Both methods use maximum likelihood approach. For most contingency table problems, the minimum discrimination information approach yields maximum likelihood estimates. The iterative proportional fitting (IPF) algorithm due to Deming and Stephan (1940) is currently widely used to estimate model parameters. This is to ensure that expected values are obtained iteratively for model whose expected are not directly obtainable (from marginal totals of observed values).

1.2. Model Selection

Many techniques exist in the literature for selecting models. These include: forward selection; backward selection; stepwise procedure; selection based on saturated parameters; selection based on marginal and partial association due to Brown (1976) and Aitkin (1979) method. The backward selection method is however commonly used.

In backward selection, we usually start with the most complex model. Terms are then sequentially deleted from the model. G^2 is computed for each of the current and the reduced model (model resulting from deletion) and using a cut off of predetermined α, say 0.05, we delete the term for which p-value is least significant (term with highest p-value). The process continues until further deletion would lead to significantly poorer fit.

2. Methods

2.1. Data

Data on gender, type of primary school attended, KCPE mathematics grades and KCSE mathematics grades of 1161 students from four of the twelve national schools, who sat their KCPE in 2010 and their KCSE in 2014 formed the sample size.

A four- dimensional contingency table consisting of four variables was constructed.

The variables were classified as follows:

- Variable 1: Gender which has two categories: i =1 and 2. 1= Male, 2= Female.
- Variable 2: Type of primary school attended for KCPE which has two categories:
- j= 1 and 2. 1= public, 2= private.
- Variable 3: KCPE mathematics results which has three categories: k = 1, 2 and 3.
- 1= Below grade C-, 2= Grade C to B+ and 3 = Above grade B+.
- Variable 4: KCSE mathematics results which has three categories: l = 1, 2 and 3. 1=below grade C-, 2= Grade C to B+ and 3 = Above grade B+.

2.2. Introduction to the Model

The log linear model to be considered is the one with four dimensions since four variables are involved. The general log linear model for a contingency table with four variables is given as:

$$\log_e E_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} + u_{12(ij)} + u_{13(ik)} + u_{14(il)} + u_{23(jk)} + u_{24(jl)} + u_{34(kl)} + u_{123(ijk)} + u_{124(ijl)} + u_{234(jkl)} \quad (2.1)$$

for all i, j, k and l.

Where

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_k u_{3(k)} = \sum_l u_{4(l)} = 0$$

$$\sum_i u_{12(ij)} = \sum_j u_{12(ij)} = \sum_i u_{13(ik)} = \sum_k u_{13(ik)} = \sum_i u_{14(il)} = \sum_l u_{14(il)} = 0$$

$$\sum_i u_{123(ijk)} = \sum_j u_{123(ijk)} \sum_k u_{123(ijk)} = \sum_i u_{124(ijl)} = \sum_j u_{124(ijl)} = \sum_l u_{124(ijl)} = 0$$

$$\sum_j u_{234(jkl)} = \sum_k u_{234(jkl)} \sum_l u_{234(jkl)} = \sum_i u_{134(ikl)} = \sum_k u_{134(ikl)} = \sum_l u_{134(ikl)} = 0$$

$$\sum_i u_{1234(ijkl)} = \sum_j u_{1234(ijjkl)} \sum_k u_{1234(ijkl)} = \sum_l u_{1234(ijkl)} = 0$$

and

μ is the overall mean

$u_{1(i)}$is the i[th] level of gender

$u_{2(j)}$ is the j[th] level of school attended for KCPE

$u_{3(k)}$ is the k[th] level of KCPE mathematics grade

$u_{4(l)}$ is the l[th] level of KCSE mathematics grade

$u_{12(ij)}$is the interaction between i[th] level of gender and j[th] level of school attended for KCPE.

$u_{123(ijk)}$ is the interaction between i[th] level of gender, j[th] level of school attended for KCPE and k[th]level of KCPE mathematics grade.

Other interactions are similarly defined.

The "sum to zero" constraints on the parameters are to ensure that the model contains as many the number of cells in the table; suchmodel is called the saturated model. We shall consider only hierarchical models. According to Everitt (1977) the hierarchical principal emphasizes that whenever a higher order effect is included in the model; all the lower order effects composed from variables in the higher efforts are also included. Non-hierarchical models should not be entertained because non-hierarchical modeling does not provide statistical procedure for choosing among potential models Jeansonne (2002).

2.3. Fitting the Model

In this section we show how the log-linear model is fitted and how to estimate the parameters in the model.

2.3.1. Nomenclature of 4 – Dimensional Table

For a 2×2×3×3 contingency table with 2 rows, 2 columns, 3 layer categories of variable 3 and 3 layer categories of variable 4, the observed frequency in the $ijkl^{th}$ cell of the table is represented by

n_{ijkl} $i = 1, 2$

$j = 1, 2$

$k = 1, 2, 3$

$l = 1, 2, 3$

By summing the n_{ijkl} over any single subscript gives the three variable marginal totals. Thus

$$n_{ijk.} = \sum_{l=1}^{3} n_{ijkl}$$

$$n_{ij.l} = \sum_{k=1}^{3} n_{ijkl}$$

$$n_{i.kl} = \sum_{j=1}^{2} n_{ijkl}$$

and

$$n_{.jkl} = \sum_{i=1}^{2} n_{ijkl}$$

2.3.2. Testing the Hypothesis of Mutual Independence

H_o: Gender, type of KCPE School attended, KCPE and KCSE Mathematics results are independent.

The marginal totals and grand totals are obtained as follows:

$$n_{i...} = \sum_{j=1}^{2}\sum_{k=1}^{3}\sum_{l=1}^{3} n_{ijkl} \quad (2.2)$$

$$n_{.j..} = \sum_{i=1}^{2}\sum_{k=1}^{3}\sum_{l=1}^{3} n_{ijkl} \quad (2.3)$$

$$n_{..k.} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{l=1}^{3} n_{ijkl} \quad (2.4)$$

$$n_{...l} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{3} n_{ijkl} \quad (2.5)$$

$$n_{....} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{3}\sum_{l=1}^{3} n_{ijkl} \quad (2.6)$$

The hypothesis to be tested is:

$H_o: P_{ijkl} = P_{i...} \times P_{.j..} \times P_{..k.} \times P_{...l}$ (2.7)

Where

$$P_{i...} = \frac{n_{i...}}{n},\ P_{.j..} = \frac{n_{.j..}}{n}, P_{..k.} = \frac{n_{..k.}}{n}, P_{...l} = \frac{n_{...l}}{n}$$

The expected cell frequencies are obtained as:

$$E_{ijkl} = n \times P_{i...} \times P_{.j..} \times P_{..k.} \times P_{...l}$$

$$= n \times \frac{n_{i...}}{n} \times \frac{n_{.j..}}{n} \times \frac{n_{..k.}}{n} \times \frac{n_{...l}}{n}$$

$$= \frac{n_{i...} \times n_{.j..} \times n_{..k.} \times n_{...l}}{n^3} \quad (2.8)$$

We shall use the Chi-square test statistic:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{3}\sum_{l=1}^{3} \frac{\left(n_{ijkl} - E_{ijkl}\right)^2}{E_{ijkl}} \quad (2.9)$$

Taking natural logarithms of (2.8) we get:

$$\log_e E_{ijkl} = \log_e n_{i...} + \log_e n_{.j..} + \log_e n_{..k.} + \log_e n_{...l} - 3\log_e n \quad (2.10)$$

If the null hypothesis H_0 will be accepted, the following log-linear model will be fitted:

$$\log_e E_{ijkl} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} \quad (2.11)$$

Where;

u is grand mean effect

$u_{1(i)}$ is the main effectof the ith category of variable 1

$u_{2(j)}$ is the main effectof the jth category of variable 2

$u_{3(k)}$ is the main effectof the kthcategory of variable 3

$u_{4(l)}$ is the main effectof the l^{th} category of variable 4

Then the estimates of the main effect parameters in (2.11) are obtained as;

$$\hat{u} = \frac{1}{36}\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k}^{3}\sum_{l=1}^{3} \log_e E_{ijkl} = Z_{....}$$

$$\hat{u}_{1(i)} = \frac{1}{18}\sum_{j=1}^{2}\sum_{k=1}^{3}\sum_{l=1}^{3} \log_e E_{ijkl} - \hat{u} = \bar{z}_{i...} - \bar{z}_{....}$$

$$\hat{u}_{2(j)} = \frac{1}{18}\sum_{i=1}^{2}\sum_{k=1}^{3}\sum_{l=1}^{3} \log_e E_{ijkl} - \hat{u} = \bar{z}_{.j..} - \bar{z}_{....}$$

$$\hat{u}_{3(k)} = \frac{1}{12}\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{l=1}^{3} \log_e E_{ijkl} - \hat{u} = \bar{z}_{..k.} - \bar{z}_{....}z$$

$$\hat{u}_{4(l)} = \frac{1}{12}\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{3} \log_e E_{ijkl} - \hat{u} = \bar{z}_{...l} - \bar{z}_{....}$$

Hence testing for mutual independence is equivalent to testing that there is no interaction between the four variables;

Thus

$H_0: u_{1234(ijkl)} = 0$

$H_1: u_{1234(ijkl)} \neq 0$

2.4. Goodness of Fit

After fitting the model, we assess the goodness of its fit, to determine if it gives the summing of how a log-linear model will fit the data. This is done by comparing the expected frequencies to the observed cell frequencies for each model. With four variables in place, the numbers of models which can fit the data are reasonably many. Hence some procedures are needed to indicate which models may prove reasonable for the data. The likelihood ratio statistic test (G^2) due to Wilks (1938) and Pearson Chi-square (χ^2) due to Pearson (1900) will be used to test the model fitness.

3. Results and Discussion

1. Objective One: To establish if there is any association among gender, type of primary school attended, KCPE mathematics results and KCSE mathematics results.

Table 4.1 shows the model tested, the computed chi-square, degrees of freedom and table chi-square at $\alpha = 0.05$ level of significance.

Model	Factors tested in the model	Deviance χ^2	df	Table χ^2 at $\alpha = 0.05$
1	G + T + P + S	68.32	29	42.56
2	(G.T.P) + S	61.47	22	33.92
3	(G.T.S) + P	56.62	22	33.92
4	(T.P.S) + G	16.83	17	27.59
5	(G.T.S) +T	6.632	17	27.59
6	(G + T + P). S	8.773	21	32.67
7	(G + T + S).P	13.020	21	32.67
8	(T + P + S). G	53.799	24	36.415
9	(G + P + S). T	65.878	24	36.415
10	G.T+G.P+G.S+T.P+T.S+P.S	6.239	16	26.30
Saturated	G.T.P.S	0	0	

Table 1: Deviances for log-linear models fitted to the data on gender, type of school, KCPE and KCSE Mathematics performance

From Table 1 above, we consider the hypothesis that gender (G), type of primary school attended (T), KCPE mathematics results (P) and KCSE mathematics results (S) are independent. The model is

$$G + T + P + S.$$

The log-linear model that corresponds to this is

$\text{Log } \mu_{ijkl} = \log\pi_{i...} + \log\pi_{.j..} + \log\pi_{..k.} + \log\pi_{...l} - \log n(3.1)$

As can be seen in Table 4.1 the deviance for this model is 68.32 which is highly significant compared to table $\chi^2 = 42.56$ at $\alpha = 0.05$ level of significance. Thus we reject the null hypothesis and conclude that gender, type of primary school attended, KCSE mathematics results and KCSE mathematics results are dependent (associated).

2. Objective Two: To establish if there is any significant gender disparity in KCPE and KCSE mathematics performance.

Model	Factors in the model	Deviance χ^2	df	Table χ^2 at $\alpha = 0.05$
1	G + P + S	61.487	12	21.026
2	G.S + P	57.005	10	18.307
3	G.P + S	51.448	10	18.307
4	P.S + G	12.547	8	15.507
5	(P + S). G	46.966	8	15.507
6	(G + S). P	2.508	6	12.592
7	(G + P). S	2.508	6	12.592
8	(G.S)+(G.P) + (P.S)	1.502	4	9.488
Saturated		0	0	

Table 2: Deviances for log-linear models fitted to the data on gender, KCPE and KCSE mathematics performance.

From Table 2 the model 7 has a deviance 2.508 which is insignificant as compared to table $\chi.^2 = 12.592$ at $\alpha = 0.05$ level of significance. Showing that gender is independent of KCPE performance in mathematics, but both jointly influence the performance of mathematics in KCSE. This model indicates that gender has no association with KCPE and KCSE performance in mathematics, but regardless of either boy or girl, a student with grade A in mathematics in KCPE will definitely score grade A in KCSE. This objective is clearly fulfilled by this model.

3. Objective Three: To establish if there is any effect between type of primary school attended (whether public or private), KCPE and KCSE grades in mathematics.

Model	Factors tested in the model	Devianceχ^2	df	Table χ^2 at $\alpha = 0.05$
1		51.492	12	12.026
2	(T.S) + P	50.925	10	18.307
3	(T.P) + S	49.617	10	18.307
4	(P.S) + T	2.551	8	15.507
5	(P + S).T	49.050	8	15.507
6	(T + S).T	0.6766	6	12.592
7	(T+P).S	0.6766	6	12.592
8	T.S + T.P + P.S	0.4963	4	9.488
Saturated		0	0	

. Table 3: Deviances for log-linear models fitted to the data on type of primary school, KCPE and KCSE mathematics performance.

For model P.S + T in Table 3 we test:

$H_o: \pi_{jkl} = \pi_{.kl}\pi_{j..}$(3.2)

Thus, KCPE mathematics results (P) and KCSE mathematics results (S) are associated, and are jointly independent with type of primary school (T).

As we can see from the table the deviance under this hypothesis is 2.551which is highly insignificant. Therefore KCPE performance in mathematics and KCSE performance are not independent and are jointly independent with type of primary school. The findings here show that although there is an association between KCPE mathematics grades and KCSE mathematics grades, there is no association with type of primary school attended. The objective is met.

4. Objective Four: To establish if primary school performance in mathematics influences the performance of mathematics in secondary school.

We test

$H_{o:}\ \pi_{kl} = \pi_{k.}\pi_{.l}$(3.3)

Which has a corresponding log-linear model

$\text{Log}\ \mu_{kl} = \log \pi_{k.} + \log \pi_{.l}$(3.4)

The chi-square test statistic is 48.940with 4 degrees of freedom which is highly significant compared to table $\chi^2 = 6.0084$ at $\alpha = 0.05$ level of significance. Thus, since $48.94 > 6.0084$, we reject the null hypothesis of mutual independent and conclude that KCPE performance in mathematics has an effect on the KCSE performance in mathematics. The objective is hereby fulfilled.

4. Conclusion

Although KCPE mathematics grade seem to be associated with KCSE mathematics grade, this association does not depend on gender. Association between gender and KCPE mathematics grade does not depend on KCSE mathematics grade, just as association between gender and KCSE mathematics grade does not depend on KCPE mathematics grade. There is gender imbalance in both KCPE and KCSE grades.

KCSE mathematics grade is found not to be associated with type of primary school attended. It is however associated with the KCPE mathematics grade. The major determinant of KCSE mathematics grade is the KCPE mathematics grade, regardless of primary school attended for KCPE. This is an indication that the standard does not significantly vary from school to school. It also indicates that there is a significant association between the mathematics grades students obtain at KCPE level and the mathematics grades the same students obtain at KCSE level.

Although previous studies have shown gender disparity in performance of mathematics at secondary school level with boys ahead of girls, this study shows that there is a negligible gender disparity in performance of mathematics at the National school level.

The better the mathematics grade a student scores at KCPE, the better the grade the same student will obtain at KCSE level. Conversely, the lower the grade attained by a student at KCPE level, the lower will be the grade the same student obtains at KCSE level. The current situation in national schools is that majority of the students admitted had excellent grades at KCPE level which demonstrates strong background in basic mathematics content, skills and abilities. This in turn promotes their capacity to attain high mathematics grades at KCSE level.

The study shows that all the three-factor interactions but gender, KCPE mathematics grade and KCSE mathematics grade are significant. Although other studies have indicated that gender contributes significantly towards disparity in mathematics at various levels, the picture portrayed by this study for national schools does not reflect significant difference in performance. However, gender exists in both KCPE and KCSE grades- an issue that deserves attention by the society.

5. References

i. Agresti, A. (1996), An introduction to Categorical Data Analysis, New York: John Wiley &Sons.
ii. Agresti, A. (2002), Categorical Data analysis, Wiley & sons, Hoboken, New Jersey.
iii. Aitkin, M. (1979). A simultaneous test procedure for contingency tables. Applied Statistics, 28, 233-242.
iv. Barlett, M.S (1935), "Contingency table interactions," J. Roy statist. Soc, supplement, 2,248-252.
v. Bishop, Y.M., Fienbcrg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis, Cambridge: MIT press.
vi. Brown, M.B. (1976). Screeing effects in multi-dimensional contingency tables. Applied Statics, 25, 37-46.
vii. Deming, W.E. and Stephen, F.F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal are known. Annas of mathematical Statistics, 11, 427-444.
viii. Everitt, B.S. (1977). The Analysis of contingency Tables. John Wiley & sons, Inc. New York, USA.
ix. Goodman, L.A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. Journal of America Statistical Association, 65,226-256.
x. Goodman, L.A. (1971). The analysis of multi-dimensional tables: stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics,13,33-61.
xi. Grizzle, J.E, Starmer, C.F and Koch, G.G (1969). Analysis of categorical Data by Linear Models. Biometrics 25, 489-505.
xii. Haberman, S.J. (1978).Analysis of qualitative data. New York: Academic Press.
xiii. Jeansonne, A. (2002). Loglinearmodels. Retrieved October 25,2009 from

http://userwww.sfsu.edu/efc/classes/bio1710/loglinnear/log%20Models.htm

xiv. Jibasen, D. (2004). Application of log linear model to prison data. Journal of Nigerian statistical Association, 17, 49-58 Stat and Operations Research.

xv. Kastenbaum, M. A. and Roy, S. N. (1956). On the hypothesis of no interaction in multiway contingency table. The Annals of Mathematical Statistics, 27,749-757.

xvi. Knoke, D., and P.J. Burke. 1980. Log- Linear Models. Sage publications Inc. Newberry Park, California, USA.

xvii. Lawal, H.B. (2003). Categorical data analysis with SAS and SPSS applications. New Jersey: Lawrence Erlbaum.

xviii. Wrigley, N. (1985). Using log-linear models to analyze categorical data. Journal of the Royal statistics society, series B, 3-40.