Statistical Analysis on Students' Performance

Elepo, Tayo Afusat[1] & Balogun, Oluwafemi Samson[*2]

[1]Department of Statistics, Kwara State Polytechnics,
Ilorin, Kwara State, Nigeria
[1]elepotayo1@yahoo.com

[*2]Department of Statistics and Operations Research,
Modibbo Adama University of Technology,
P.M.B. 2076, Yola, Adamawa State, Nigeria.
[*2]balogun.os@mautech.edu.ng

Abstract: This research uses Cohen's Kappa to examine the performance of students in the Faculty of Science, University of Ilorin. The data was collected from eight departments in the faculty and it covers the performance of students measured by their Grade Point Average (GPA) and Cumulative Grade Point Average (CGPA) in both their first and final year between 2000-2006 academic sessions. It is of interest to determine the proportion of students that improved on their performance, dropped from the class of grade point which they started with and those that maintained their performance using psychometrics approach. Also, the strength of agreement that exist between the first and the final year was examined.

***Keywords**:* Cohen's Kappa, Intra-class Kappa, Agreement, Raters.

1. Introduction

Education in a broad sense is the process of exposing the individuals to concepts and activities which physically, mentally, morally and spiritually help equip him/her with the knowledge of things around him. Education also exposes the individual to further knowledge by means of books, mass media and academic institutions.

From the foregoing, thousands of people normally apply into the Nigeria universities-the peak of tertiary institutions. The search for knowledge and the little recognition of certificates of the lower tertiary institution in Nigeria labor market has subjected the university into an over-crowded community with many still outside, eager to add to the congestion. In order to bring about fair play and to exercise justice in the admission of candidates into the universities, the National University Commission (NUC) was set up to look into the affairs of the universities. The commission was established in 1978 and it has since embarked on some policies so as to ensure that admission processes are conducted without duplication of admission.

The information used for this research was obtained by the method of transcription from records and they are all secondary type of data. Data were collected from eight departments in Faculty of Science: Biochemistry, Chemistry, Physics, Geology, Computer Science, Mathematics, Statistics and Microbiology. The data include students' performance measured by their GPA and CGPA as appropriate, in both their first and final year.

The aim of this research is to use Cohen's Kappa to study students' performance of some departments in the Faculty of Science, University of Ilorin and objectives are: to know the strength of agreement that exist between student grade point in both their first and final year, to know the proportion of students that maintained their CGPA (i.e. those that maintained what they started with), to know the proportion of students that dropped from the class of grade point they started with and to know the proportion of students that improved on their performance.

In a study conducted by (Akinrefon & Balogun, 2014), control chart was used to monitor students' performances, the causes underlying the charting statistics that are less than the lower control limits were identified which indicate a negative shift in students CGPA. Also, the reason for charting statistics falling above the upper control limit, which indicate the positive shift in student CGPA was identified. Then, a solution to correct students' poor performance and was suggested. If the charting statistics for all the semester fall within the control limits, the student has maintained the desire target GPA value.

According to (Balogun *et al.,* 2014) the main focus of their research is to develop models that can be used to study the trend of graduate emigration in Nigeria using log-linear modeling based on the results from of likelihood ratio (G^2), Akaike Information Criteria (AIC) , Saturated model has a perfect fit for modeling graduate emigration in Nigeria. This implies all the three factors involved (discipline, year and sex) has to be included in the model in order to have an appropriate result.

Since its introduction Kappa statistics, several authors have applied the concept in different field, for instance; (Zeeshan *et al.,* 2015) carried out an initial audit for evaluating the case notes for each team against the TONK score. In order to evaluate the producibility of this score, the Cohen's kappa was used and substantial agreement was noted. The article by (Viera & Garret, 2015) provided a basic overview of Kappa statistics as one measure of inter-observer agreement. They concluded that "Kappa is affected by prevalence but nonetheless kappa can provide more information than a simple calculation of the raw proportion of agreement".

(Kilem, 2002) established that the unpredictable behavior of the PI and Kappa statistics is due to a wrong method of computing the chance agreement probability. (Warrens, 2015) reviewed five ways to look at Cohen's kappa. Nevertheless, the five approaches illustrate the diversity of interpretations available to researchers who use kappa. In (Wang *et al.*, 2015) Cohen's kappa coefficient was used to assess between raters agreement, which has the desirable property of correcting for chance agreement. It was concluded that despite the limitations, the kappa coefficient is an informative measure of agreement in most circumstance that is widely used in clinical research.

1.1 Categorical Response Data

A categorical variable is one which the measurement scale consists of set of categories. For instance, political philosophy may be measured as "liberal", "moderate", or "conservative" also smoking status might be measured using categories "never smoked", "former smoker" and "current smoker" etc. Though categorical scales are common in the social and biomedical science, they occur frequently in the behavioral sciences, public health, ecology education and marketing. They even occur in highly quantitative field such as engineering science and industrial quality control.

1.2 Categorical Data Analysis

These are data consisting of a classification of the behavior or subjects into a number of mutually exclusive and exhaustive corresponding categories. A multivariate quantitative data is one in which each individual is described by a number of attributes. All individuals with the same description are enumerated and count is entered into a cell of the resulting contingency table

1.3 Contingency Tables

The multidimensional table in which each dimension is specified by discrete variable or grouped continuous (range) variable gives a basic summary for multivariate discrete and grouped continuous data. If the cell of the table are number of observation in the corresponding values of the discrete variables then it is CONTINGENCY TABLES. The discrete or grouped continuous variables that can be used to classify a table are known as FACTORS. Examples include Sex (Male or Female), religion (Christianity, Islam, Traditional etc.).

Types of Contingency table:

One dimensional $\left(1\times J\right)$ tables

Two dimensional (I×J) tables

Square tables (I×I)

Multidimensional tables

1.4 Measures of Agreement

Agreement is a special case of association which reflects the extent to which observers classify a given subject identically into the same category. In order to assess the psychometric integrity of different ratings, inter-raters agreement is computed. Inter-rater reliability coefficients reveal the similarity or consistency of the pattern of responses, or the ranking-ordering of responses between two or more raters (or two or more rating sources), independent of the level or magnitude of those ratings. For example, let us consider the table 1.

Ratings of three subjects by three raters, one observes from the table 1 that all the raters were consistent in their ratings, rater 2 maintained his leading ratings followed by rater 1 and rater 3 respectively.

Inter-rater agreement on the other hand is to measure that ratings are similar in level or magnitude. It pertains to the extent to which the raters classify a given subject identically into the same category. (Kozlowski & Haltrup, 1992) noted that an inert-rater agreement index is designed to "reference the interchangeability among raters: it addresses the extent to which raters makes essentially the same ratings". Thus, theoretically, obtaining high levels of agreement should be more difficult than obtaining high levels of reliability or consistency.

2. Materials and Method

2.1 *Kappa Statistics*

There is wide disagreement about the usefulness of Kappa statistics to assess rater agreement. At the least, it can be said that: kappa statistics should not be viewed as the unequivocal standard or default way to quantity agreement, it should be concerned about using a statistics that is the source of so much controversy and it should consider alternatives and make an informed choice.

One can distinguish between two possible uses of kappa as a way to test rater independence (that is, as a test statistic), and as a way to qualify the level of agreement (that is, as an effect-size measure).

2.2 Cohen's Kappa Coefficient

Cohen's kappa is one of the most commonly used statistic for assessing the nominal agreement between two raters (Warrens, 2010; 2011).

(Cohen, 1960) proposed a standardized coefficient of raw agreement for nominal scales in terms of the proportion of the subjects classified into the same category by the two observers. However, the idea of having an agreement measure was anticipated before 1960. For example, decades earlier Corrado Gini already considered measures for assessing agreement on a nominal scale (Warrens, 2013). The proportion is estimated as;

$$\pi_i = \sum_{i=1}^{I} \pi_{ii} \qquad (1)$$

And under the baseline constraints of complete independence between ratings by the two observers, which is the expected agreement proportion estimated as;

$$\pi_0 = \sum_{i=1}^{I} \pi_{i.}\pi_{.j} \qquad (2)$$

The kappa statistics can now be written as;

$$K_c = \frac{\pi_i - \pi_0}{1-\pi_0} \qquad (3)$$

$$where\, \pi_i \text{ and } \pi_0 \text{ are as defined above}$$

(Landis & Koch, 1977a) have characterized different ranges of values for kappa with respect to the degree of agreement they suggest. Although, these original suggestions were admitted to be "clearly arbitrary", they have become incorporated into the literature as standards for the interpretation of the kappa values. For most purposes, values greater than 0.75 or so may be taken to represent excellent agreement beyond chance, values below 0.40 or so may be taken to represent poor agreement beyond chance, values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance and this is clearly shown in table 2. Bias of one rater relative to another refers to discrepancies between these marginal distributions. Bias decreases as the marginal distributions becomes more nearly equivalent. The effect of rater bias on kappa has been investigated by (Feinstein & Ciccheti 1990) and (Bryt *et al.*, 1993).

Early approaches to this problem have focused on the observed proportion of agreement; see (Goodman & Kruskal, 1954), this suggests the chance that the agreement can be ignored. Later, Cohen's kappa was introduced for measuring nominal scale chance-corrected agreement. (Scott, 1955) defined π_e using the underlying assumption that the distribution of proportion over the Ith categories for the population is known, and is equal for the two raters. Therefore, if the two raters are interchangeable, in the sense that the marginal distributions are identical, then Cohen's and Scott's measures are equivalent because Cohen's kappa is an extension of Scott's index of chance-corrected measures. To determine whether K differs significantly from zero, one could use the asymptotic variance formulae given by (Fleiss *et al.*, 1969) for the general $I \times I$ table. For large n, Fleiss' formulae is practically equivalent to the exact variance derived by (Everitt, 1968) based on the central hypergeometric distribution. Under the hypothesis of only chance agreement, the estimated large-sample variance of K is given by;

$$\tilde{V}ar_0(K_e) = \frac{\pi_e - \pi_e{}^2 - \sum_{i=1}^{I} \pi_{i.}\pi_{.i}(\pi_{i.} + \pi_{.i})}{n(1-\pi_e)^2} \ldots..(4)$$

Assuming that

$$\frac{\tilde{K}}{\sqrt{\tilde{V}ar_0(\tilde{K})}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Follows a normal distribution, we can test the hypothesis of the chance agreement with reference to the standard normal distribution

2.3 *Intraclass Kappa*

Intraclass kappa was defined for data consisting of blind dichotomous ratings on each of n subjects by two fixed raters. It is assumed that the ratings on a subject are interchangeable, that is, in the population of subject; the two ratings for each subject have a distribution that is invariant under permutations of the raters to ensure that there is no rater bias (Scott, 1955), (Bloch & Kraemer, 1988), (Donner & Eliasziw, 1992) and (Banergee *et al.,* 1999). The Intraclass is estimated as;

$$K_i = \frac{\pi_0^e - \pi_e^*}{1 - \pi_e^*} \qquad (6)$$

$$where$$

$$\pi_e^* = \sum \pi_{ii}$$

$$\pi_0^e = \sum \left[\frac{\frac{\pi_{i.} + \pi_{.j}}{\pi_{..}}}{2} \right]^2$$

Furthermore, to obtain the proportion of those that maintained the performance or grade they started with, the proportion of those that improved on their performance and also the proportion of those that dropped from the class of grade point they started with. Let

P_1= the proportion of those that maintained what they started with, that is, the diagonal table

P_2 = the proportion of those that improved on their performance, that is, those below the diagonal table

P_3= the proportion of those that dropped from the class of grade point they started with, that is, those above the diagonal table.

3. Data Analysis

This section shows the analysis on Cohen kappa, Intra-class kappa statistic and the proportion of students who maintained, dropped and improved on their performance as shown in table 3 and 4 below.

This calculation is done to demonstrate the percentage of students who maintained, improved and dropped in the CGPA they started with

$$P_1 = \frac{4.3538}{7.9998} \times 100 = 54.42$$

$$P_2 = \frac{2.7513}{7.9998} \times 100 = 34.39$$

$$P_3 = \frac{0.8947}{7.9998} \times 100 = 11.18$$

Discussion, Summary and Conclusion

4.1 *Discussion*

For Physics Department, 55.56% of the students were able to maintain their grade point, 35.56% of the students improved while 8.89% dropped from the class of grade point they started with.

For Statistics Department, 74.04% of the students were able to maintain their grade point, 18.52% of the students improved while 7.4% dropped from the class of grade point they started with.

For Microbiology Department, 48.98% of the students were able to maintain their grade point, 44.89% of the students improved while 6.12dropped from the class of grade point they started with.

For Mathematics Department, 36% of the students were able to maintain their grade point, 64% of the students improved while none dropped from the class of grade point they started with.

For Geology Department, 46.15% of the students were able to maintain their grade point, 49.23% of the students improved while 4.62% dropped from the class of grade point they started with.

For Computer Science Department, 51.49% of the students were able to maintain their grade point, 1.79% of the students improved while 46.71% dropped from the class of grade point they started with.

For Biochemistry Department, 60.17% of the students were able to maintain their grade point, 29.66% of the students improved while 10.17% dropped from the class of grade point they started with.

For Chemistry Department, 62.96% of the students were able to maintain their grade point, 31.46% of the students improved while 5.56% dropped from the class of grade point they started with.

4.2.1 Summary

From the above interpretation, we could see that 54.42% of the students were able to maintain their CGPA that they started with, 34.39% of the students improved and 11.18% of the students dropped from the class of grade point they started with. Also, the strength of agreement between the first and the final year result is on the 0.40%.

4.3 *Conclusion*

It can be observed that Mathematics department has the highest number of students that improved on their performance, Statistics department had the highest number of students that maintained their grade point and Computer Science department had the highest number of students that dropped from the grade point they started with. Also, the strength of agreement that exist between the first and the final year result is on Average, that is, “fair”.

References

Akinrefon, A.A and Balogun, O.S. (2014). Use of Shewart control chart technique in monitoring student performance. Bulgarian Journal of Science and Education Policy, 8(2), 311-324.

Banergee, M., Capozzoli, M., Mcsweeney, I.J. and Sinha, D. (1999). Beyond Kappa: A review of interrater agreement measures. The Canadian journal of Statistics, 20(1), 3-23.

Bloch, D.A., Kraemer, H.C. (1988). 2×2 Kappa coefficients: Measures of agreement or association. Biometrics, 45, 269-287.

Balogun, O.S., Bright, D.E., Akinrefon, A.A. and

Abdulkadir, S.S. (2014). Modeling Graduate Emmigration Nigeria using Log-Linear Approach. Bulgarian Journal of Science and Education Policy, 8(2), 375-391.

Bryt, T., Bishop, J. and Carlin, J.B. (1993). Bias, prevalence and kappa. J. Clin. Epidemiol., 46, 423-429.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Edu. and Psych. Meas., 20, 37-46.

Donner, A. and Eliasziw, M. (1992). A goodness of fit approach to inference procedures for kappa statistic: confidence interval construction, significance testing and sample size estimation. Statist. Med., 11, 1511-1519.

Everitt, B.S. (1968). Moments of the statistics kappa and weighted kappa. British J. Math. Statist. Psych., 21, 97-103.

Feinstein, A.R. and Cicchetti, D.V. (1990). High agreement but low kappa I: the problems of two paradoxes. J. Clin. Epidemiol., 43, 543-548.

Fleiss, J.L., Cohen, J. and Everitt, B.S. (1969). Large sample standards errors of kappa and weighted kappa. Psych. Bull., 72, 323-327.

Goodman, L.A. and Kruskal, W.H. (1954). Measuring of association for cross classifications. J. Amer. Statist. Assoc., 49, 732-768.

Kilem Gwett (2002). Kappa Statistic is not satisfactory for assessing the extent of a agreement between raters. Series: Statistical Methods for Inter-Rater Reliability Assessment, No. 1: 1-5.

Kozlowski, S.W.J. and Hattrup, K. (1992). A disagreement about within group agreement: Disentangling issues of consistency versus consensus. J. Applied Psych., 77(2), 161-167.

Landis, R.J. and Koch, G.G. (1977a). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Scott, W.A. (1955). Reliability of content analysis: The case of nominal scale coding. Public Opinion. Quart., 19, 321-325.

Khan, Z., Sayers, A.E., Khattak, M.U. and Chamber, I.R. (2015). The TONK score: a tool for assessing quality in Trauma and Orthopaedic note-keeping. SICOT J., 1(29):1-4.

Viera, A. J., Garraett, J.M. (2015). Understanding Interobserver Agreement: The Kappa Statistic. Research Series (Family Medecine), 37(5): 360-363.

Warren, M. J. (2010). Inequalities between kappa and kappa like statistc for $k \times k$ tables. Psychometrika 75: 176-185.

Warren, M. J. (2011). Cohen's kappa is weighted average.

Statistical Methodology 8: 473-484.

Warren, M. J. (2013). A comparison of Cohen's kappa and agreement coefficients by Corrado Gini. International Journal of Research and Reviews in applied Sciences 16: 345-351.

Warren, M. J. (2015). Five ways to look at Cohen's Kappa. J. Psychol. Psychother., 5(4):1-4.

Wan Tang, Jun Hu, Hui Zhang, Pan Wu and Hua H.E. (2015). Kappa coefficient: a popular measure of rater agreement. Shanghai Archive of psychiatry, 27(1): 62-67.

Appendix

Table 1: Example of Raters

Subject	Rater 1	Rater 2	Rater 3
1	5	6	2
2	3	4	2
3	1	2	1

Table 2: The Range Of Kappa Statistic with the Respective Strength of Agreement

Kappa statistic	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Table 3: Cohen's and Intra-Class Kappa Estimates

S/No	**Department**	**Cohen's Kappa**	**Intra-class Kappa**
1	Physics	0.3410	0.3280
2	Statistics	0.6291	0.6279
3	Microbiology	0.2409	0.2214

4	Mathematics	0.0315	0.035
5	Computer Science	0.2861	0.2615
6	Geology	0.1955	0.1710
7	Biochemistry	0.4169	0.6017
8	Chemistry	0.3865	0.3822

Table 4: The proportion of Students that improved, maintained and dropped

S/No	**Department**	P_1	P_2	P_3	**Sum**
1	Physics	0.5556	0.3556	0.0889	1.0001=1
2	Statistics	0.7407	0.1852	0.074	0.9999=1
3	Microbiology	0.4898	0.4489	0.0612	0.9999=1
4	Mathematics	0.3600	0.6400	0	1.0000=1
5	Computer Science	0.5149	0.0179	0.4671	0.9999=1
6	Geology	0.4615	0.4923	0.0462	0.9999=1
7	Biochemistry	0.6017	0.2966	0.1017	0.9999=1
8	Chemistry	0.6296	0.3148	0.0556	0.9999=1

Table 5: The Strength of Agreement for each Department

S/No	**Department**	**Strength of Agreement**
1	Physics	Fair
2	Statistics	Substantial
3	Microbiology	Fair
4	Mathematics	Slight
5	Computer Science	Fair
6	Geology	Moderate
7	Biochemistry	Slight
8	Chemistry	Fair